公益財団法人笹川平和財団

海洋政策研究所　御中

テキストマイニングによる
海洋関連白書分析に関する業務

成果報告書

2022 年 3 月

ジョルダン株式会社

企画営業本部

目　　次

1. はじめに

本報告書は、公益財団法人笹川平和財団海洋政策研究所（以下「OPRI」と記す）からの委託業務「テキストマイニングによる海洋関連白書分析に関する業務」において実施した内容を取りまとめたものである。

2. 業務概要

2.1 本業務の目的

海洋には、海洋温暖化、海洋酸性化、富栄養化、海洋ごみ汚染、干潟藻場の減少、密漁など多くの問題が存在する一方、水産資源や海底鉱物資源、再生可能エネルギーの開発、生物多様性の維持、海上輸送による貿易、二酸化炭素の貯留など様々な価値も有し、人類にとって重要な財産である。これら海洋を巡る現状分析や将来展望について、水産白書、環境白書、海洋白書等の公開文書に記載されている。これら海洋関連白書について、現時点で入手可能な過去の公開文書類のテキスト情報を抽出し、テキストマイニングにより政府省庁、海洋関連組織における関心事や動向の分析と情報の可視化を行うとともに、その作業手順の定式化を図った。

2.2 データベースの作成

2021 年 10 月末時点で下記 URL から取得可能な公開文書について、手操作により電子ファイル（PDF 文書ファイル）を取得した。

水産白書 https://www.jfa.maff.go.jp/j/kikaku/wpaper/
海洋基本計画 https://www8.cao.go.jp/ocean/policies/plan/plan.html
海洋白書 https://www.spf.org/OPRI/projects/information~white-paper.html
環境白書 https://www.env.go.jp/policy/hakusyo/past_index.html

1）PDF 文書ファイルからのテキスト抽出

PDF 文書ファイルからテキスト情報を抽出する際、図や表、脚注、フッター、句読点の無い章・節・項の見出し文字列などを除き、本文のテキスト情報だけを抽出することとした。当初、PDF 文書ファイル閲覧ツール（Web ブラウザ、Acrobat Reader など）で手操作によるテキスト抽出作業を行ったが、作業効率が悪く、PDF 文書ファイルによっては意図通りにテキストが抽出されない場合もあった。そこで、Windows マシンの Python 環境下で動作する「PDF 文書からのテキスト抽出用ツール」を準備した。本業務の成果物として、その操作手順書「PDF 文書からのテキスト抽出手順」を作成した。

実作業では、PDF 文書ファイルとテキスト抽出結果との整合性が高く、精査作業が効率的な Apache Tika を利用している。しかし、最終的には PDF 文書ファイルとテキストファイルを見比べ、テキストマイニングの入力として適切な本文テキスト以外を削除する操作は必須である。その精査作業では、「コラム」や「事例」など文書中の囲み記事をテキスト化対象とした。

2) テキスト・クリーニング

PDF 文書ファイルとテキスト抽出結果の精査作業段階で下記テキスト・クリーニング操作を行った。

- 1)、ア）、〇、●、・ など箇条書きの文字の削除
- 本文中の「（以下、□□と記す。）」との但し書きの削除
- 文章の最後は「。¥n」で統一

KH Coder で前処理を実行し、Chasen での形態素解析処理に不適切な文字コード（例：①、㎏、㎢）が検出された場合には、KH Coder のテキストの自動修正機能を実行している。

本作業による各白書の PDF 文書ファイルを年度毎に統合化したテキストファイルを作成し、先頭行に KH Coder での文書識別用 h5 タグを挿入した。そのテキスト抽出結果ファイルの一覧を表 2-1 に示す。

表 2-1 テキスト抽出結果ファイル一覧

文書名	期間	テキストファイル名
水産白書（施策編）	2007 年～2020 年	Fisherie_Policy_ally.txt
水産白書	2007 年～2020 年	水産白書 2007-2020b.txt
海洋基本計画	1 次、2 次、3 次	海洋基本計画 plan123.txt
海洋白書	2004 年～2020 年	海洋白書 2004-2020b.txt
環境白書	2008 年～2020 年	環境 2008-2020bz.txt
環境・水産・海洋白書	2008 年～2020 年	海洋・水産・環境 2008-2020bz.txt

（注） 環境白書が 2008 年～2020 年の期間であるため、水産白書と海洋白書も同期間に限定して 3 白書を 1 ファイルとした。

2.3 テキストマイニングによる分析

テキスト抽出結果ファイル（表 2-1）を入力データとし、KH Coder での前処理結果である抽出語リストを目視確認し、分析に必要な複合語や不適切な単語を選別した。環境・水産・海洋白書（2008 年～2020 年）について KH Coder の前処理メニューにある複合語の検出（茶筌を利用）を行い、強制抽出する複合語 316 を選出した。また、分析対象から除外する単語 54 を規定した除外語リストを作成した。KH Coder の前処理メニューで分析に使用する語の取捨選択で、これらテキストファイルを指定している。

表 2- 2 「分析に使用する語の取捨選択」での設定ファイル

機能	テキストファイル名	備 考
強制抽出する語の指定	環境・水産・海洋-複合語 min.txt	付録 1 に掲載
使用しない語の指定	除外語_015_OPRI.txt	付録 2 に掲載

2.3.1 KH Coder による分析

テキスト抽出結果ファイル（表 2-1）ごとに KH Coder を用いて、共起ネットワーク、対応分析、トピック分析の処理を行った。その処理結果は、共通的な記述構成の報告書として白書別に OPRI 殿に提出した。その提出文書を付録に掲載した。（表 2-3）

表 2-3 KH Coder による処理結果報告書の一覧

分析対象文書	期間	備 考
水産白書（施策編）	2007 年～2020 年	付録 3 に掲載
水産白書	2007 年～2020 年	付録 4 に掲載
海洋基本計画	1 次、2 次、3 次	付録 5 に掲載
海洋白書	2004 年～2020 年	付録 6 に掲載
環境白書	2008 年～2020 年	付録 7 に掲載
環境・水産・海洋白書	2008 年～2020 年	付録 8 に掲載

KH Coder を利用した各白書の処理結果報告書では、年度毎に全文章を 1 文書として扱い、入力テキストファイルに対応して KH Coder が自動設定する分析対象語数と可視化された分析結果の判読の上限値とされる 150 語付近での 2 ケースを実施している。

共起ネットワーク分析では、まず各年度の抽出された単語を集約した全年度分について単語（node）間の共起関係（edge）ネットワークを描画し、次に各年度での単語（node）間の共起関係（edge）を描画している。前者を「共起ネットワーク（語・語）」、後者を「共起ネットワーク（語・年）」と題している。各 2 ケースの分析対象語数について実施した結果、分析対象語数 150 前後では単語（node）が多すぎて「共起ネットワーク（語・語）」ネットワーク図の目視判読には適さない。

対応分析では、主成分の累積寄与率 70%を目途に、複数主成分での対応分析結果を可視化している。年度別の影響を可視化するため「語・年」で描画すると共に、原点付近の単語描画を省略し、特徴的な単語を描画している。目視での特徴判読には、KH Coder で自動設定される分析対象語数よりも多くの対象語では分析結果の判読が容易ではないことが判明した。

トピック分析（LDA）では予めトピック数を設定する必要があるので、KH Coder のトピック数推定（ldatuning）の出力結果図を参考に設定した。この出力結果図に描画される 4 種の評価パラメータ（Giffiths2004、Arun2010、CaoJuan2009、Deveaud2014）は、トピック分析用 R 言語 stm パッケージのトピック数推定関数（searchK）で使用される評価パラメータ（heldout likelihood、lbound、residual dispersion、semantic coherence）とは異なる。トピック数推定（ldatuning）では R 言語の ldatuning パッケージを利用しているが、そのマニュアルには評価パラメータの詳細は記載されていない。公開されているソースコードから処理を推察すると、指定したトピック数で topicmodels パッケージの LDA 処理を起動し、その結果について CaoJuan2009 では 2 トピック間の単語からコサイン類似度を求め、その総当たり平均が最小になるようなトピック数を求めている。Deveaud2014 では、2 トピック間での単語出現分布の Jensen-Shannon ダイバージェンスに似た式で算出し、その総当たり平

均が最大になるようなトピック数を求めている。Arun2010 では Kullback-Leibler ダイバージェンスを算出している。Griffiths2004 では、topicmodels パッケージの LDA で算出している対数尤度（logLiks）と同じように調和平均で評価している。Giffiths2004 では、KH Coder のトピック数推定（perplexity）と同様、トピック数の増加に対して滑らかな曲線となり、トピック数が多い程良い評価となることが多く、最適なトピック数の決定には窮する。一方、CaoJuan2009 の最大値と Deveaud2014 最小値によるトピック数の判定が容易なことが多く、本作業では主としてこれらの評価パラメータからトピック数を決定した。

KH Coder での LDA 処理結果は、各トピックの出現確率の上位 10 単語が一覧表で表示される。その一覧表でトピックを選び、そのトピック比率を可視化（折れ線グラフとヒートマップ）すると、そのトピックの経年変化の様子を把握できる。提出文書に掲載した折れ線グラフ以外でのトピック比率の変化比較ができるよう、このトピック比率表を Excel データとして保存して別途提供した。この「トピック出現確率」と「トピック比率」の算出方法は、KH Coder マニュアルには明記されていない。

KH Coder のトピック分析（LDA）では、R 言語の topicmodels パッケージを利用しているが、同じ環境下で異なる分析結果とならぬよう、処理パラメータ（乱数初期値、Gibbs サンプリング数、空評価数）を固定化して処理を実行している。トピックが安定化する Gibbs サンプリング数は、分析対象語数とトピック数に依存するが、KH Coder では Gibbs サンプリング数：2000、空評価数：1000 として全 3000 サンプルと固定しており、この値を変更できない。但し、KH Coder でのトピック分析で自動設定される値は 150 以下であり、トピック数 20 程度であれば Gibbs サンプリング数：2000 で問題ないと推察する。しかし、分析対象語数を増やした場合には、Gibbs サンプリング数：2000 では抽出トピックが安定化していない懸念がある。

そこで、分析対象文が最も多い環境・水産・海洋白書（2008～2020 年）について、KH Coder での前処理結果である抽出語リスト（Document Term Matrix に相当）を csv 形式で出力し、それを topicmodels パッケージの LDA の入力データとして R-Studio で処理した。本作業では、年度毎に全文章を 1 文書とした入力テキストファイルとしており、KH Coder の「トピックの推定結果での確率」は topicmodels の LDA 処理結果に格納されている「beta（per-document-per-word probability）」であり,「トピック比率」は「gamma（parameters of the posterior topic distribution for each document）」と一致することを確認した。 一方、「環境・水産・海洋白書（2008～2020 年）」で分析対象語数:75、トピック数:16 とした場合、各トピックの出現確率上位 5 の単語群は Gibbs サンプリング数:2,000 と 100,000 で顕著な差異はなかった。しかし、分析対象語数:154、トピック数:16 の場合、各トピックの出現確率上位 5 の単語群が安定化するのは、Gibbs サンプリング数:50,000 以上であることが判明した。これら調査結果の詳細は、下記文書で報告した。

- KHCoder での LDA 分析結果と R パッケージでの処理結果の比較（付録 9 に掲載）
- 環境・水産・海洋白書の LDA 分析での Gibbs サンプリング数について（付録 10 に掲載）

2.3.2 教師付き LDA 分析

OPRI 殿より提示された「関心の高いキーワード」を軸に LDA 分析でトピックを抽出するためには、R 言語の seededlda パッケージを利用した。この seededlda パッケージには、toplicmodels パッケージの LDA と同様な機能（textmodel_lda）が準備されているので、まず topicmodels パッケージの LDA 処理結果と比較した。但し、topicmodels パッケージ LDA への入力データは DTM（Document Term Matrix）或いは tm パッケージ用入力データの dtf 形式であるが、seededlda での入力データは quanteda パッケージ用入力データの dfm 形式である。具体的には、tidyverse パッケージの関数を利用して DTM から dtf、dfm へのフォーマット変換処理が必要である。

KH Coder 処理結果と topicmodels パッケージの処理結果の整合性は確認済であるが、その前提は入力データが DTM 形式の場合である。DTM から dtf フォーマット変換後に topicmodels で処理した場合には、KH Coder の処理結果と topicmodels の処理結果が完全には一致しない。DTM から dtf フォーマット変換結果と dfm フォーマット変換結果が等価であることを確認した後、topicmodels パッケージと seededlda パッケージでの処理結果を比較した。seededlda パッケージでの処理結果に格納されている「phi」が topicmodels の「beta」（「トピックでの単語出現確率」）に、「theta」が「gamma」（「トピック比率」）に相当する。両パッケージでの処理結果を比較すると、Gibbs サンプリングが十分に大きい場合には、トピックの単語出現確率の上位単語に顕著な差異は無い。しかし、「theta」と「gamma」のパターンは似ているものの、その数値の差異は大きい。本業務では、「theta」や「gamma」の定性的（パターン）評価に着目しているので、両者の処理結果に顕著な差異は無いと判断した。その詳細手順と比較結果は、下記文書で報告した。

- 夏目漱石(1914)著『こころ』で KHCoder と R-Studio の LDA 処理結果を比較（付録 11 に掲載）

環境・水産・海洋白書（2008～2020 年）の分析対象語数：154 での topicmodels の LDA 処理結果は、「2.3.1 KH Coder による分析作業」に記載した。KH Coder の「外部変数と見出し」メニューから各白書の上位 10 個の特徴語を一覧表形式で Excel に出力できる。その出力結果を図 2-1 に示す。この一覧表から、各白書で共起している単語や Jaccard 係数の高い単語をトピックのキーワード候補として絞り込める。

OPRI 殿から提示された「関心の高いキーワード」を表 2-4 に示す。この表中の 5 類辞書の単語は、「再生可能エネルギー」を除き、KH Coder による LDA 分析での分析対象だった 154 語に含まれている。分析対象語数を 226 とすれば、5 類辞書の全単語が含まれるので、分析対象語数を 154 と 226 とした 2 ケースについて 5 類辞書ファイルで教師付き LDA 分析を実施した。8 類辞書の全単語を網羅する分析対象語数は 4727 となり、「2.3.1 KH Coder による分析作業」での LDA 処理で対象とした単語数の 30 倍である。これについても、5 類辞書と 8 類辞書の 2 ケースを実施した。

辞書として指定した「関心の高いキーワード」を軸に、seededlda パッケージで環境・水産・海洋白書（2008～2020 年）についてトピックを抽出し、白書ごとに年度単位でのトピックの出現確率を可視化した。また、8 類辞書をベースに KHCoder のコーディングファイル（分類を規定したキーワード群、表 2-

5）を規定し、KH Coder のクロス集計機能を用いて、これらキーワードの各白書での出現状況を可視化したので、これらから省庁横断的な関心事項の経年変化の状況を把握できる。8 分類クロス集計結果を図 2-2、そのヒートマップを図 2-3、バブルプロットを図 2-4 に示す。

辞書ファイルを指定した教師付き LDA の処理結果の詳細は、通常の LDA 分析結果と対比して下記文書で報告した。

- 環境・水産・海洋白書（2008～2020） 教師付き LDA 分析処理結果（付録 12 に掲載）
- 環境・水産・海洋白書（2008～2020） 教師付き LDA 分析処理結果（その２）（付録 13 に掲載）

海2008		海2009		海2010		海2011		海2012		海2013		海2014	
政策	.080	島	.045	海賊	.040	沿岸域	.062	津波	.059	沿岸域	.053	北極	.045
基本	.074	教育	.045	探査	.036	海域	.045	教育	.038	海域	.043	基本	.042
法案	.058	離島	.043	情報	.034	保護区	.034	ロシア	.038	大陸棚	.042	産業	.037
規定	.047	海域	.035	海底	.033	ゴミ	.033	被害	.037	教育	.038	船舶	.037
総合的	.045	総合的	.033	科学技術	.030	管理	.032	風力発電	.035	津波	.037	観測	.036
本部	.038	大陸棚	.028	申請	.030	離島	.032	洋上	.034	総合的	.037	EEZ	.035
国連海洋法条約	.035	国際	.027	大陸棚	.028	海岸	.030	中国	.033	離島	.036	総合的	.032
秩序	.035	管理	.027	船舶	.026	漂着	.030	南シナ海	.032	基本	.031	計画	.032
施策	.035	問題	.027	運	.025	海賊	.030	復興	.029	管理	.030	施策	.030
海域	.031	EEZ	.026	海底熱水鉱床	.024	排他的経済水域	.030	海域	.027	総合	.029	調査	.030

水2008		水2009		水2010		水2011		水2012		水2013		水2014	
子ども	.064	漁村	.049	水産資源	.044	水産物	.053	水産物	.079	養殖	.150	漁船	.066
魚	.051	魚	.032	資源管理	.043	水産	.051	魚	.064	養殖業	.070	生産量	.066
水産物	.049	時代	.029	漁業者	.041	復興	.051	消費者	.061	水産物	.061	漁獲	.066
小売	.033	水産物	.025	漁獲	.034	漁港	.050	漁協	.049	漁船	.044	漁業資源	.061
食	.033	漁業者	.024	水産物	.030	養殖業	.048	販売	.046	種苗	.044	資源管理	.057
消費者	.031	江戸	.023	マグロ	.029	被災地	.048	魚介類	.040	経営	.036	漁業者	.053
価格	.028	水産資源	.022	漁場	.027	宮城	.047	調理	.037	水産	.035	操業	.047
経営	.027	魚介類	.022	漁協	.024	津波	.045	養殖業	.036	餌	.033	水産物	.044
調理	.026	漁労	.020	水産	.023	東日本	.045	加工	.036	天然	.033	魚種	.043
産地	.025	漁場	.019	漁船	.023	再開	.045	水産	.035	魚	.033	経営	.038

環2008		環2009		環2010		環2011		環2012		環2013		環2014	
廃棄物	.060	環境	.061	環境	.056	環境	.055	環境	.052	環境	.051	環境	.056
環境	.052	推進	.043	生物多様性	.045	生物多様性	.050	実施	.044	実施	.043	実施	.048
処理	.050	廃棄物	.042	地球	.044	廃棄物	.041	地域	.044	推進	.037	地域	.043
社会	.047	利用	.042	利用	.037	利用	.039	処理	.044	処理	.036	推進	.039
推進	.043	削減	.042	推進	.036	実施	.039	利用	.037	廃棄物	.035	技術	.039
循環型	.043	対策	.040	対策	.036	社会	.037	対策	.036	地域	.034	対策	.038
利用	.040	排出量	.040	社会	.036	推進	.037	自然	.036	対策	.034	自然	.035
実施	.038	生物多様	.038	温暖化	.035	処理	.035	推進	.036	保全	.033	保全	.035
リサイクル	.038	実施	.037	水	.035	保全	.035	生物多様性	.035	生物多様	.033	廃棄物	.035
対策	.037	排出	.035	廃棄物	.034	地球	.034	社会	.035	自然	.031	処理	.035

海2015		海2016		海2017		海2018		海2019		海2020	
EEZ	.045	北極	.053	中国	.045	安全保障	.039	海底	.038	日本	.033
海域	.044	沿岸域	.037	海域	.036	中国	.035	深海	.037	北極	.028
沿岸域	.037	政策	.032	EEZ	.033	海域	.034	衛星	.032	島嶼国	.028
教育	.031	国際	.032	国連海洋法条約	.031	会議	.034	北極	.028	女性	.027
管理	.030	開発	.032	観測	.031	観測	.031	洋上	.027	能力	.024
総合	.027	総合	.031	開発	.029	議論	.030	プラスチック	.026	運航	.023
地震	.026	海域	.031	計画	.028	政策	.027	AUV	.025	研究	.022
総合的	.026	中国	.029	沿岸域	.027	日本	.026	日本	.025	安全	.022
酸性化	.025	国連	.026	議論	.027	国連	.025	科学	.025	教育	.022
考察	.025	開催	.026	基本	.026	産業	.025	開発	.024	ArCS	.021

水2015		水2016		水2017		水2018		水2019		水2020	
漁村	.056	漁獲	.059	漁業者	.044	水産	.058	漁船	.053	水産	.057
漁獲	.049	漁船	.043	漁獲	.043	生徒	.040	漁業者	.049	漁業者	.040
漁業者	.044	資源管理	.042	漁船	.043	漁業者	.040	水産物	.047	生徒	.040
経営	.041	マグロ	.037	魚	.040	水産業	.037	特	.046	水産業	.037
水産物	.040	漁業者	.035	水産物	.040	漁協	.036	水産	.043	漁協	.035
漁船	.039	水域	.035	水温	.034	水産高校	.035	漁獲	.041	水産高校	.035
操業	.039	魚	.034	操業	.032	漁船	.034	資源管理	.040	漁船	.034
魚	.034	水産物	.034	水産	.029	魚	.033	減少	.037	魚	.032
漁協	.029	操業	.033	漁場	.029	資源管理	.033	経営	.037	資源管理	.032
加工	.029	管理	.033	資源管理	.029	漁獲	.033	魚種	.034	漁獲	.032

環2015		環2016		環2017		環2018		環2019		環2020	
地域	.061	実施	.051	実施	.049	実施	.042	気候変動	.054	地域	.051
環境	.050	対策	.048	環境	.046	環境	.042	適応	.045	環境	.045
実施	.044	環境	.047	推進	.042	地域	.040	実施	.044	気候変動	.043
推進	.040	推進	.045	対策	.038	推進	.038	推進	.043	実施	.041
活用	.036	廃棄物	.043	廃棄物	.038	活用	.033	地域	.043	対策	.039
自然	.035	処理	.040	開催	.036	支援	.033	環境	.042	社会	.039
廃棄物	.035	開催	.036	処理	.035	廃棄物	.031	影響	.041	支援	.037
保全	.034	排出	.035	施設	.032	対策	.031	プラスチック	.040	活用	.036
処理	.034	規制	.034	地域	.032	促進	.030	対策	.036	推進	.036
開催	.031	評価	.034	支援	.031	処理	.029	支援	.034	影響	.033

図 2-1 各白書で共起している上位 10 単語（Jaccard 係数）

表 2-4 「関心の高いキーワード」

	キーワードを規定したR言語のコード
5類辞書	dict5 <- dictionary(list("気候変動"=c("気候変動"), "生物多様性"=c("生物多様性"), "温暖化"=c("温暖化","温室効果ガス"), "水産物"=c("水産物"), "エネルギー"=c("エネルギー","再生可能エネルギー")))
8類辞書	dict8 <- dictionary(list("気候変動" = c("気候変動"), "プラスチック" = c("海洋プラスチック","プラスチックごみ","プラスチック", "マイクロプラスチック"), "生物多様性" = c("生物多様性"), "水産資源" = c("水産資源","漁業資源","水産物"), "地球温暖化" = c("温暖化","二酸化炭素","温室効果ガス","CO2","低炭素"), "震災" = c("地震","津波","防災","復興","被災","災害","大震災"), "再可エネルギー" = c("再生可能エネルギー","風力発電","風車"), "北極域" = c("北極","北極圏")))

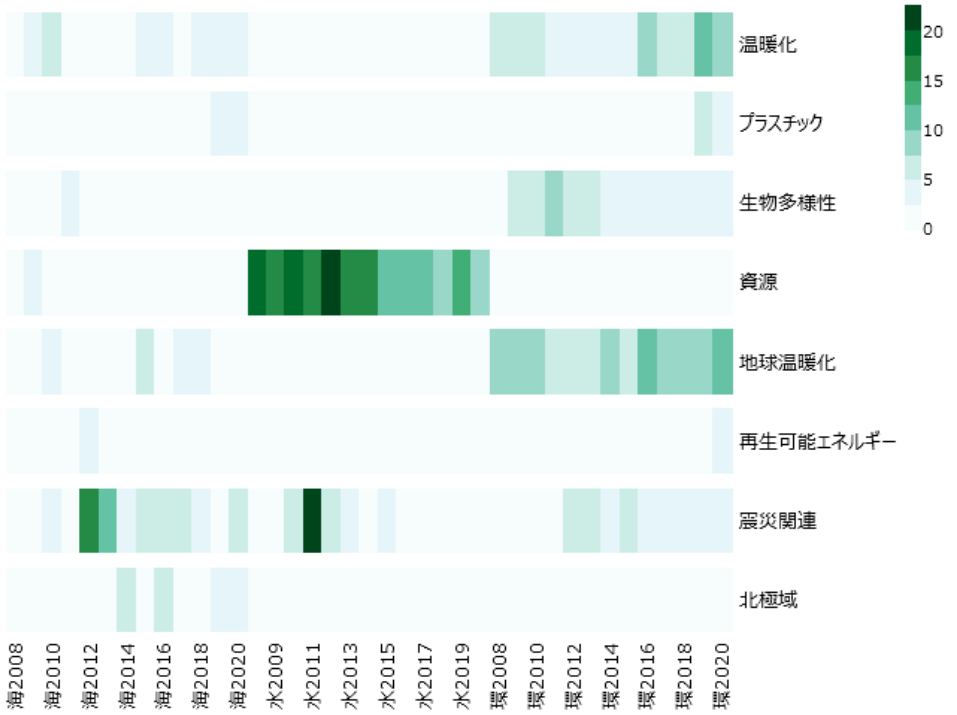
表 2-5　8分類クロス集計用 KH Coder コーディングファイル

ファイル名：OPRI_keywords_8.txt

```
1  ＊温暖化
2  気候変動 or 温暖化 or 地球温暖化
3
4  ＊プラスチック
5  海洋プラスチック  or プラスチックごみ or プラスチック or マイクロプラスチック
6
7  ＊生物多様性
8  生物多様性 or 海洋生物多様性 or 海洋生態系
9
10 ＊資源
11 水産資源 or 漁業資源 or 水産物
12
13 ＊地球温暖化
14 温暖化  or "二酸化炭素 or 温室効果ガス or CO2 or 低炭素
15
16 ＊再生可能エネルギー
17 再生可能エネルギー or 洋上風力発電 or 風車
18
19 ＊震災関連
20 地震 or 津波 or 防災 or 復興 or 被災 or 災害 or 大震災
21
22 ＊北極域
23 北極 or 北極圏
24 [EOF]
```

	＊温暖化	＊プラスチック	＊生物多様性	＊資源	＊地球温暖化	＊再可エネルギー	＊震災関連	＊北極域	ケース数
海2008	11 (0.91%)	0 (0.00%)	25 (2.08%)	18 (1.50%)	10 (0.83%)	2 (0.17%)	12 (1.00%)	0 (0.00%)	1204
海2009	40 (3.26%)	0 (0.00%)	12 (0.98%)	32 (2.61%)	26 (2.12%)	0 (0.00%)	17 (1.38%)	25 (2.04%)	1228
海2010	69 (7.01%)	1 (0.10%)	4 (0.41%)	6 (0.61%)	32 (3.25%)	5 (0.51%)	27 (2.74%)	2 (0.20%)	985
海2011	9 (0.58%)	6 (0.38%)	78 (4.98%)	26 (1.66%)	9 (0.58%)	10 (0.64%)	8 (0.51%)	1 (0.06%)	1565
海2012	17 (0.95%)	0 (0.00%)	4 (0.22%)	19 (1.06%)	9 (0.50%)	59 (3.30%)	279 (15.62%)	37 (2.07%)	1786
海2013	22 (1.27%)	0 (0.00%)	26 (1.50%)	12 (0.69%)	15 (0.86%)	42 (2.42%)	208 (11.97%)	43 (2.48%)	1737
海2014	37 (2.39%)	0 (0.00%)	2 (0.13%)	27 (1.74%)	28 (1.81%)	31 (2.00%)	65 (4.20%)	92 (5.94%)	1549
海2015	63 (4.43%)	0 (0.00%)	3 (0.21%)	24 (1.69%)	83 (5.83%)	21 (1.48%)	99 (6.96%)	12 (0.84%)	1423
海2016	59 (3.97%)	0 (0.00%)	22 (1.48%)	20 (1.34%)	13 (0.87%)	30 (2.02%)	85 (5.71%)	106 (7.12%	1488
海2017	41 (2.28%)	13 (0.72%)	38 (2.11%)	29 (1.61%)	60 (3.33%)	17 (0.94%)	97 (5.39%)	34 (1.89%	1801
海2018	72 (4.17%)	14 (0.81%)	33 (1.91%)	26 (1.51%)	45 (2.61%)	29 (1.68%)	68 (3.94%)	35 (2.03%	1727
海2019	46 (3.49%)	66 (5.00%)	29 (2.20%)	18 (1.36%)	30 (2.27%)	23 (1.74%)	16 (1.21%)	50 (3.79%	1319
海2020	67 (4.02%)	61 (3.66%)	34 (2.04%)	15 (0.90%)	24 (1.44%)	7 (0.42%)	85 (5.10%)	62 (3.72%	1667
水2008	10 (1.31%)	0 (0.00%)	2 (0.26%)	138 (18.11%	11 (1.44%)	0 (0.00%)	6 (0.79%)	1 (0.13%)	762
水2009	3 (0.57%)	0 (0.00%)	4 (0.76%)	84 (15.88%)	3 (0.57%)	0 (0.00%)	3 (0.57%)	1 (0.19%)	529
水2010	2 (0.28%)	0 (0.00%)	6 (0.85%)	128 (18.05%	2 (0.28%)	0 (0.00%)	36 (5.08%)	0 (0.00%)	709
水2011	2 (0.20%)	0 (0.00%)	1 (0.10%)	151 (15.38%	2 (0.20%)	0 (0.00%)	223 (22.71%)	0 (0.00%)	982
水2012	1 (0.08%)	0 (0.00%)	0 (0.00%)	254 (20.53%	2 (0.16%)	8 (0.65%)	76 (6.14%)	0 (0.00%)	1237
水2013	7 (0.47%)	0 (0.00%)	0 (0.00%)	228 (15.37%	4 (0.27%)	6 (0.40%)	57 (3.84%)	0 (0.00%)	1483
水2014	12 (0.82%)	1 (0.07%)	2 (0.14%)	237 (16.17%	12 (0.82%)	2 (0.14%)	34 (2.32%)	1 (0.07%)	1466
水2015	7 (0.49%)	4 (0.28%)	3 (0.21%)	164 (11.47%	5 (0.35%)	0 (0.00%)	40 (2.80%)	0 (0.00%)	1430
水2016	12 (0.76%)	7 (0.44%)	9 (0.57%)	173 (10.89%	4 (0.25%)	0 (0.00%)	35 (2.20%)	0 (0.00%)	1589
水2017	13 (1.09%)	7 (0.58%)	2 (0.17%)	139 (11.60%	6 (0.50%)	0 (0.00%)	25 (2.09%)	0 (0.00%)	1198
水2018	15 (1.11%)	7 (0.52%)	4 (0.30%)	136 (10.06%	6 (0.44%)	0 (0.00%)	20 (1.48%)	0 (0.00%)	1352
水2019	16 (1.18%)	13 (0.96%)	5 (0.37%)	199 (14.69%	7 (0.52%)	0 (0.00%)	29 (2.14%)	2 (0.15%)	1355
水2020	15 (1.06%)	7 (0.50%)	4 (0.28%)	136 (9.65%)	6 (0.43%)	0 (0.00%)	21 (1.49%)	0 (0.00%)	1410
環2008	221 (6.29%)	31 (0.88%)	37 (1.05%)	6 (0.17%)	329 (9.37%)	23 (0.65%)	20 (0.57%)	0 (0.00%)	3512
環2009	213 (5.64%)	35 (0.93%)	229 (6.06%	11 (0.29%)	343 (9.08%)	25 (0.66%)	22 (0.58%)	0 (0.00%)	3778
環2010	280 (6.88%)	24 (0.59%)	278 (6.83%	16 (0.39%)	342 (8.41%)	29 (0.71%)	18 (0.44%)	15 (0.37%	4069
環2011	161 (4.12%)	42 (1.08%)	301 (7.71%	15 (0.38%)	245 (6.28%)	19 (0.49%)	63 (1.61%)	0 (0.00%)	3904
環2012	151 (3.87%)	18 (0.46%)	217 (5.57%	16 (0.41%)	241 (6.18%)	63 (1.62%)	222 (5.69%)	2 (0.05%)	3899
環2013	173 (4.88%)	16 (0.45%)	190 (5.36%	10 (0.28%)	249 (7.02%)	37 (1.04%)	212 (5.98%)	4 (0.11%)	3545
環2014	191 (4.66%)	13 (0.32%)	169 (4.12%	9 (0.22%)	344 (8.39%)	67 (1.63%)	163 (3.98%)	6 (0.15%)	4099
環2015	158 (4.13%)	11 (0.29%)	138 (3.61%	10 (0.26%)	280 (7.33%)	74 (1.94%)	204 (5.34%)	2 (0.05%)	3822
環2016	289 (8.07%)	12 (0.34%)	92 (2.57%)	10 (0.28%)	428 (11.96%)	49 (1.37%)	129 (3.60%)	2 (0.06%)	3580
環2017	230 (6.68%)	17 (0.49%)	94 (2.73%)	11 (0.32%)	330 (9.59%)	84 (2.44%)	119 (3.46%)	2 (0.06%)	3442
環2018	152 (5.27%)	28 (0.97%)	87 (3.02%)	17 (0.59%)	278 (9.64%)	69 (2.39%)	89 (3.09%)	1 (0.03%)	2883
環2019	333 (10.66%)	184 (5.89%)	109 (3.49%	9 (0.29%)	312 (9.98%)	61 (1.95%)	109 (3.49%)	4 (0.13%)	3125
環2020	335 (8.89%)	126 (3.34%)	158 (4.19%	13 (0.35%)	403 (10.70%)	107 (2.84%)	140 (3.72%)	6 (0.16%)	3767
合計	3555 (4.31%	764 (0.93%)	2451 (2.97%	2592 (3.15%	4578 (5.56%)	999 (1.21%)	3181 (3.86%	548 (0.67%	82406
カイ2乗値	1507.196**	1723.006**	1524.547**	7983.803**	2466.344**	566.761**	2751.184**	2643.776**	

図 2-2 8 分類クロス集計結果（出現頻度）



（注）x 軸のラベル表記は隔年。 数値は図 2-2 中の%。

図 2-3 8 分類クロス集計ヒートマップ

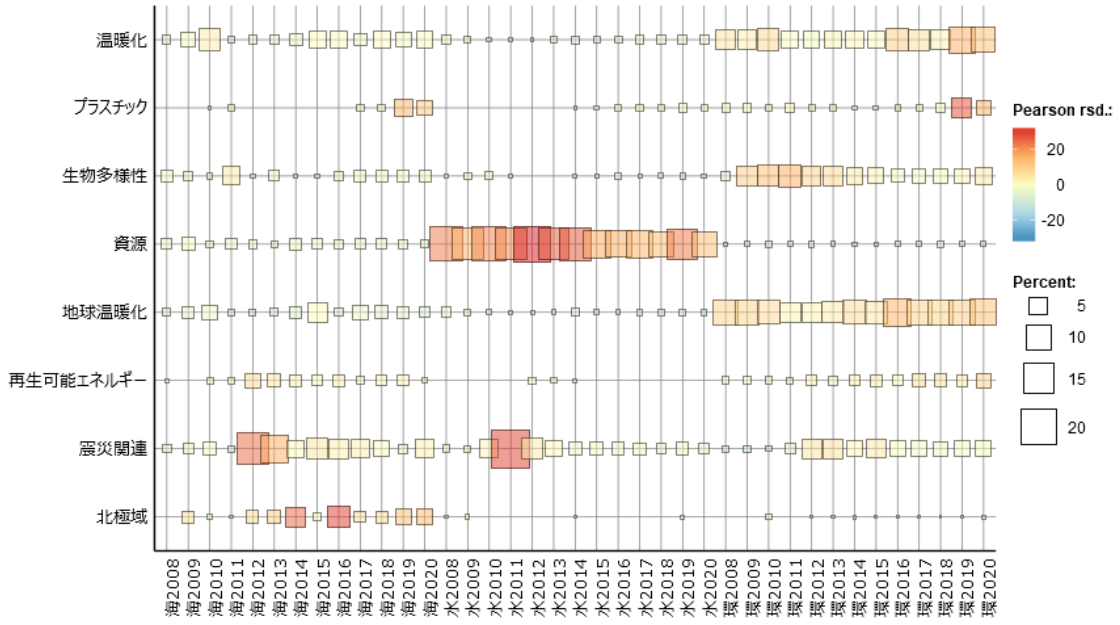


図 2-4 8 分類クロス集計バブルプロット

3. 分析ツールについて

本業務ではフリーツールの KH Coder を利用し、形態素解析から今日にネットワーク分析、対応分析、LDA トピック分析を行い、その後 R-Studio で R 言語の seededlda パッケージで教師付き LDA 分析を実施した。KH Coder では、形態素解析結果をデータベースに登録しており、源泉テキスト文との対比（KWCI コンコーダンス）機能による精査作業、強制抽出語や抽出対象からの除外（削除語）設定が容易で、分析対象語の絞り込みの試行錯誤を短時間で効率的に行える。

年次白書と云う文書の性格上、記載内容や記述様式が各年度で踏襲されていることがあり、テキストマイニングの手法解説で紹介される小説やニュース記事の分析とは少々異なる視点での分析が求められる。実際、環境・水産・海洋白書（2008～2020）について、集計単位を年、分析対象語数：75、トピック数:20 で KH Coder のトピック分析（LDA）を行い、KWIC コンコーダンス機能を利用してトピックの特徴語の源泉文章を確認すると、同じ文章が異なる年次白書中にあり、しかも複数のトピックに出現した。その結果を図 3-1 に示す。踏襲されている定型的な文章は、集計単位を文として KH Coder のトピック分析（LDA）を行うと検出できる。その結果を図 3-2 に示す。特に、水産白書施策編では、3～8 年にわたり使用された文章が多数ある。なお、これらの年次白書で共通な文章が及ぼす影響については、本作業では調査していない。

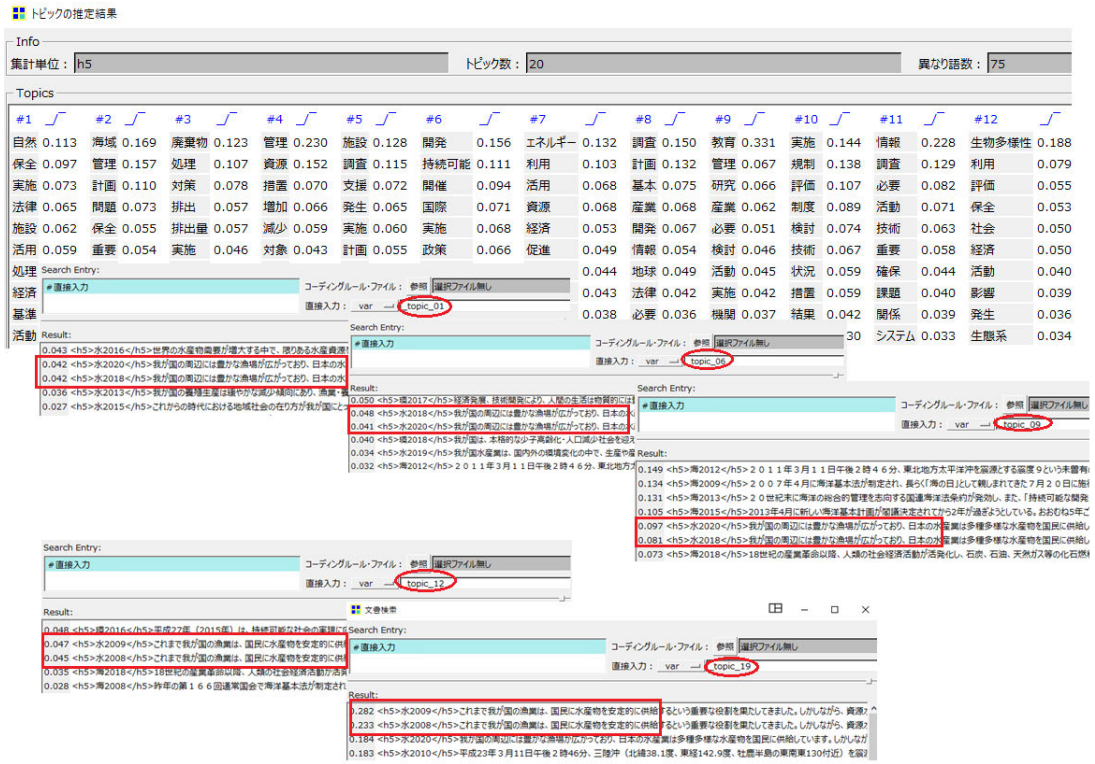


図 3-1 LDA 分析結果の特徴語の源泉文章で確認された定型文

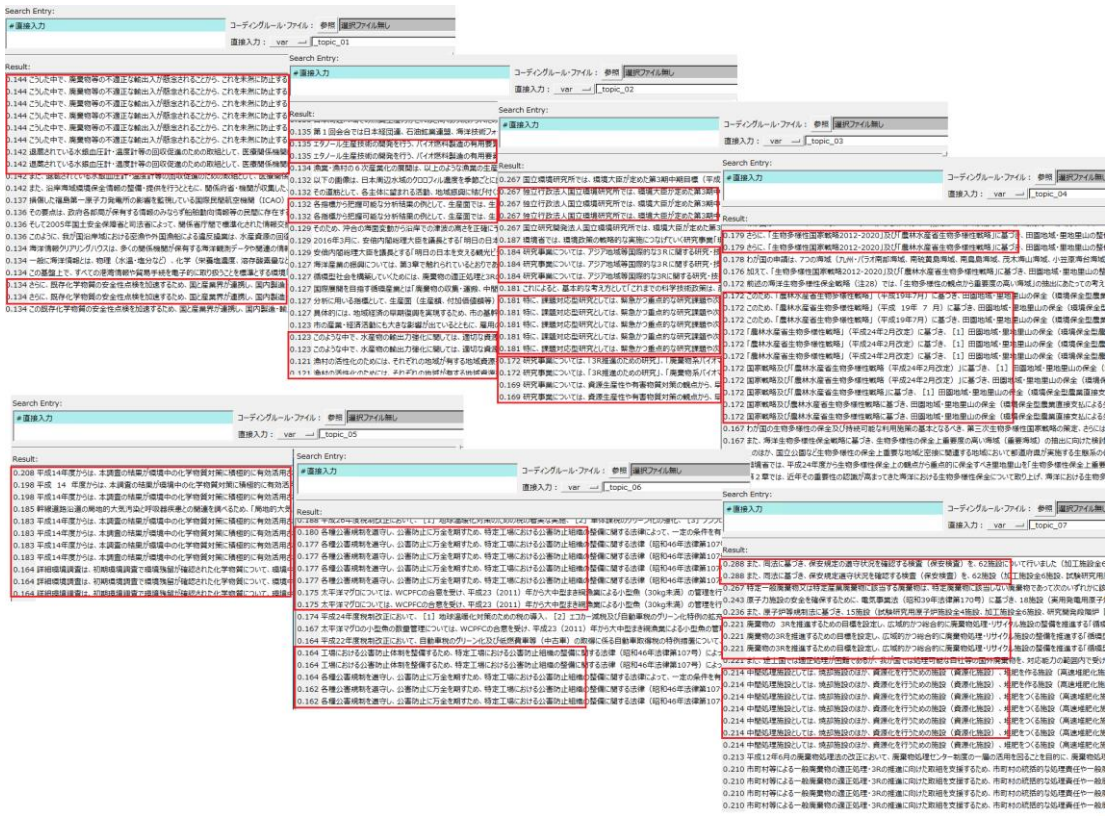


図 3-2 白書で踏襲されていた定型的な文章

Gephi は大きなネットワークを分析する場合に便利なツールであり、ネットワークの概要情報（平均次数、グラフ密度、モジュラリティ、固有ベクトル中心性など）を算出する機能は、KH Coder には無い機能であるから、その有効性を調査した。

年次白書の文章から形態素解析により抽出された単語について、他の文章中で出現する頻度の高い単語を KH Coder の共起ネットワーク図により、語・語間の共起状況とクラスタ状況を可視化する（図 3-3）。デフォルトの「サブグラフ検出（modularity）」で描画し、pajek（拡張子.net）で保存し、ネットワーク分析ツール gephi で ForceAtlas2 レイアウト表示すると、固定ノードサイズだが文字サイズ可変のグラフが表示される（図 3-4）。本作業の初期段階でネットワーク分析ツールとして検討したが、可視化操作が複雑であるため実作業では使用しなかった。

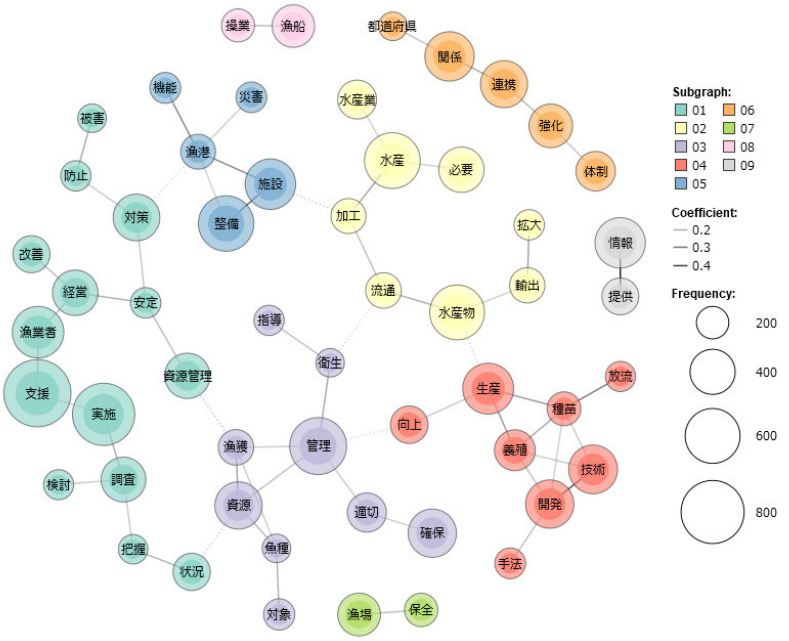


図 3-3 KH Coder での共起ネットワーク図

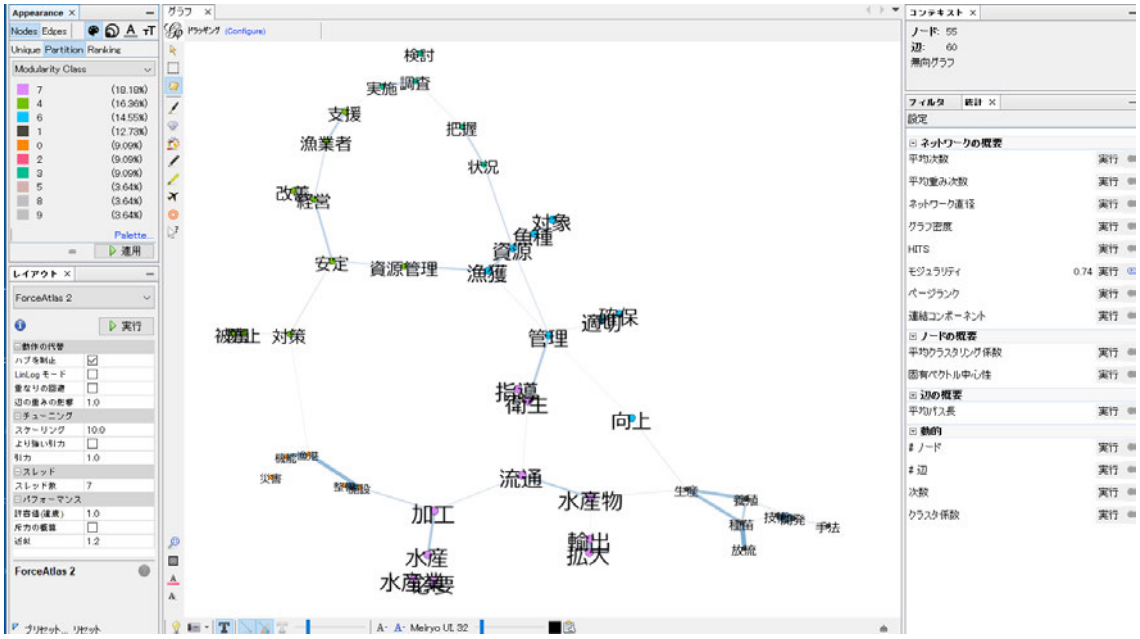


図 3-4 共起ネットワークを gephi で可視化

KH Coder のトピック推定では、抽出語の頻度行列（DTM:Document Term Matrix）を topicmodels パッケージの LDA への入力データとしている。これを R 言語の wordcloud2 パッケージで可視化することもできる（図 3-5）。但し、描画結果の再現性（同じ図を描画）がないので、本業務では使用しなかった。

海洋白書 2019【75語】

図 3-5 環境・水産・海洋白書（2008～2020 年）の抽出語結果（75 語）
から海洋白書 2019 年を wordcloud2 で可視化

KH Coder の「トピック数の探索（ldatuning）」の評価パラメータを調査した結果、topicmodels パッケージの LDA 処理結果にテキストモデルの評価パラメータ（対数尤度）が格納されており、サンプリング過程での格納方法が判明した。（図 3-6）

> **loglikelihood:** Object of class "numeric"; the log-likelihood of each document given the parameters for the topic distribution and for the word distribution of each topic is approximated using the variational parameters and underestimates the log-likelihood by the Kullback-Leibler divergence between the variational posterior probability and the true posterior probability.

図 3-6 topicmodels パッケージ マニュアルでの記載

そこで環境・水産・海洋白書（2008～2020 年）の抽出語結果（75 語）の DTM データを用いて、KH Coder での LDA 分析で設定したトピック数 16 での Gibbs サンプリング数の範囲を 2,000～100,000 として対数尤度（赤）と Gibbs サンプリング過程での対数尤度の調和平均（青）とし、各近似曲線も併せて可視化した。その結果を図 3-7 に示す。この図では、30,000 ステップまでは評価が悪化し、50,000～60,000 ステップ付近で安定化した後やや悪化し、80,000 ステップ以降では改善傾向がみられる。

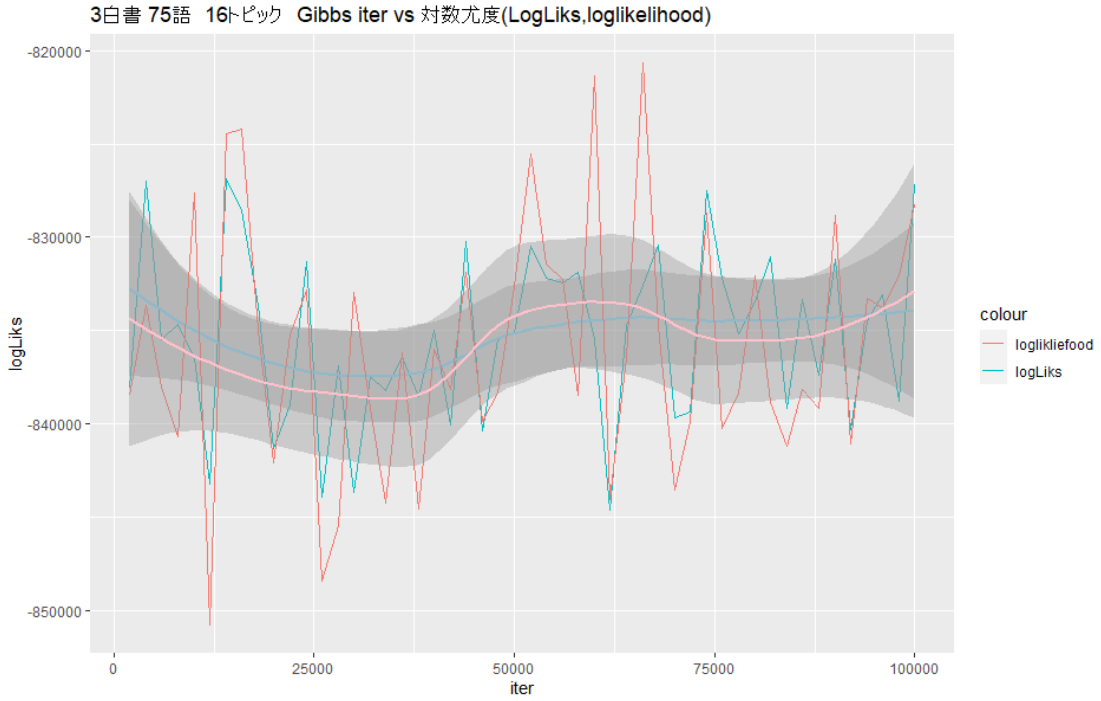


図 3-7 抽出語数 75、トピック数 16 での Gibbs サンプル数と
LDA トピックモデル評価パラメータの関係

「2.3.1 KH Coder による分析作業」での報告書で示したように、抽出語数が 75 の場合は Gibbs サンプリング数：10,000 で各トピックの特徴語が安定化していた。そこで、トピック数を 2～40 の範囲で対数尤度（赤）と Gibbs サンプリング過程での対数尤度の調和平均（青）とし、各近似曲線も併せて可視化した。その結果を図 3-8 に示す。この図では、トピック数の増加に対して評価パラメータ値の改善傾向が続き、トピック数 40 でも安定化していない。KH Code には「トピック数の探索」では評価パラメータ（perplexity）を選択できるので、トピック数を 2～70 と設定して実行した。R 言語の ldatuning パッケージのコードでは、topicmodels:LDA のデフォルト設定からパラメータを変更していないので、VEM でサンプリング数 2000 と推察される。その結果を図 3-9 に示す。この図からは、トピック数は 16 付近と判断される。

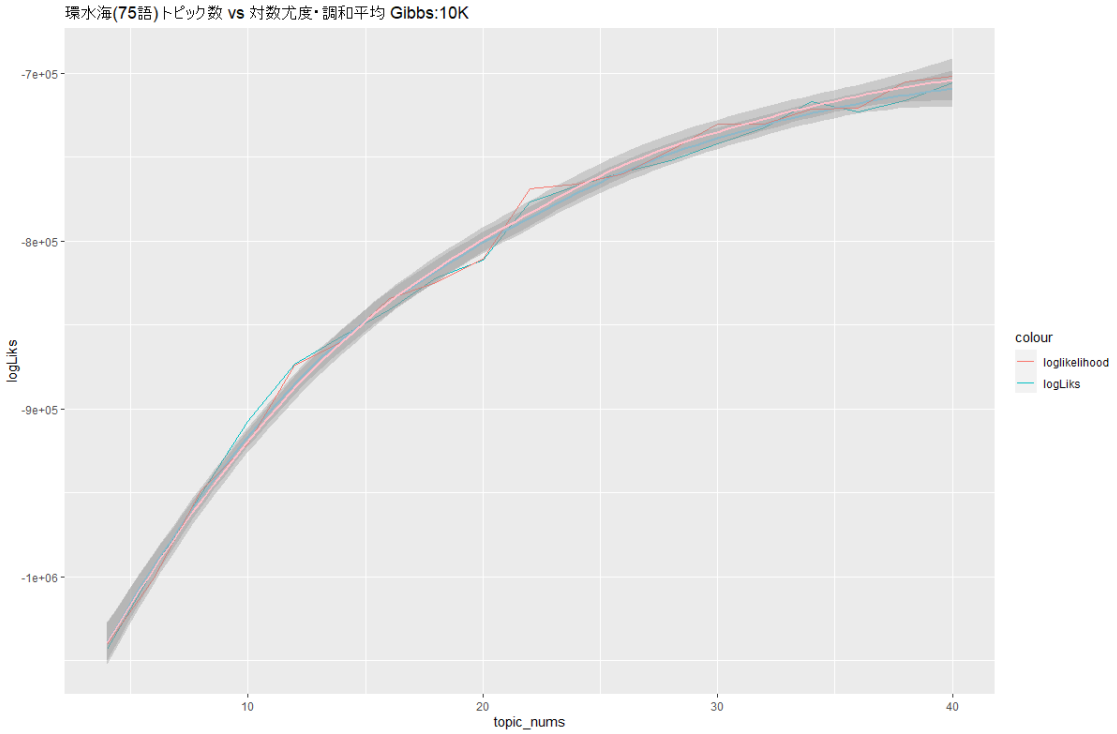


図 3-8 抽出語数 75、Gibbs サンプル数 10,000 でのトピック数とLDA トピックモデル評価パラメータの関係

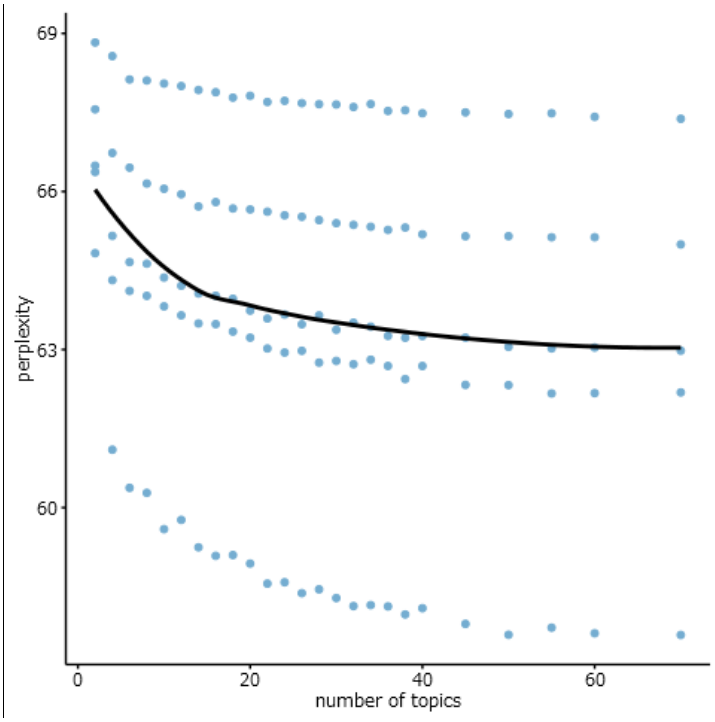


図 3-9 抽出語数 75 でのトピック数と perplexity の関係

4. 今後の展望

本業務初期段階の手法の検討・調査で日本語テキストの形態素解析ツールでは辞書データが重要で、形態素解析結果がその後の分析や解釈に大きく影響することを確認していた。また、形態素解析ツールを含めテキストマイニングと可視化処理環境の選択肢としては Python と R 言語があり、インターネット上に公開されているテキストマイニング処理関連情報の多くは Python であった。しかし、Python での処理結果の可視化には描画処理コードを matplotlib で、R でも ggplot2 で作成する必要があり、分析過程での処理手順の試行錯誤への課題が判明した。テキストマイニングツール KH Coder の紹介記事と開発者の書籍「社会調査のための計量テキスト分析」を参考に、水産白書施策編について形態素解析から、共起ネットワーク分析、対応分析、トピック分析、処理結果の可視化といった一連の作業が KH Coder を利用して効率的に実施できることを確認し、一連の作業手順を確定して各白書の分析処理を行った。この作業手順を他の白書に適用することで、共通的な品質での分析が期待できる。

本業務では pdf 文書からの本文テキスト抽出手順ツールを整備したので、pdf 文書が公開されている「国土交通白書」や「エネルギー白書」については、本業務での作業手順を踏襲して各白書本文のテキスト化が可能である。しかし、「海上保安白書」は pdf 文書が公開されておらず、web ブラウザでの閲覧に制限されているので、階層化された HTML ページからのテキスト抽出作業は容易ではない。白書の年次単位でのテキスト化作業が終われば、本業務で確定した KH Coder を利用した作業手順での分析作業を実施できる。

KH Coder の処理では、形態素解析は Chasen、トピック分析や可視化処理には R 言語のパッケージを利用しており、GitHub で処理コードが公開されている。KH Coder では、入力されたテキストデータや形態素解析結果である抽出語をデータベース（MySQL）へ登録し、Perl 言語で記述された処理結果の可視化画面でのインタラクティブな操作を実現しており、プログラム設計書資料が未公開な状況下では、処理コードの内容確認は容易ではない。また、KH Coder での処理に使用している R 言語は 3.1 版と古く、CRAN での保守対象外となっているが、トピックの推定で使用されている topicmodels パッケージの処理結果については R 言語の現行バージョン（3.6.3、4.1）との整合性を確認した。しかし、KH Coder では、Gibbs サンプリング数：2,000 に固定されており、分析対象語やトピック数が大きい場合に Gibbs サンプリング数を変更設定できないことが問題である。

本作業での成果として、KH Coder の形態素解析結果である DTM（Document Term Matrix）データを最新版 R 言語の各種テキストマイニング用パッケージで処理する手順を確定した。したがって、KH Coder の「トピックの推定」での Gibbs サンプリング数を変更した処理結果は、R-Srtudio で topicmodels パッケージを実行すれば得られるが、KWIC コンコーダンスによるトピック特徴語を含む文章の検索ができないので、特徴語からのトピック類推・判読や処理結果の吟味作業の非効率化が懸念される。その代替えとしては、R 言語環境下での LDA トピック分析結果の可視化ツール（LDAvis）を利用すれば、特徴語からのトピック類推・判読作業や処理結果の吟味・精査作業に有効である。また、LDA 処理結果のトピックの特徴語リストから排他的構成のコーディングルールを作成し、KH Coder のクロス集計機能により文書毎のトピック出現率を可視化し（図 2-2～2-4 参照）、更に共起ネットワーク分析を行い、モデル処理結果の判読・吟味を行う。

最適トピック数の推定のための LDA トピックモデル評価パラメータについては、「3. 分析ツールについて」で報告したが、マルコフ連鎖モンテカルロ法である Gibbs サンプリング法を使用した topicmodels パッケージでの LDA 処理では、安定的な処理結果を得るには Gibbs サンプリング数として十分に大きな値を設定する必要がある。しかも最適トピック数を決定するためには、反復して LDA 処理を実行するので、膨大な計算機リソースを要することが問題である。本業務では ldatuning パッケージの処理結果図から目視でトピック数を判定したが、topicmodels パッケージの処理結果からトピックモデルの評価パラメータ（logliklihood、perplexity、coherence 等）の定量的評価により、トピック数と Gibbs サンプリング数を決定する手法を確定する。

形態素解析ツールの辞書データが重要であることは手法検討段階で判明しており、辞書データが古い Mecab は使用せず、Python 環境で janome または spacy/Ginza を使用する予定であった。Windows 環境下では R 言語から Mecab を使用する予定であったので、KH Coder で使用されている Chasen は検討対象外であった。よって、Chasen による形態素解析の性能確認は未着手である。

本業務で使用した R 言語のパッケージよりも多くの使用実績が報告されている Python のテキストマイニング・パッケージ（gensim、sickit-learn、GuidedLDA）との比較も肝要である。

また、本業務で使用した KH Coder には、「ベイズ学習による分類」が準備されており、排他的な分類ないしはカテゴリ分けに利用できる。学習した分類基準は，分類「見本」から自動的に生成した「ある種のコーディングルール」と KH Coder のマニュアルに記されており、コーディングルールによるクロス集計結果と同様、各白書間での関心事項・重要課題の把握への活用が期待できる。

年次白書には全く同じ文章が数年にわたり記載されている（図 3-1、3-2 参照）が、その文章を除去した際の分析結果への影響については未検討である。異なる白書間での省庁横断的な関心事項の経年変化の状況把握には影響しないと予想されるが、同一白書での共起ネットワーク分析や関心事項の経年変化の判読には影響する可能性がある。

5. 参考文献と参照資料

1) 社会調査のための計量テキスト分析 樋口耕一
2) R によるテキストマイニング 石田基広
3) トピックモデルによる統計的潜在意味解析 奥村学
4) 対応分析入門 原理から応用まで 藤本一男（訳）
5) Pajek を活用した社会ネットワーク分析 安田由紀（訳）
6) 確率的トピックモデル 統計数理研究所 H24 年度公開講座
7) https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf
8) https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf
9) https://cran.r-project.org/web/packages/seededlda/seededlda.pdf
10) https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
11) https://cran.r-project.org/web/packages/ldatuning/ldatuning.pdf

12) the website for *Text Mining with R*
https://www.tidytextmining.com/index.html

付録１　強制抽出する語の設定テキストファイル

強制抽出語（316 語）
CCSBT、CMA1、CMP、COP、D.Waste-Net、DHA、EEZ、EPA、FAO、FTA、GOTS、GSSI、IATCC、ICCAT、IGES、IOTC、IUU 漁業、IWC、MEL、MSC、NPFC、PCB、POPs 条約、RCEP、RDF、RMP、RMS、TAC、TAC 法、TPP、WCPFC、ZEB、ZEH、アジア太平洋地域、アジェンダ２１、エコツーリズム、エルニーニョ現象、オゾン層、カルタヘナ議定書、グリーン購入法、クルーズ船、クルックフィールズ、グローバル化、コンビニエンスストア、シップ・リサイクル条約、スーパーマーケット、ステークホルダー、ダイオキシン、バーゼル条約、バーゼル法、バイオマス、はえ縄、パリ協定、ヒートショック、プラスチックごみ、ブルーエコノミー、フルオロカーボン、フロン類、マイクロプラスチック、まき網、マラッカ海峡、マンガン団塊、メタンハイドレート、モントリオール議定書、リオ地球サミット、リユース、レッドデータブック、安全確保、安全性、安全保障、安定供給、安定的、委員会、意見交換、意見書、意思決定、一体的、一般局、一般的、沿岸域、沿岸海域、沿岸漁業、沿岸国、沿岸部、遠洋漁業、沖合漁業、卸売市場、温室効果ガス、温暖化、化学物質、加工業者、加工品、加盟国、可能性、科学技術、科学的、科学的知見、科学的調査、課題解決、海域管理、海事活動、海事産業、海上テロ、海上交通、海上輸送、海賊行為、海底ゴミ、海底資源、海底熱水鉱床、海氷、海面上昇、海洋ゴミ、海洋プラスチック、海洋汚染、海洋研究開発機構、海洋政策研究財団、外国漁船、外来種、各国政府、確信度、閣議決定、学校給食、学校教育、活性化、観測データ、基準値、基本的、気候変動、記録的、技術開発、漁獲量、漁業権、漁業資源、漁業者、魚介類、魚種、競争力、近隣諸国、具体化、具体的、経済産業省、経済的、経済発展、継続的、計画的、鯨類、研究開発、研究機関、研究者、顕在化、減少傾向、効果的、効率的、好循環、高度化、高齢化、合理的、国際会議、国際機関、国際的、国際法、国土交通省、国内法、国民生活、国立公園、国連海洋法条約、再生可能エネルギー、最終処分場、最終的、災害廃棄物、削減目標、笹川平和財団海洋政策研究所、産業廃棄物、酸性化、資源開発、資源管理、資源循環、資源評価、資源量、事業活動、事業者、事例、持続可能、持続的、自主的、自然環境、自然災害、自排局、実効性、実施計画、実用化、社会的、主体的、取組事例、取締船、種苗、周辺海域、集中的、重要性、循環型、巡視船、諸外国、諸問題、商業化、消費者、省エネ、浄化槽、食品リサイクル、食品ロス、新型コロナウイルス、深海底、人工衛星、人材育成、水環境、水産エコラベル、水産業、水産高校、水産資源、水産大学校、世界的、生産者、生産量、生態系、生物資源、生物多様性、西日本、積極的、接続水域、先進国、先進的、先進的事例、専門家、戦略的、全国的、総合的、藻場、多様化、多様性、太平洋島嶼国、代表的、大気汚染、大規模、大型化、大部分、脱炭素、地下水、地球温暖化、地球環境、地方公共団体、中国海軍、中国側、調査研究、長期的、低炭素、低潮線、定期的、定置網、締約国、適応策、天然ガス、天然資源、伝統的、電子レンジ、途上国、島嶼国、東京 2020 大会、東電福島第一原発、東日本、東日本

大震災、統合的、内閣総理大臣、内水面、日本近海、日本財団、日本政府、日本籍船、農林水産業、農林水産大臣、廃棄物、排出削減、排出抑制、排出量、排他的経済水域、被災地、必要性、付加価値、普及啓発、浮体式、賦存、風力発電、文部科学省、閉鎖性海域、米軍、保護区、包括的、報告書、放射性セシウム、放射性物質、方向性、法制度、法整備、本格的、民間企業、明確化、問題点、有機農業、洋上風力、養殖技術、養殖業、養殖業者、利活用、利用者、陸域、流通業、領有権、令和

付録 2 強制抽出する語の設定テキストファイル

除外語（54 語）
３月、こと、コラム、さまざま、はじめ、ほか、わが国、以降、引き続き、円滑、億円、下、我が国、海、海洋、各国、各地、環境、漁業、魚、近年、月、現在、国、今後、施策、事業、事業、事例、時点、取組、取組事例、食、新た、図、推進、水、税、先進的事例、前年、全国、全体、地域、注、等、特に、年、年度、表、平成、様々、令和

付録 3「水産白書施策編（2007～2020 年）の分析結果」

文書番号：JRDN-21-023

1. 前処理結果

Chanse での前処理の結果、総抽出語数：160414、異なり語数：3341 のうち 2578 語が分析処理で使用された。抽出語出現数の頻度分布を図 1-1、抽出語リスト（上位 200 語）を図 1 -2 に示す。

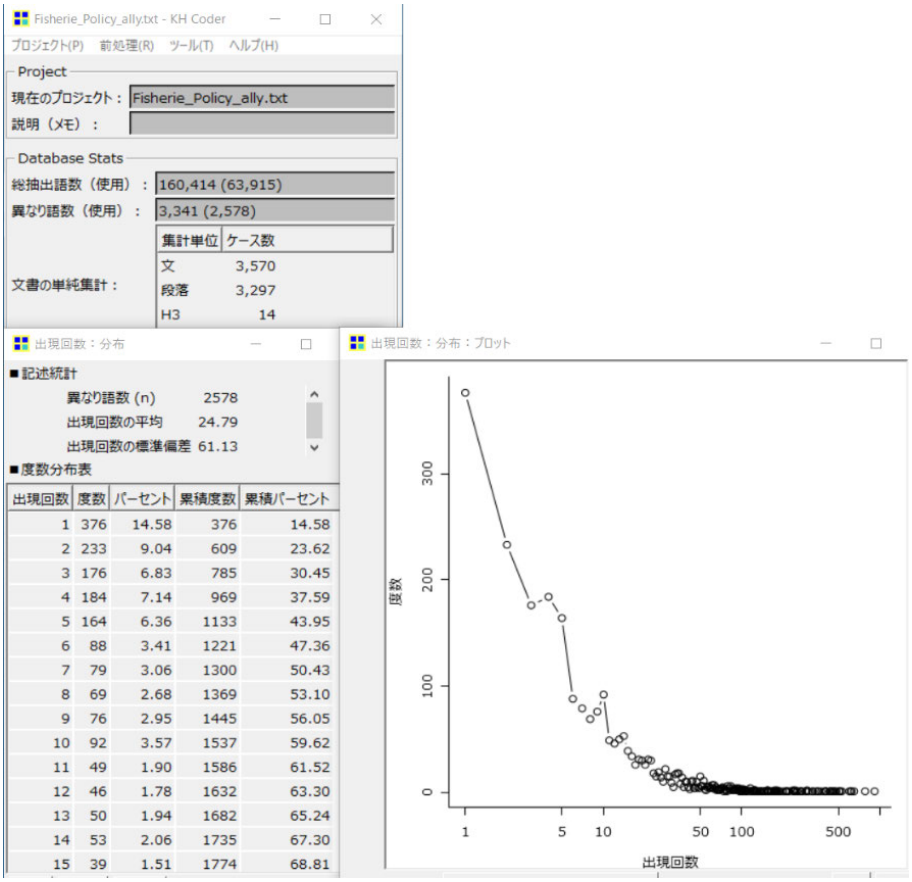


図 1-1 抽出語出現数の頻度分布

#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度
1	支援	サ変名詞	912	36	改善	サ変名詞	266				
2	実施	サ変名詞	779	37	提供	サ変名詞	263				
3	管理	サ変名詞	648	38	食品	名詞	252				
4	水産	名詞	621	39	安全	形容動詞	250				
5	整備	サ変名詞	615	40	漁獲	サ変名詞	243				
6	水産物	名詞	598	41	流通	サ変名詞	243	71	魚種	タグ	154
7	漁業者	タグ	529	42	漁村	名詞	242	72	被害	名詞	154
8	生産	サ変名詞	515	43	輸出	サ変名詞	239	73	衛生	名詞	153
9	施設	サ変名詞	507	44	加工	サ変名詞	235	74	普及	サ変名詞	153
10	情報	名詞	504	45	漁港	名詞	234	75	産地	名詞	152
11	関係	サ変名詞	482	46	導入	サ変名詞	234	76	団体	名詞	147
12	技術	名詞	472	47	種苗	タグ	219	77	維持	サ変名詞	145
13	開発	サ変名詞	462	48	保全	サ変名詞	213	78	産業	名詞	143
14	確保	サ変名詞	461	49	活動	サ変名詞	212	79	農林	名詞	139
15	促進	サ変名詞	447	50	操業	サ変名詞	208	80	養殖業	タグ	138
16	資源	名詞	440	51	消費者	タグ	207	81	政策	名詞	137
17	連携	サ変名詞	437	52	効率的	タグ	198	82	市場	名詞	135
18	対策	サ変名詞	430	53	対応	サ変名詞	188	83	生物	名詞	133
19	経営	サ変名詞	407	54	対象	名詞	186	84	規制	サ変名詞	129
20	資源管理	タグ	405	55	機能	サ変名詞	182	85	就業	サ変名詞	128
21	必要	形容動詞	404	56	資金	名詞	180	86	機関	名詞	126
22	調査	サ変名詞	394	57	手法	名詞	180	87	回復	サ変名詞	125
23	強化	サ変名詞	374	58	災害	名詞	179	88	広域	名詞	125
24	漁場	名詞	358	59	安定	形容動詞	177	89	漁協	名詞	124
25	漁船	名詞	353	60	防止	サ変名詞	175	90	復興	サ変名詞	124
26	活用	サ変名詞	342	61	拡大	サ変名詞	174	91	コスト	名詞	123
27	養殖	サ変名詞	326	62	指導	サ変名詞	173	92	影響	サ変名詞	123
28	利用	サ変名詞	324	63	放流	サ変名詞	173	93	観点	名詞	122
29	体制	名詞	304	64	水産資源	タグ	172	94	発生	サ変名詞	122
30	計画	サ変名詞	296	65	検討	サ変名詞	171	95	持続的	タグ	121
31	適切	形容動詞	296	66	法律	名詞	170	96	協議	サ変名詞	118
32	水産業	タグ	293	67	制度	名詞	165	97	効果的	タグ	116
33	措置	サ変名詞	289	68	把握	サ変名詞	164	98	育成	サ変名詞	115
34	向上	サ変名詞	273	69	評価	サ変名詞	157	99	海域	名詞	115
35	状況	名詞	269	70	都道府県	名詞	155	100	供給	サ変名詞	115

#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度
101	国際	名詞	115	136	適正	形容動詞	91				
102	実態	名詞	112	137	開催	サ変名詞	90				
103	輸入	サ変名詞	112	138	調整	サ変名詞	90				
104	内水面	タグ	109	139	東日本	タグ	90				
105	基本	名詞	108	140	消費	サ変名詞	89				
106	可能	形容動詞	107	141	協定	サ変名詞	88	171	作成	サ変名詞	77
107	被災	サ変名詞	107	142	有効	形容動詞	88	172	運用	サ変名詞	76
108	効果	名詞	106	143	価格	名詞	87	173	発揮	サ変名詞	76
109	ニーズ	名詞	104	144	科学的	タグ	87	174	ノリ	名詞	75
110	削減	サ変名詞	104	145	改革	サ変名詞	86	175	委員会	タグ	75
111	省庁	名詞	103	146	大震災	名詞	86	176	金融	名詞	75
112	組合	名詞	103	147	高い	形容詞	85	177	予測	サ変名詞	74
113	漁場	タグ	103	148	収集	サ変名詞	85	178	計画的	タグ	73
114	事業者	タグ	102	149	場合	副詞可能	85	179	充実	サ変名詞	73
115	構造	名詞	101	150	予算	名詞	85	180	積極的	タグ	73
116	重要	形容動詞	101	151	安定供給	タグ	83	181	地方	名詞	73
117	品質	名詞	101	152	関連	サ変名詞	83	182	理解	サ変名詞	73
118	ウナギ	名詞	100	153	機器	名詞	83	183	資源評価	タグ	72
119	共同	サ変名詞	99	154	研究	サ変名詞	83	184	特性	名詞	72
120	行政	名詞	99	155	実現	サ変名詞	83	185	保証	サ変名詞	72
121	協同	サ変名詞	98	156	配慮	サ変名詞	83	186	放射性物質	タグ	72
122	策定	サ変名詞	97	157	事故	名詞	82	187	沿岸漁業	タグ	71
123	設定	サ変名詞	97	158	収益	名詞	82	188	緊急	形容動詞	71
124	周辺	名詞	96	159	改良	サ変名詞	81	189	遵守	サ変名詞	71
125	省エネ	タグ	95	160	確立	サ変名詞	81	190	処理	サ変名詞	71
126	水域	名詞	95	161	早期	名詞	81	191	結果	副詞可能	70
127	日本	地名	95	162	保険	名詞	81	192	国産	名詞	70
128	変化	サ変名詞	95	163	見直し	名詞	79	193	津波	名詞	70
129	構築	サ変名詞	94	164	高度化	タグ	79	194	栄養	名詞	69
130	表示	サ変名詞	94	165	実証	サ変名詞	79	195	気象	名詞	69
131	安全性	タグ	92	166	助成	サ変名詞	79	196	再生	サ変名詞	68
132	協力	サ変名詞	92	167	浜	名詞C	79	197	赤潮	名詞	68
133	安定的	タグ	91	168	変動	サ変名詞	79	198	保存	サ変名詞	68
134	監視	サ変名詞	91	169	技術開発	タグ	78	199	啓発	サ変名詞	67
135	国民	名詞	91	170	基盤	名詞	77	200	生活	サ変名詞	67

図 1-2 抽出語リスト（上位 200 語）

2. 共起ネットワーク

KHCoder により自動設定される最小出現頻度：150、分析対象語数：75 とした場合と、最小出現頻度：83、分析対象語数：156 の結果を示す。

分析対象語数：75 での共起ネットワークの描画条件を以下に、語・語での共起ネットワークを図 2-1、語・年での共起ネットワークを図 2-2 に示す。

また、分析対象語数：156 での共起ネットワークの描画条件を以下に、語・語での共起ネットワークを図 2-3、語・年での共起ネットワークを図 2-4 に示す。

	最小出現頻度	分析対象語数	描画結果
語・語	150	75	ノード数：60　エッジ数：75
語・年	150	75	ノード数：36　エッジ数：75

	最小出現頻度	分析対象語数	描画結果
語・語	83	156	ノード数：123　エッジ数：156
語・年	83	156	ノード数：50　エッジ数：156

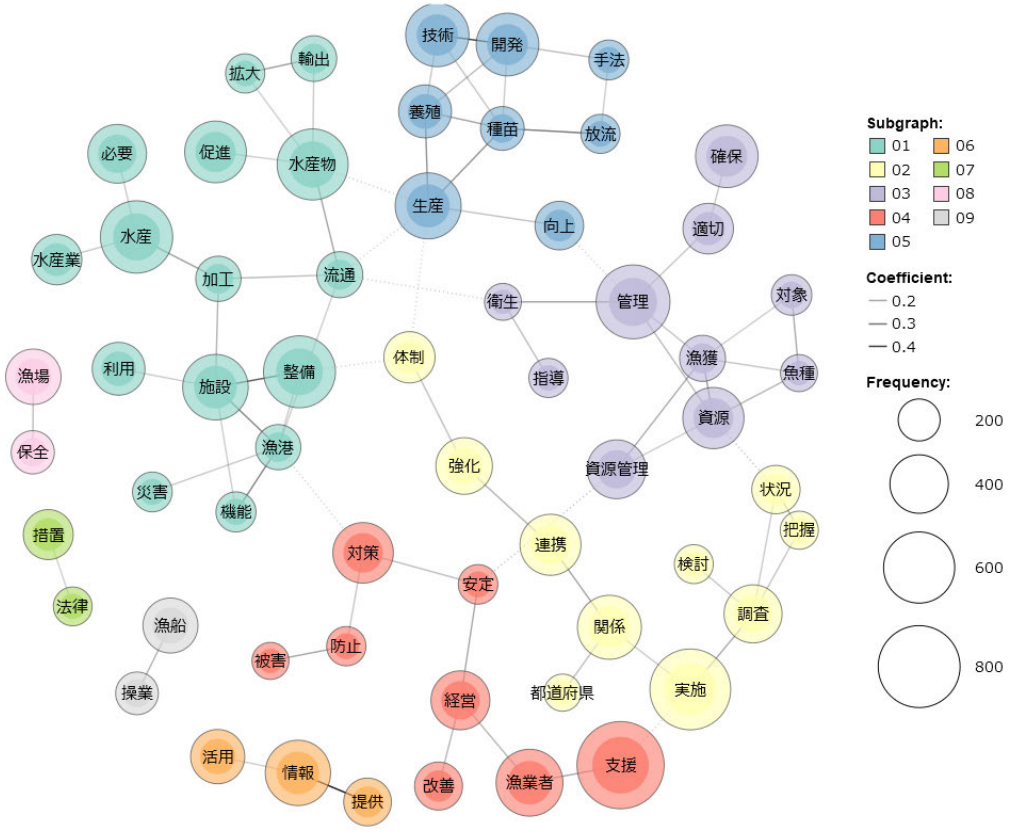


図 2-1　共起ネットワーク（語・語）（分析対象語数：75）

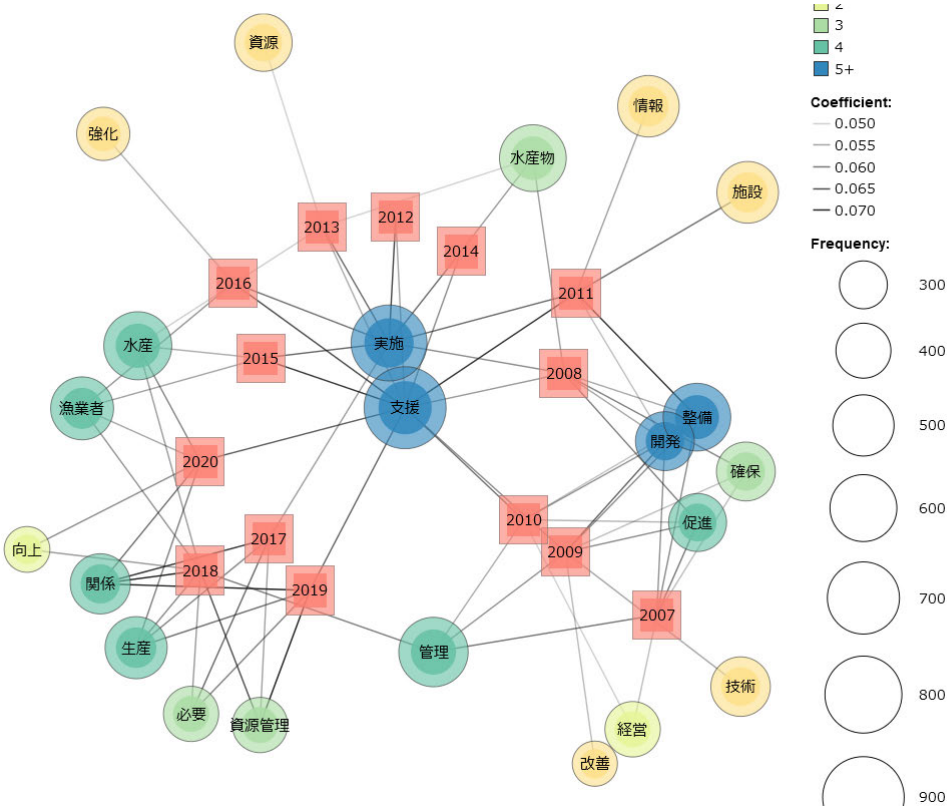


図 2-2 共起ネットワーク（語・年）（分析対象語数：75）

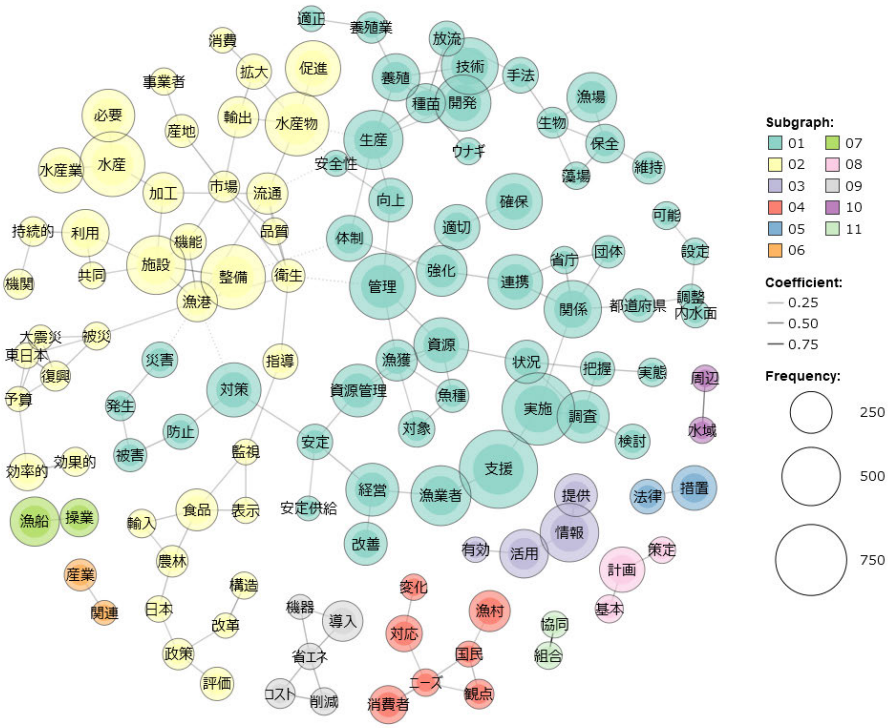


図 2-3 共起ネットワーク（語・語）（分析対象語数：156）

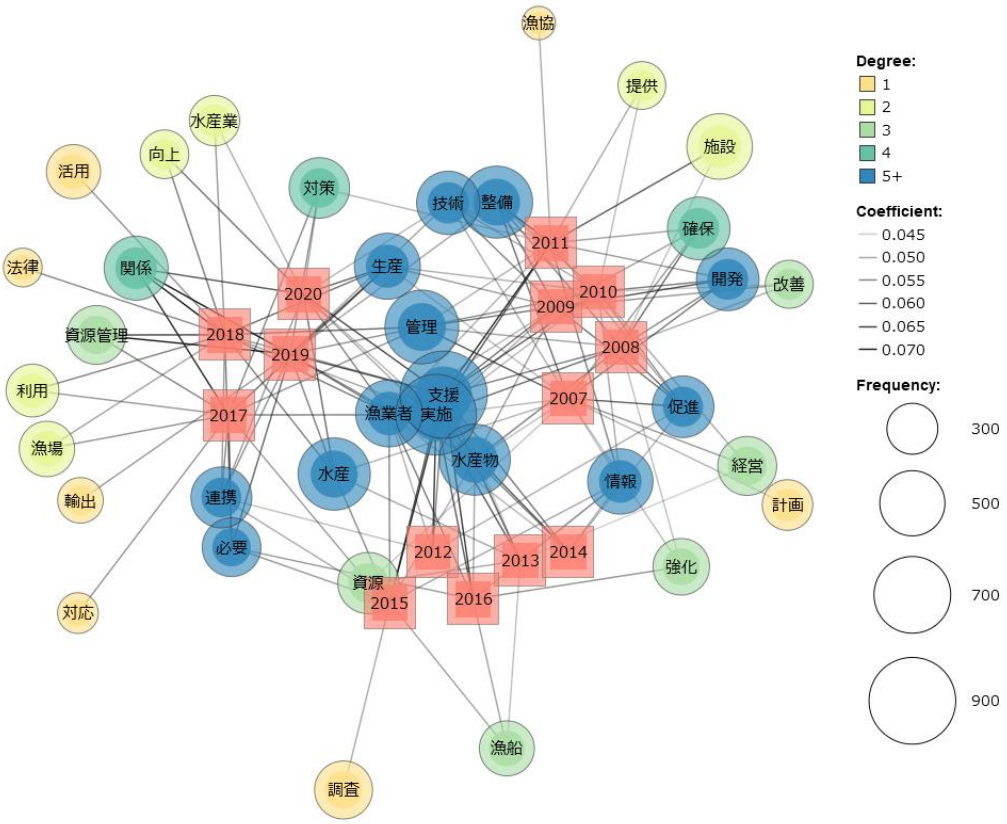


図 2-4 共起ネットワーク（語・年）（分析対象語数：156）

3. 対応分析

KHCoder により自動設定される最小出現頻度：150、分析対象語数：75 とした場合と、最小出現頻度：83、分析対象語数：156 で対応分析処理を行った。

分析対象語数：75 の累積寄与率は成分 1 と 2 で 77.41%であり、その結果を図 3-1 に示す。

分析対象語数：156 の累積寄与率は成分 1 と 2 で 72.83%であり、その結果を図 3-2 に示す。

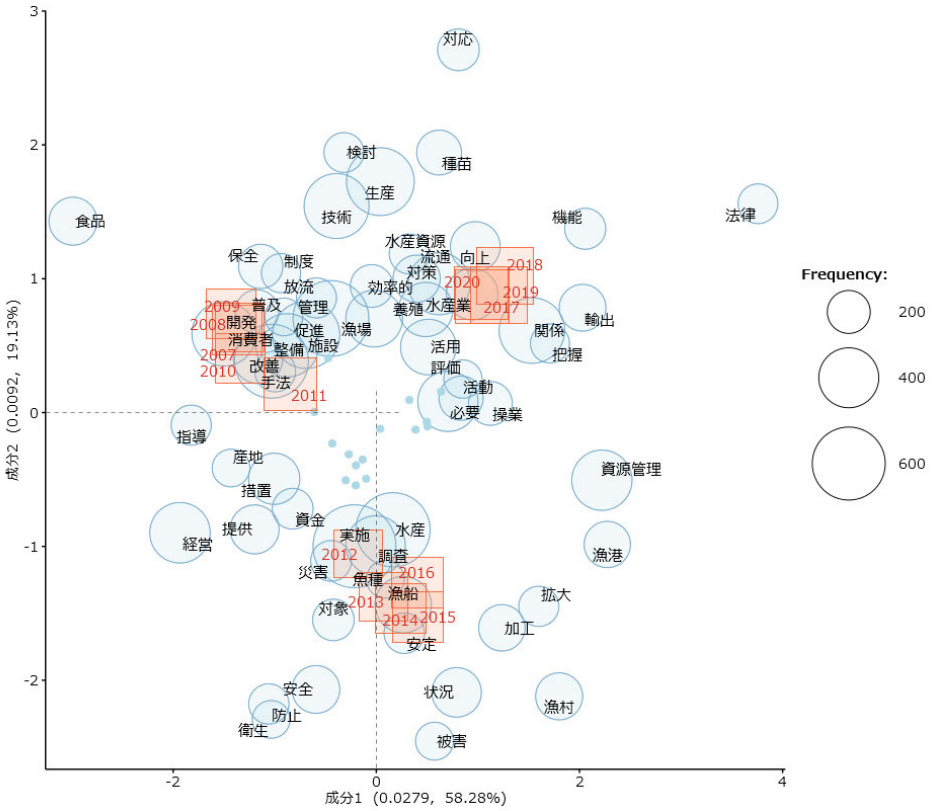


図 3-1 語･年の対応分析結果（分析対象語数：75）

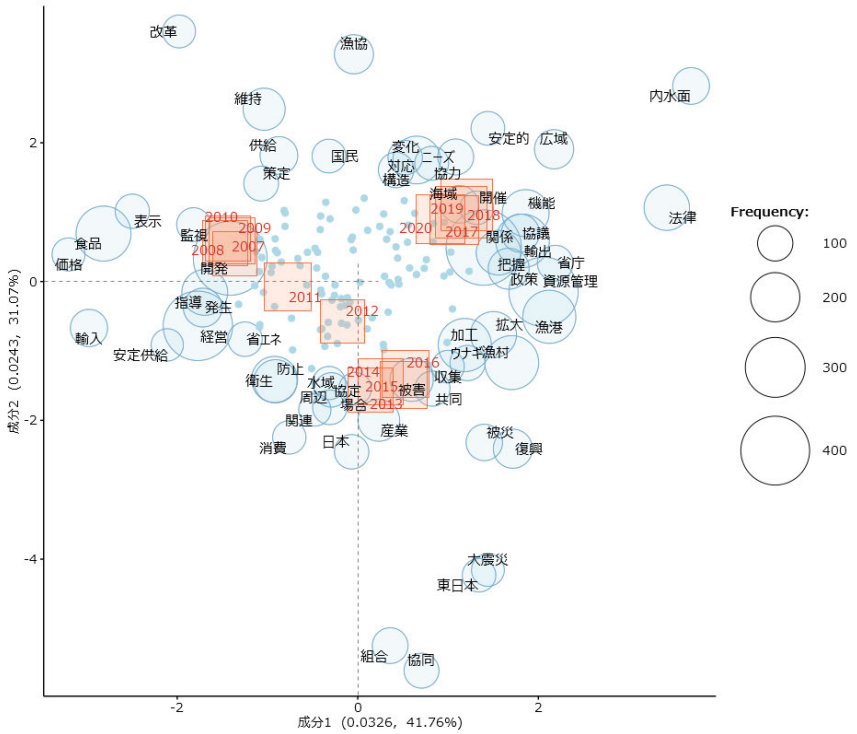


図 3-2 語･年の対応分析結果（分析対象語数：156）

4. LDA 分析

KHCoder により自動設定される最小出現数：150、分析対象語数：75 で集計単位を H5（年）とした場合と、分析対象語数：156 での LDA 分析を行った。

分析対象語数：75 での LDA トピック数推定結果を図 4-1、分析対象語数：156 での結果を図 4-2 に示す。これらの図から 75 語でのトピック数は 6 、156 語では 14 と推察した。

分析対象語数：75、トピック数：6 での LDA 処理結果を表 4-1、そのヒートマップを図 4-3、ヒートマップ樹形図を図 4-4、*トピック比率集計表を表 4-2、トピック比率を図 4-5～6 に示す。

また、分析対象語数：156、トピック数：14 でのその LDA 処理結果を表 4-3、そのヒートマップを図 4-7、ヒートマップ樹形図を図 4-8、*トピック比率集計表を表 4-4、トピック比率を図 4-9～12 に示す。

*は別途 excel 形式で提供。

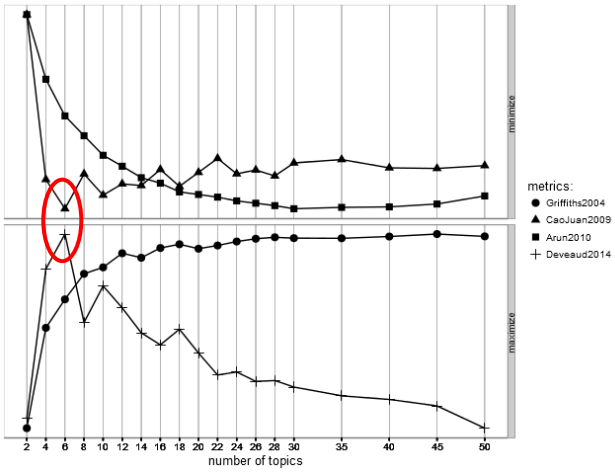


図 4-1 LDAtuning 実行結果（分析対象語数：75）

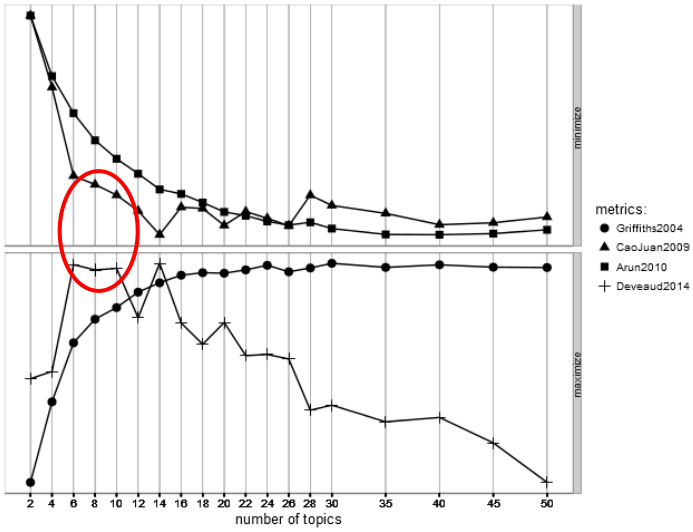


図 4-2 LDAtuning 実行結果（分析対象語数：156）

表 4-1 LDA 処理結果（6 トピックス、分析対象語数：75）

Topics

#1		#2		#3		#4		#5		#6	
支援	0.127	実施	0.139	経営	0.095	水産	0.153	生産	0.120	資源管理	0.100
確保	0.117	管理	0.139	整備	0.089	支援	0.097	対策	0.098	関係	0.091
連携	0.105	情報	0.108	開発	0.070	水産物	0.081	養殖	0.088	漁業者	0.073
資源	0.094	強化	0.087	措置	0.062	漁船	0.070	技術	0.087	漁港	0.060
施設	0.069	水産物	0.066	食品	0.058	状況	0.065	向上	0.078	漁村	0.052
必要	0.045	適切	0.062	漁業者	0.057	漁獲	0.059	効率的	0.058	漁場	0.051
流通	0.045	調査	0.051	促進	0.051	実施	0.055	種苗	0.057	輸出	0.050
手法	0.044	整備	0.046	提供	0.048	安全	0.054	活用	0.056	法律	0.043
水産業	0.039	利用	0.044	施設	0.044	調査	0.047	対応	0.054	機能	0.042
都道府県	0.038	放流	0.043	技術	0.043	災害	0.036	促進	0.047	必要	0.042

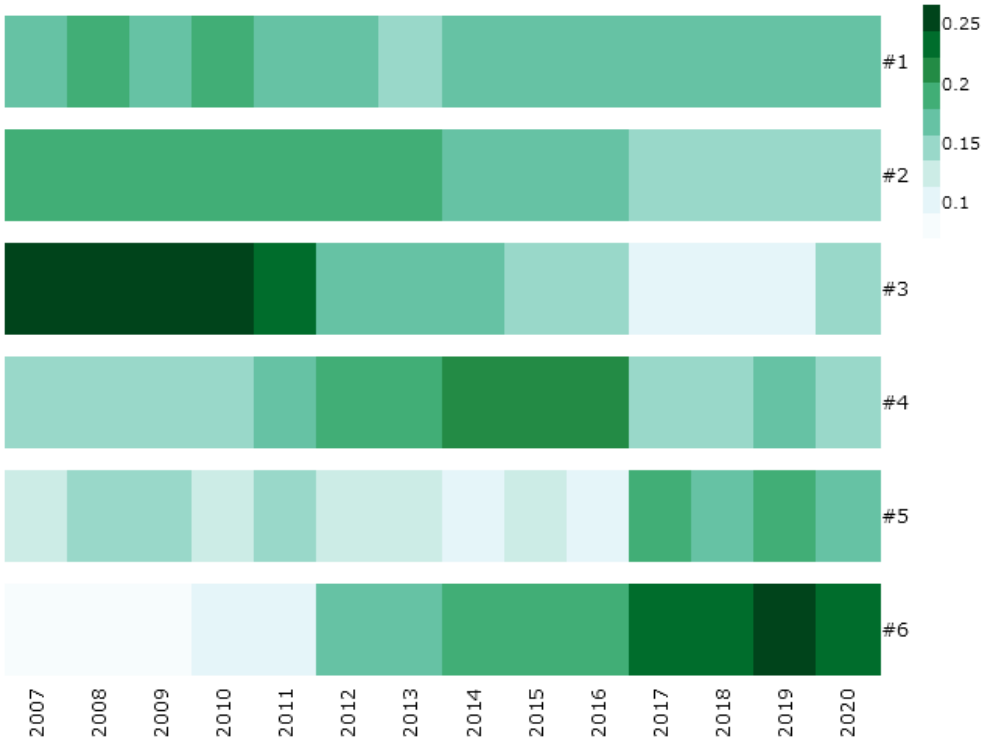


図 4-3　LDA ヒートマップ（6 トピックス、分析対象語数：75）

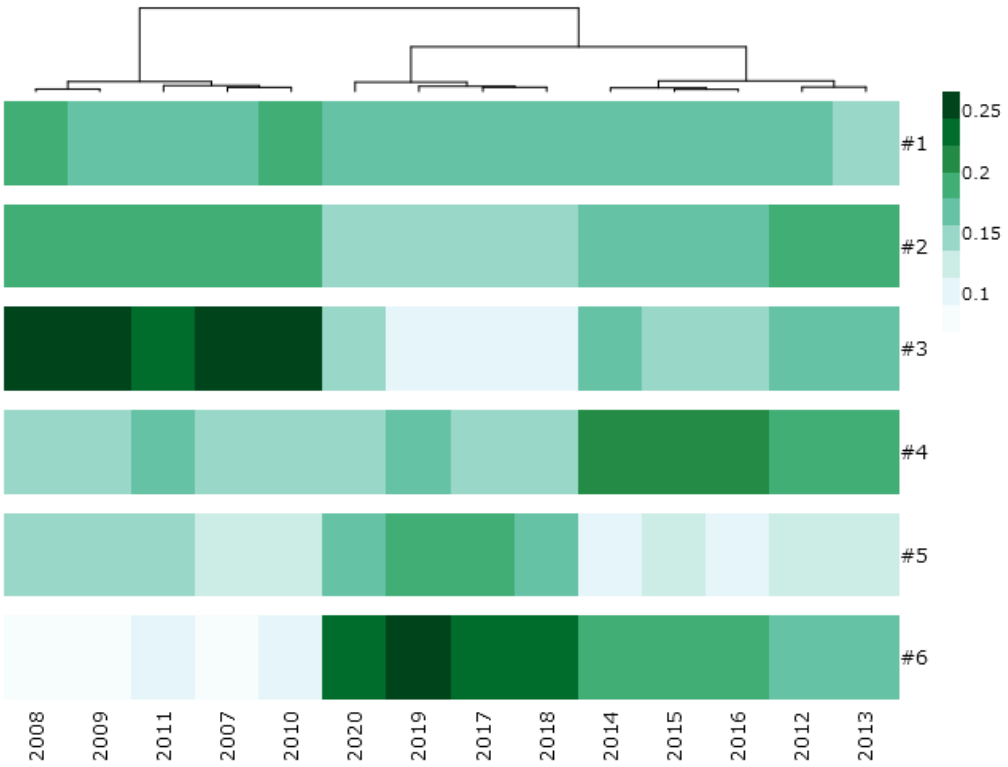


図 4-4　LDA ヒートマップ樹形図（6 トピックス、分析対象語数：75）

表 4-2　トピック比率集計表（6 トピックス、分析対象語数：75）

	#1	#2	#3	#4	#5	#6	ケース数
2007	0.171	0.185	0.266	0.156	0.131	0.091	1
2008	0.182	0.188	0.26	0.149	0.147	0.074	1
2009	0.175	0.186	0.265	0.156	0.147	0.07	1
2010	0.181	0.192	0.249	0.153	0.132	0.093	1
2011	0.167	0.186	0.242	0.159	0.136	0.109	1
2012	0.17	0.179	0.172	0.193	0.119	0.167	1
2013	0.151	0.19	0.174	0.194	0.118	0.172	1
2014	0.167	0.173	0.16	0.207	0.111	0.181	1
2015	0.16	0.163	0.154	0.216	0.117	0.189	1
2016	0.162	0.171	0.154	0.212	0.111	0.191	1
2017	0.162	0.156	0.113	0.155	0.186	0.228	1
2018	0.179	0.153	0.113	0.156	0.172	0.227	1
2019	0.162	0.144	0.111	0.157	0.18	0.246	1
2020	0.163	0.148	0.147	0.148	0.167	0.228	1

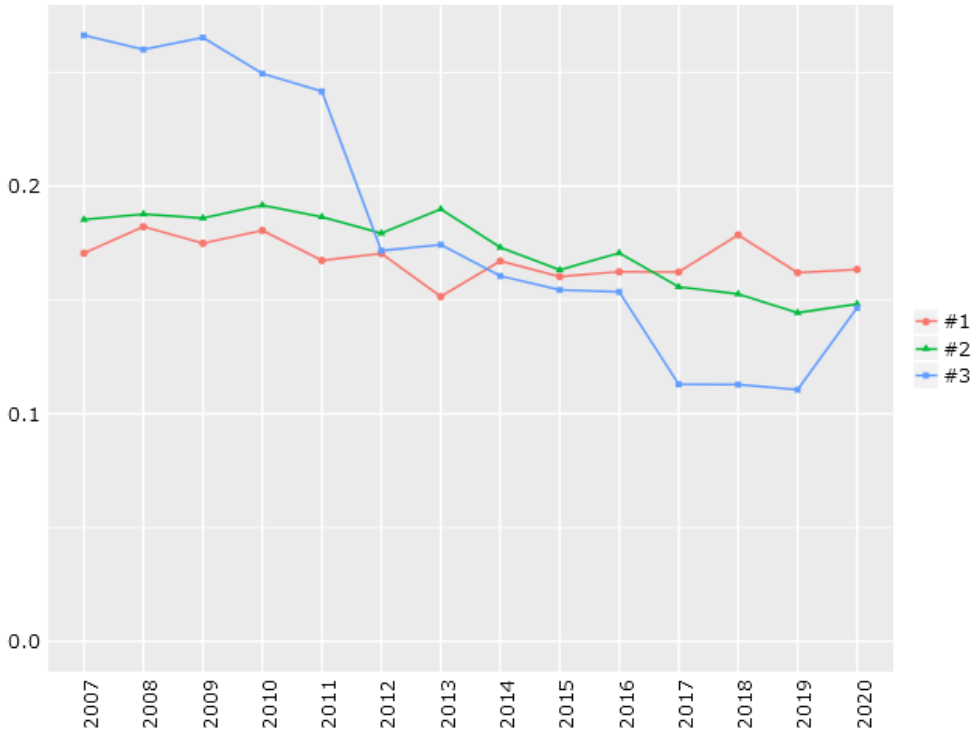


図 4-5　1～3 トピックの比率（6 トピックス、分析対象語数：75）

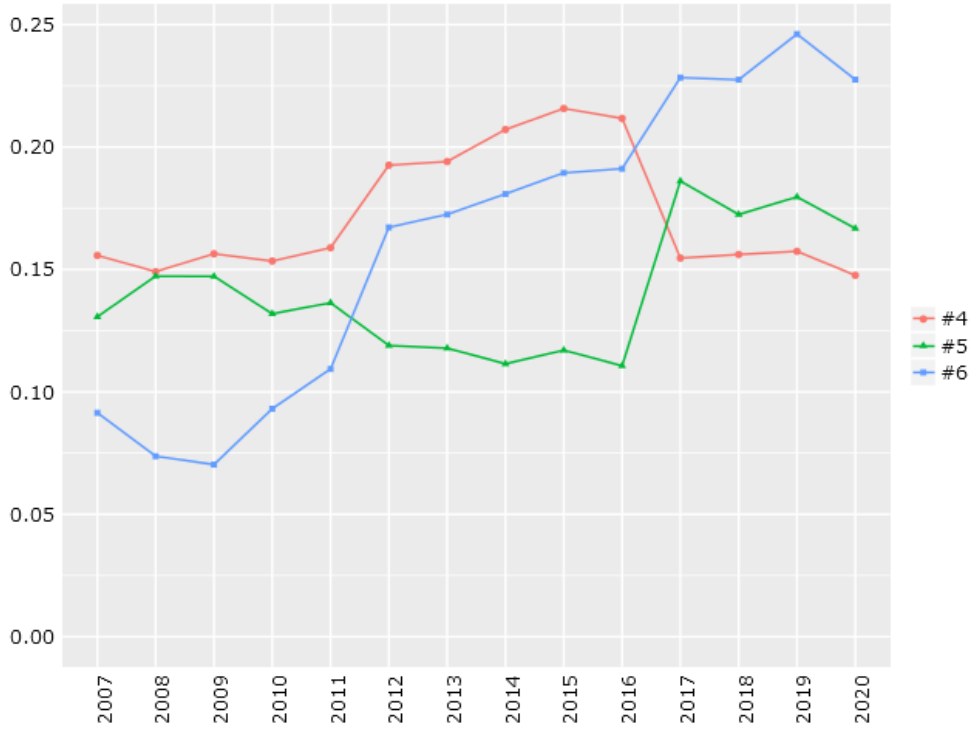


図 4-6　4～6 トピックの比率（6 トピックス、分析対象語数：75）

表 4-3 LDA 処理結果（14 トピックス、分析対象語数：156）

Topics

#1		#2		#3		#4		#5		#6		#7	
資源	0.123	関係	0.109	漁場	0.126	経営	0.135	支援	0.095	資源管理	0.110	水産業	0.067
確保	0.096	法律	0.069	必要	0.110	食品	0.100	生産	0.076	復興	0.074	水産	0.066
技術	0.068	輸出	0.053	体制	0.110	開発	0.096	対策	0.076	加工	0.072	機能	0.064
操業	0.058	資源管理	0.052	適切	0.109	促進	0.065	管理	0.072	支援	0.059	導入	0.060
放流	0.049	向上	0.050	連携	0.065	保全	0.059	技術	0.061	被災	0.049	広域	0.058
制度	0.043	内水面	0.043	強化	0.053	輸入	0.037	漁協	0.051	対策	0.049	把握	0.054
促進	0.042	漁業者	0.040	調査	0.051	価格	0.033	維持	0.050	東日本	0.049	漁村	0.047
水産資源	0.041	養殖	0.040	実態	0.038	導入	0.033	対応	0.046	大震災	0.048	連携	0.041
養殖	0.034	漁港	0.038	技術	0.037	表示	0.030	供給	0.040	漁港	0.045	水産資源	0.037
情報	0.033	拡大	0.038	機器	0.026	安全	0.030	改革	0.036	被害	0.043	実施	0.037

#8		#9		#10		#11		#12		#13		#14	
管理	0.061	活用	0.110	水産	0.126	管理	0.082	情報	0.105	整備	0.209	実施	0.236
支援	0.048	調査	0.099	漁船	0.092	検討	0.073	漁業者	0.072	確保	0.108	水産物	0.154
水産	0.047	支援	0.093	関係	0.069	生産	0.072	提供	0.071	措置	0.091	利用	0.057
組合	0.043	開発	0.059	漁獲	0.055	漁業者	0.070	施設	0.061	施設	0.072	資源	0.047
協同	0.040	強化	0.057	指導	0.054	効率的	0.069	計画	0.059	回復	0.042	規制	0.037
産業	0.038	手法	0.048	産地	0.049	水産物	0.065	水産	0.046	計画	0.041	流通	0.030
漁業者	0.038	団体	0.042	市場	0.047	種苗	0.049	防止	0.042	管理	0.040	対象	0.029
情報	0.035	消費者	0.038	水産業	0.044	政策	0.048	促進	0.041	設定	0.038	支援	0.029
状況	0.035	安全性	0.036	基本	0.041	構造	0.047	連携	0.041	育成	0.037	資金	0.028
経営	0.033	状況	0.035	安定供給	0.037	持続的	0.040	機関	0.040	資金	0.035	都道府県	0.028

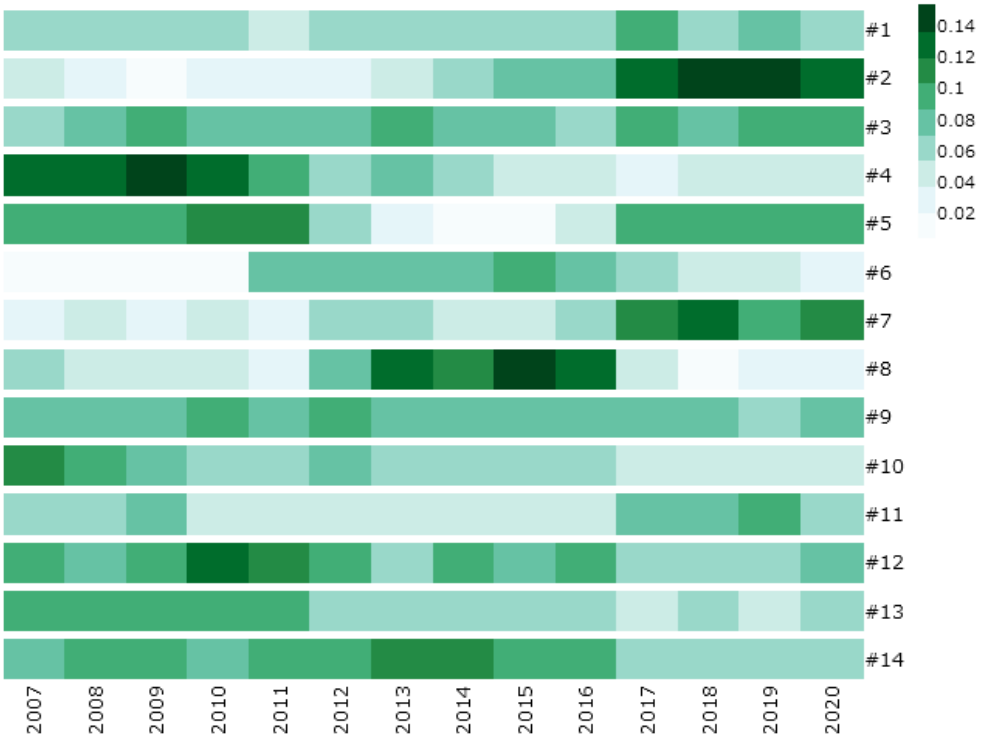


図 4-7 LDA ヒートマップ（14 トピックス、分析対象語数：156）

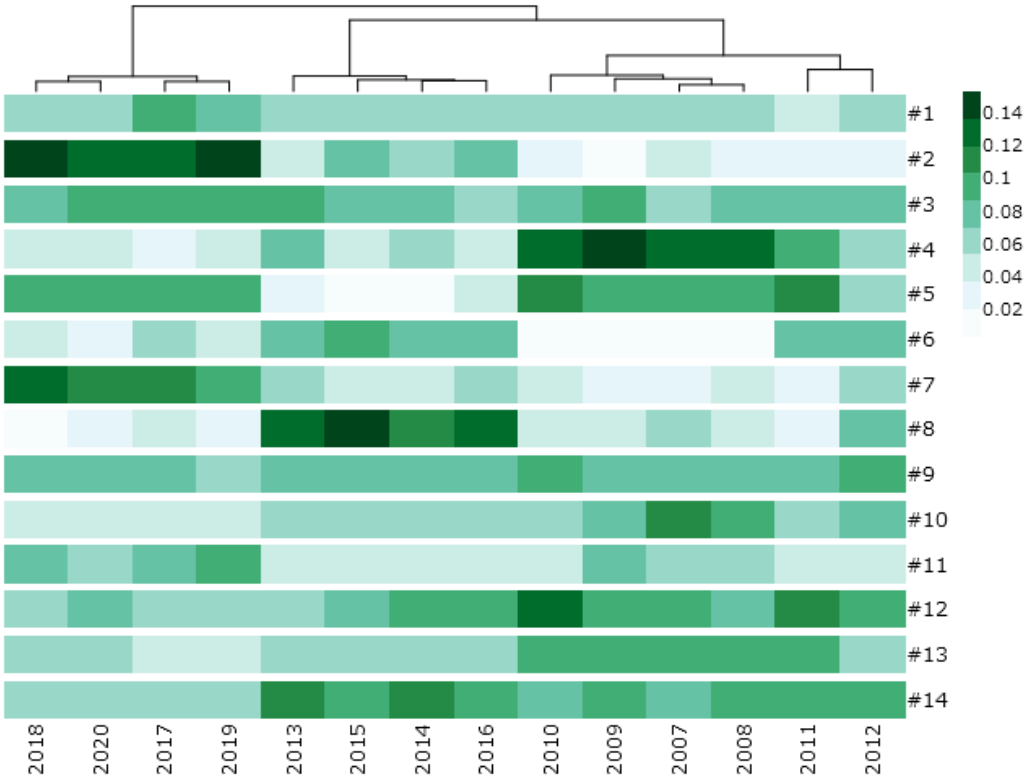


図 4-8　LDA ヒートマップ樹形図（14 トピックス、分析対象語数：156）

表 4-4　トピック比率集計表（6 トピックス、分析対象語数：75）

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	ケース数
2007	0.066	0.043	0.069	0.126	0.087	0.009	0.033	0.06	0.075	0.11	0.061	0.087	0.087	0.087	1
2008	0.063	0.032	0.073	0.125	0.092	0.004	0.045	0.052	0.079	0.095	0.061	0.086	0.096	0.098	1
2009	0.067	0.014	0.092	0.143	0.088	0.008	0.031	0.041	0.08	0.082	0.071	0.089	0.099	0.094	1
2010	0.066	0.029	0.085	0.13	0.104	0.011	0.038	0.038	0.09	0.065	0.049	0.121	0.096	0.077	1
2011	0.047	0.032	0.083	0.09	0.112	0.081	0.034	0.03	0.073	0.068	0.044	0.118	0.095	0.093	1
2012	0.064	0.033	0.083	0.07	0.057	0.078	0.055	0.077	0.098	0.078	0.048	0.09	0.069	0.1	1
2013	0.06	0.045	0.09	0.083	0.029	0.078	0.054	0.131	0.072	0.068	0.048	0.07	0.064	0.107	1
2014	0.063	0.064	0.083	0.055	0.019	0.087	0.045	0.111	0.083	0.067	0.048	0.09	0.067	0.118	1
2015	0.065	0.076	0.085	0.046	0.014	0.093	0.054	0.147	0.082	0.054	0.042	0.081	0.058	0.103	1
2016	0.06	0.081	0.07	0.044	0.038	0.083	0.054	0.122	0.082	0.069	0.052	0.096	0.059	0.09	1
2017	0.091	0.124	0.092	0.035	0.096	0.061	0.106	0.04	0.072	0.038	0.083	0.057	0.04	0.066	1
2018	0.069	0.136	0.085	0.039	0.092	0.047	0.122	0.02	0.077	0.037	0.083	0.068	0.059	0.067	1
2019	0.076	0.153	0.089	0.039	0.099	0.051	0.102	0.035	0.056	0.049	0.088	0.056	0.049	0.059	1
2020	0.062	0.133	0.087	0.053	0.099	0.032	0.11	0.034	0.079	0.044	0.058	0.083	0.062	0.065	1

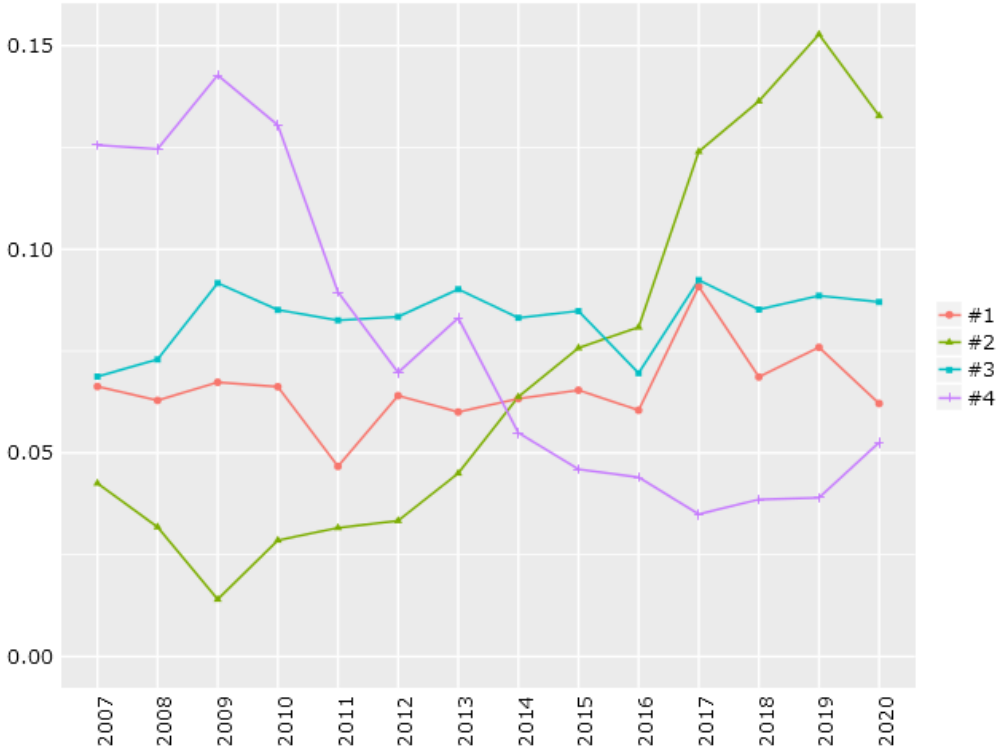


図 4-9　1～4 トピックの比率（14 トピックス、分析対象語数：156）

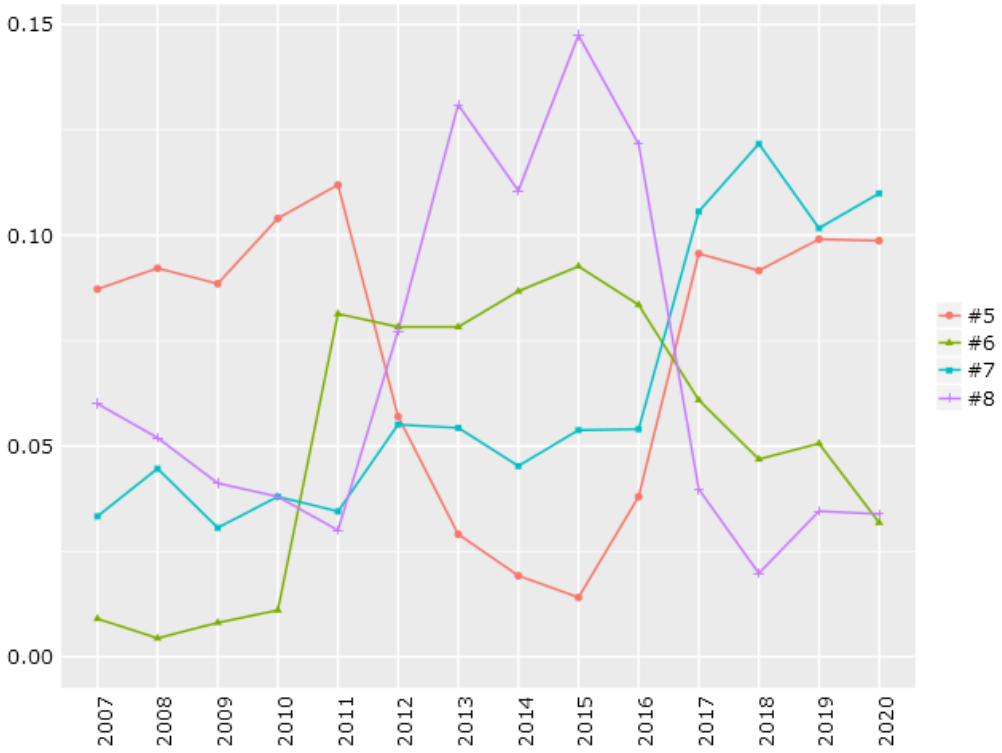


図 4-10　5～8 トピックの比率（14 トピックス、分析対象語数：156）

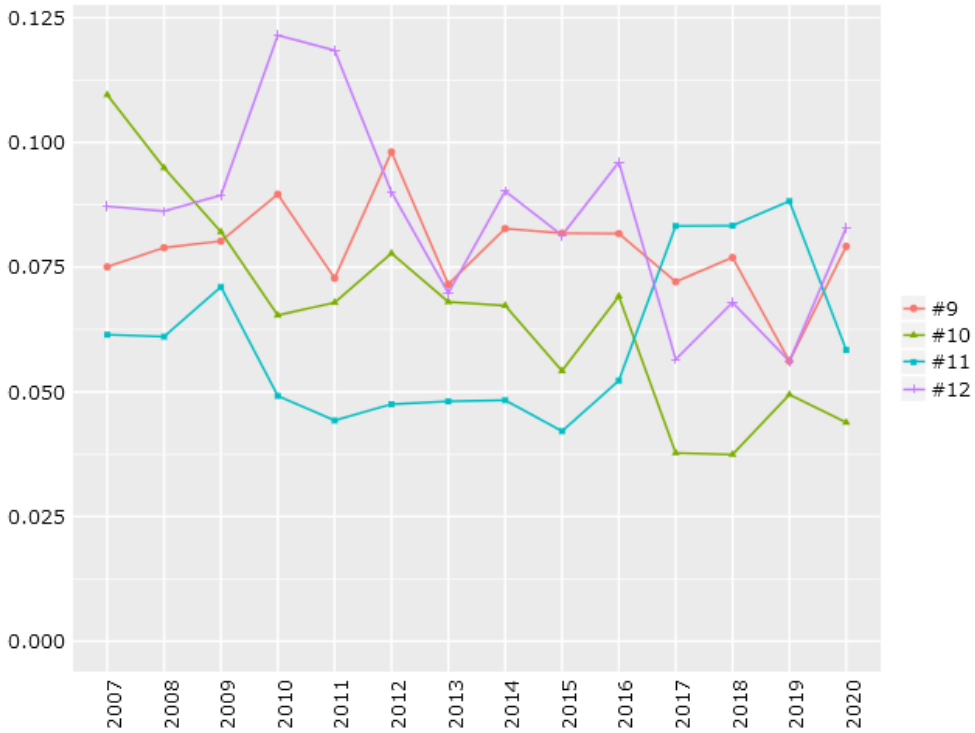


図 4-11　9～12 トピックの比率（14 トピックス、分析対象語数：156）

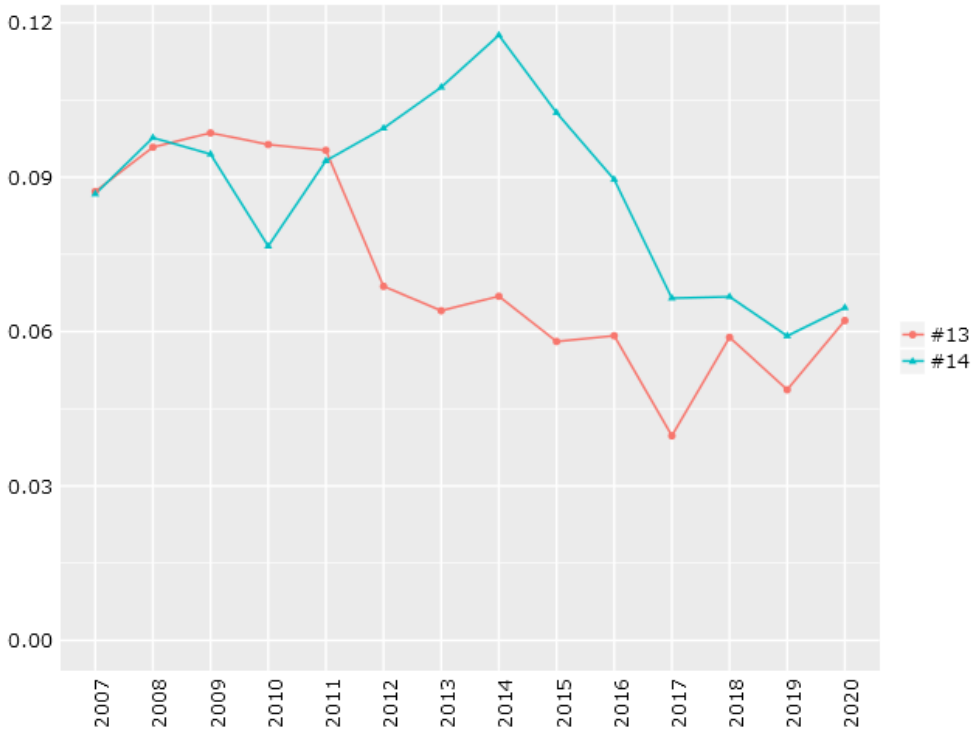


図 4-12　13～14 トピックの比率（14 トピックス、分析対象語数：156）

付録 4 水産白書本編（2007～2020 年）の分析結果

文書番号：JRDN-21-024

1. 前処理結果

環境白書･海洋白書･水産白書（2008～2020 年）の分析で設定した強制抽出語（316 語）と 54 語の除外語を設定し、「動詞、感動詞、動詞 B、副詞 B」を除外して前処理を実行した。

Chanse での前処理の結果、総抽出語数：733,418、異なり語数：13,972 のうち 9,993 語が分析処理で使用された。抽出語出現数の頻度分布を図 1-1、抽出語リスト（上位 200 語）を図 1-2 に示す。

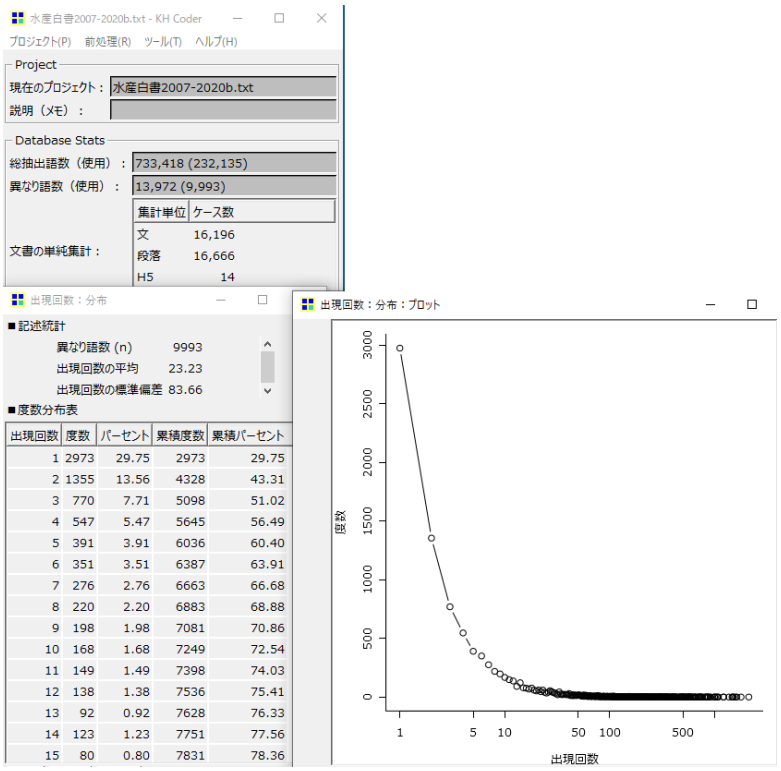


図 1-1 抽出語出現数の頻度分布

#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度
1	水産物	名詞	2111	36	施設	サ変名詞	649				
2	漁船	名詞	1780	37	販売	サ変名詞	647				
3	漁獲	サ変名詞	1623	38	技術	名詞	640				
4	管理	サ変名詞	1572	39	多い	形容詞	621				
5	水産	名詞	1508	40	影響	サ変名詞	610				
6	漁業者	タグ	1461	41	水産業	タグ	605	71	日本	地名	458
7	資源	名詞	1365	42	世界	名詞	589	72	機関	名詞	450
8	養殖	サ変名詞	1215	43	向上	サ変名詞	585	73	商品	名詞	450
9	減少	サ変名詞	1211	44	流通	サ変名詞	584	74	漁港	名詞	432
10	操業	サ変名詞	1084	45	水域	名詞	573	75	産業	名詞	431
11	生産	サ変名詞	1059	46	水産資源	タグ	571	76	回復	サ変名詞	424
12	増加	サ変名詞	1049	47	価格	名詞	570	77	活用	サ変名詞	423
13	資源管理	タグ	1012	48	開発	サ変名詞	570	78	制度	名詞	414
14	経営	サ変名詞	995	49	輸出	サ変名詞	563	79	強化	サ変名詞	412
15	実施	サ変名詞	989	50	確保	サ変名詞	558	80	漁労	名詞	411
16	利用	サ変名詞	909	51	支援	サ変名詞	553	81	海域	名詞	403
17	関係	サ変名詞	907	52	食品	名詞	542	82	期待	サ変名詞	403
18	漁協	名詞	869	53	結果	副詞可能	537	83	割合	名詞	399
19	必要	形容動詞	859	54	対策	サ変名詞	528	84	中心	名詞	399
20	重要	形容動詞	856	55	機能	サ変名詞	527	85	可能	形容動詞	398
21	加工	サ変名詞	844	56	漁獲量	タグ	515	86	計画	サ変名詞	395
22	マグロ	名詞	831	57	高い	形容詞	513	87	開催	サ変名詞	385
23	消費者	タグ	815	58	情報	名詞	510	88	傾向	名詞	385
24	養殖業	タグ	804	59	就業	サ変名詞	503	89	地元	名詞	381
25	活動	サ変名詞	766	60	連携	サ変名詞	493	90	変化	サ変名詞	373
26	魚種	タグ	754	61	市場	名詞	490	91	種苗	タグ	369
27	状況	名詞	738	62	大きい	形容詞	488	92	提供	サ変名詞	363
28	生産量	タグ	735	63	輸入	サ変名詞	488	93	課題	名詞	360
29	消費	サ変名詞	729	64	発生	サ変名詞	487	94	周辺	名詞	357
30	漁場	名詞	722	65	拡大	サ変名詞	485	95	適切	形容動詞	357
31	調査	サ変名詞	717	66	多く	副詞可能	484	96	保全	サ変名詞	356
32	対象	名詞	700	67	規制	サ変名詞	481	97	供給	サ変名詞	355
33	措置	サ変名詞	696	68	被害	名詞	477	98	中国	地名	348
34	漁村	名詞	692	69	導入	サ変名詞	464	99	生物	名詞	347
35	魚介類	タグ	657	70	太平洋	地名	463	100	委員会	タグ	343

#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度
101	水準	名詞	343	136	協議	サ変名詞	277				
102	産地	名詞	341	137	目的	名詞	276				
103	持続的	タグ	340	138	問題	ナイ形容	276				
104	制限	サ変名詞	338	139	普及	サ変名詞	270				
105	昭和	固有名詞	336	140	効果	名詞	268				
106	復興	サ変名詞	330	141	国際	名詞	268	171	保護	サ変名詞	229
107	国内	名詞	327	142	促進	サ変名詞	267	172	人	名詞C	228
108	整備	サ変名詞	327	143	品質	名詞	267	173	安定	形容動詞	227
109	調理	サ変名詞	327	144	許可	サ変名詞	266	174	教育	サ変名詞	226
110	女性	名詞	321	145	改善	サ変名詞	264	175	基本	名詞	225
111	水揚げ	名詞	320	146	カツオ	名詞	263	176	検討	サ変名詞	225
112	研究	サ変名詞	317	147	策定	サ変名詞	263	177	文化	名詞	225
113	上昇	サ変名詞	316	148	料理	サ変名詞	261	178	防止	サ変名詞	224
114	出荷	サ変名詞	315	149	TAC	タグ	260	179	開始	サ変名詞	223
115	沿岸	名詞	314	150	共同	サ変名詞	260	180	活性化	タグ	222
116	福島	地名	310	151	捕鯨	名詞	256	181	体験	サ変名詞	222
117	役割	名詞	310	152	主	形容動詞	253	182	評価	サ変名詞	222
118	協定	サ変名詞	309	153	海面	名詞	251	183	育成	サ変名詞	220
119	都道府県	名詞	308	154	使用	サ変名詞	250	184	ブリ	名詞	219
120	対応	サ変名詞	306	155	体制	名詞	249	185	放射性物質	タグ	219
121	協力	サ変名詞	304	156	卸売市場	タグ	248	186	期間	名詞	218
122	実習	サ変名詞	304	157	購入	サ変名詞	248	187	目標	名詞	218
123	需要	名詞	302	158	サバ	名詞	246	188	安全	形容動詞	217
124	保存	サ変名詞	297	159	種類	名詞	246	189	事故	名詞	217
125	場合	副詞可能	296	160	生活	サ変名詞	245	190	参加	サ変名詞	216
126	組合	名詞	295	161	発展	サ変名詞	242	191	年間	名詞	216
127	維持	サ変名詞	294	162	合意	サ変名詞	239	192	宮城	地名	215
128	団体	名詞	291	163	設定	サ変名詞	239	193	サンマ	名詞	214
129	設置	サ変名詞	290	164	低下	サ変名詞	236	194	推移	サ変名詞	214
130	それぞれ	副詞可能	289	165	海外	名詞	235	195	ニーズ	名詞	213
131	サケ	名詞	288	166	水産庁	組織名	235	196	東日本	タグ	213
132	内水面	タグ	286	167	方法	名詞	232	197	摂取	サ変名詞	212
133	所得	名詞	284	168	一部	副詞可能	229	198	当たり	副詞可能	210
134	作業	サ変名詞	283	169	小型	名詞	229	199	国民	名詞	209
135	放流	サ変名詞	283	170	生態系	タグ	229	200	栄養	名詞	208

図 1-2 抽出語リスト（上位 200 語）

2. 共起ネットワーク

KHCoder により自動設定される最小出現頻度：430、分析対象語数：75 とした場合と、最小出現頻度：248、分析対象語数：157 の結果を示す。

分析対象語数：75 での共起ネットワークの描画条件を以下に、語・語での共起ネットワークを図 2-1、語・年での共起ネットワークを図 2-2 に示す。

また、分析対象語数：157 での共起ネットワークの描画条件を以下に、語・語での共起ネットワークを図 2-3、語・年での共起ネットワークを図 2-4 に示す。

	最小出現頻度	分析対象語数	描画結果
語・語	430	75	ノード数：55　エッジ数：75
語・年	430	75	ノード数：51　エッジ数：75

	最小出現頻度	分析対象語数	描画結果
語・語	248	157	ノード数：112　エッジ数：157
語・年	248	157	ノード数：73　エッジ数：157

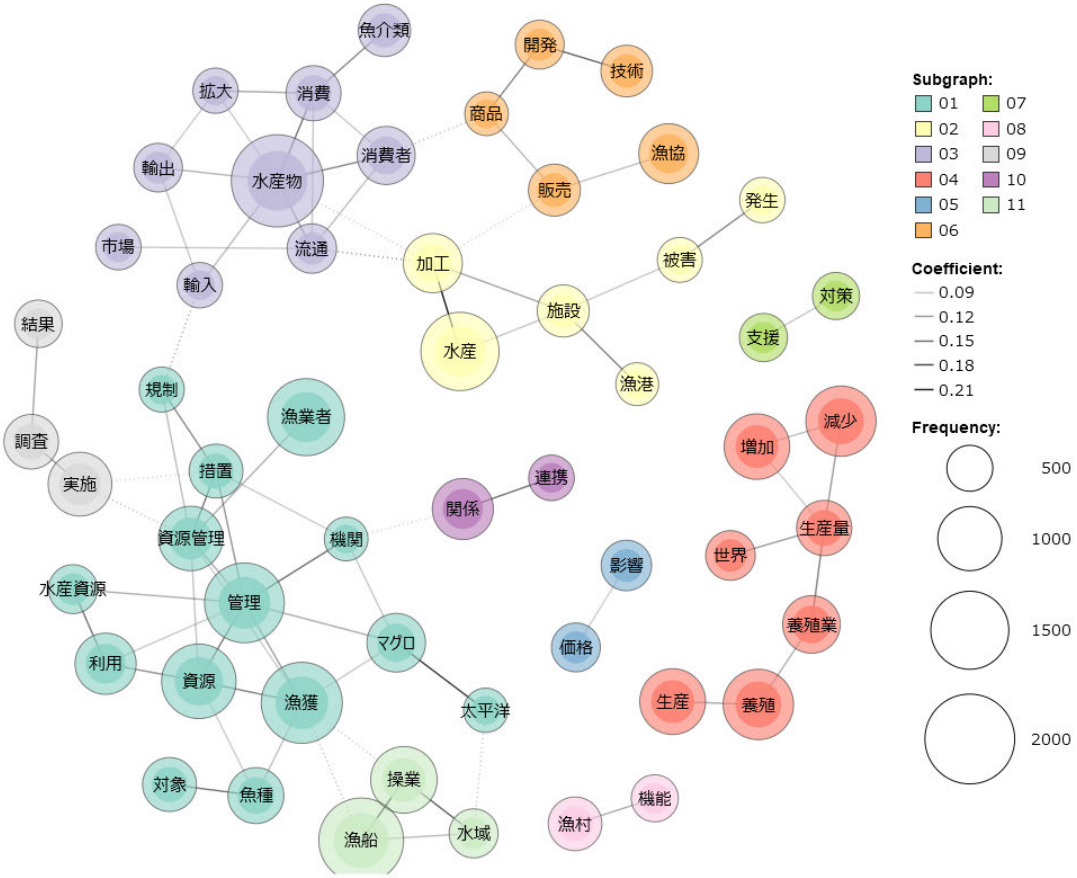


図 2-1　共起ネットワーク（語・語）（分析対象語数：75）

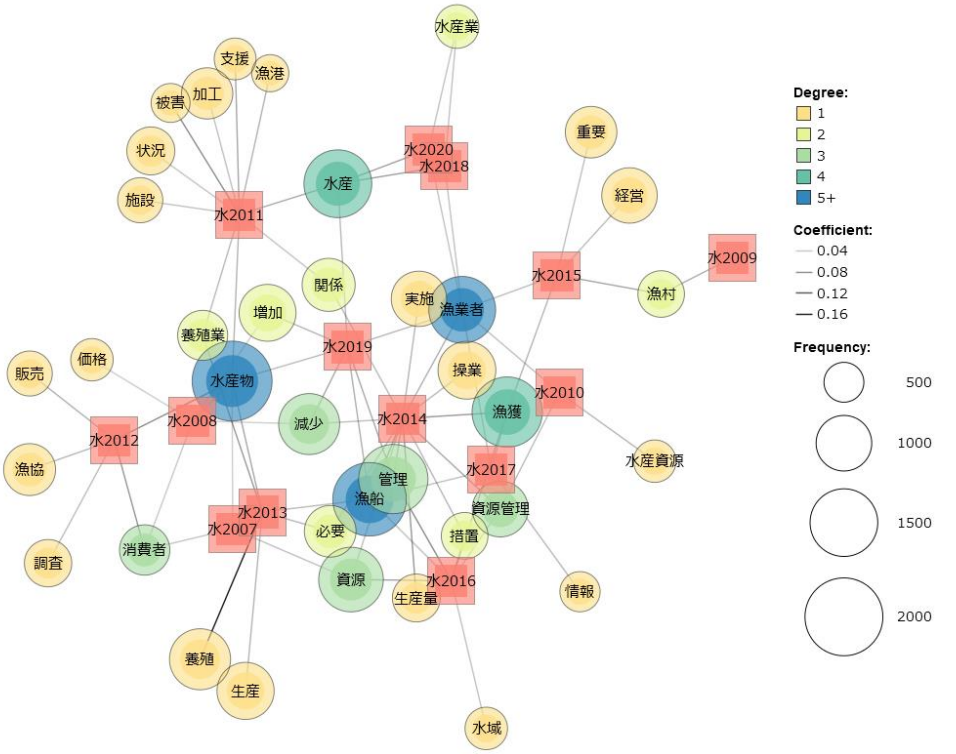


図 2-2 共起ネットワーク（語・年）（分析対象語数：75）

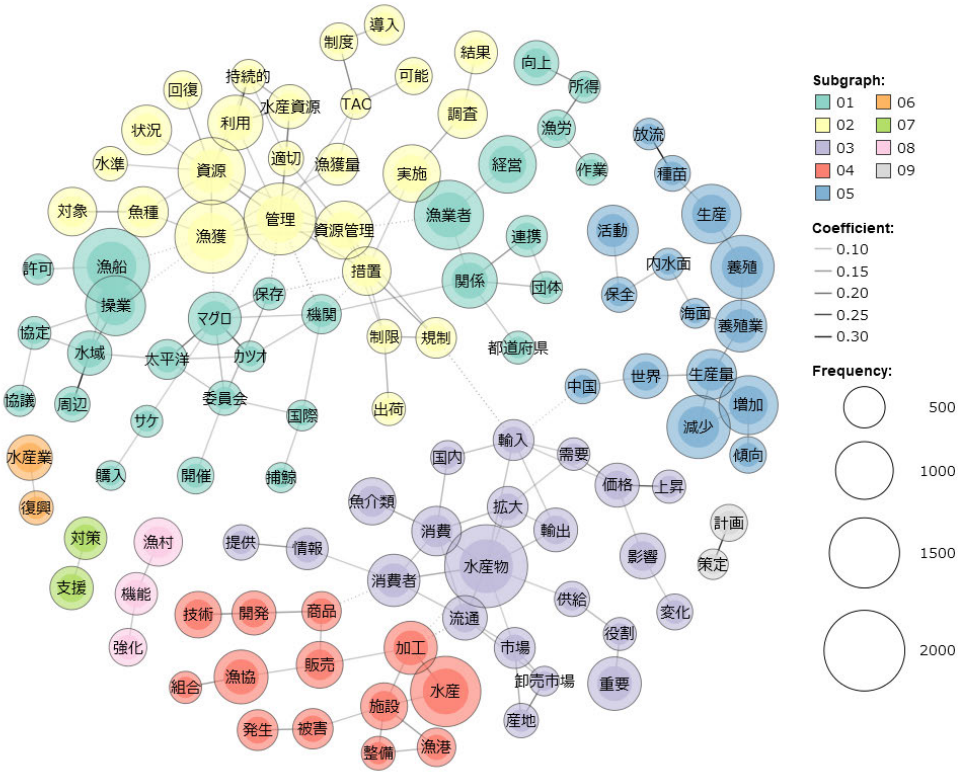


図 2-3 共起ネットワーク（語・語）（分析対象語数：157）

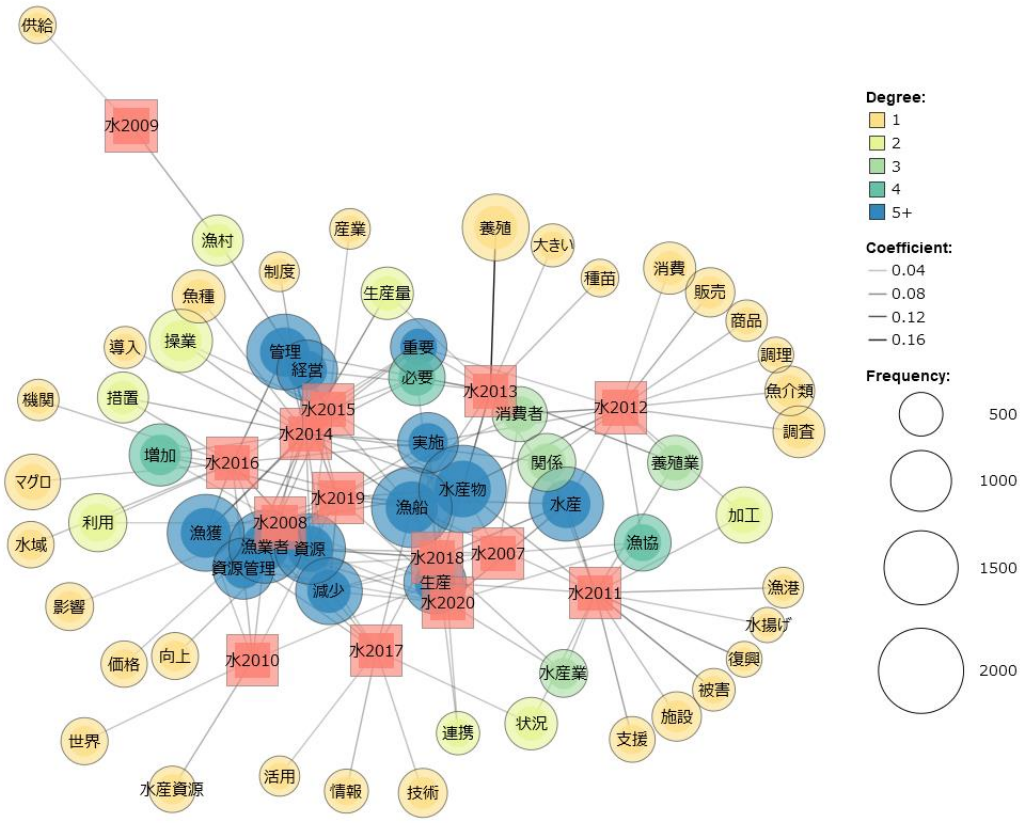


図 2-4 共起ネットワーク（語・年）（分析対象語数：157）

3. 対応分析

KHCoder により自動設定される最小出現頻度：430、分析対象語数：75 とした場合と、最小出現頻度：248、分析対象語数：157 で対応分析処理を行った。

分析対象語数：75 の累積寄与率は成分 1 と 2 で 47.78%であり、成分 3 と 4 の累積寄与率は約２３%である。その結果を図 3-1,3-2 に示す。

分析対象語数：157 の累積寄与率は成分 1 と 2 で 43.63%であり、成分 3 と 4 の累積寄与率は約２5%である。その結果を図 3-3,3-4 に示す。

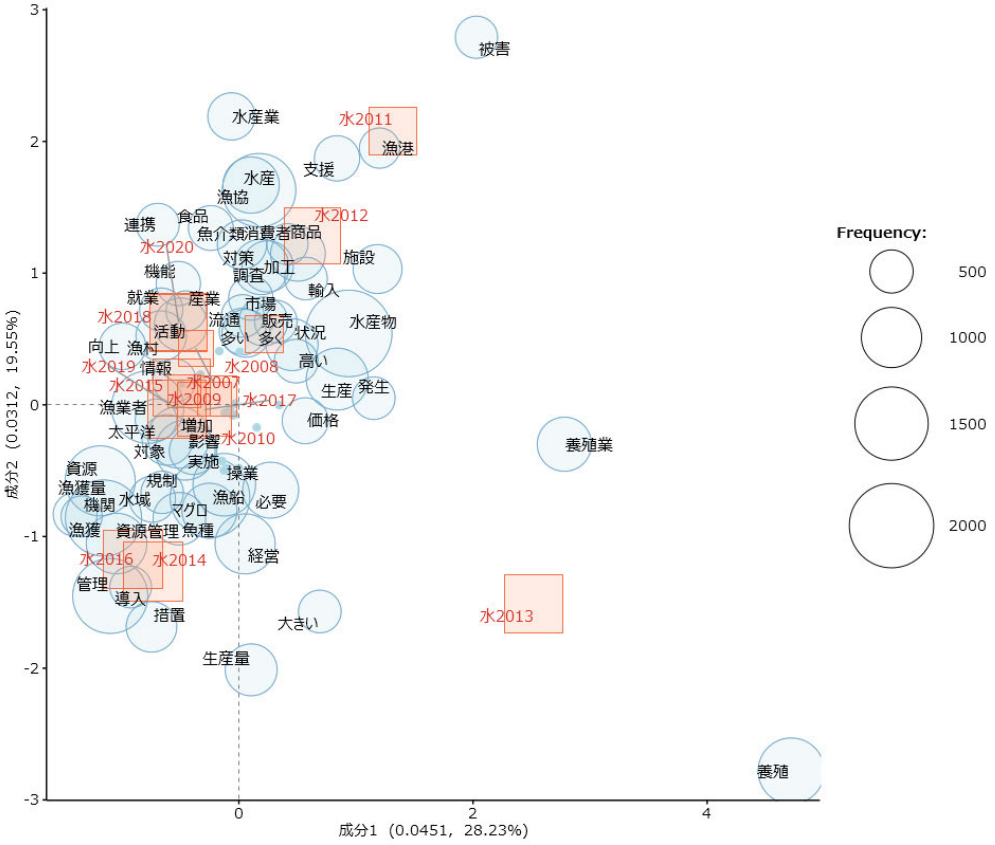


図 3-1 語･年の対応分析結果 （成分 1 と 2）（分析対象語数：75）

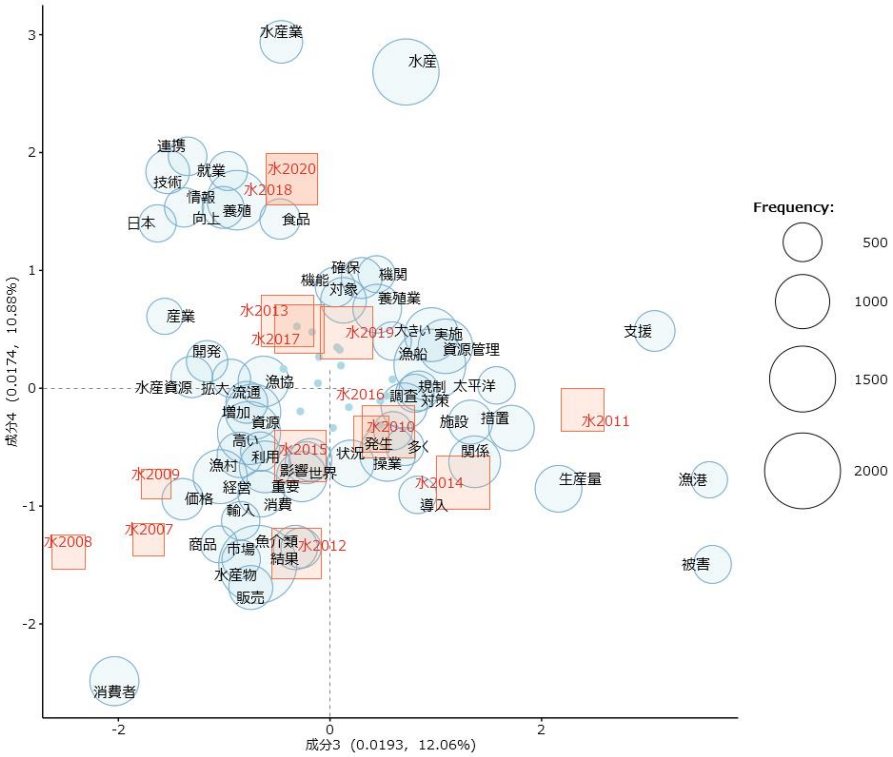


図 3-2 語･年の対応分析結果 （成分 3 と 4）（分析対象語数：75）

図 3-3 語･年の対応分析結果（成分 1と2）（分析対象語数：157）

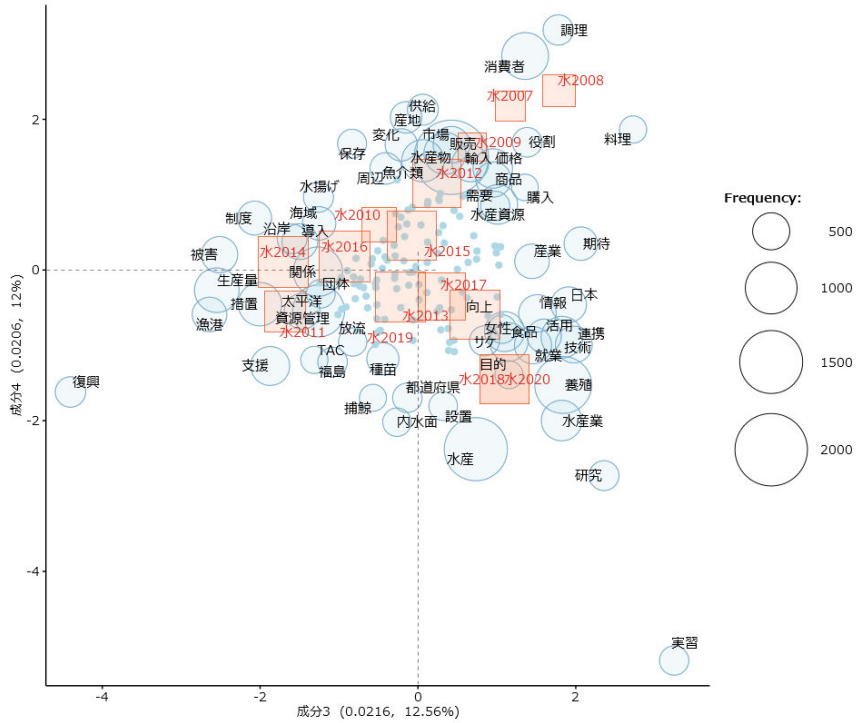


図 3-4 語･年の対応分析結果（成 3と4）（分析対象語数：157）

4. LDA 分析

KHCoder により自動設定される最小出現数：430、分析対象語数：75 で集計単位を H5（年）とした場合と、分析対象語数：157 での LDA 分析を行った。

分析対象語数：75 での LDA トピック数推定結果を図 4-1、分析対象語数：157 での結果を図 4-2 に示す。これらの図から 75 語でのトピック数は 12 と 20 、157 語では 12 と 20 と推察した。

分析対象語数：75、トピック数：12 での LDA 処理結果を表 4-1、そのヒートマップを図 4-3、ヒートマップ樹形図を図 4-4、*トピック比率集計表を表 4-2、トピック比率を図 4-5～7 に示す。また、トピック数：20 での LDA 処理結果を表 4-3、そのヒートマップを図 4-8、ヒートマップ樹形図を図 4-9、*トピック比率集計表を表 4-4、トピック比率を図 4-10～14 に示す。

分析対象語数：157、トピック数：12 での LDA 処理結果を表 4-5、そのヒートマップを図 4-15、ヒートマップ樹形図を図 4-16、*トピック比率集計表を表 4-6、トピック比率を図 4-17～19 に示す。また、トピック数：20 での LDA 処理結果を表 4-7、そのヒートマップを図 4-20、ヒートマップ樹形図を図 4-21、*トピック比率集計表を表 4-8、トピック比率を図 4-22～26 に示す。

*は別途 excel 形式で提供。

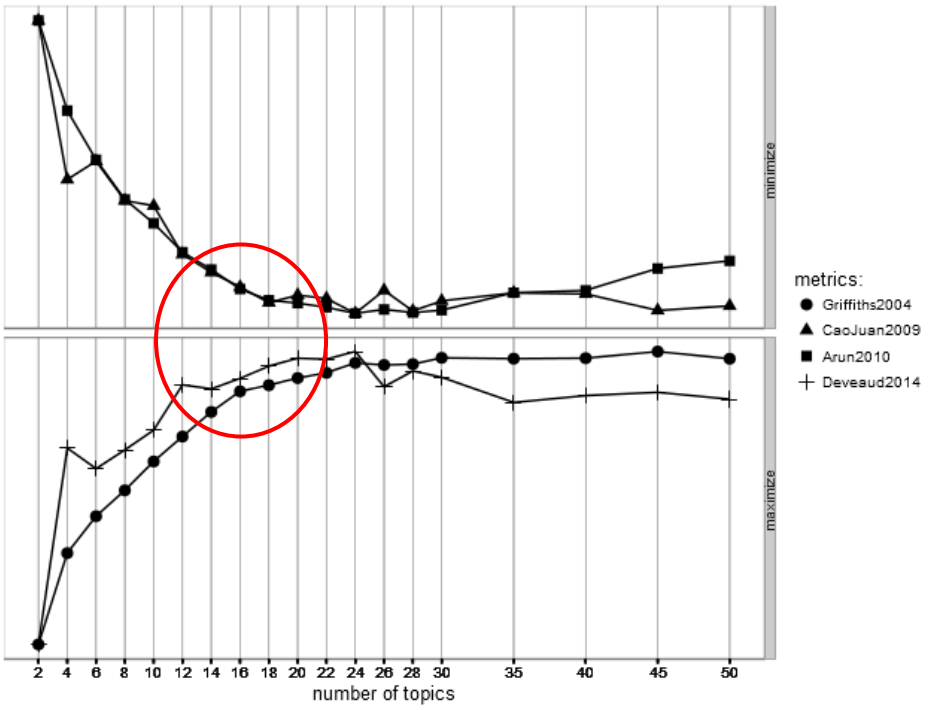


図 4-1 LDAtuning 実行結果（分析対象語数：75）

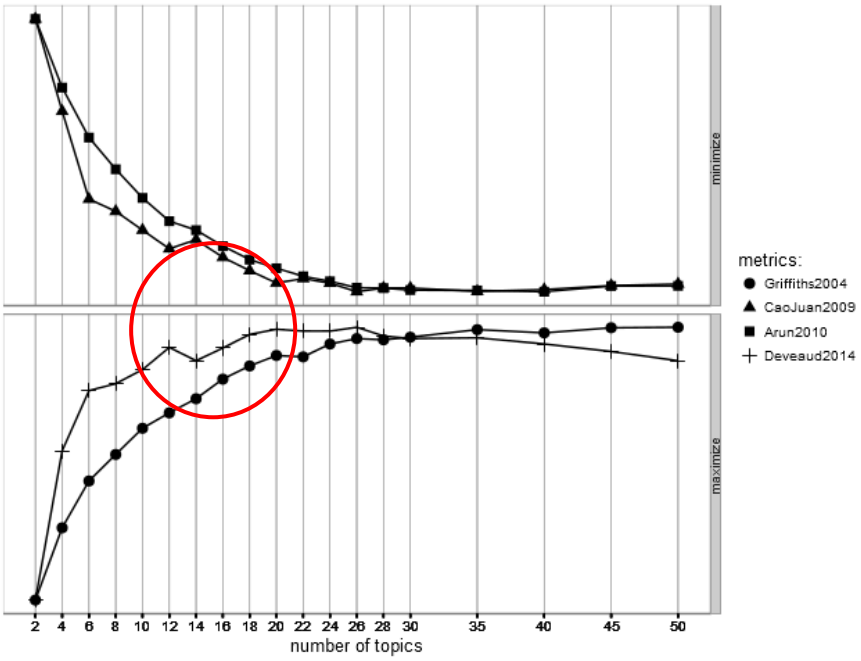


図 4-2 LDAtuning 実行結果（分析対象語数：157）

表 4-1 LDA 処理結果（12 トピックス、分析対象語数：75）

Topics

#1		#2		#3		#4		#5		#6	
水産	0.213	向上	0.112	漁村	0.205	管理	0.165	水産物	0.137	操業	0.194
水産業	0.107	漁船	0.086	経営	0.138	漁獲	0.128	消費者	0.125	重要	0.133
連携	0.067	管理	0.079	産業	0.085	資源	0.106	漁協	0.076	必要	0.116
実施	0.063	流通	0.061	生産	0.077	水域	0.055	魚介類	0.068	利用	0.090
漁協	0.061	機能	0.057	漁獲	0.072	措置	0.055	販売	0.063	結果	0.058
就業	0.061	拡大	0.057	消費	0.062	機関	0.053	商品	0.061	価格	0.047
減少	0.056	確保	0.054	活動	0.053	マグロ	0.050	活動	0.045	消費	0.037
食品	0.041	対策	0.052	輸出	0.041	実施	0.043	消費	0.042	施設	0.035
漁業者	0.035	増加	0.051	加工	0.027	漁船	0.042	多い	0.039	販売	0.033
多い	0.035	減少	0.048	管理	0.026	関係	0.038	調査	0.032	輸出	0.031

#7		#8		#9		#10		#11		#12	
漁業者	0.144	水産物	0.144	養殖	0.307	水産資源	0.121	被害	0.085	漁船	0.111
技術	0.108	減少	0.095	養殖業	0.100	資源管理	0.067	水産	0.085	生産量	0.097
情報	0.105	増加	0.093	水産物	0.056	世界	0.067	支援	0.082	資源管理	0.083
資源	0.085	価格	0.063	漁船	0.047	漁場	0.053	漁港	0.074	漁業者	0.065
開発	0.071	資源	0.054	管理	0.043	利用	0.053	施設	0.066	措置	0.058
調査	0.058	経営	0.049	生産	0.039	マグロ	0.051	加工	0.059	導入	0.053
魚種	0.056	生産	0.048	大きい	0.036	高い	0.048	養殖業	0.055	魚種	0.050
日本	0.051	市場	0.047	経営	0.035	多い	0.043	関係	0.047	実施	0.047
漁獲	0.044	消費者	0.044	生産量	0.032	魚介類	0.039	調査	0.045	増加	0.042
漁獲量	0.038	利用	0.039	マグロ	0.031	連携	0.038	対策	0.042	関係	0.041

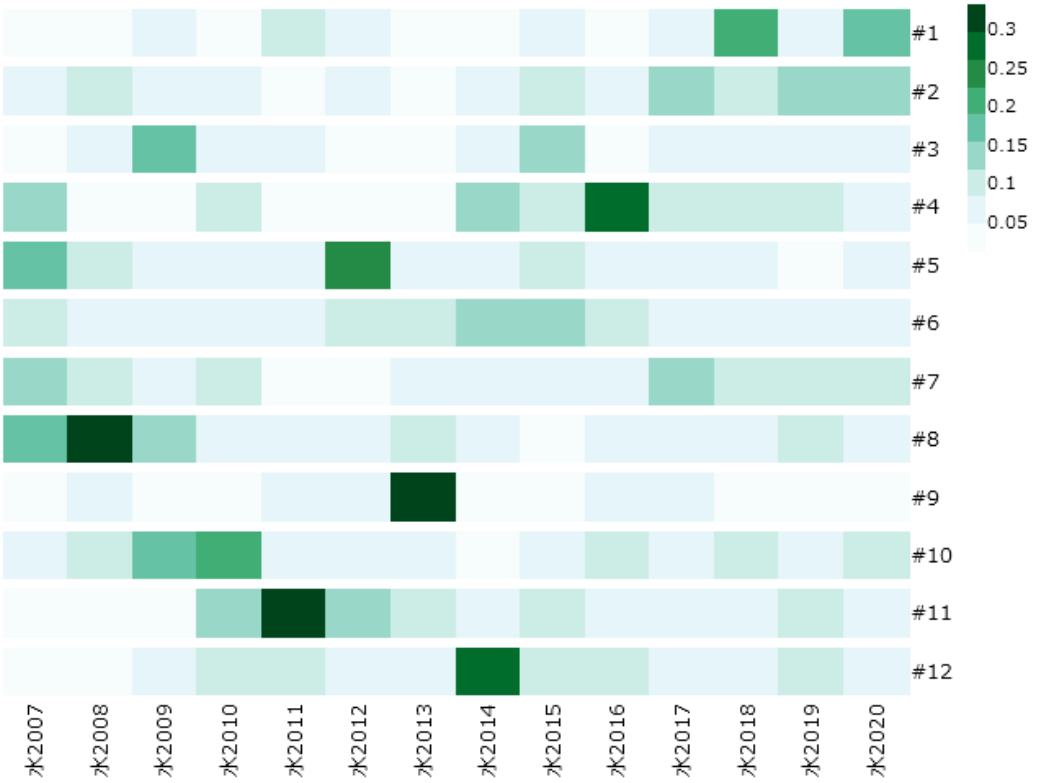


図 4-3　LDA ヒートマップ（12 トピックス、分析対象語数：75）

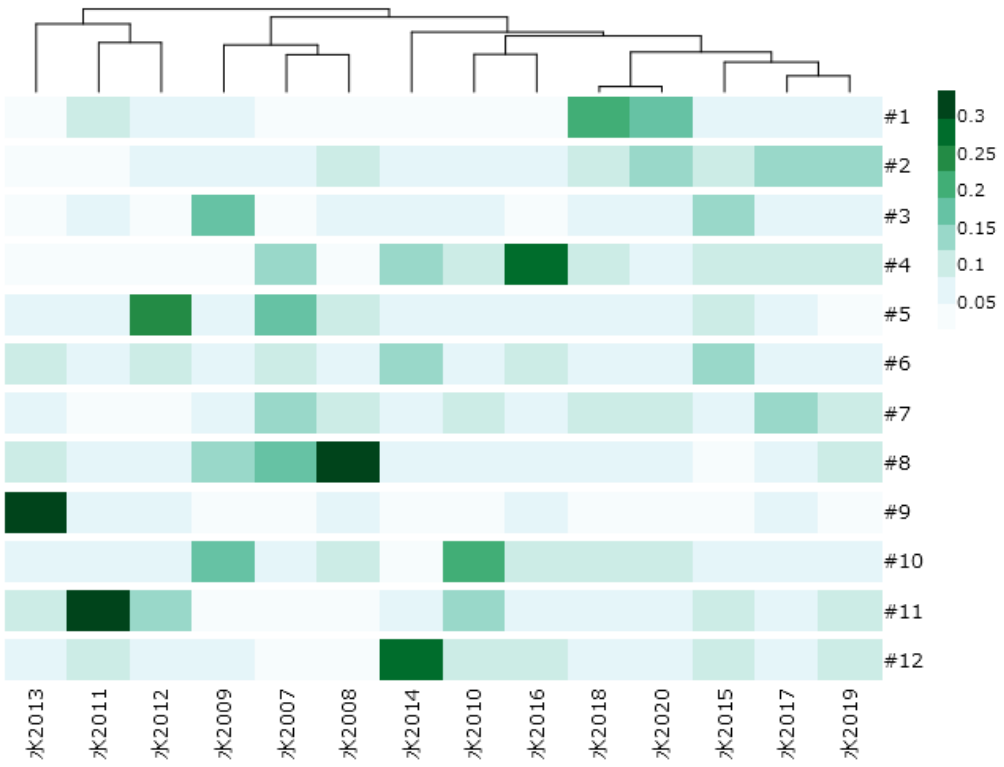


図 4-4　LDA ヒートマップ樹形図（12 トピックス、分析対象語数：75）

表 4-2 トピック比率集計表（12 トピックス、分析対象語数：75）

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	ケース数
水2007	0.046	0.065	0.042	0.131	0.163	0.091	0.144	0.167	0.026	0.061	0.046	0.017	1
水2008	0.042	0.095	0.052	0.031	0.098	0.082	0.093	0.303	0.057	0.087	0.041	0.019	1
水2009	0.055	0.081	0.179	0.037	0.06	0.068	0.067	0.152	0.039	0.158	0.026	0.078	1
水2010	0.035	0.053	0.053	0.111	0.055	0.068	0.092	0.078	0.045	0.201	0.121	0.088	1
水2011	0.097	0.049	0.05	0.026	0.081	0.08	0.032	0.059	0.051	0.052	0.333	0.09	1
水2012	0.064	0.071	0.031	0.044	0.235	0.099	0.048	0.083	0.059	0.067	0.137	0.061	1
水2013	0.032	0.042	0.043	0.042	0.054	0.11	0.052	0.098	0.304	0.06	0.103	0.06	1
水2014	0.043	0.059	0.053	0.125	0.061	0.124	0.065	0.061	0.033	0.014	0.072	0.29	1
水2015	0.049	0.118	0.145	0.089	0.089	0.128	0.075	0.041	0.022	0.069	0.085	0.089	1
水2016	0.016	0.077	0.03	0.269	0.05	0.101	0.067	0.059	0.049	0.111	0.075	0.094	1
水2017	0.066	0.13	0.054	0.092	0.068	0.077	0.153	0.068	0.059	0.072	0.075	0.084	1
水2018	0.192	0.116	0.05	0.091	0.058	0.07	0.103	0.063	0.037	0.087	0.055	0.076	1
水2019	0.078	0.152	0.05	0.103	0.046	0.07	0.091	0.088	0.04	0.079	0.089	0.113	1
水2020	0.18	0.124	0.056	0.072	0.058	0.073	0.11	0.07	0.04	0.093	0.052	0.072	1

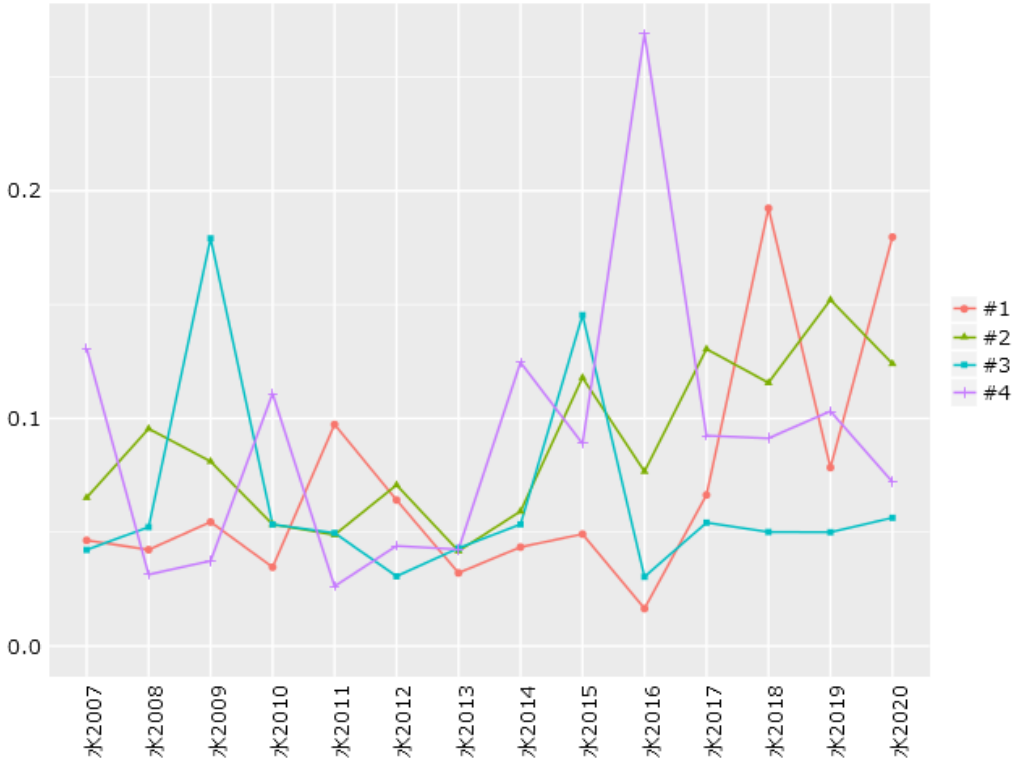


図 4-5 1～4 トピックの比率（12 トピックス、分析対象語数：75）

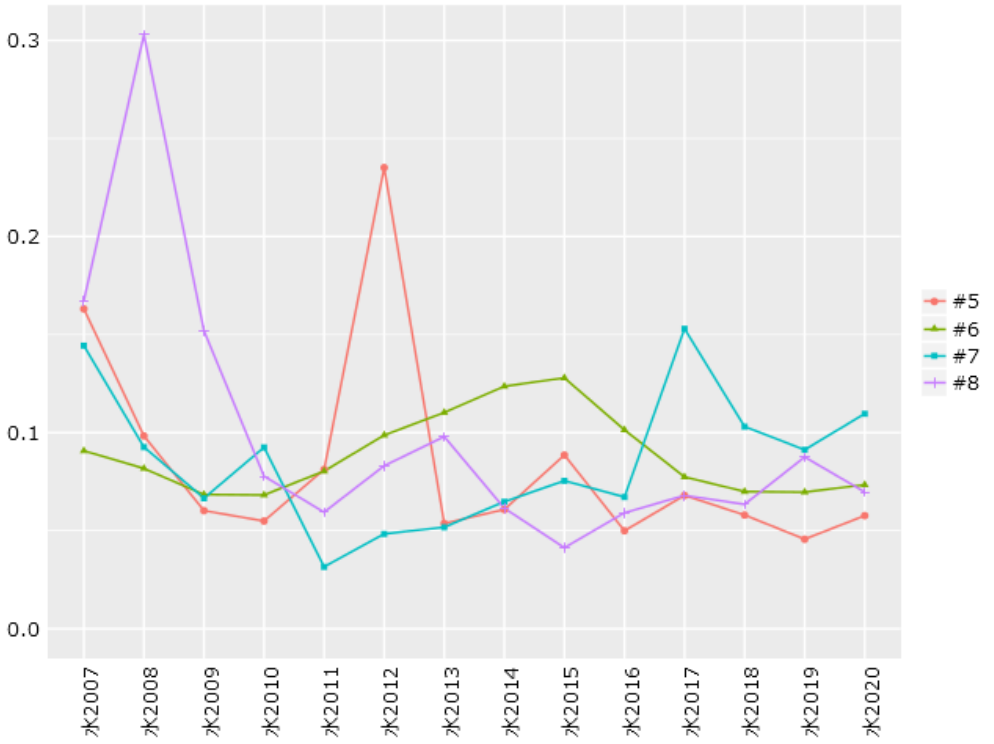


図 4-6 5～8 トピックの比率（12 トピックス、分析対象語数：75）

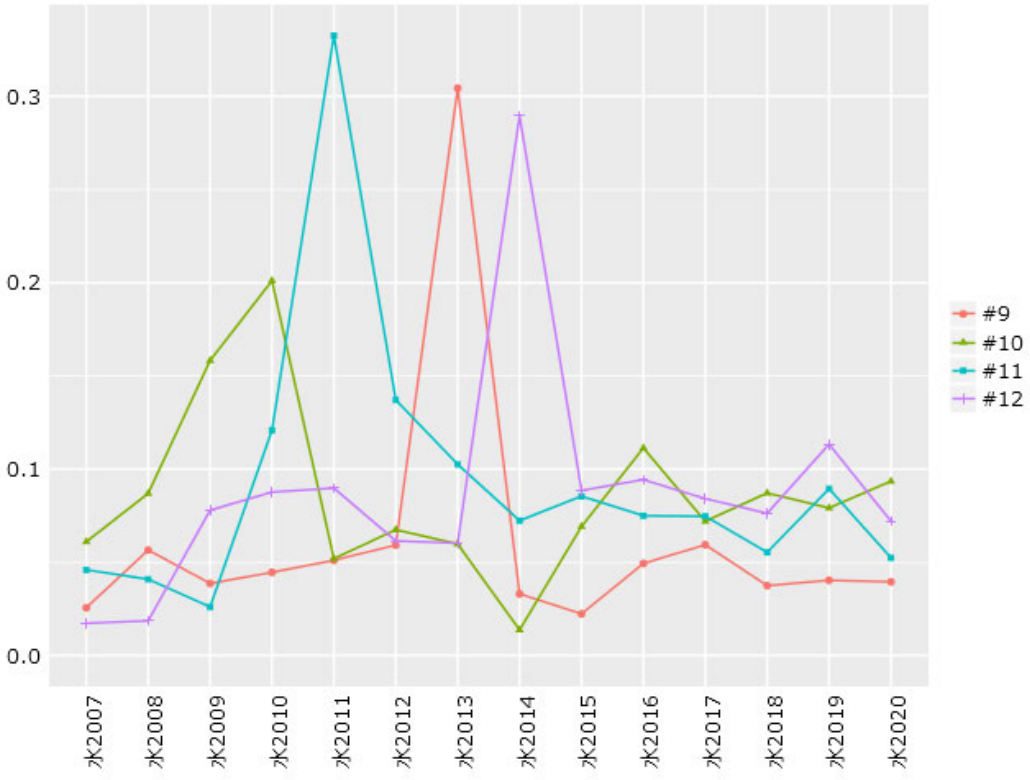


図 4-7 9～12 トピックの比率（12 トピックス、分析対象語数：75）

表 4-3 LDA 処理結果（20 トピックス、分析対象語数：75）

Topics

#1		#2		#3		#4		#5		#6		#7	
水産物	0.197	向上	0.133	漁村	0.280	管理	0.187	経営	0.200	資源	0.338	重要	0.154
消費者	0.127	拡大	0.077	産業	0.121	漁船	0.106	漁船	0.164	漁獲	0.162	実施	0.131
商品	0.112	漁業者	0.076	経営	0.089	機関	0.098	管理	0.127	管理	0.155	活動	0.099
販売	0.105	減少	0.071	多く	0.072	増加	0.075	生産	0.069	マグロ	0.089	必要	0.093
消費	0.093	食品	0.057	高い	0.057	水域	0.073	対象	0.061	漁獲量	0.053	確保	0.064
食品	0.040	漁獲量	0.057	漁協	0.055	措置	0.064	結果	0.051	水産物	0.035	流通	0.055
措置	0.039	加工	0.056	管理	0.052	対象	0.049	必要	0.049	情報	0.030	操業	0.054
魚介類	0.039	機能	0.055	生産量	0.047	太平洋	0.043	水産	0.027	発生	0.025	開発	0.047
向上	0.031	状況	0.046	操業	0.046	影響	0.042	輸出	0.027	水域	0.017	輸出	0.037
連携	0.028	漁獲	0.046	加工	0.036	結果	0.040	市場	0.024	実施	0.017	消費	0.034

#8		#9		#10		#11		#12		#13		#14	
利用	0.141	水産	0.289	情報	0.144	調査	0.147	就業	0.172	漁獲	0.260	漁場	0.200
価格	0.138	水産業	0.124	技術	0.126	増加	0.121	資源	0.114	魚種	0.151	利用	0.086
生産	0.101	漁協	0.075	減少	0.119	対策	0.085	日本	0.102	操業	0.128	消費	0.076
水域	0.074	食品	0.066	漁業者	0.112	高い	0.075	利用	0.098	導入	0.090	多い	0.075
影響	0.068	必要	0.065	漁船	0.098	操業	0.074	活動	0.098	水産物	0.055	魚種	0.073
輸入	0.059	実施	0.060	開発	0.081	マグロ	0.067	産業	0.068	資源管理	0.036	生産	0.066
重要	0.058	連携	0.058	向上	0.043	関係	0.065	多い	0.062	漁業者	0.036	多く	0.057
状況	0.052	生産	0.043	機能	0.043	規制	0.050	漁獲量	0.051	販売	0.033	規制	0.050
技術	0.051	支援	0.035	状況	0.041	多く	0.045	価格	0.038	関係	0.027	魚介類	0.049
世界	0.050	施設	0.033	関係	0.030	加工	0.040	連携	0.038	対象	0.025	水産	0.048

#15		#16		#17		#18		#19		#20	
養殖	0.357	漁協	0.166	水産資源	0.209	被害	0.103	水産物	0.186	生産量	0.151
養殖業	0.127	市場	0.150	資源管理	0.145	支援	0.093	減少	0.144	漁船	0.131
水産物	0.114	消費者	0.134	漁業者	0.141	漁港	0.076	増加	0.128	資源管理	0.117
大きい	0.057	結果	0.099	世界	0.130	水産	0.075	水産業	0.086	措置	0.100
発生	0.040	輸入	0.067	連携	0.057	漁船	0.067	流通	0.062	漁業者	0.089
操業	0.039	魚介類	0.051	水産	0.044	水産物	0.064	経営	0.055	減少	0.073
生産量	0.038	活動	0.040	水産物	0.024	加工	0.062	影響	0.050	関係	0.059
日本	0.025	加工	0.039	消費	0.022	施設	0.057	生産	0.045	実施	0.045
必要	0.024	水産物	0.037	漁協	0.022	養殖業	0.054	販売	0.041	大きい	0.041
マグロ	0.024	漁業者	0.034	実施	0.022	関係	0.049	開発	0.027	規制	0.035

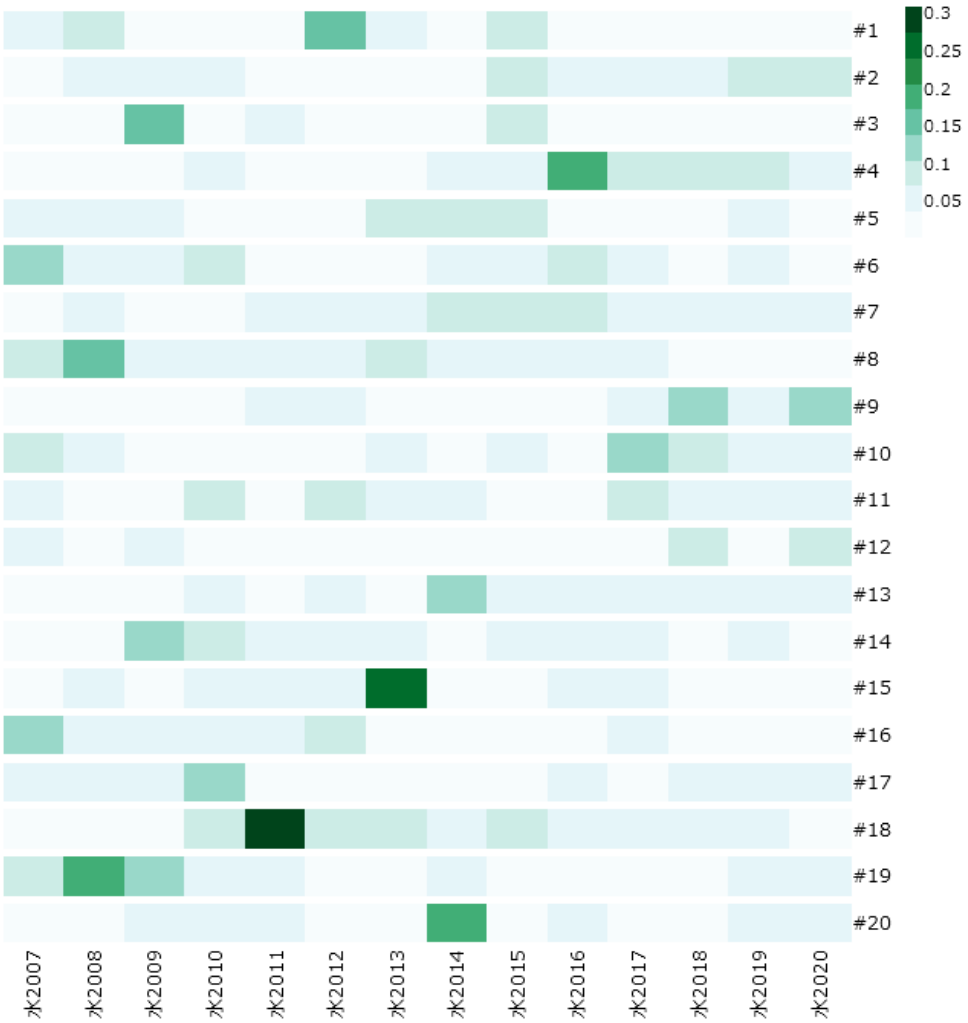


図 4-8 LDA ヒートマップ（20 トピックス、分析対象語数：75）

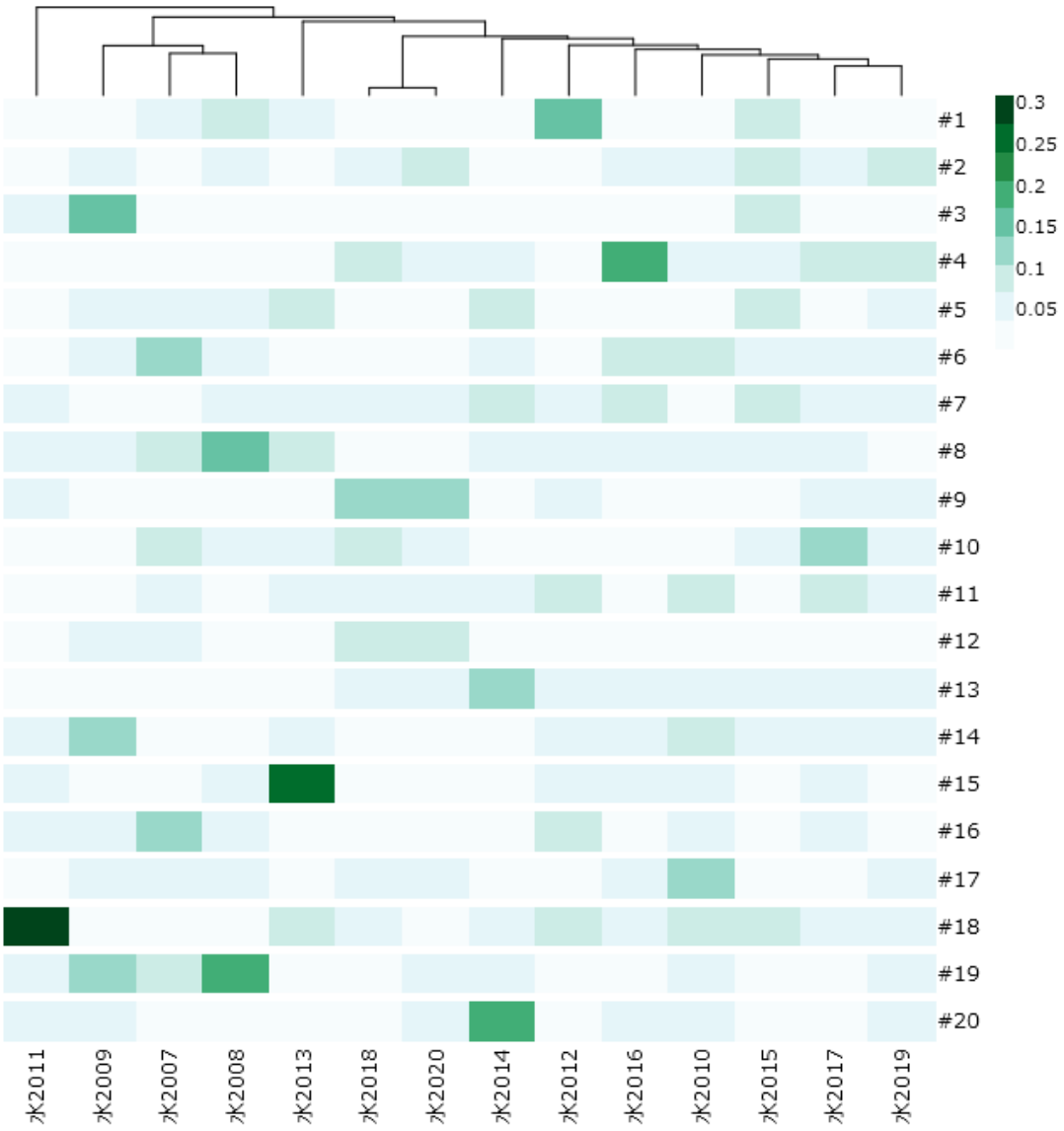


図 4-9　LDA ヒートマップ樹形図（20 トピックス、分析対象語数：75）

表 4-4 トピック比率集計表（20 トピックス、分析対象語数：75）

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	
水2007	0.062	0.015	0.027	0.03	0.057	0.121	0.038	0.079	0.028	0.088	
水2008	0.092	0.049	0.025	0.023	0.04	0.04	0.048	0.141	0.005	0.044	
水2009	0.035	0.058	0.146	0.017	0.041	0.067	0.018	0.061	0.011	0.036	
水2010	0.022	0.04	0.026	0.041	0.025	0.079	0.031	0.055	0.016	0.036	
水2011	0.019	0.032	0.048	0.01	0.006	0.02	0.057	0.052	0.066	0.012	
水2012	0.144	0.029	0.023	0.017	0.019	0.027	0.071	0.049	0.044	0.034	
水2013	0.047	0.015	0.026	0.027	0.082	0.03	0.06	0.072	0.025	0.041	
水2014	0.032	0.02	0.031	0.049	0.08	0.047	0.076	0.054	0.027	0.02	
水2015	0.074	0.073	0.098	0.053	0.081	0.042	0.097	0.059	0.024	0.04	
水2016	0.037	0.047	0.02	0.183	0.016	0.104	0.076	0.052	0.013	0.015	
水2017	0.037	0.061	0.031	0.074	0.024	0.06	0.052	0.042	0.047	0.126	
水2018	0.031	0.067	0.035	0.081	0.026	0.025	0.051	0.023	0.138	0.076	
水2019	0.024	0.098	0.02	0.079	0.056	0.048	0.062	0.028	0.053	0.043	
水2020	0.02	0.075	0.03	0.067	0.032	0.035	0.046	0.025	0.136	0.069	
	#11	#12	#13	#14	#15	#16	#17	#18	#19	#20	ケース数
水2007	0.046	0.055	0.037	0.016	0.023	0.128	0.038	0.035	0.073	0.004	1
水2008	0.035	0.037	0.016	0.024	0.042	0.066	0.048	0.028	0.179	0.017	1
水2009	0.015	0.052	0.013	0.109	0.027	0.055	0.043	0.024	0.119	0.053	1
水2010	0.084	0.026	0.039	0.085	0.043	0.04	0.118	0.094	0.041	0.058	1
水2011	0.028	0.007	0.03	0.055	0.054	0.053	0.02	0.309	0.07	0.051	1
水2012	0.094	0.033	0.039	0.049	0.047	0.086	0.029	0.105	0.027	0.034	1
水2013	0.045	0.021	0.03	0.052	0.245	0.021	0.034	0.072	0.019	0.035	1
水2014	0.056	0.028	0.113	0.034	0.028	0.023	0.012	0.041	0.04	0.19	1
水2015	0.032	0.03	0.071	0.039	0.018	0.021	0.037	0.076	0.019	0.017	1
水2016	0.036	0.03	0.063	0.057	0.04	0.03	0.055	0.049	0.013	0.063	1
水2017	0.074	0.035	0.057	0.051	0.049	0.042	0.03	0.055	0.023	0.028	1
水2018	0.054	0.083	0.053	0.035	0.032	0.03	0.05	0.04	0.032	0.036	1
水2019	0.042	0.036	0.058	0.064	0.024	0.028	0.04	0.069	0.063	0.066	1
水2020	0.047	0.091	0.059	0.03	0.037	0.031	0.049	0.034	0.043	0.044	1

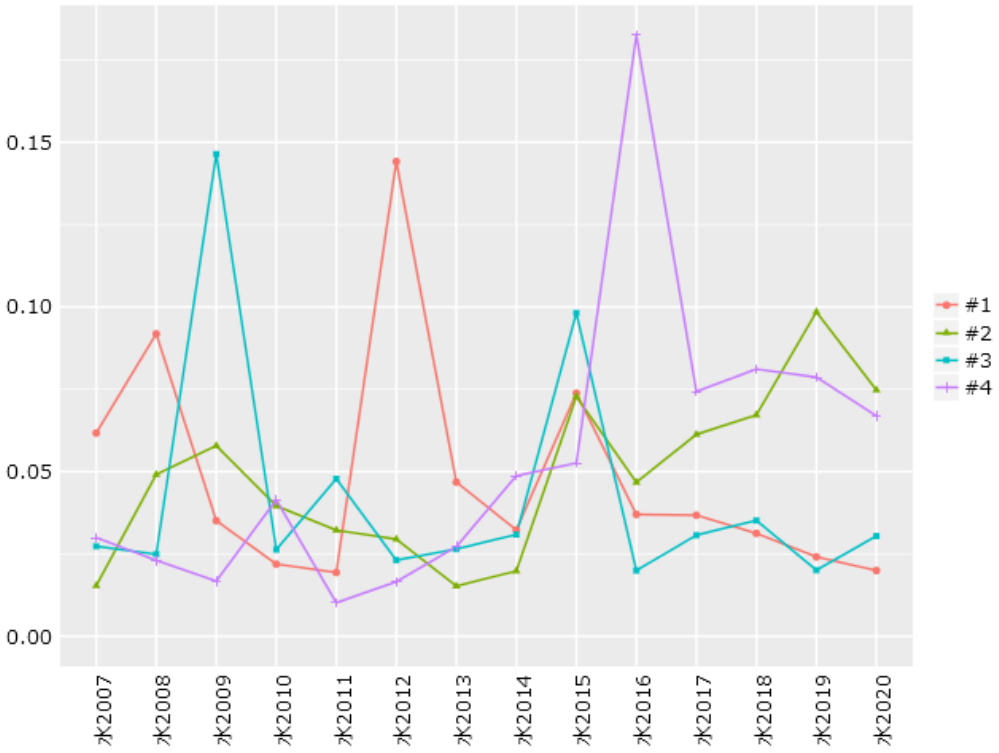


図 4-10 1～4 トピックの比率（20 トピックス、分析対象語数：75）

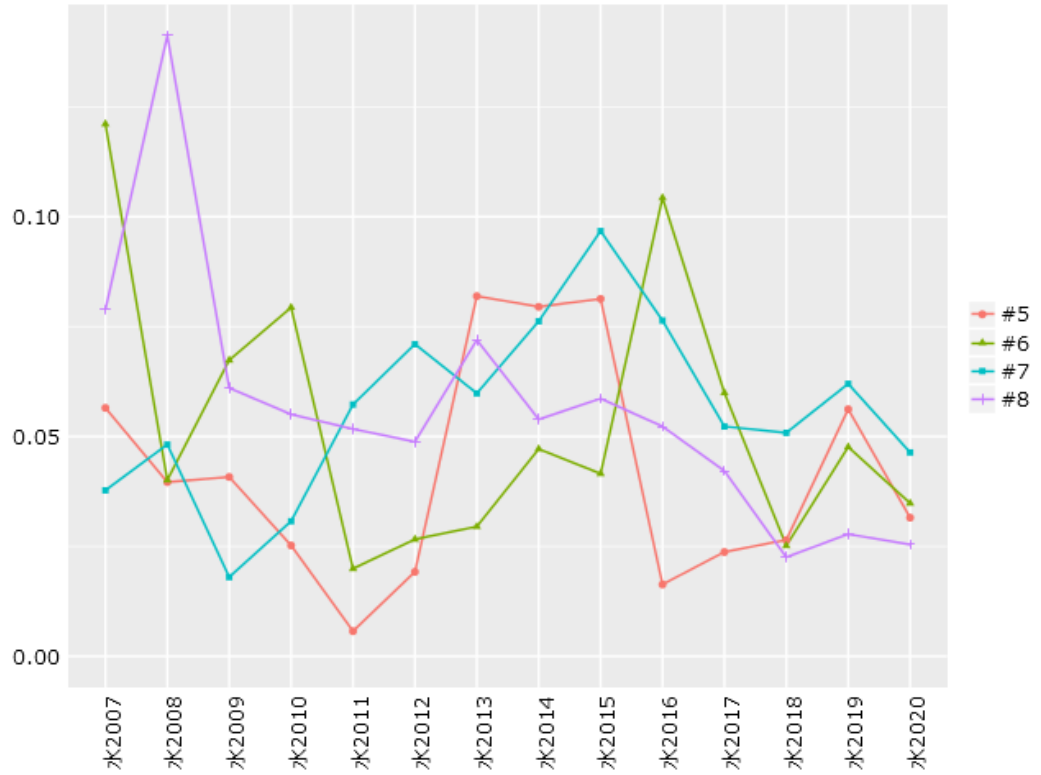


図 4-11　5～8 トピックの比率（20 トピックス、分析対象語数：75）

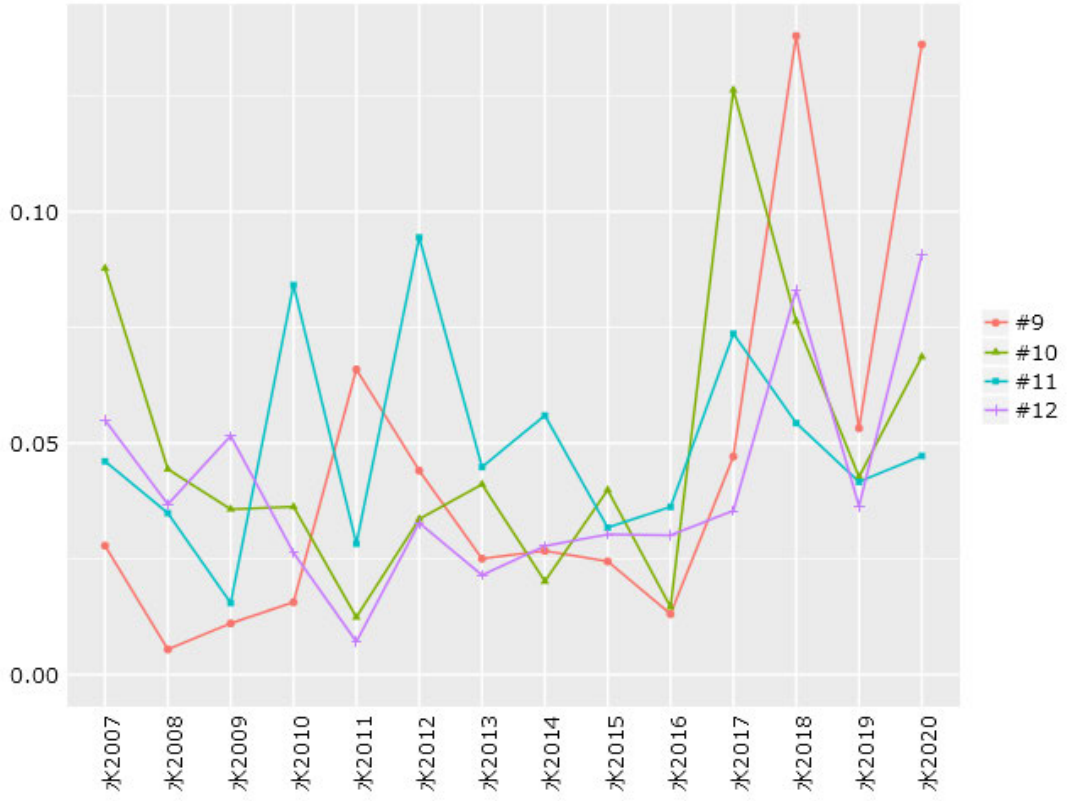


図 4-12　9～12 トピックの比率（20 トピックス、分析対象語数：75）

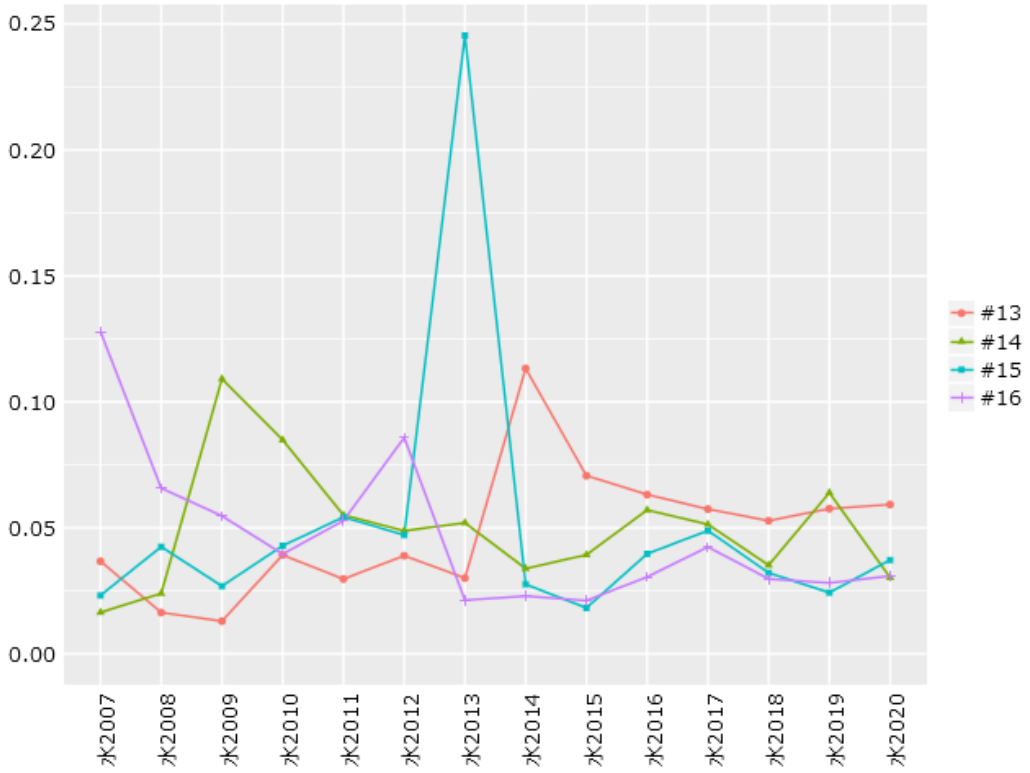


図 4-13 13～16 トピックの比率（20 トピックス、分析対象語数：75）

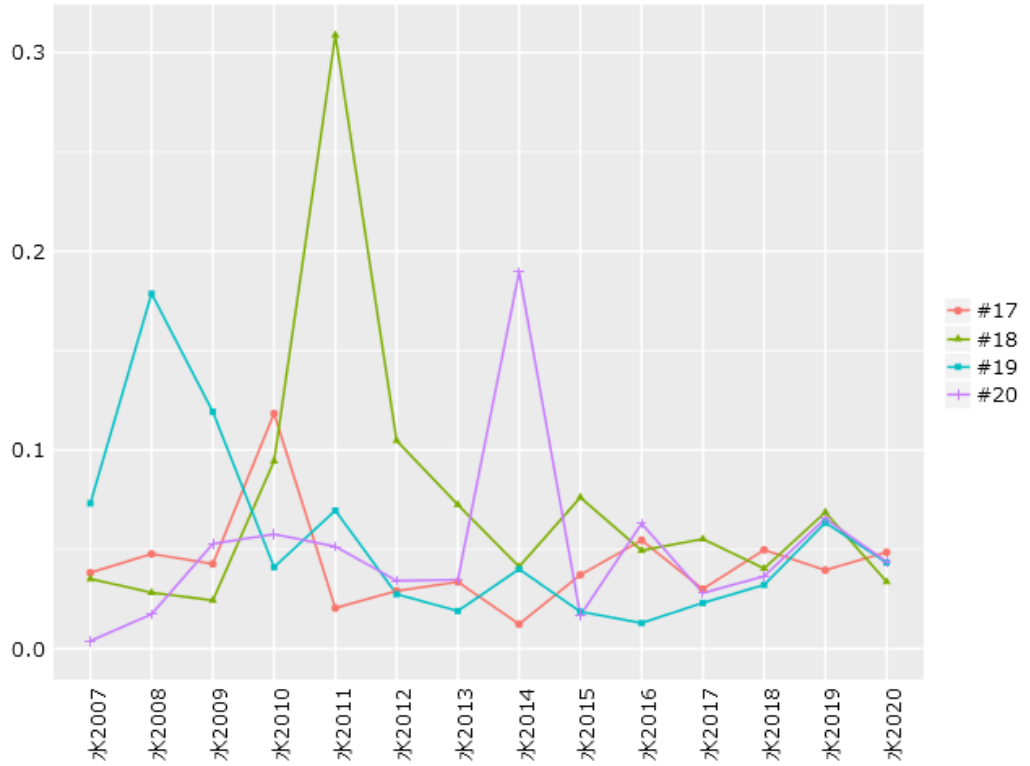


図 4-14 17～20 トピックの比率（20 トピックス、分析対象語数：75）

表 4-5 LDA 処理結果（12 トピックス、分析対象語数：157）

Topics

#1		#2		#3		#4		#5		#6	
実施	0.077	水産	0.075	管理	0.102	漁村	0.093	養殖	0.218	水産資源	0.082
操業	0.074	被害	0.067	資源	0.082	経営	0.083	養殖業	0.071	漁業者	0.074
漁獲	0.053	支援	0.057	漁獲	0.074	重要	0.048	種苗	0.036	資源管理	0.058
重要	0.049	復興	0.054	機関	0.044	産業	0.046	大きい	0.035	生産	0.054
対象	0.037	養殖業	0.044	水域	0.038	生産	0.041	水産物	0.034	世界	0.046
確保	0.033	漁港	0.040	マグロ	0.035	地元	0.036	技術	0.028	消費	0.046
関係	0.031	漁船	0.033	漁獲量	0.034	消費	0.034	生産	0.025	魚介類	0.040
活動	0.031	加工	0.032	増加	0.031	操業	0.033	経営	0.025	漁場	0.036
施設	0.029	施設	0.029	回復	0.029	供給	0.033	生産量	0.024	購入	0.036
調査	0.029	状況	0.026	生物	0.029	利用	0.031	漁場	0.023	割合	0.034

#7		#8		#9		#10		#11		#12	
減少	0.109	水産物	0.138	管理	0.070	漁船	0.098	漁船	0.081	水産	0.149
増加	0.105	消費者	0.100	向上	0.055	生産量	0.078	資源	0.074	実習	0.062
水産物	0.085	販売	0.054	拡大	0.044	資源管理	0.069	情報	0.064	水産業	0.057
流通	0.065	漁協	0.046	機能	0.039	漁業者	0.053	期待	0.055	連携	0.047
影響	0.056	商品	0.042	加工	0.038	魚種	0.042	状況	0.052	漁協	0.047
水産業	0.045	魚介類	0.038	対象	0.035	制度	0.041	技術	0.051	就業	0.046
連携	0.041	調理	0.038	協定	0.033	減少	0.036	開発	0.045	研究	0.039
生産	0.037	利用	0.035	都道府県	0.032	措置	0.031	漁業者	0.043	漁業者	0.038
対策	0.034	結果	0.032	女性	0.030	導入	0.029	減少	0.039	必要	0.034
輸入	0.031	市場	0.028	整備	0.028	昭和	0.029	活用	0.038	産業	0.033

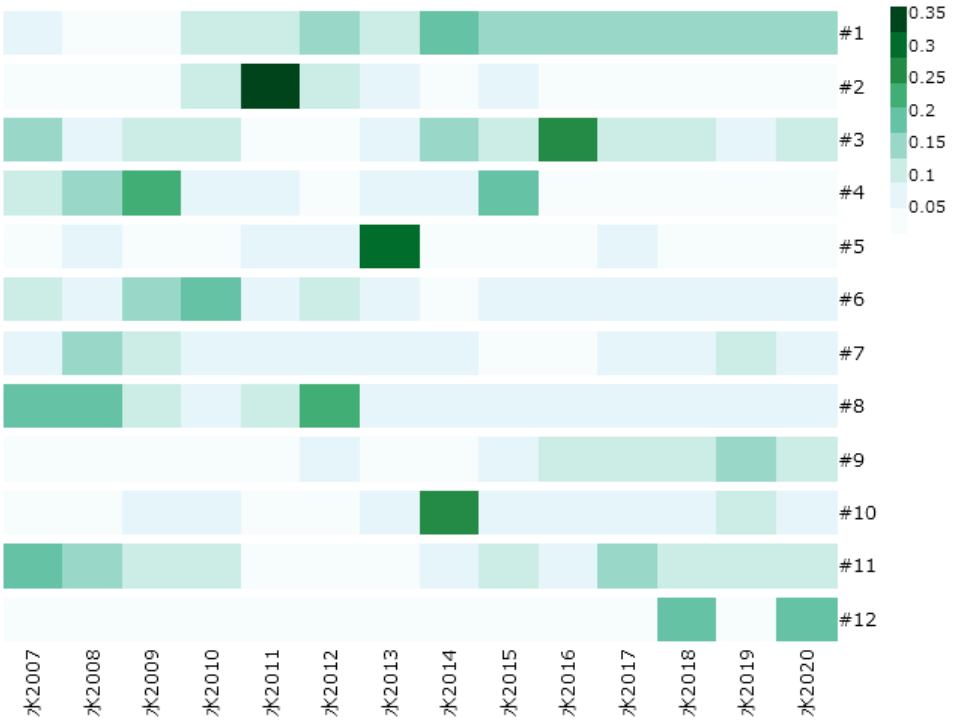


図 4-15 LDA ヒートマップ（12 トピックス、分析対象語数：157）

図 4-16 LDA ヒートマップ樹形図（12 トピックス、分析対象語数：157）

表 4-6 トピック比率集計表（12 トピックス、分析対象語数：157）

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	ケース数
水2007	0.086	0.029	0.154	0.099	0.028	0.097	0.068	0.194	0.03	0.012	0.172	0.032	1
水2008	0.048	0.014	0.073	0.145	0.052	0.066	0.162	0.196	0.042	0.043	0.133	0.027	1
水2009	0.036	0.017	0.122	0.211	0.043	0.143	0.091	0.095	0.049	0.063	0.09	0.039	1
水2010	0.125	0.109	0.105	0.063	0.047	0.201	0.066	0.075	0.035	0.064	0.089	0.02	1
水2011	0.106	0.36	0.048	0.081	0.054	0.058	0.089	0.092	0.024	0.041	0.031	0.016	1
水2012	0.162	0.109	0.04	0.04	0.061	0.099	0.075	0.232	0.072	0.028	0.046	0.036	1
水2013	0.116	0.074	0.071	0.068	0.29	0.054	0.062	0.085	0.05	0.073	0.034	0.023	1
水2014	0.181	0.034	0.134	0.069	0.038	0.028	0.056	0.074	0.045	0.254	0.058	0.028	1
水2015	0.158	0.066	0.1	0.182	0.023	0.056	0.04	0.088	0.069	0.083	0.092	0.044	1
水2016	0.163	0.043	0.255	0.035	0.04	0.076	0.044	0.062	0.095	0.08	0.081	0.026	1
水2017	0.16	0.041	0.095	0.034	0.055	0.081	0.072	0.085	0.124	0.051	0.156	0.045	1
水2018	0.151	0.039	0.111	0.037	0.031	0.06	0.073	0.058	0.1	0.056	0.112	0.171	1
水2019	0.137	0.041	0.088	0.036	0.033	0.085	0.11	0.059	0.149	0.117	0.1	0.045	1
水2020	0.146	0.03	0.098	0.032	0.041	0.069	0.072	0.062	0.1	0.063	0.104	0.184	1

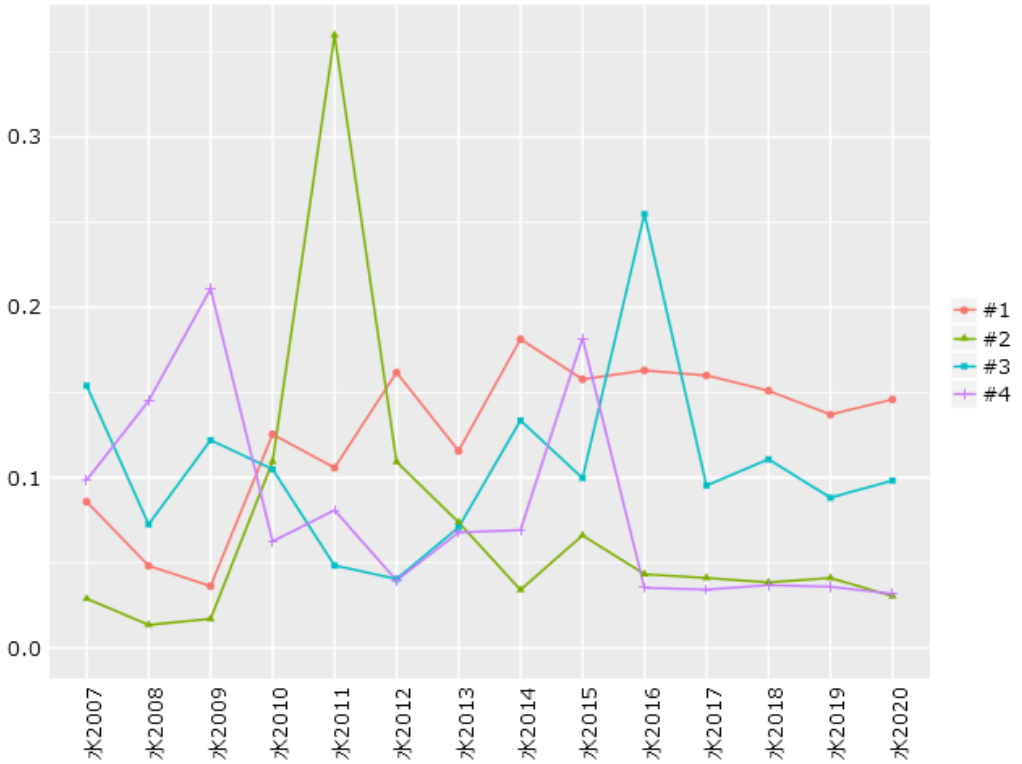


図 4-17 1～4 トピックの比率（12 トピックス、分析対象語数：157）

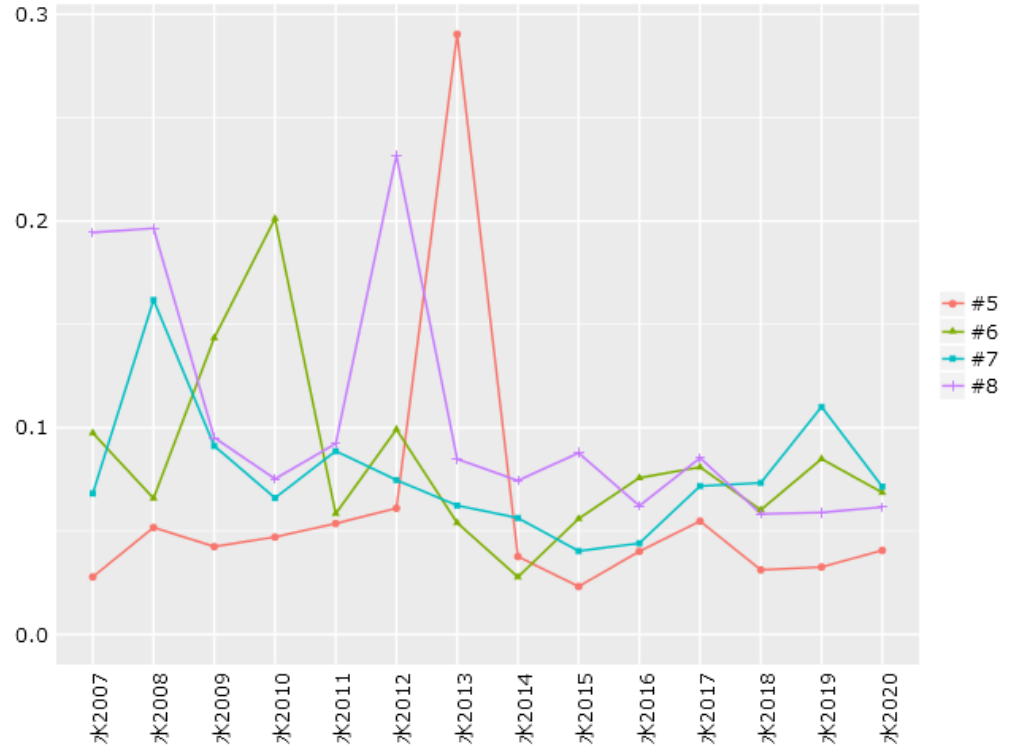


図 4-18 5～8 トピックの比率（12 トピックス、分析対象語数：157）

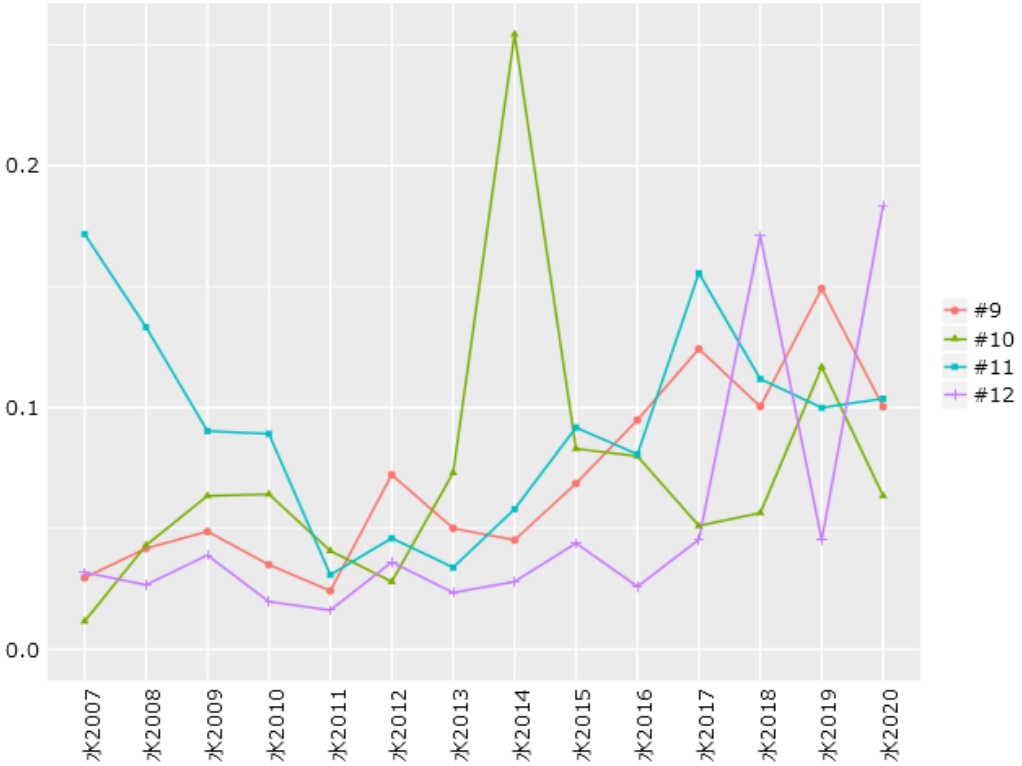


図 4-19 9～12 トピックの比率（12 トピックス、分析対象語数：157）

表 4-7 LDA 処理結果（20 トピックス、分析対象語数：157）

Topics

#1		#2		#3		#4		#5		#6		#7	
必要	0.145	漁村	0.157	減少	0.131	生産	0.106	管理	0.143	経営	0.220	漁獲	0.150
利用	0.083	漁獲	0.065	水産業	0.099	対策	0.059	資源	0.108	減少	0.075	資源管理	0.149
重要	0.082	漁業者	0.058	水産	0.079	発生	0.056	機関	0.079	管理	0.065	規制	0.059
操業	0.075	産業	0.057	漁協	0.061	食品	0.050	水域	0.062	水産物	0.061	大きい	0.039
確保	0.061	地元	0.052	対応	0.046	消費	0.039	中心	0.043	漁労	0.054	漁場	0.039
場合	0.041	向上	0.051	多い	0.046	設置	0.039	生物	0.042	昭和	0.038	影響	0.037
開催	0.040	漁場	0.046	所得	0.033	策定	0.037	太平洋	0.041	漁船	0.033	適切	0.035
多く	0.037	消費	0.038	サケ	0.029	計画	0.037	漁獲量	0.040	魚種	0.032	海域	0.035
価格	0.037	重要	0.035	輸出	0.028	整備	0.035	カツオ	0.037	拡大	0.031	漁獲量	0.030
産地	0.033	高い	0.027	傾向	0.027	関係	0.035	中国	0.036	加工	0.026	割合	0.026

#8		#9		#10		#11		#12		#13	
養殖	0.298	水産	0.109	水産資源	0.123	魚介類	0.116	水産物	0.130	生産量	0.158
養殖業	0.106	被害	0.101	漁業者	0.099	利用	0.057	増加	0.112	漁業者	0.113
種苗	0.051	支援	0.088	資源	0.082	マグロ	0.054	生産	0.092	制度	0.083
水産物	0.047	復興	0.088	保全	0.064	多い	0.053	供給	0.078	TAC	0.044
大きい	0.039	漁港	0.057	利用	0.060	高い	0.047	影響	0.065	資源	0.043
生産	0.036	漁船	0.051	世界	0.050	多く	0.044	状況	0.064	導入	0.041
生産量	0.034	水産物	0.050	マグロ	0.038	購入	0.042	価格	0.060	水準	0.041
品質	0.020	加工	0.049	周辺	0.036	消費	0.034	役割	0.038	販売	0.041
高い	0.020	施設	0.039	変化	0.035	水域	0.033	回復	0.031	水産物	0.040
漁場	0.019	養殖業	0.036	就業	0.034	漁協	0.033	輸入	0.030	関係	0.037

#14		#15		#16		#17		#18		#19		#20	
向上	0.082	情報	0.120	関係	0.087	漁船	0.166	資源	0.118	水産物	0.152	水産	0.152
増加	0.072	状況	0.061	実施	0.086	操業	0.069	期待	0.091	消費者	0.146	実習	0.070
加工	0.056	消費者	0.060	活動	0.076	措置	0.068	管理	0.077	販売	0.072	連携	0.054
水産物	0.053	技術	0.059	福島	0.055	マグロ	0.065	活動	0.056	商品	0.071	水産業	0.052
機能	0.049	漁業者	0.057	調査	0.052	魚種	0.055	漁獲	0.049	漁協	0.065	就業	0.047
確保	0.042	開発	0.054	太平洋	0.045	増加	0.049	制度	0.037	料理	0.048	研究	0.047
流通	0.040	調査	0.050	販売	0.044	結果	0.045	計画	0.033	結果	0.044	漁業者	0.045
拡大	0.036	活用	0.042	女性	0.033	管理	0.045	開発	0.032	調理	0.037	食品	0.042
必要	0.035	課題	0.039	状況	0.031	実施	0.044	日本	0.028	加工	0.031	漁協	0.038
支援	0.032	漁獲量	0.029	出荷	0.028	減少	0.037	役割	0.026	市場	0.031	産業	0.026

図 4-20 LDA ヒートマップ（20 トピックス、分析対象語数：157）

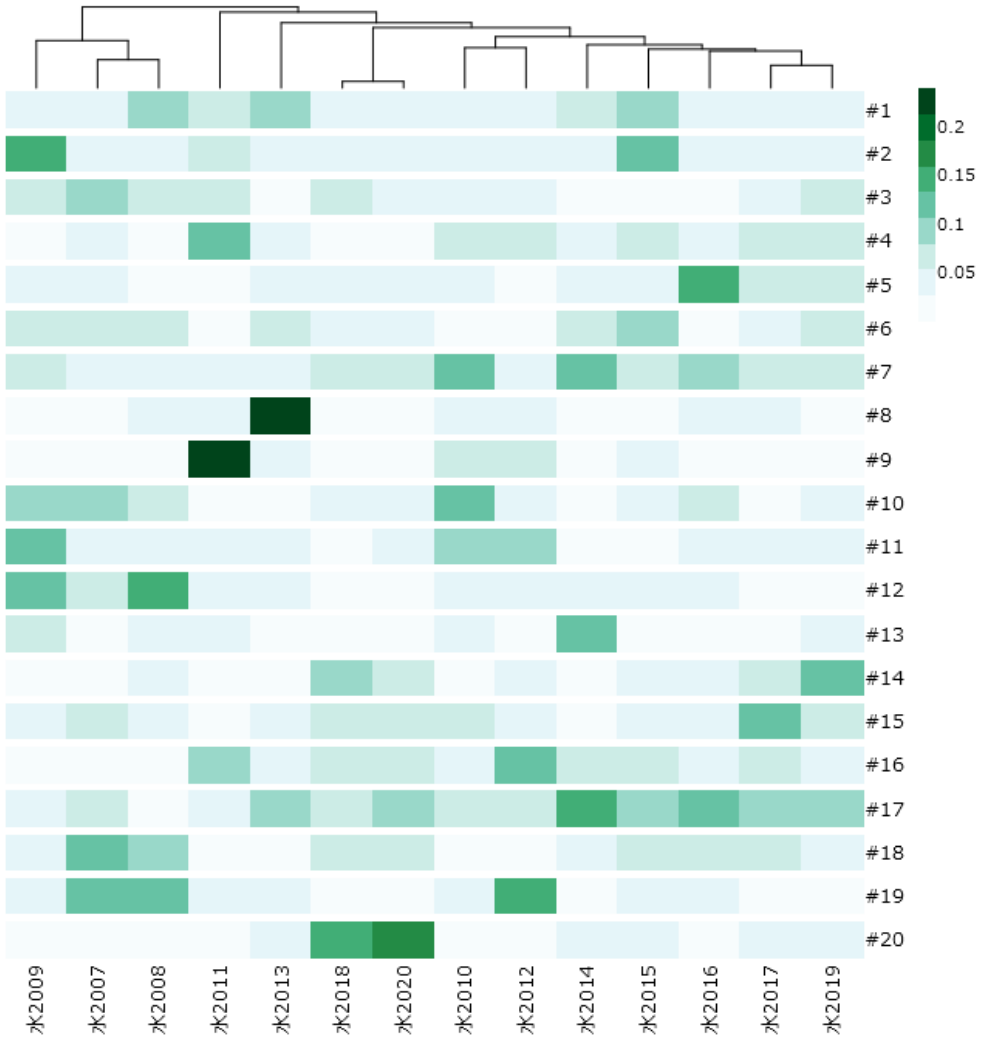


図 4-21 LDA ヒートマップ樹形図（20 トピックス、分析対象語数：157）

表 4-8 トピック比率集計表（20 トピックス、分析対象語数：157）

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	
水2007	0.05	0.033	0.082	0.052	0.053	0.061	0.033	0.016	0.01	0.092	
水2008	0.082	0.051	0.074	0.026	0.018	0.061	0.032	0.031	0.005	0.063	
水2009	0.032	0.146	0.066	0.016	0.05	0.061	0.056	0.02	0.006	0.082	
水2010	0.047	0.046	0.043	0.064	0.034	0.024	0.131	0.03	0.07	0.112	
水2011	0.056	0.055	0.066	0.11	0.023	0.021	0.037	0.033	0.239	0.025	
水2012	0.053	0.031	0.044	0.076	0.021	0.019	0.042	0.041	0.064	0.038	
水2013	0.082	0.034	0.018	0.047	0.04	0.072	0.039	0.221	0.041	0.01	
水2014	0.08	0.031	0.019	0.034	0.053	0.072	0.123	0.023	0.022	0.014	
水2015	0.086	0.111	0.009	0.055	0.04	0.085	0.063	0.021	0.032	0.049	
水2016	0.053	0.03	0.013	0.047	0.152	0.024	0.103	0.032	0.02	0.066	
水2017	0.043	0.046	0.052	0.065	0.055	0.029	0.073	0.04	0.025	0.022	
水2018	0.051	0.046	0.058	0.019	0.047	0.038	0.066	0.024	0.018	0.041	
水2019	0.033	0.028	0.07	0.064	0.058	0.058	0.08	0.024	0.026	0.047	
水2020	0.047	0.038	0.052	0.027	0.048	0.03	0.074	0.023	0.017	0.036	
	#11	#12	#13	#14	#15	#16	#17	#18	#19	#20	ケース数
水2007	0.045	0.063	0.006	0.017	0.059	0.007	0.069	0.125	0.114	0.01	1
水2008	0.033	0.144	0.037	0.031	0.053	0.001	0.019	0.096	0.121	0.023	1
水2009	0.118	0.115	0.054	0.014	0.029	0.012	0.036	0.041	0.035	0.011	1
水2010	0.082	0.044	0.037	0.021	0.059	0.037	0.062	0.014	0.034	0.007	1
水2011	0.052	0.053	0.03	0.012	0.019	0.094	0.029	0.014	0.03	0.003	1
水2012	0.085	0.03	0.021	0.038	0.041	0.111	0.063	0.021	0.14	0.022	1
水2013	0.033	0.045	0.022	0.021	0.045	0.05	0.089	0.018	0.043	0.031	1
水2014	0.015	0.038	0.132	0.022	0.026	0.061	0.147	0.032	0.027	0.028	1
水2015	0.017	0.031	0.02	0.03	0.038	0.075	0.086	0.059	0.049	0.043	1
水2016	0.04	0.029	0.022	0.052	0.032	0.05	0.129	0.055	0.028	0.025	1
水2017	0.042	0.014	0.024	0.069	0.122	0.062	0.092	0.061	0.024	0.041	1
水2018	0.023	0.02	0.023	0.09	0.058	0.073	0.076	0.062	0.012	0.157	1
水2019	0.033	0.015	0.049	0.111	0.054	0.052	0.101	0.04	0.016	0.041	1
水2020	0.032	0.023	0.024	0.076	0.057	0.07	0.082	0.058	0.024	0.163	1

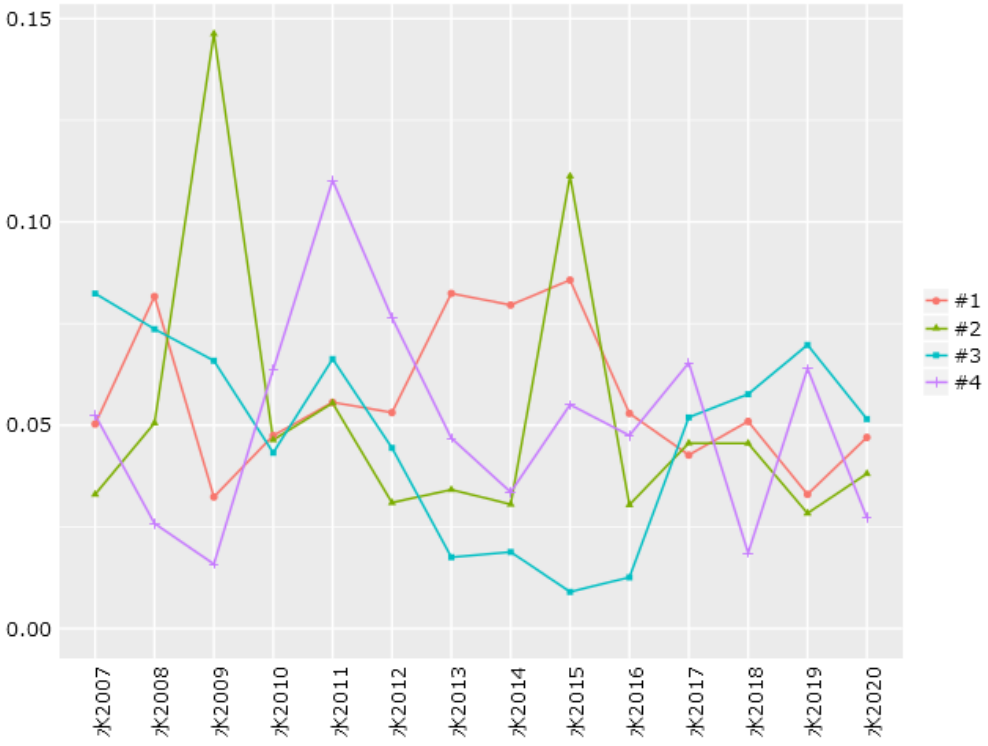


図 4-22 1～4 トピックの比率（20 トピックス、分析対象語数：157）

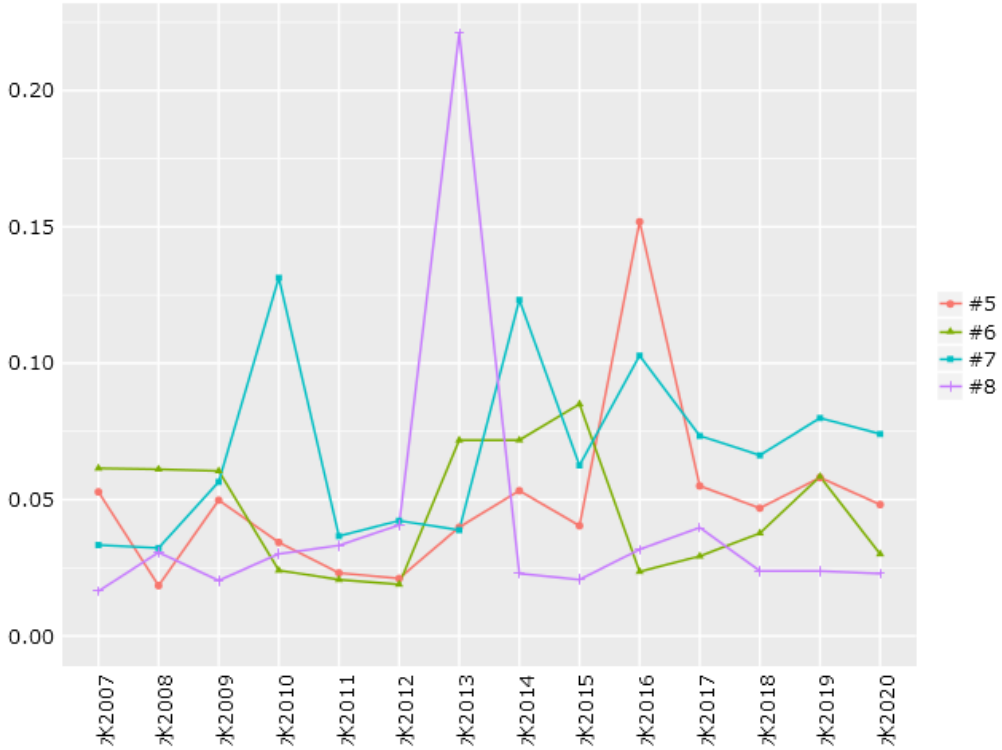


図 4-23 5～8 トピックの比率（20 トピックス、分析対象語数：157）

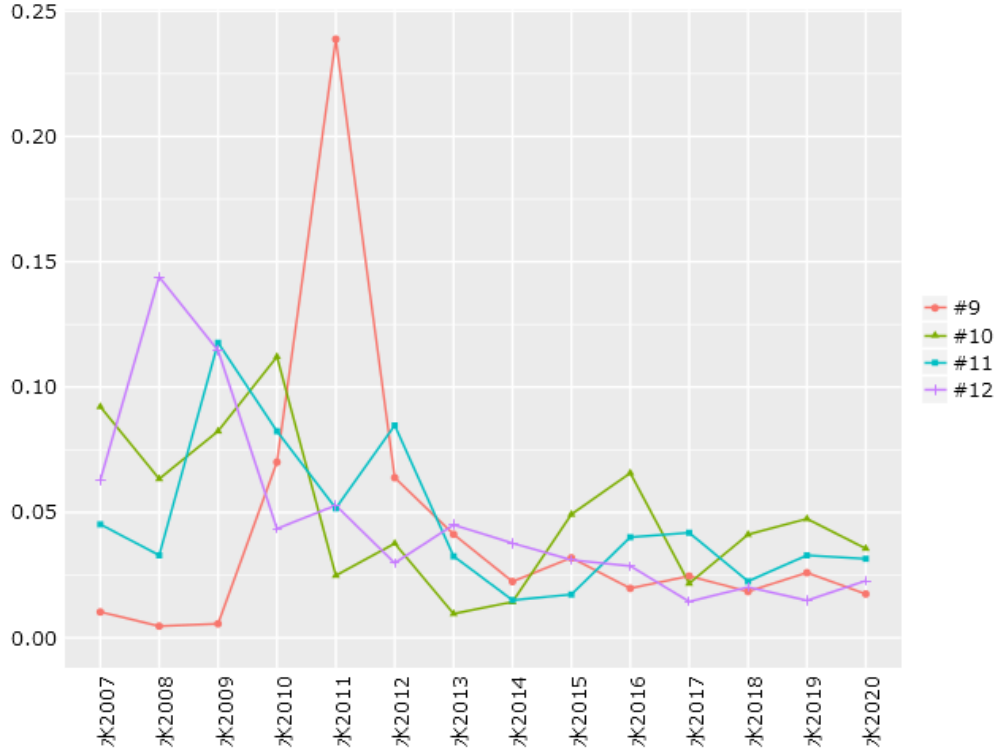


図 4-24 9～12 トピックの比率（20 トピックス、分析対象語数：157）

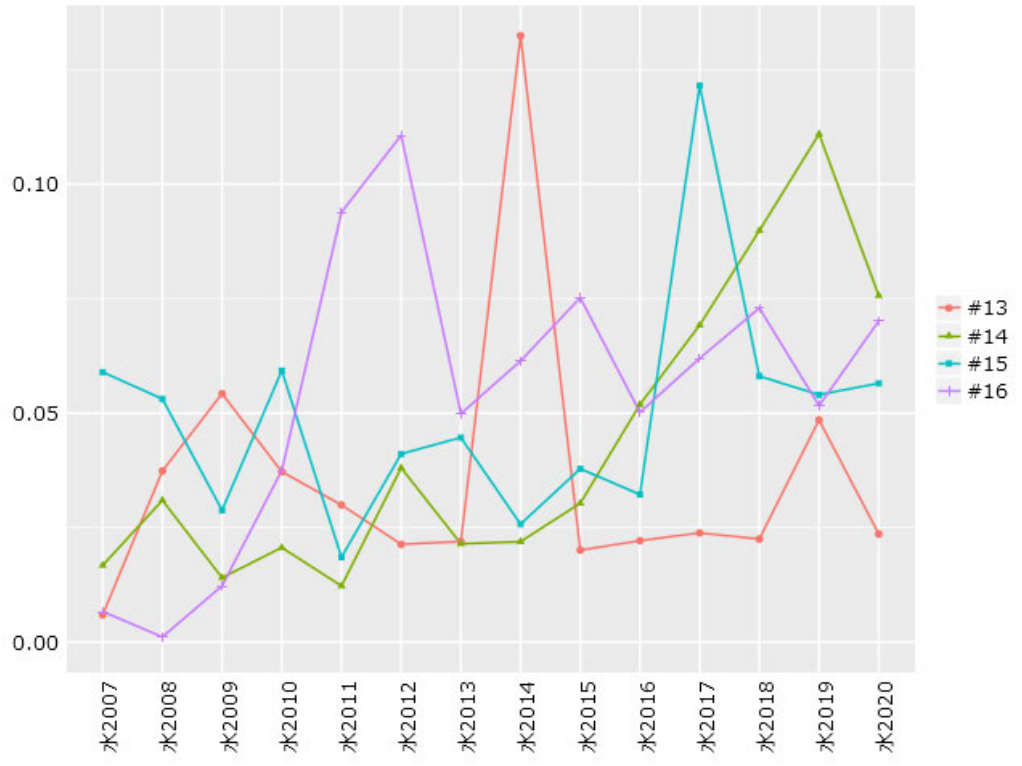


図 4-25 13～16 トピックの比率（20 トピックス、分析対象語数：157）

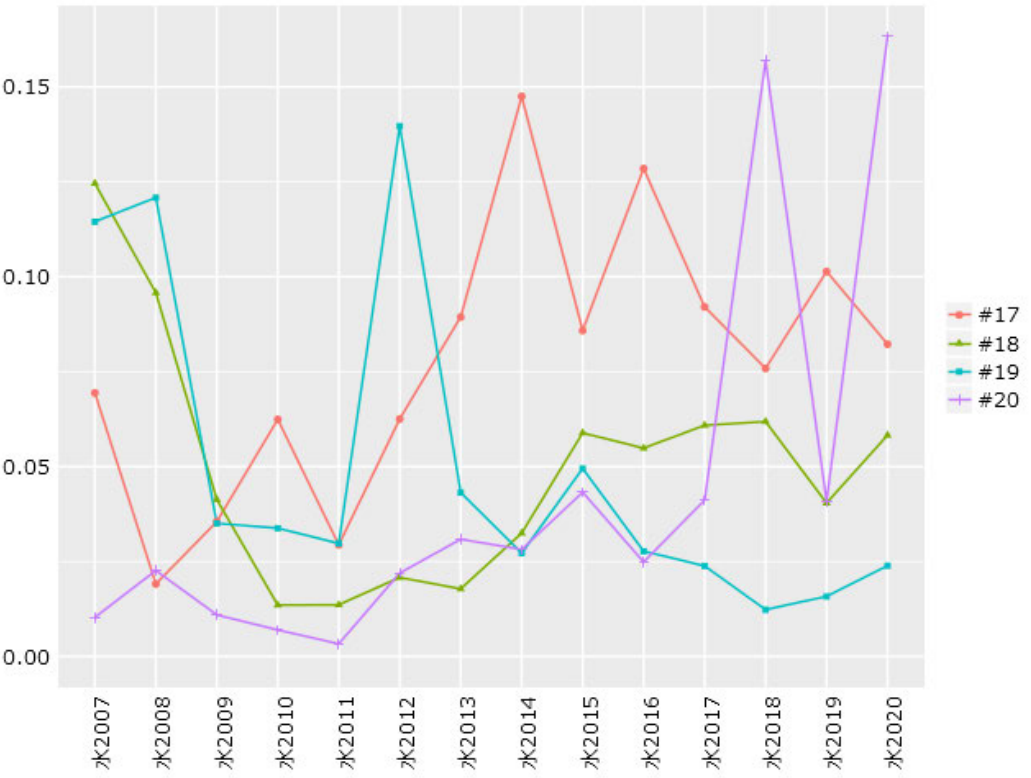


図 4-26 17～20 トピックの比率（20 トピックス、分析対象語数：157）

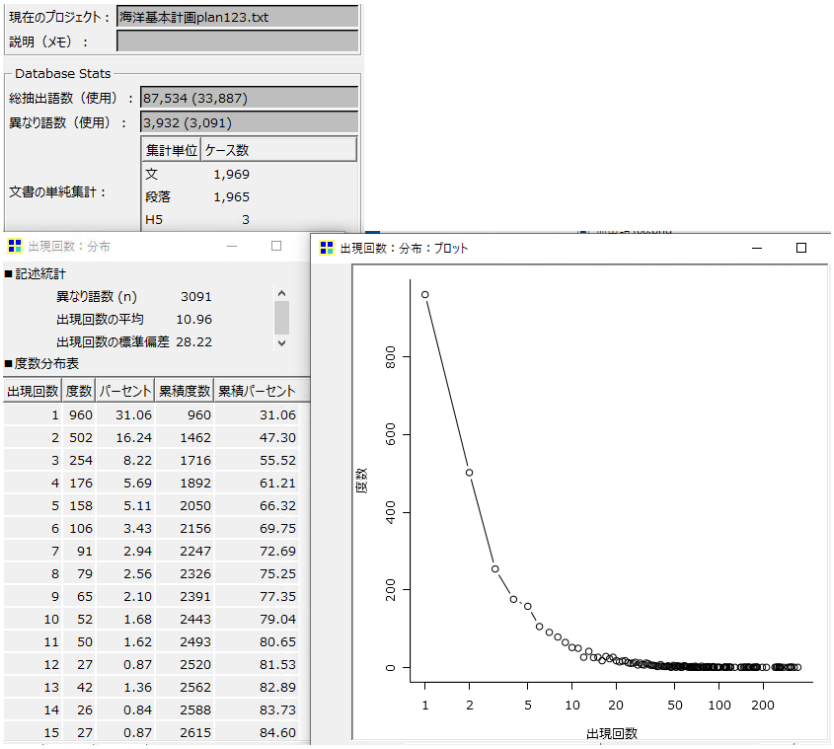
付録 5「海洋基本計画の分析結果」

文書番号：JRDN-21-021

1. 前処理結果

環境白書･海洋白書･水産白書（2008～2020 年）の分析で設定した強制抽出語（131 語）と 54 語の除外語を設定し、「動詞、感動詞、動詞 B、副詞 B」を除外して前処理を実行した。

Chanse での前処理の結果、総抽出語数：87534、異なり語数：3932 のうち 3091 語が分析処理で使用された。抽出語出現数の頻度分布を図 1-1、抽出語リスト（上位 200 語）を図 1 -2 に示す。



出現回数	度数	パーセント	累積度数	累積パーセント
1	960	31.06	960	31.06
2	502	16.24	1462	47.30
3	254	8.22	1716	55.52
4	176	5.69	1892	61.21
5	158	5.11	2050	66.32
6	106	3.43	2156	69.75
7	91	2.94	2247	72.69
8	79	2.56	2326	75.25
9	65	2.10	2391	77.35
10	52	1.68	2443	79.04
11	50	1.62	2493	80.65
12	27	0.87	2520	81.53
13	42	1.36	2562	82.89
14	26	0.84	2588	83.73
15	27	0.87	2615	84.60

図 1-1 抽出語出現数の頻度分布

#	抽出語	品詞/活用	頻度
1	実施	サ変名詞	344
2	情報	名詞	325
3	国際	名詞	312
4	整備	サ変名詞	312
5	強化	サ変名詞	307
6	連携	サ変名詞	303
7	利用	サ変名詞	294
8	保全	サ変名詞	275
9	確保	サ変名詞	264
10	必要	形容動詞	260
11	調査	サ変名詞	258
12	開発	サ変名詞	257
13	関係	サ変名詞	255
14	管理	サ変名詞	249
15	技術	名詞	249
16	資源	名詞	245
17	産業	名詞	240
18	海域	名詞	210
19	活用	サ変名詞	197
20	活動	サ変名詞	182
21	計画	サ変名詞	181
22	船舶	名詞	181
23	重要	形容動詞	179
24	観測	サ変名詞	176
25	体制	名詞	175
26	協力	サ変名詞	172
27	適切	形容動詞	170
28	研究	サ変名詞	162
29	促進	サ変名詞	159
30	対応	サ変名詞	159
31	向上	サ変名詞	158
32	対策	サ変名詞	153
33	検討	サ変名詞	150
34	政策	名詞	150
35	安全	形容動詞	147
36	機関	名詞	142
37	支援	サ変名詞	140
38	状況	名詞	128
39	離島	サ変名詞	120
40	観点	名詞	119
41	基本	名詞	119
42	施設	サ変名詞	116
43	充実	サ変名詞	116
44	基盤	名詞	113
45	津波	名詞	112
46	地球	名詞	111
47	影響	サ変名詞	110
48	分野	名詞	110
49	積極的	タグ	109
50	維持	サ変名詞	108
51	課題	名詞	101
52	国際的	タグ	100
53	社会	名詞	98
54	海上	名詞	97
55	把握	サ変名詞	93
56	エネルギー	名詞	92
57	研究開発	タグ	92
58	実現	サ変名詞	91
59	発生	サ変名詞	91
60	防止	サ変名詞	91
61	国民	名詞	90
62	総合的	タグ	90
63	海岸	名詞	89
64	関連	サ変名詞	86
65	港湾	名詞	86
66	構築	サ変名詞	85
67	育成	サ変名詞	84
68	探査	サ変名詞	84
69	北極	地名	84
70	教育	サ変名詞	83
66	構築	サ変名詞	85
67	育成	サ変名詞	84
68	探査	サ変名詞	84
69	北極	地名	84
70	教育	サ変名詞	83
71	経済	名詞	82
72	システム	名詞	81
73	評価	サ変名詞	81
74	生産	サ変名詞	79
75	安全保障	タグ	78
76	貢献	サ変名詞	76
77	策定	サ変名詞	76
78	提供	サ変名詞	76
79	問題	ナイ形容	76
80	海底	名詞	74
81	機能	サ変名詞	72
82	世界	名詞	71
83	海上輸送	タグ	70
84	共有	サ変名詞	69
85	鉱物	名詞	69
86	収集	サ変名詞	69
87	生態系	タグ	69
88	可能	形容動詞	68
89	人材	名詞	68
90	導入	サ変名詞	67
91	航行	サ変名詞	66
92	地震	名詞	66
93	着実	形容動詞	66
94	形成	サ変名詞	65
95	拡大	サ変名詞	64
96	能力	名詞	64
97	排他的経済水域	タグ	64
98	災害	名詞	63
99	科学技術	タグ	61
100	発展	サ変名詞	61
101	効果的	タグ	60
102	将来	副詞可能	60
103	生物多様性	タグ	60
104	ルール	名詞	59
105	衛星	名詞	59
106	水産資源	タグ	59
107	理解	サ変名詞	59
108	監視	サ変名詞	58
109	基礎	名詞	58
110	効率的	タグ	58
111	生物	名詞	58
112	変化	サ変名詞	58
113	データ	名詞	57
114	振興	サ変名詞	57
115	総合	サ変名詞	57
116	輸送	サ変名詞	57
117	船員	名詞	56
118	周辺海域	タグ	55
119	気候変動	タグ	54
120	措置	サ変名詞	54
121	造船	サ変名詞	54
122	予測	サ変名詞	54
123	科学的	タグ	53
124	技術開発	タグ	53
125	事業者	タグ	53
126	計画的	タグ	52
127	持続可能	タグ	52
128	処理	サ変名詞	52
129	多様	形容動詞	52
130	秩序	名詞	52
131	制度	名詞	51
132	安定的	タグ	50
133	解決	サ変名詞	50
134	再生	サ変名詞	50
135	大学	名詞	50
136	海運	名詞	49
137	競争力	タグ	49
138	高い	形容詞	49
139	政府	名詞	49
140	創出	サ変名詞	49
141	役割	名詞	49
142	手法	名詞	48
143	日本	地名	48
144	各種	名詞	47
145	海難	名詞	46
146	成果	名詞	46
147	陸域	タグ	46
148	拠点	名詞	45
149	権益	名詞	45
150	航路	名詞	45
151	資源開発	タグ	45
152	府	名詞C	45
153	モニタリング	名詞	44
154	参画	サ変名詞	44
155	調整	サ変名詞	44
156	アジア	地名	43
157	水産業	タグ	43
158	知見	名詞	43
159	改善	サ変名詞	42
160	普及	サ変名詞	42
161	民間企業	タグ	42
162	沿岸域	タグ	41
163	戦略	名詞	41
164	方針	名詞	41
165	洋上	名詞	41
166	会議	サ変名詞	40
167	確立	サ変名詞	40
168	国家	名詞	40
169	自然	形容動詞	40
170	地方公共団体	タグ	40
171	展開	サ変名詞	40
172	防災	サ変名詞	40
173	本部	名詞	40
174	海賊	名詞	39
175	航海	サ変名詞	39
176	漂着	サ変名詞	39
177	目的	名詞	39
178	安定	形容動詞	38
179	可能性	タグ	38
180	目標	名詞	38
181	あり方	名詞	37
182	温暖化	タグ	37
183	気象	名詞	37
184	場	名詞C	37
185	負荷	サ変名詞	37
186	プロジェクト	名詞	36
187	一層	副詞	36
188	訓練	サ変名詞	36
189	継続的	タグ	36
190	被害	名詞	36
191	再生可能エネルギー	タグ	35
192	実証	サ変名詞	35
193	水質	名詞	35
194	設定	サ変名詞	35
195	法律	名詞	35
196	油	名詞C	35
197	規模	名詞	34
198	広域	名詞	34
199	高潮	名詞	34
200	市場	名詞	34

図 1-2 抽出語リスト（上位 200 語）

2. 共起ネットワーク

KHCoder による自動設定値は、最小出現頻度：70、分析対象語数：73 であった共起ネットワークの描画条件を以下に、語・語での共起ネットワークを図 2-1、語・年での共起ネットワークを図 2-2 に示す。

	最小出現頻度	分析対象語数	描画結果
語・語	80	73	ノード数：57　エッジ数：73
語・年	80	73	ノード数：39　エッジ数：73

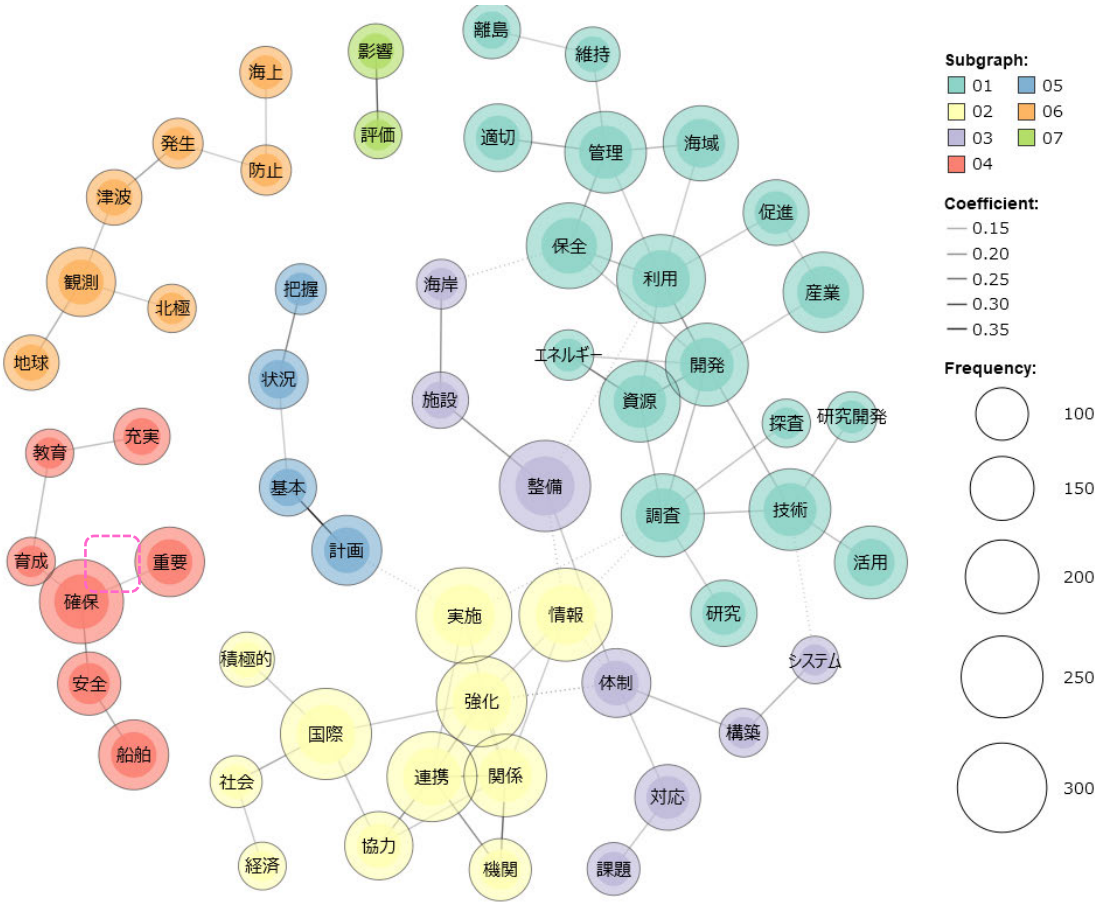


図 2-1　共起ネットワーク（語・語）

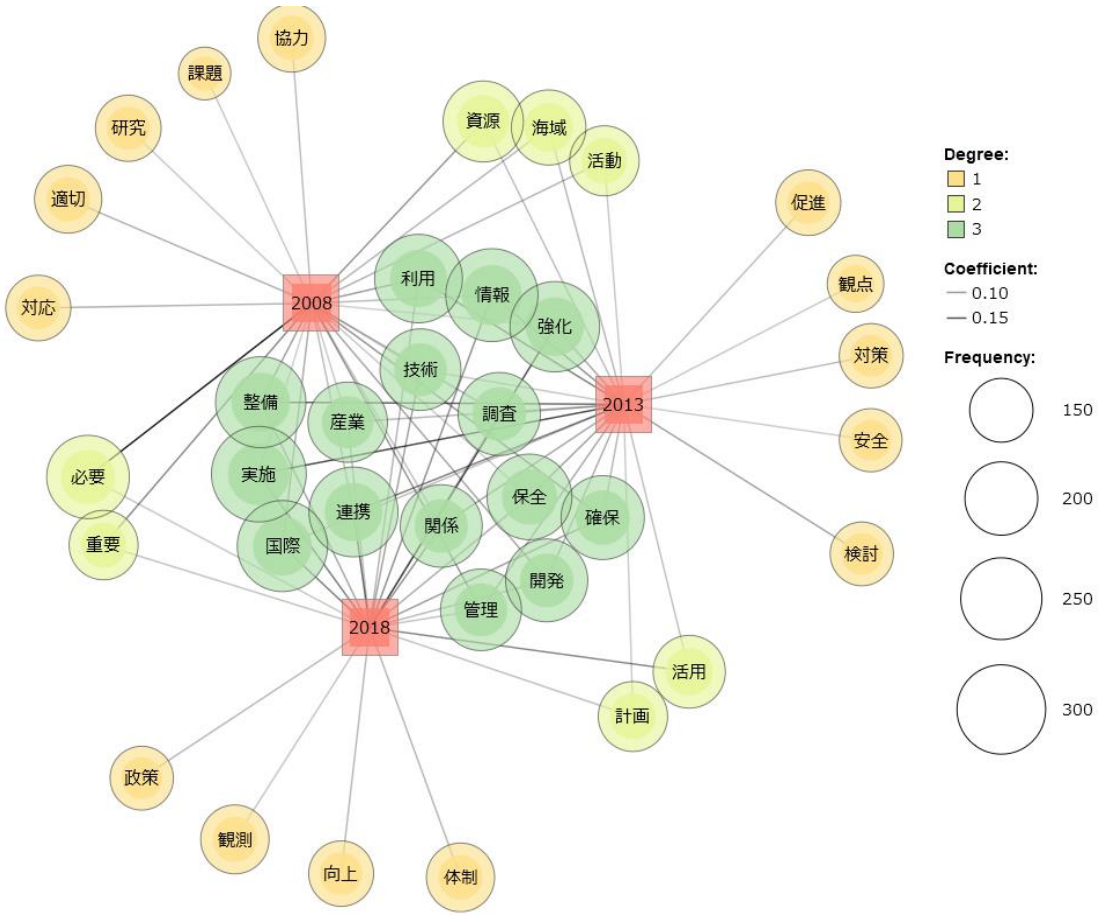


図 2-2 共起ネットワーク（語・年）

3. 対応分析

KHCoder により自動設定される最小出現数：80、分析対象語数：73 で対応分析処理を行った。累積寄与率は成分 1 と 2 で 100%であり、その結果を図 3-1 に示す。更に、分析対象語数：152 でも対応分析処理を行った。その累積寄与率は成分 1 と 2 で 100%であり、その結果を図 3-2 に示す。

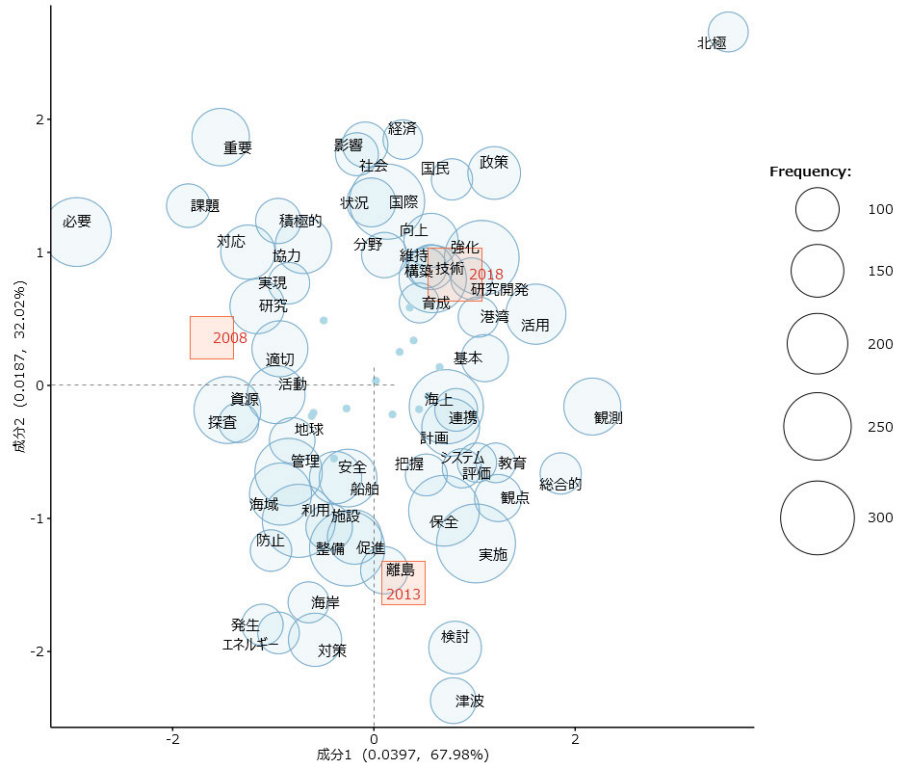


図 3-1 語･年の対応分析結果（分析対象語数：73）

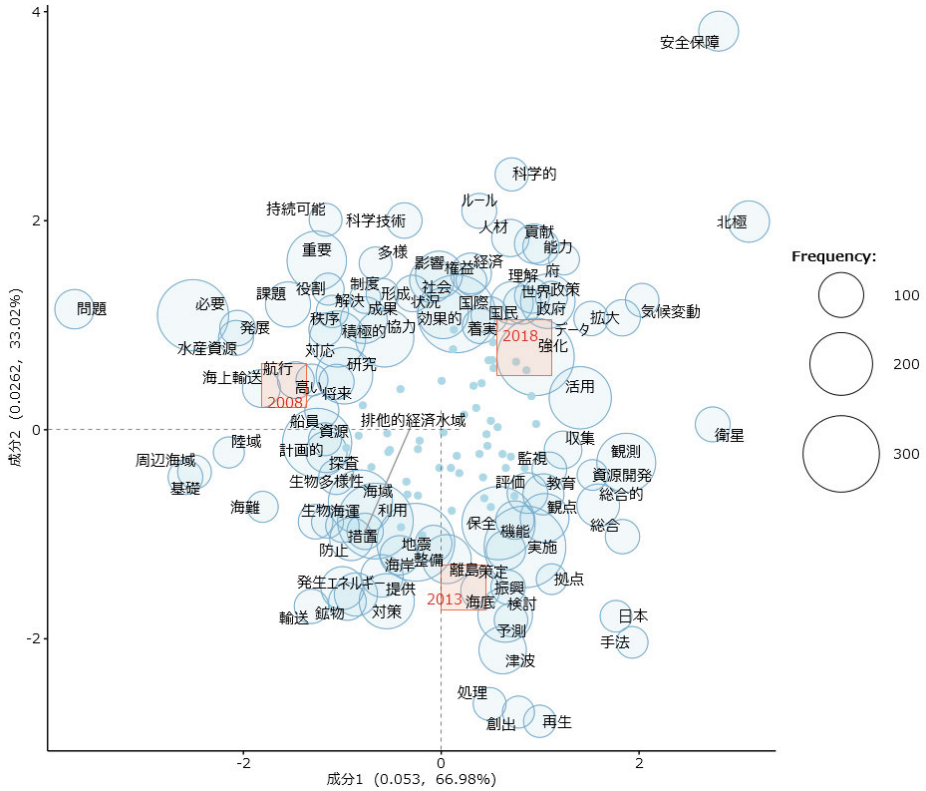


図 3-2 語･年の対応分析結果（分析対象語数：152）

4. LDA 分析

KHCoder により自動設定される最小出現数：80、分析対象語数：73 で集計単位を H5（年）とした場合と、分析対象語数：152 での LDA 分析を行った。

分析対象語数：73 での LDA トピック数推定結果を図 4-1、分析対象語数：152 での結果を図 4-2 に示す。これらの図から 73 語でのトピック数は 10、152 語では 8 と推察した。

分析対象語数：73、トピック数：10 での LDA 処理結果を表 4-1、そのヒートマップを図 4-3、トピック比率を図 4-4～6 に示す。

また、分析対象語数：152、トピック数：8 でのその LDA 処理結果を表 4-2、そのヒートマップを図 4-7、トピック比率を図 4-8～9 に示す。

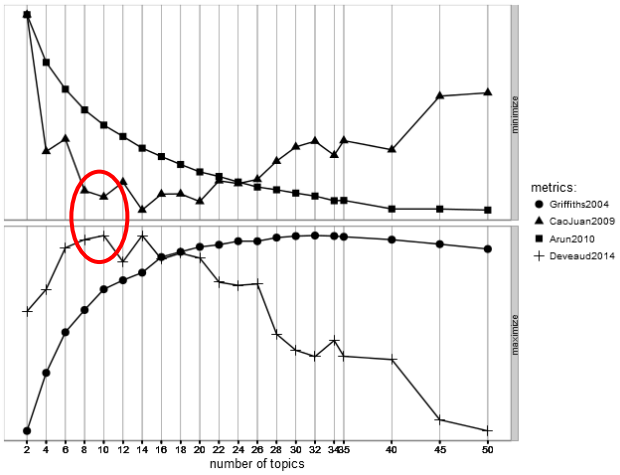


図 4-1 LDAtuning 実行結果（分析対象語数：73）

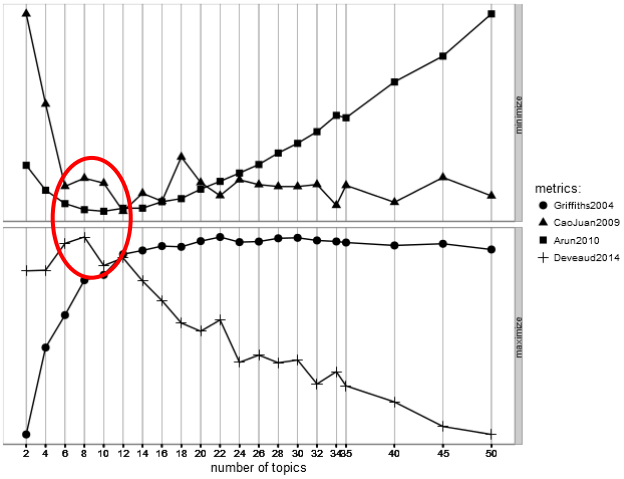


図 4-2 LDAtuning 実行結果（分析対象語数：152）

表 4-1 LDA 処理結果（10 トピックス、分析対象語数：73）

Topics

#1		#2		#3		#4		#5	
体制	0.128	管理	0.148	強化	0.163	保全	0.183	情報	0.208
関係	0.095	対策	0.089	活用	0.140	観測	0.088	連携	0.153
産業	0.093	関係	0.086	国際	0.123	総合的	0.078	状況	0.085
技術	0.084	安全	0.069	政策	0.076	津波	0.075	支援	0.081
確保	0.080	施設	0.068	北極	0.073	基盤	0.067	計画	0.080
向上	0.078	適切	0.054	国民	0.057	調査	0.064	開発	0.076
経済	0.071	把握	0.046	機関	0.036	研究	0.061	充実	0.049
分野	0.048	防止	0.044	評価	0.036	システム	0.060	政策	0.047
国際	0.046	エネルギー	0.036	構築	0.034	観点	0.058	向上	0.043
研究開発	0.040	発生	0.029	観測	0.032	充実	0.041	海上	0.040

#6		#7		#8		#9		#10	
調査	0.125	技術	0.124	必要	0.204	利用	0.128	実施	0.250
資源	0.108	基本	0.111	重要	0.119	確保	0.126	整備	0.132
協力	0.106	維持	0.110	国際	0.106	開発	0.112	産業	0.120
対応	0.083	育成	0.084	研究	0.082	海域	0.109	検討	0.114
強化	0.059	促進	0.083	積極的	0.068	整備	0.086	教育	0.054
利用	0.056	連携	0.073	課題	0.060	活動	0.076	保全	0.041
実現	0.048	港湾	0.058	適切	0.046	地球	0.064	離島	0.039
適切	0.032	探査	0.058	影響	0.038	資源	0.052	情報	0.033
必要	0.030	影響	0.055	国際的	0.038	機関	0.047	関連	0.030
船舶	0.028	観測	0.047	国民	0.030	促進	0.040	船舶	0.026

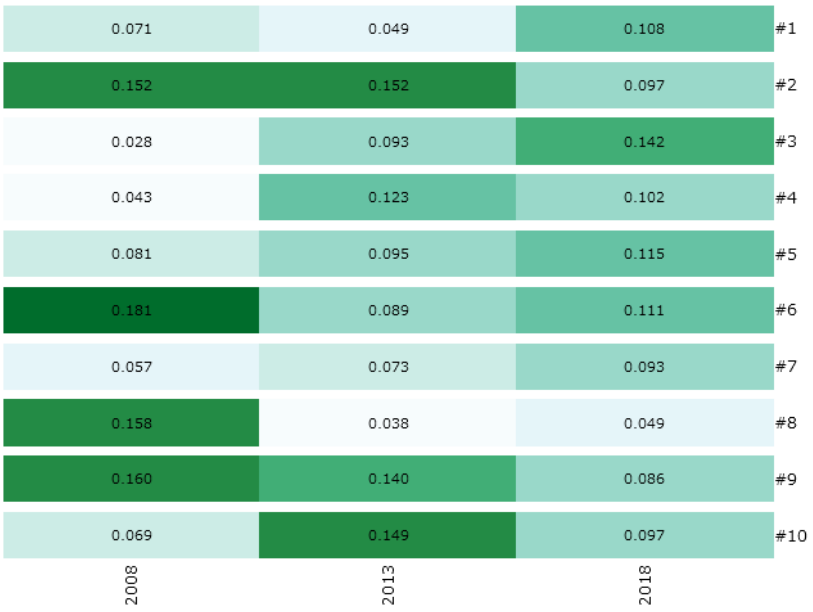


図 4-3 LDA ヒートマップ（10 トピックス、分析対象語数：73）

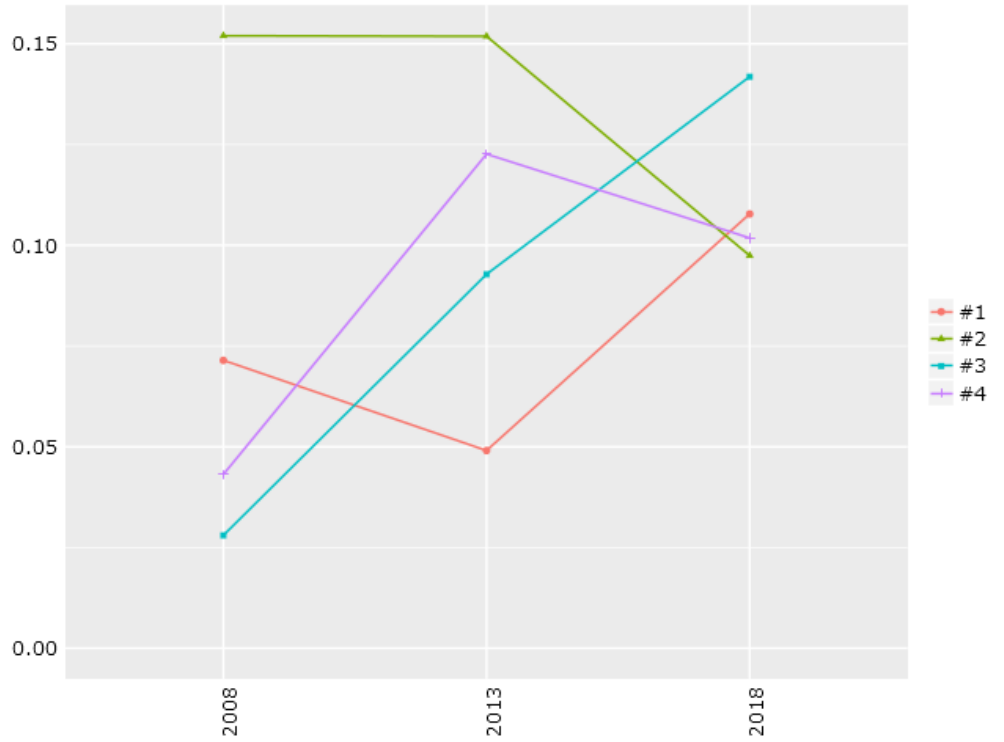


図 4-4　1～4 トピックの比率（10 トピックス、分析対象語数：73）

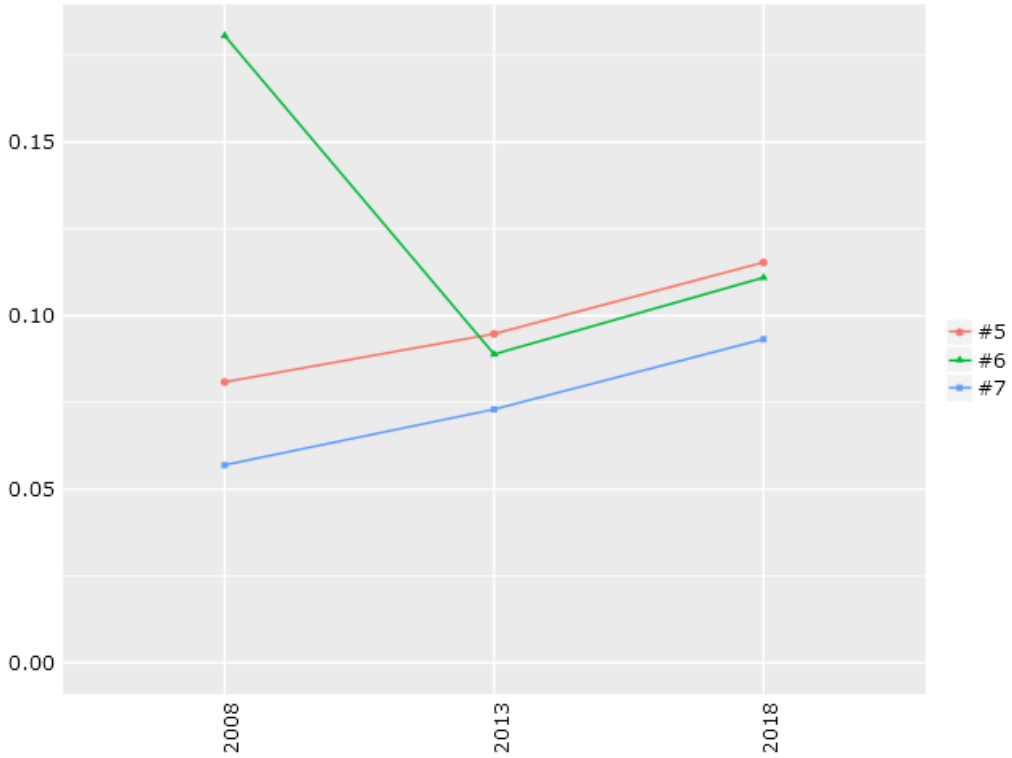


図 4-5　5～7 トピックの比率（10 トピックス、分析対象語数：73）

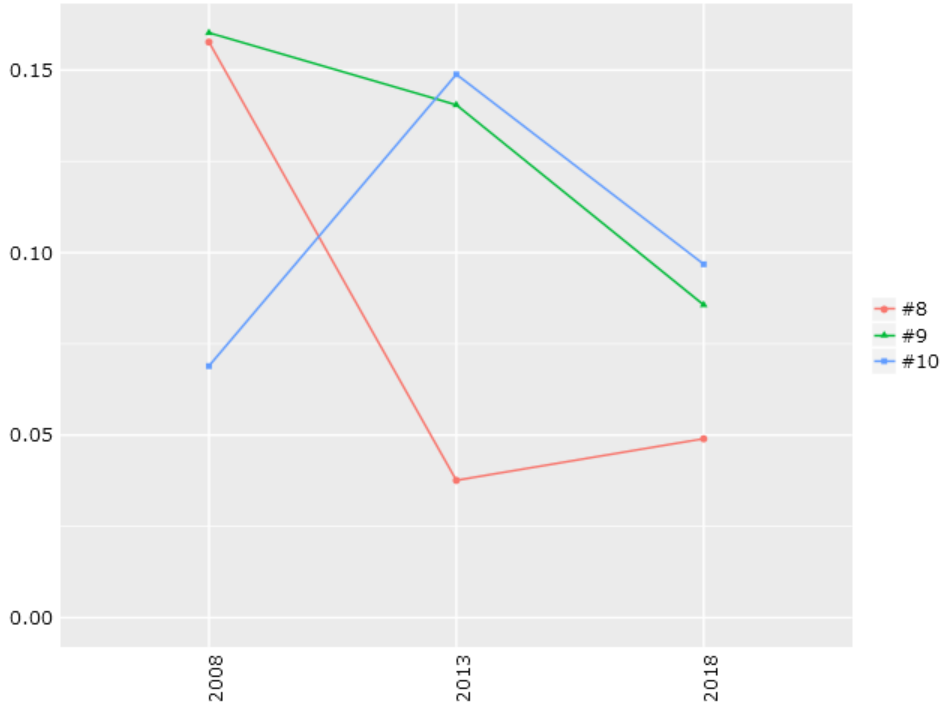


図 4-6 8～10 トピックの比率（10 トピックス、分析対象語数：73）

表 4-2 LDA 処理結果（8 トピックス、分析対象語数：152）

Topics

#1		#2		#3		#4	
連携	0.117	国際	0.114	資源	0.080	活用	0.078
計画	0.070	強化	0.110	調査	0.076	関係	0.076
情報	0.069	政策	0.052	管理	0.065	観測	0.063
産業	0.058	維持	0.044	船舶	0.044	体制	0.050
充実	0.047	社会	0.038	対策	0.041	調査	0.049
支援	0.042	研究開発	0.034	防止	0.040	向上	0.049
把握	0.039	国民	0.034	活動	0.037	北極	0.049
管理	0.038	影響	0.034	地球	0.037	安全保障	0.045
確保	0.038	構築	0.033	発生	0.033	技術	0.042
海上	0.037	育成	0.033	海岸	0.032	状況	0.042

#5		#6		#7		#8	
実施	0.138	利用	0.147	適切	0.081	必要	0.124
整備	0.093	開発	0.086	技術	0.059	重要	0.075
保全	0.086	促進	0.081	情報	0.056	確保	0.065
検討	0.054	海域	0.075	協力	0.042	整備	0.044
津波	0.048	産業	0.055	基盤	0.035	体制	0.040
強化	0.037	安全	0.045	保全	0.034	問題	0.037
観測	0.033	向上	0.036	活動	0.034	対応	0.034
機能	0.031	生態系	0.031	基本	0.033	海上輸送	0.034
離島	0.028	離島	0.030	対応	0.033	課題	0.033
基本	0.026	国際的	0.030	生産	0.030	研究	0.033

図 4-7 LDA ヒートマップ（8 トピックス、分析対象語数：152）

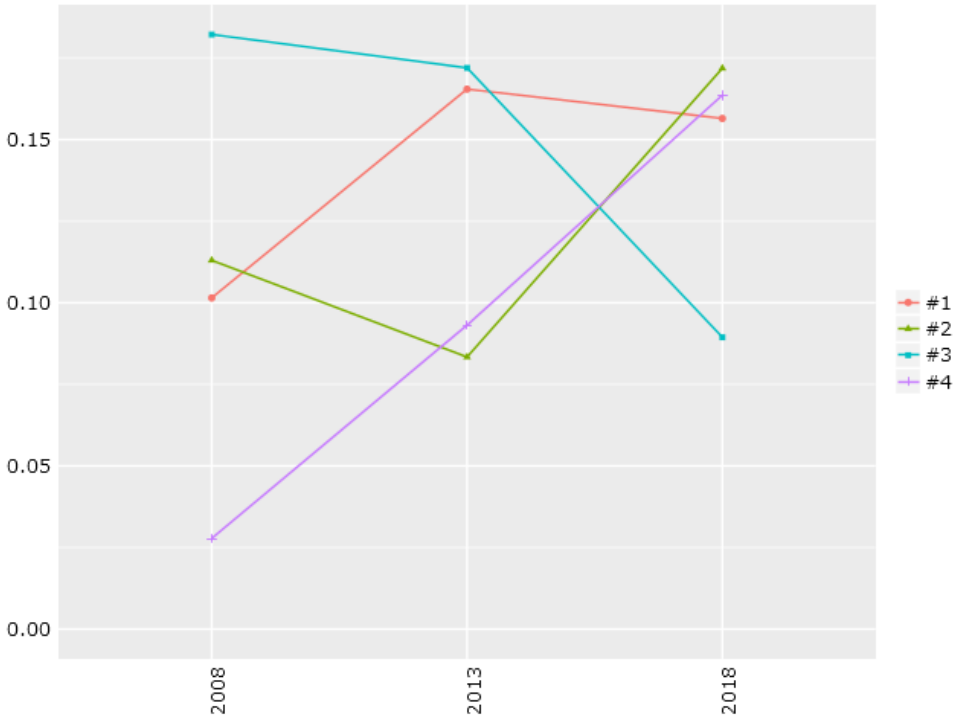


図 4-8 1～4 トピックの比率（8 トピックス、分析対象語数：152）

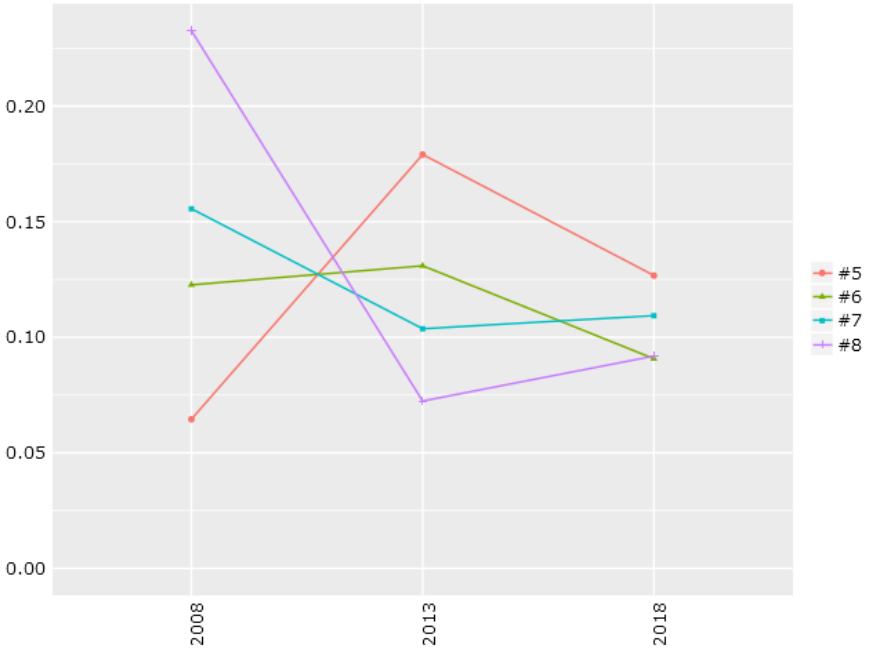


図 4-8 5～8 トピックの比率（8 トピックス、分析対象語数：152）

付録 6「海洋白書（2004～2020 年）の分析結果」

文書番号：JRDN-21-025

1. 前処理結果

環境白書･海洋白書･水産白書（2008～2020 年）の分析で設定した強制抽出語（316 語）と 54 語の除外語を設定し、「動詞、感動詞、動詞 B、副詞 B」を除外して前処理を実行した。

Chanse での前処理の結果、総抽出語数：1,056,793、異なり語数：20,563 のうち 15,864 語が分析処理で使用された。抽出語出現数の頻度分布を図 1-1、抽出語リスト（上位 200 語）を図 1 -2 に示す。

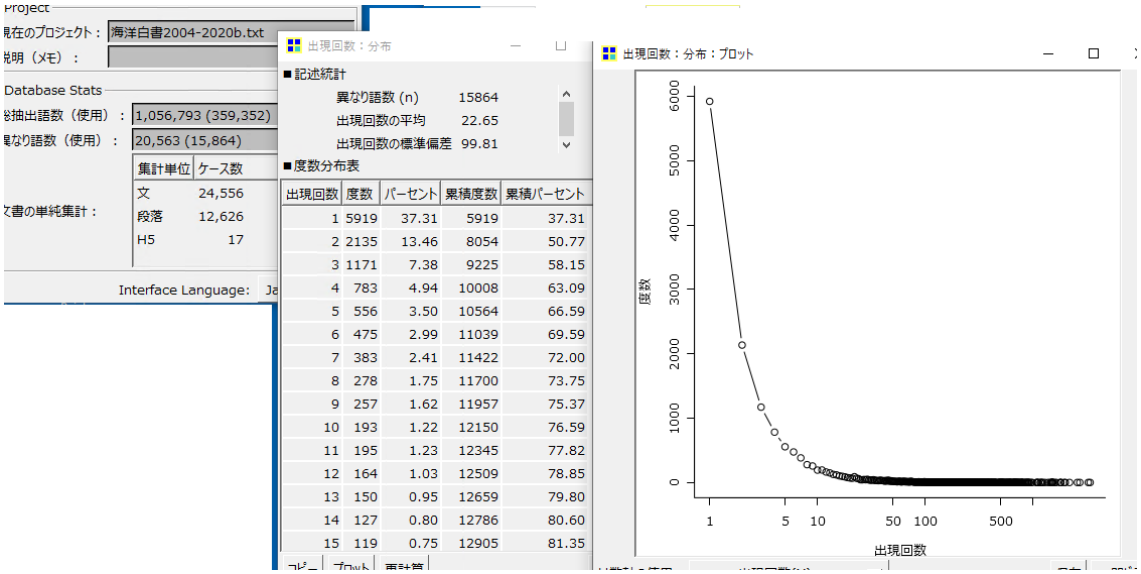


図 1-1 抽出語出現数の頻度分布

#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度
1	管理	サ変名詞	3245	36	社会	名詞	884				
2	開発	サ変名詞	2712	37	協力	サ変名詞	880				
3	計画	サ変名詞	2570	38	対策	サ変名詞	859				
4	海域	名詞	2175	39	総合	サ変名詞	853				
5	基本	名詞	1988	40	策定	サ変名詞	839				
6	利用	サ変名詞	1941	41	持続可能	タグ	821	71	国際的	タグ	621
7	必要	形容動詞	1807	42	状況	名詞	802	72	制度	名詞	618
8	日本	地名	1800	43	機関	名詞	792	73	行動	サ変名詞	616
9	政策	名詞	1787	44	対応	サ変名詞	788	74	議論	サ変名詞	615
10	実施	サ変名詞	1642	45	条約	名詞	787	75	目的	名詞	612
11	情報	名詞	1584	46	分野	名詞	785	76	構築	サ変名詞	602
12	関係	サ変名詞	1525	47	開催	サ変名詞	775	77	委員会	タグ	600
13	保全	サ変名詞	1518	48	北極	地名	766	78	地震	名詞	594
14	問題	ナイ形容	1510	49	離島	サ変名詞	752	79	海上	名詞	590
15	調査	サ変名詞	1509	50	設置	サ変名詞	735	80	大きい	形容詞	582
16	沿岸域	タグ	1496	51	地球	名詞	734	81	変化	サ変名詞	581
17	資源	名詞	1495	52	支援	サ変名詞	723	82	データ	名詞	570
18	教育	サ変名詞	1471	53	津波	名詞	723	83	具体的	タグ	564
19	国際	名詞	1441	54	目標	名詞	709	84	参加	サ変名詞	563
20	活動	サ変名詞	1377	55	海底	名詞	707	85	生物	名詞	554
21	研究	サ変名詞	1286	56	エネルギー	名詞	700	86	国連	組織名	550
22	総合的	タグ	1220	57	確保	サ変名詞	691	87	施設	サ変名詞	536
23	世界	名詞	1214	58	EEZ	タグ	688	88	発展	サ変名詞	535
24	中国	地名	1175	59	生態系	タグ	686	89	促進	サ変名詞	532
25	産業	名詞	1169	60	連携	サ変名詞	686	90	高い	形容詞	528
26	重要	形容動詞	1113	61	強化	サ変名詞	680	91	期待	サ変名詞	519
27	技術	名詞	1106	62	国家	名詞	670	92	結果	副詞可能	517
28	船舶	名詞	1082	63	評価	サ変名詞	668	93	体制	名詞	517
29	観測	サ変名詞	1074	64	安全	形容動詞	666	94	可能	形容動詞	515
30	会議	サ変名詞	981	65	経済	名詞	665	95	存在	サ変名詞	515
31	検討	サ変名詞	959	66	大陸棚	名詞	663	96	生産	サ変名詞	512
32	課題	名詞	922	67	発生	サ変名詞	660	97	太平洋	地名	511
33	影響	サ変名詞	920	68	沿岸	名詞	658	98	場合	副詞可能	509
34	整備	サ変名詞	904	69	システム	名詞	646	99	排他的経済水域	タグ	506
35	政府	名詞	893	70	関連	サ変名詞	630	100	実現	サ変名詞	503

#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度
101	中心	名詞	503	136	提供	サ変名詞	390				
102	多く	副詞可能	502	137	沖	名詞C	387				
103	保護	サ変名詞	498	138	水産	名詞	386				
104	採択	サ変名詞	496	139	被害	名詞	385				
105	科学	名詞	493	140	基盤	名詞	384				
106	対象	名詞	490	141	米国	地名	383	171	開始	サ変名詞	333
107	活用	サ変名詞	484	142	科学的	タグ	382	172	再生	サ変名詞	333
108	自然	形容動詞	481	143	観光	サ変名詞	381	173	積極的	タグ	333
109	海岸	名詞	474	144	規制	サ変名詞	379	174	提言	サ変名詞	331
110	島	名詞C	470	145	ひとつ	副詞可能	378	175	鉱物	名詞	328
111	戦略	名詞	452	146	役割	名詞	377	176	会合	サ変名詞	327
112	可能性	タグ	449	147	輸送	サ変名詞	372	177	発表	サ変名詞	326
113	気候変動	タグ	447	148	機能	サ変名詞	368	178	防災	サ変名詞	326
114	国連海洋法条約	タグ	447	149	能力	名詞	367	179	成果	名詞	325
115	本部	名詞	447	150	導入	サ変名詞	364	180	動き	名詞	325
116	内容	名詞	445	151	維持	サ変名詞	362	181	風力発電	タグ	324
117	航行	サ変名詞	440	152	航路	名詞	362	182	合意	サ変名詞	321
118	大学	名詞	440	153	組織	サ変名詞	360	183	多様	形容動詞	321
119	多い	形容詞	439	154	規模	名詞	356	184	育成	サ変名詞	319
120	プロジェクト	名詞	437	155	空間	名詞	356	185	決定	サ変名詞	317
121	規定	サ変名詞	430	156	枠組み	名詞	354	186	法	名詞C	317
122	措置	サ変名詞	425	157	それぞれ	副詞可能	352	187	現状	名詞	316
123	生物多様性	タグ	422	158	調整	サ変名詞	352	188	適切	形容動詞	311
124	制定	サ変名詞	421	159	温暖化	タグ	350	189	区域	名詞	309
125	海賊	名詞	418	160	形成	サ変名詞	350	190	展開	サ変名詞	309
126	洋上	名詞	417	161	アメリカ	地名	349	191	発電	サ変名詞	309
127	振興	サ変名詞	414	162	作業	サ変名詞	349	192	基礎	名詞	307
128	港湾	名詞	408	163	石油	名詞	345	193	作成	サ変名詞	305
129	領海	名詞	406	164	共同	サ変名詞	343	194	将来	副詞可能	305
130	探査	サ変名詞	405	165	減少	サ変名詞	343	195	新しい	形容詞	304
131	増加	サ変名詞	404	166	理解	サ変名詞	343	196	共有	サ変名詞	301
132	設定	サ変名詞	402	167	予測	サ変名詞	342	197	国民	名詞	300
133	拡大	サ変名詞	399	168	基本的	タグ	341	198	実証	サ変名詞	300
134	安全保障	タグ	390	169	認識	サ変名詞	339	199	把握	サ変名詞	300
135	管轄	サ変名詞	390	170	法律	名詞	339	200	人材	名詞	299

図 1-2 抽出語リスト（上位 200 語）

2. 共起ネットワーク

KHCoder により自動設定される最小出現頻度：610、分析対象語数：75 とした場合と、最小出現頻度：360、分析対象語数：153 の結果を示す。

分析対象語数：75 での共起ネットワークの描画条件を以下に、語・語での共起ネットワークを図 2-1、語・年での共起ネットワークを図 2-2 に示す。

また、分析対象語数：153 での共起ネットワークの描画条件を以下に、語・語での共起ネットワークを図 2-3、語・年での共起ネットワークを図 2-4 に示す。

	最小出現頻度	分析対象語数	描画結果
語・語	610	75	ノード数：52　エッジ数：75
語・年	610	75	ノード数：51　エッジ数：75

	最小出現頻度	分析対象語数	描画結果
語・語	360	153	ノード数：67　エッジ数：75
語・年	360	153	ノード数：52　エッジ数：75

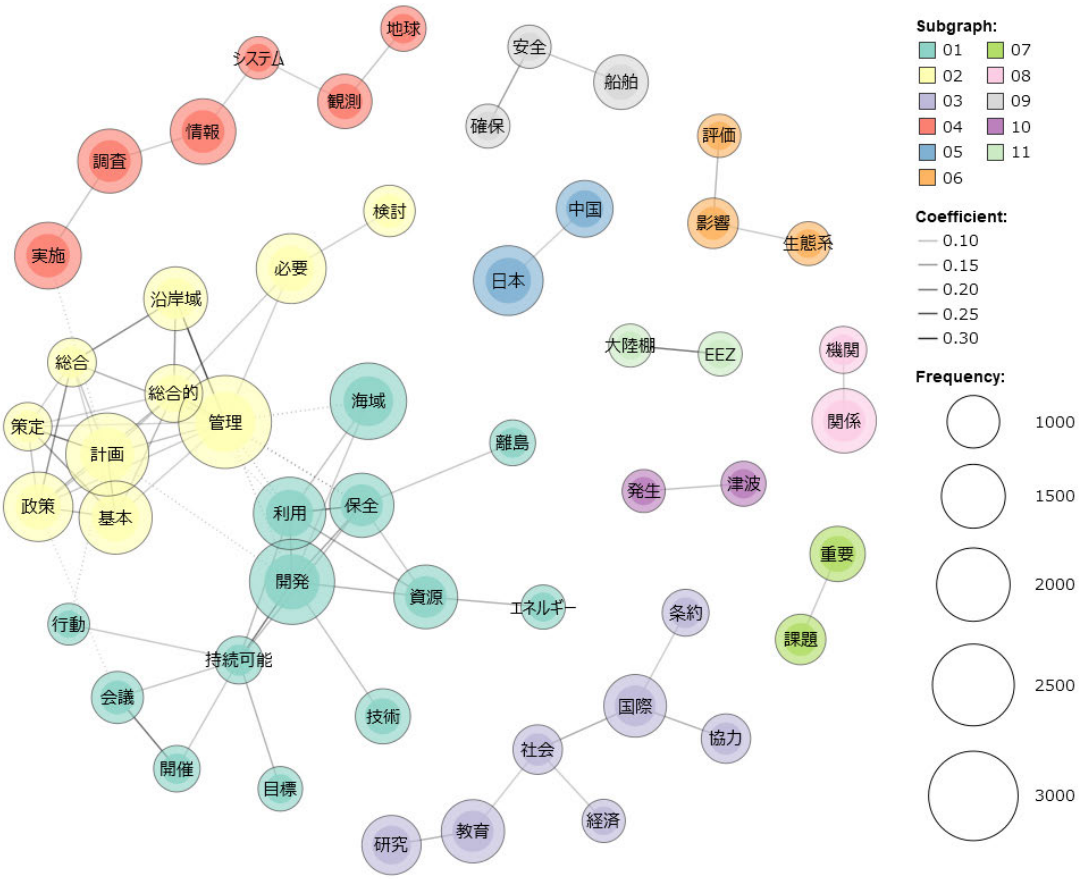


図 2-1　共起ネットワーク（語・語）(分析対象語数：75)

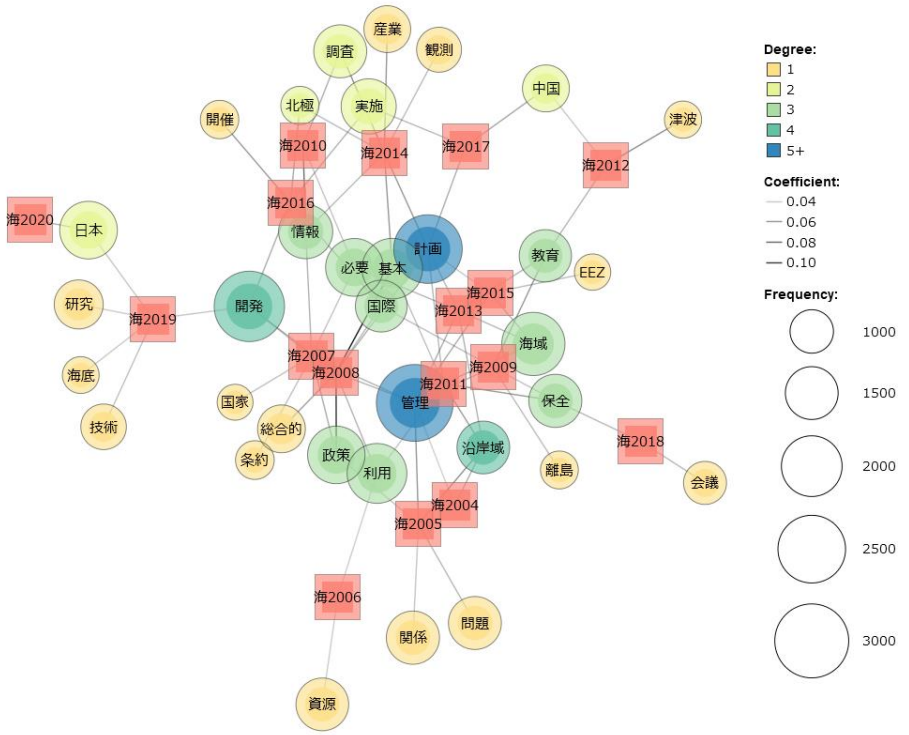


図 2-2 共起ネットワーク（語・年）(分析対象語数：75)

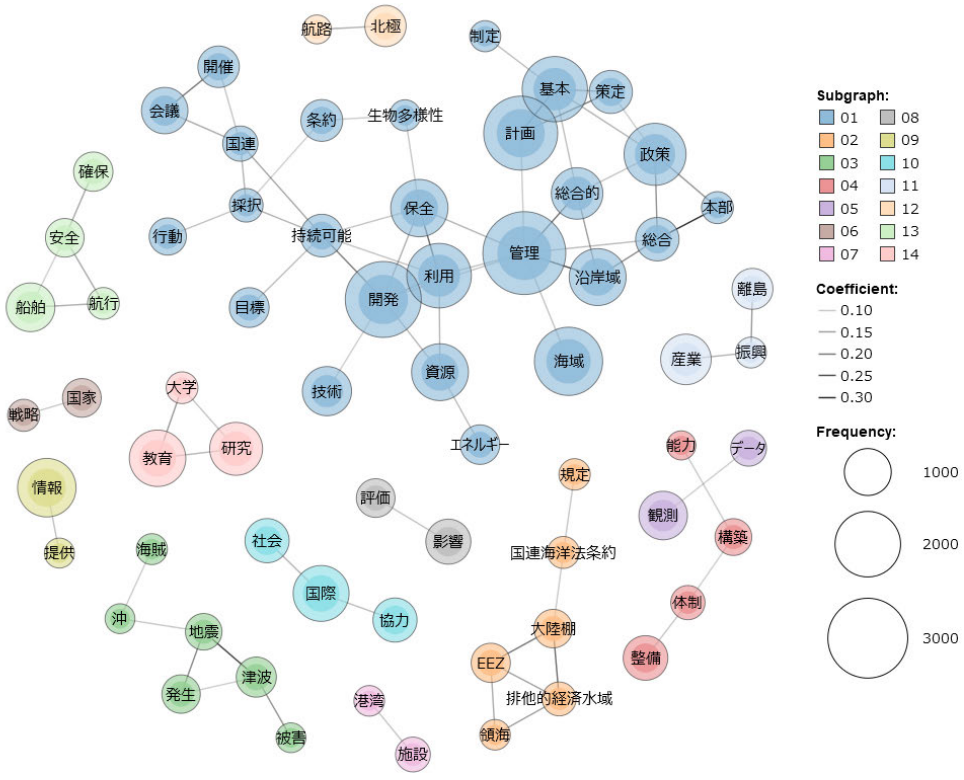


図 2-3 共起ネットワーク（語・語）（分析対象語数：153）

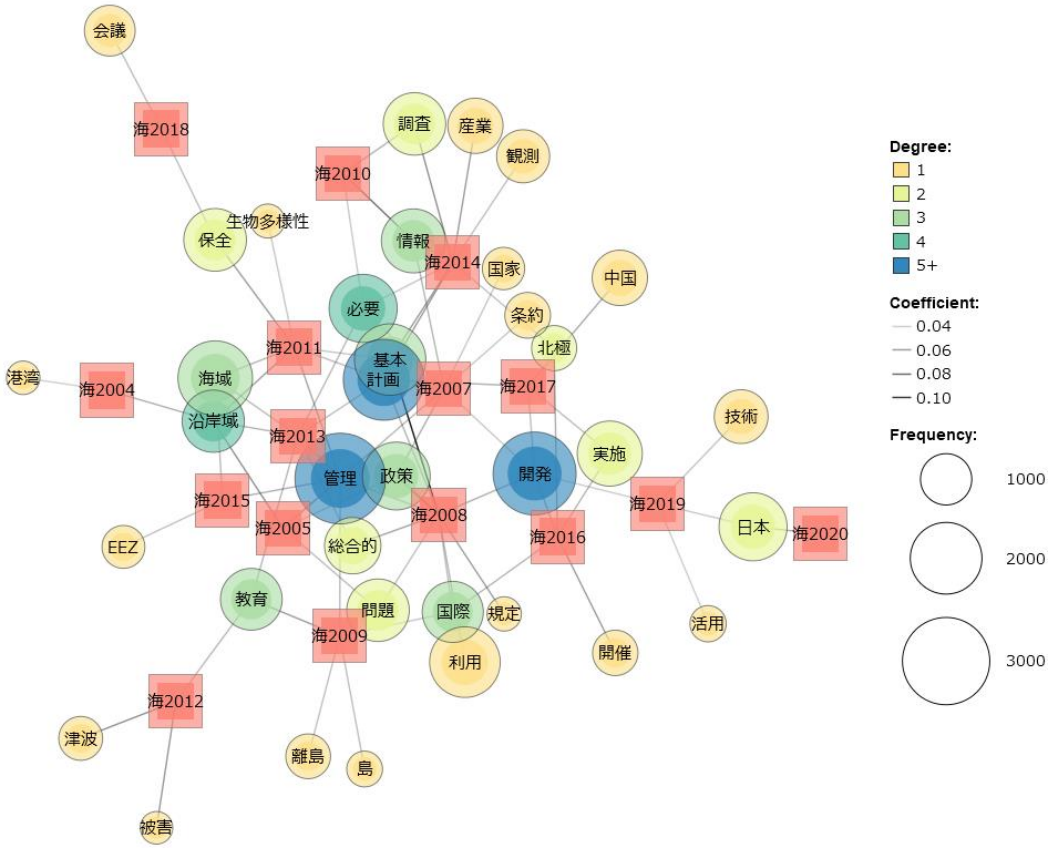


図 2-4 共起ネットワーク（語・年）（分析対象語数：153）

3. 対応分析

KHCoder により自動設定される最小出現頻度：610、分析対象語数：75 とした場合と、最小出現頻度：360、分析対象語数：153 で対応分析処理を行った。

分析対象語数：75 の累積寄与率は成分 1 と 2 で 37.28%であり、成分 3 と 4 の累積寄与率は約 25%である。その結果を図 3-1,3-2 に示す。

分析対象語数：153 の累積寄与率は成分 1 と 2 で 34.65%であり、成分 3 と 4 の累積寄与率は約 25%である。その結果を図 3-3,3-4 に示す。

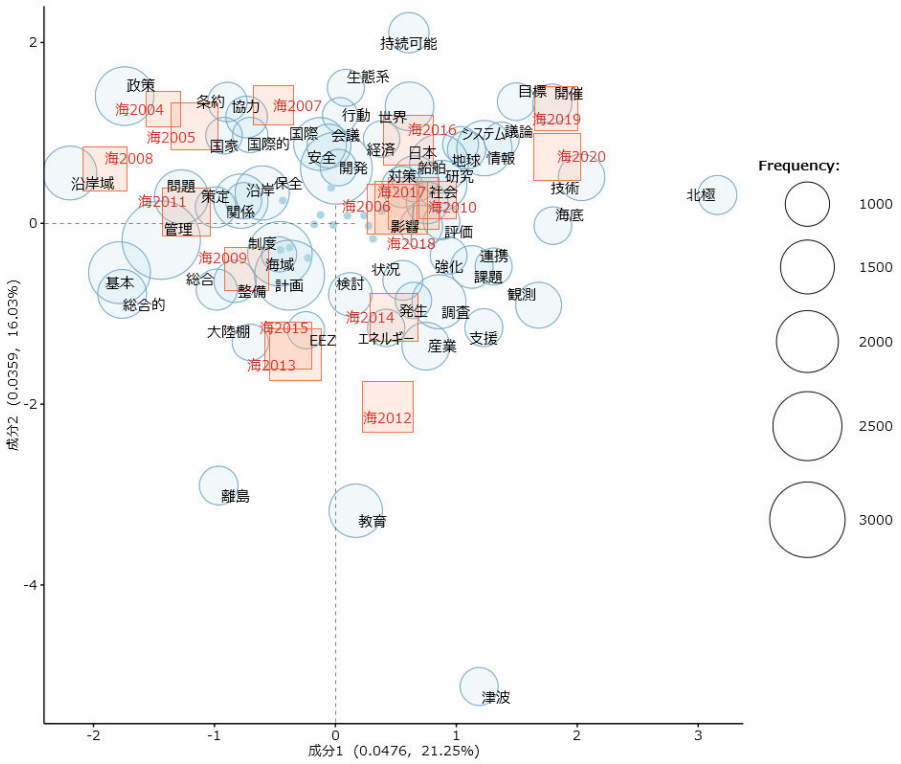


図 3-1 語・年の対応分析結果 （成分 1 と 2）（分析対象語数：75）

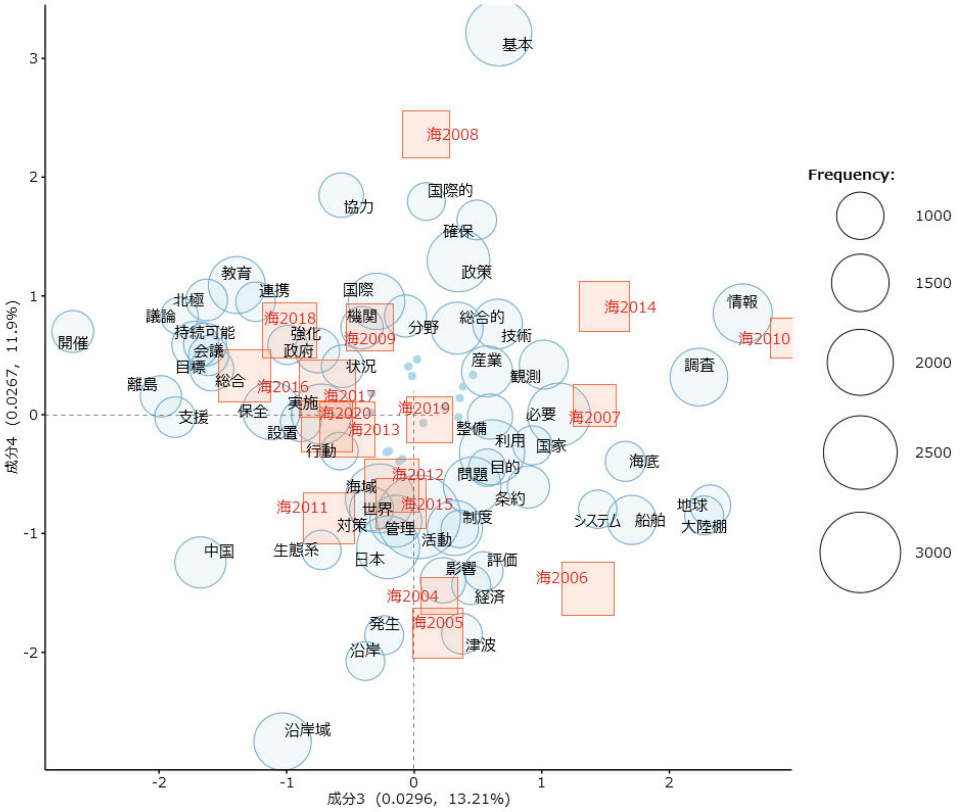


図 3-2 語・年の対応分析結果 （成分 3 と 4）（分析対象語数：75）

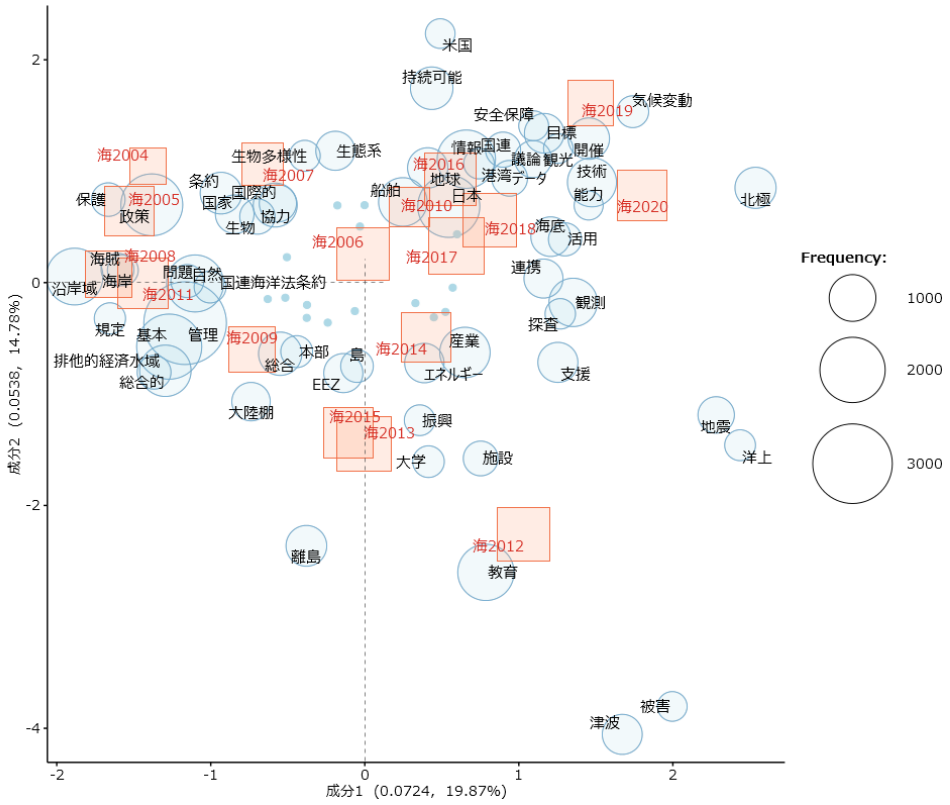


図 3-3 語･年の対応分析結果（成分 1と2）（分析対象語数：153）

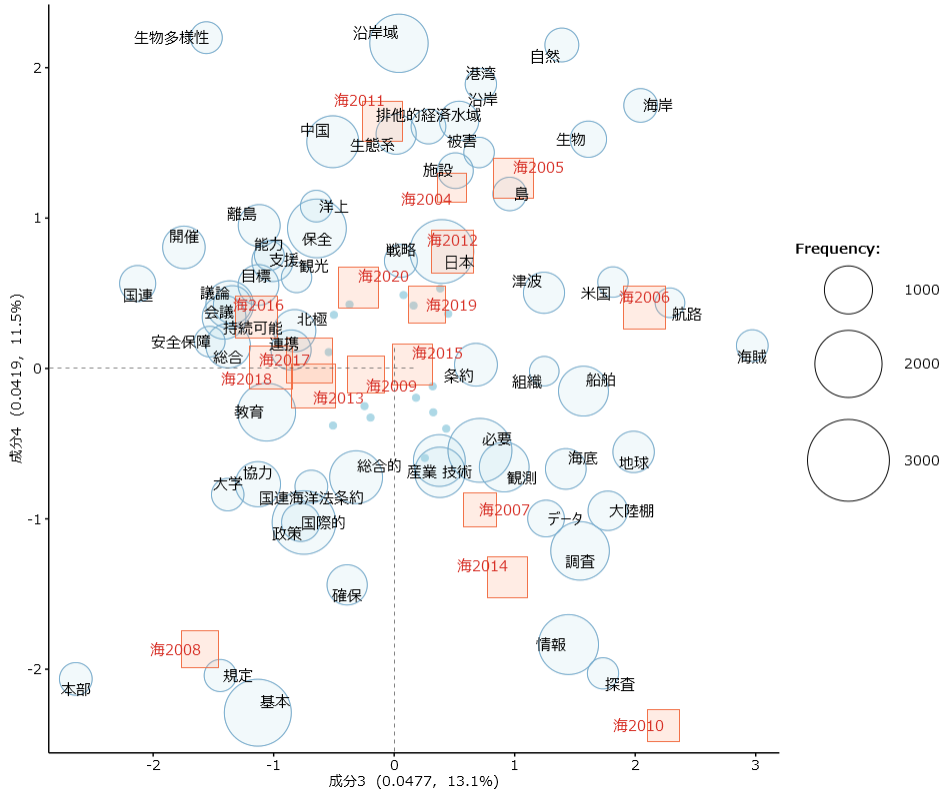


図 3-4 語･年の対応分析結果（成 3と4）（分析対象語数：153）

4. LDA 分析

KHCoder により自動設定される最小出現数：610、分析対象語数：75 で集計単位を H5（年）とした場合と、分析対象語数：153 での LDA 分析を行った。

分析対象語数：75 での LDA トピック数推定結果を図 4-1、分析対象語数：153 での結果を図 4-2 に示す。これらの図から 75 語でのトピック数は 16、153 語では 18 と推察した。

分析対象語数：69、トピック数：16 での LDA 処理結果を表 4-1、そのヒートマップを図 4-3、ヒートマップ樹形図を図 4-4、*トピック比率集計表を表 4-2、トピック比率を図 4-5～8 に示す。

また、分析対象語数：153、トピック数：18 での LDA 処理結果を表 4-3、そのヒートマップを図 4-9、ヒートマップ樹形図を図 4-10、*トピック比率集計表を表 4-4、トピック比率を図 4-11～15 に示す。

*は別途 excel 形式で提供。

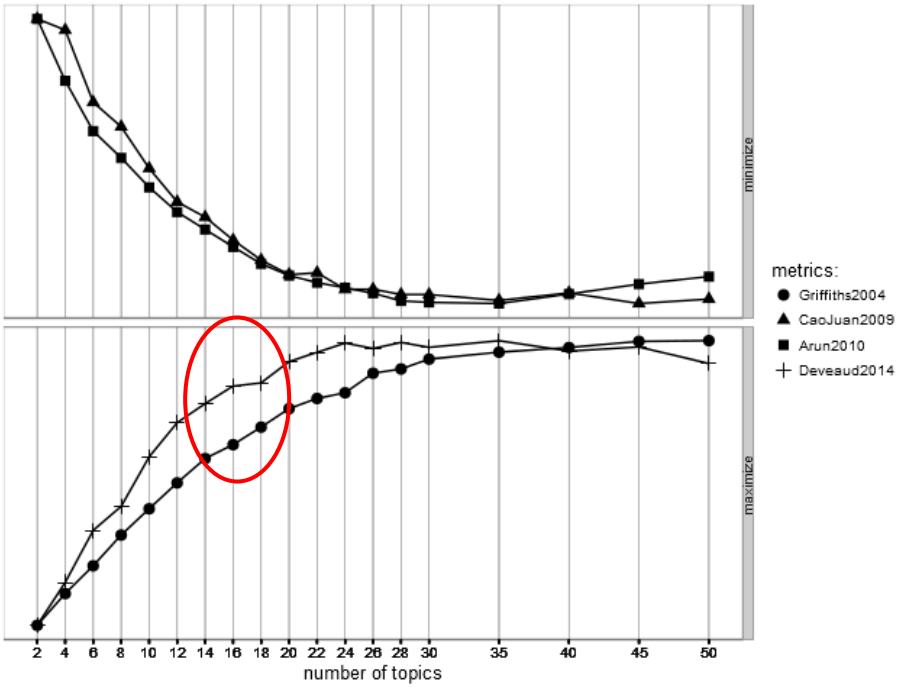


図 4-1 LDAtuning 実行結果（分析対象語数：75）

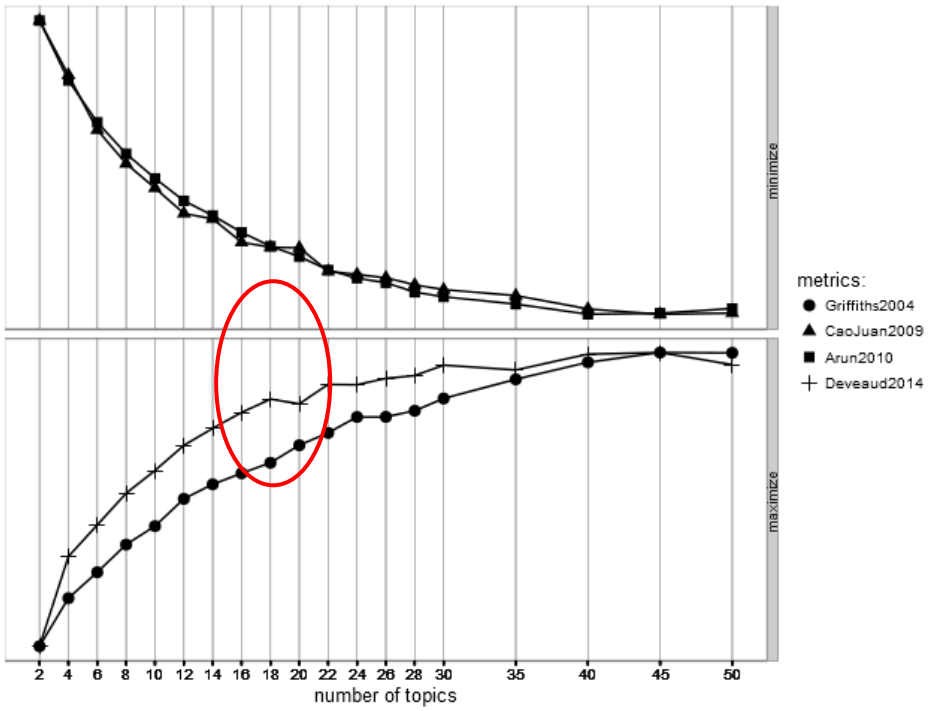


図 4-2 LDAtuning 実行結果（分析対象語数：153）

表 4-1 LDA 処理結果（16 トピックス、分析対象語数：75）

#1		#2		#3		#4		#5		#6		#7		#8	
基本	0.190	計画	0.139	EEZ	0.127	情報	0.143	沿岸域	0.193	保全	0.143	中国	0.193	離島	0.131
政策	0.123	実施	0.116	海域	0.110	調査	0.097	管理	0.160	問題	0.140	観測	0.136	管理	0.108
総合的	0.060	総合	0.091	管理	0.079	基本	0.088	日本	0.074	利用	0.122	発生	0.089	基本	0.096
計画	0.060	中国	0.075	利用	0.072	必要	0.082	開発	0.062	必要	0.093	研究	0.072	海域	0.095
開発	0.047	設置	0.060	資源	0.070	大陸棚	0.048	世界	0.054	計画	0.090	活動	0.069	活動	0.060
国際的	0.044	策定	0.053	安全	0.052	整備	0.041	政策	0.052	生態系	0.065	影響	0.065	沿岸域	0.055
研究	0.042	検討	0.040	対策	0.051	関係	0.039	制度	0.043	整備	0.064	沿岸	0.057	総合的	0.049
管理	0.037	課題	0.038	海底	0.050	確保	0.039	影響	0.042	沿岸	0.058	調査	0.056	エネルギー	0.045
国際	0.035	資源	0.033	評価	0.041	計画	0.038	総合的	0.039	調査	0.046	関係	0.052	連携	0.039
協力	0.033	強化	0.032	国家	0.041	課題	0.032	経済	0.035	対応	0.040	開発	0.042	関係	0.038

#9		#10		#11		#12		#13		#14		#15		#16	
観測	0.107	情報	0.135	日本	0.079	北極	0.182	津波	0.265	会議	0.085	管理	0.099	教育	0.314
調査	0.099	政策	0.105	資源	0.074	支援	0.068	産業	0.121	持続可能	0.083	開発	0.096	実施	0.056
計画	0.093	船舶	0.096	研究	0.070	国際	0.058	必要	0.096	開発	0.067	国際	0.074	状況	0.053
船舶	0.074	行動	0.062	技術	0.068	情報	0.056	大陸棚	0.069	保全	0.067	協力	0.061	社会	0.053
産業	0.065	国家	0.057	開発	0.059	日本	0.054	計画	0.053	議論	0.063	必要	0.057	対応	0.052
開発	0.049	条約	0.051	海底	0.058	開催	0.053	発生	0.051	目標	0.062	関係	0.056	確保	0.047
利用	0.045	利用	0.047	世界	0.048	連携	0.048	整備	0.051	開催	0.062	問題	0.049	検討	0.045
検討	0.041	活動	0.043	地球	0.045	世界	0.047	支援	0.042	政府	0.046	政府	0.047	開発	0.042
必要	0.041	分野	0.038	実施	0.038	協力	0.047	調査	0.034	技術	0.044	海域	0.046	技術	0.037
管理	0.040	対応	0.035	経済	0.037	研究	0.042	利用	0.027	日本	0.040	会議	0.045	日本	0.033

図 4-3　LDA ヒートマップ（16 トピックス、分析対象語数：75）

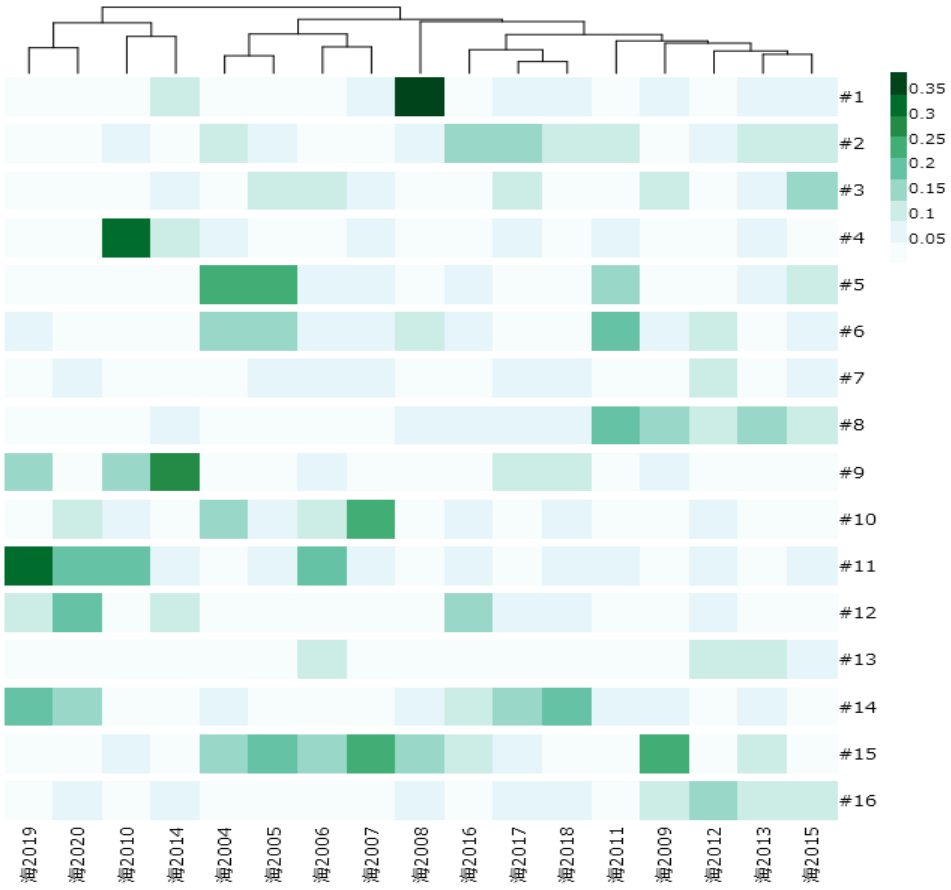


図 4-4　LDA ヒートマップ樹形図（16 トピックス、分析対象語数：75）

表 4-2 トピック比率集計表（16 トピックス、分析対象語数：75）

	#1	#2	#3	#4	#5	#6	#7	#8	
海2004	0.042	0.128	0.01	0.045	0.249	0.128	0.008	0.008	
海2005	0.034	0.061	0.086	0.017	0.249	0.131	0.069	0.006	
海2006	0.017	0.037	0.099	0.003	0.051	0.064	0.049	0.016	
海2007	0.064	0.031	0.083	0.064	0.052	0.054	0.058	0.002	
海2008	0.381	0.073	0.03	0.004	0.028	0.096	0.002	0.047	
海2009	0.064	0.026	0.088	0.015	0.023	0.078	0.016	0.14	
海2010	0.018	0.047	0.027	0.335	0.019	0.011	0.012	0.016	
海2011	0.041	0.125	0.017	0.047	0.15	0.175	0.035	0.199	
海2012	0.037	0.067	0.023	0.028	0.023	0.09	0.124	0.094	
海2013	0.052	0.097	0.046	0.059	0.081	0.025	0.009	0.153	
海2014	0.113	0.038	0.079	0.102	0.019	0.02	0.004	0.065	
海2015	0.053	0.111	0.158	0.038	0.121	0.052	0.057	0.109	
海2016	0.017	0.134	0.042	0.028	0.076	0.046	0.031	0.06	
海2017	0.045	0.136	0.088	0.052	0.041	0.019	0.08	0.067	
海2018	0.082	0.12	0.031	0.008	0.023	0.027	0.056	0.055	
海2019	0.024	0.009	0.018	0.011	0.022	0.071	0.033	0.024	
海2020	0.011	0.011	0.018	0.029	0.039	0.032	0.072	0.005	
	#9	#10	#11	#12	#13	#14	#15	#16	ケース数
海2004	0.008	0.145	0.019	0.003	0.007	0.056	0.137	0.006	1
海2005	0.012	0.047	0.044	0.014	0.005	0.039	0.179	0.008	1
海2006	0.054	0.108	0.208	0.04	0.093	0.027	0.128	0.006	1
海2007	0.032	0.218	0.051	0.007	0.01	0.026	0.224	0.024	1
海2008	0.001	0.042	0.034	0.003	0.006	0.07	0.132	0.05	1
海2009	0.078	0.028	0.013	0.035	0.008	0.044	0.225	0.118	1
海2010	0.148	0.06	0.199	0.01	0.004	0.033	0.053	0.008	1
海2011	0.015	0.037	0.046	0.004	0.003	0.075	0.026	0.004	1
海2012	0.043	0.046	0.065	0.047	0.124	0.017	0.042	0.132	1
海2013	0.031	0.013	0.038	0.031	0.105	0.068	0.095	0.095	1
海2014	0.272	0.005	0.055	0.087	0.043	0.012	0.022	0.064	1
海2015	0.034	0.006	0.046	0.016	0.057	0.028	0.007	0.108	1
海2016	0.006	0.066	0.052	0.154	0.004	0.119	0.124	0.041	1
海2017	0.087	0.031	0.031	0.048	0.006	0.169	0.045	0.055	1
海2018	0.103	0.067	0.052	0.066	0.024	0.2	0.022	0.065	1
海2019	0.137	0.026	0.326	0.092	0.004	0.183	0.008	0.012	1
海2020	0.029	0.108	0.178	0.176	0.032	0.163	0.015	0.083	1

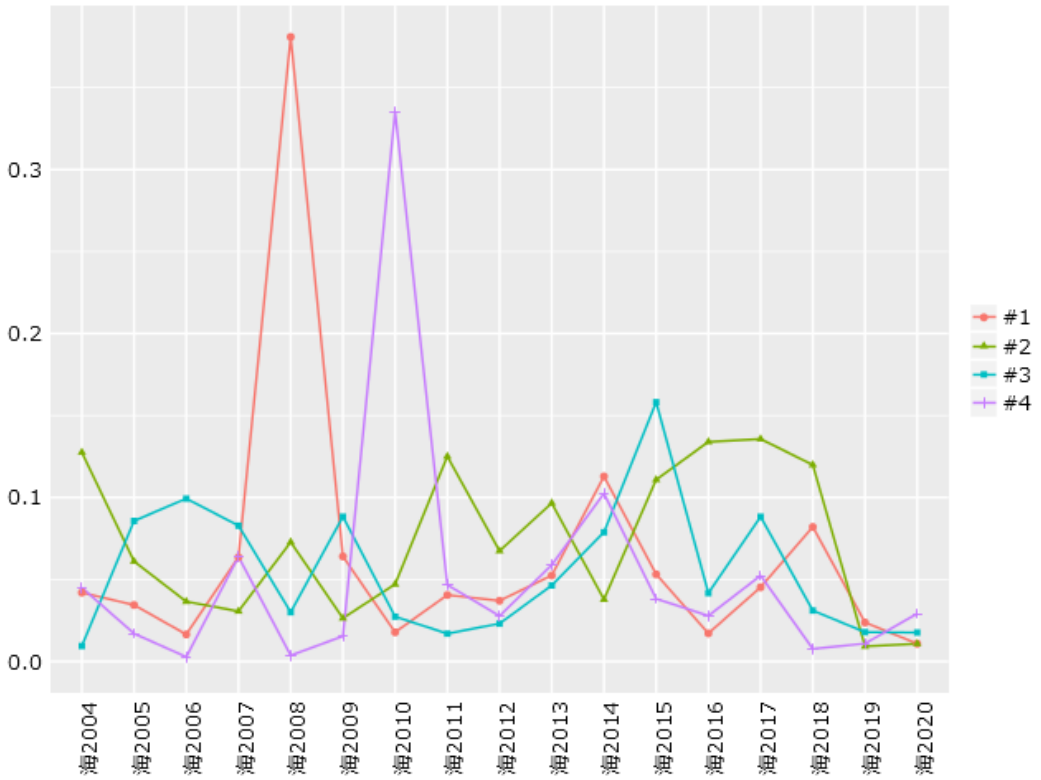


図 4-5 1～4 トピックの比率（16 トピックス、分析対象語数：75）

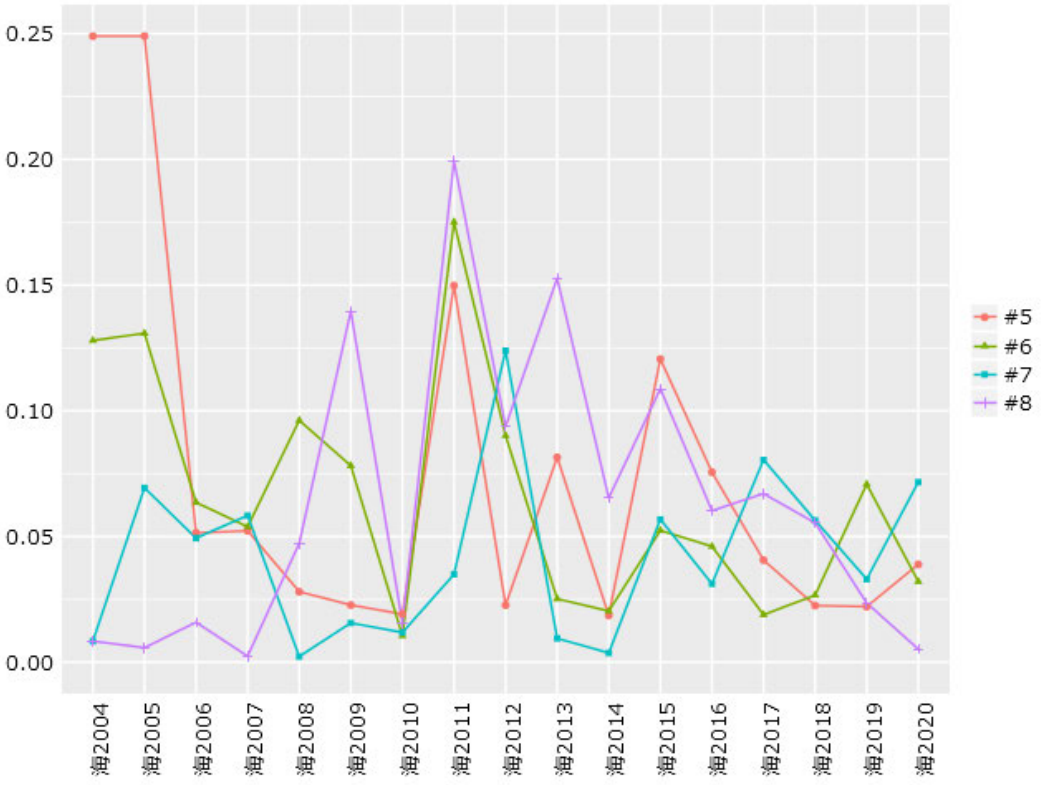


図 4-6 5～8 トピックの比率（16 トピックス、分析対象語数：75）

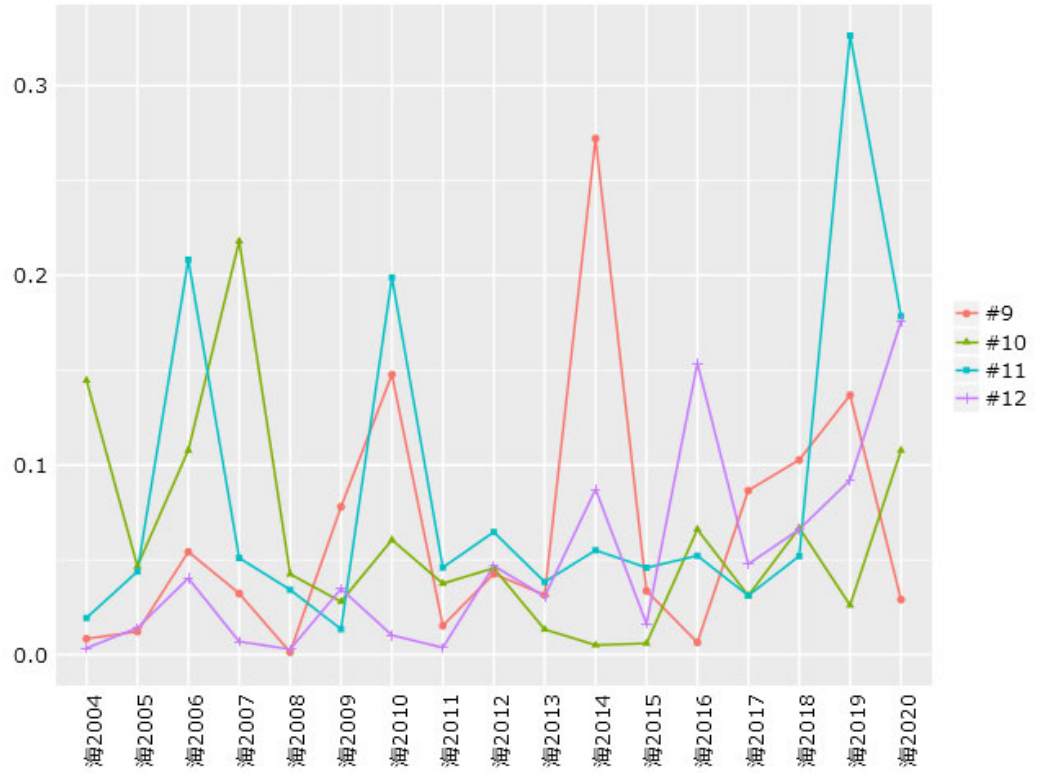


図 4-7 9～12 トピックの比率（16 トピックス、分析対象語数：75）

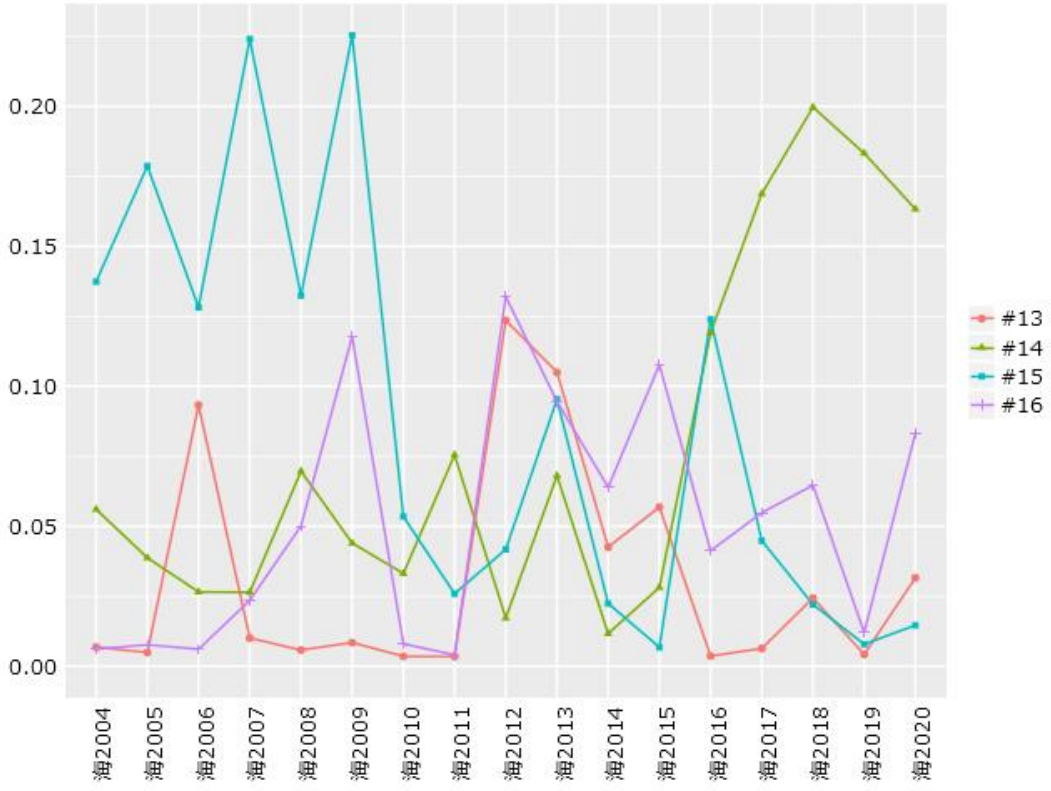


図 4-8 13～16 トピックの比率（16 トピックス、分析対象語数：75）

表 4-3 LDA 処理結果（18 トピックス、分析対象語数：153）

Topics

#1		#2		#3		#4		#5		#6	
北極	0.099	施設	0.071	保全	0.106	調査	0.079	沿岸域	0.130	基本	0.166
開発	0.096	観測	0.051	海岸	0.073	大陸棚	0.074	管理	0.079	政策	0.066
連携	0.042	沿岸	0.051	生態系	0.060	情報	0.071	日本	0.056	計画	0.056
協力	0.038	被害	0.043	重要	0.047	開発	0.061	計画	0.036	本部	0.034
国際	0.038	離島	0.043	利用	0.045	必要	0.059	対策	0.031	資源	0.033
研究	0.036	中国	0.039	生物	0.044	海賊	0.040	制度	0.026	総合的	0.032
世界	0.035	排他的経済水域	0.038	生物多様性	0.041	重要	0.034	総合的	0.026	規定	0.030
日本	0.033	必要	0.038	海域	0.037	管理	0.032	策定	0.026	確保	0.025
支援	0.033	太平洋	0.037	基本	0.037	課題	0.031	会議	0.024	開発	0.023
政策	0.031	問題	0.034	問題	0.036	関係	0.028	排他的経済水域	0.024	整備	0.023

#7		#8		#9		#10		#11		#12	
観測	0.075	持続可能	0.074	管理	0.087	海域	0.087	教育	0.167	情報	0.127
計画	0.062	開催	0.061	離島	0.060	中国	0.077	津波	0.129	船舶	0.064
調査	0.055	会議	0.061	海域	0.058	管理	0.059	被害	0.041	活動	0.057
産業	0.055	国連	0.057	計画	0.051	EEZ	0.054	実施	0.036	必要	0.048
船舶	0.051	目標	0.048	必要	0.046	計画	0.045	洋上	0.035	行動	0.040
問題	0.039	議論	0.039	エネルギー	0.039	策定	0.038	利用	0.031	関連	0.038
情報	0.038	重要	0.038	総合	0.037	状況	0.033	大学	0.030	米国	0.035
基本	0.037	生物多様性	0.029	支援	0.036	資源	0.030	地震	0.030	システム	0.032
管理	0.037	機関	0.028	関係	0.035	影響	0.028	導入	0.029	データ	0.031
利用	0.028	参加	0.028	総合的	0.034	体制	0.027	産業	0.025	安全保障	0.030

#13		#14		#15		#16		#17		#18	
技術	0.051	利用	0.053	管理	0.085	教育	0.110	保全	0.080	地震	0.087
日本	0.047	船舶	0.047	政策	0.083	保全	0.059	計画	0.079	日本	0.048
影響	0.044	評価	0.045	開発	0.074	国際	0.058	開発	0.068	対策	0.044
海底	0.044	産業	0.044	問題	0.051	協力	0.043	港湾	0.051	発生	0.043
研究	0.040	米国	0.040	関係	0.043	社会	0.039	中国	0.041	世界	0.041
資源	0.038	資源	0.040	利用	0.038	基本	0.035	観光	0.039	中国	0.035
海域	0.033	航路	0.038	必要	0.036	資源	0.033	生態系	0.039	安全	0.032
活動	0.032	津波	0.029	協力	0.035	機関	0.029	実施	0.037	観測	0.031
調査	0.029	地球	0.028	国際	0.031	島	0.028	利用	0.034	政府	0.028
エネルギー	0.027	国際	0.027	海域	0.030	国際的	0.026	行動	0.032	技術	0.028

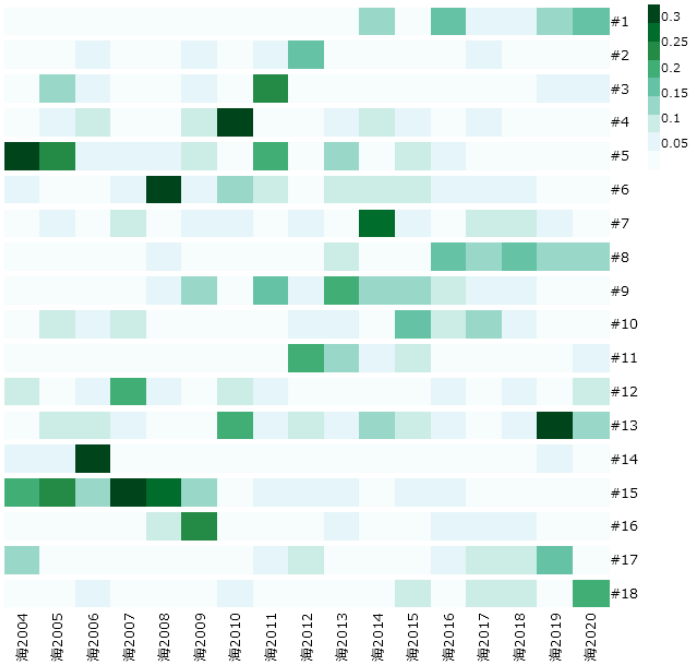


図 4-9　LDA ヒートマップ（18 トピックス、分析対象語数：153）

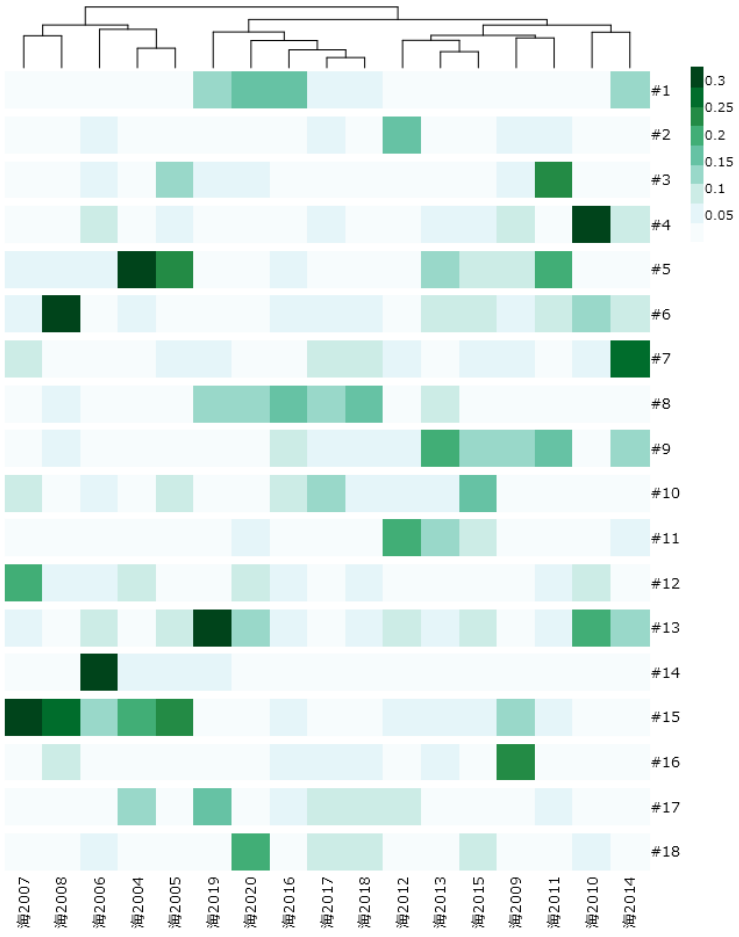


図 4-10　LDA ヒートマップ樹形図（18 トピックス、分析対象語数：153）

表 4-4 トピック比率集計表（18 トピックス、分析対象語数：153）

	#1	#2	#3	#4	#5	#6	#7	#8	#9	
海2004	0.004	0.02	0.029	0.035	0.3	0.038	0.009	0.014	0.024	
海2005	0.014	0.024	0.113	0.067	0.251	0.006	0.04	0.006	0.005	
海2006	0.036	0.051	0.044	0.082	0.048	0.007	0.017	0.012	0.009	
海2007	0.004	0.005	0.037	0.006	0.039	0.046	0.073	0.027	0.003	
海2008	0.004	0.004	0.033	0.019	0.052	0.318	0.006	0.042	0.048	
海2009	0.033	0.066	0.048	0.078	0.081	0.038	0.059	0.036	0.11	
海2010	0.005	0.006	0.021	0.304	0.03	0.136	0.049	0.014	0.005	
海2011	0.007	0.052	0.216	0.013	0.189	0.086	0.012	0.022	0.176	
海2012	0.035	0.177	0.023	0.034	0.028	0.027	0.043	0.007	0.062	
海2013	0.017	0.008	0.013	0.057	0.109	0.101	0.024	0.085	0.191	
海2014	0.114	0.012	0.034	0.074	0.035	0.083	0.281	0.003	0.109	
海2015	0.005	0.004	0.016	0.054	0.101	0.084	0.046	0.006	0.131	
海2016	0.18	0.018	0.004	0.024	0.071	0.037	0.011	0.146	0.081	
海2017	0.069	0.043	0.004	0.055	0.033	0.059	0.092	0.12	0.066	
海2018	0.072	0.017	0.006	0.006	0.008	0.064	0.091	0.163	0.061	
海2019	0.126	0.024	0.043	0.003	0.011	0.006	0.052	0.126	0.014	
海2020	0.168	0.015	0.052	0.003	0.011	0.008	0.036	0.119	0.009	
	#10	#11	#12	#13	#14	#15	#16	#17	#18	ケース数
海2004	0.012	0.003	0.106	0.006	0.067	0.189	0.003	0.129	0.011	1
海2005	0.077	0.001	0.006	0.073	0.045	0.227	0.014	0.019	0.01	1
海2006	0.065	0.01	0.04	0.093	0.303	0.121	0.003	0.014	0.043	1
海2007	0.096	0.018	0.195	0.067	0.008	0.306	0.012	0.025	0.032	1
海2008	0.03	0.006	0.041	0.009	0.012	0.257	0.092	0.024	0.003	1
海2009	0.011	0.004	0.016	0.034	0.027	0.117	0.229	0.003	0.01	1
海2010	0.009	0.005	0.103	0.193	0.015	0.023	0.025	0.009	0.048	1
海2011	0.033	0.001	0.04	0.038	0.007	0.047	0.018	0.041	0.004	1
海2012	0.064	0.193	0.021	0.094	0.018	0.053	0.018	0.082	0.022	1
海2013	0.047	0.111	0.023	0.041	0.004	0.058	0.06	0.016	0.036	1
海2014	0.003	0.055	0.017	0.112	0.014	0.016	0.009	0.021	0.008	1
海2015	0.157	0.106	0.012	0.093	0.033	0.04	0.015	0.016	0.082	1
海2016	0.089	0.007	0.066	0.039	0.02	0.055	0.052	0.068	0.03	1
海2017	0.126	0.006	0.014	0.027	0.007	0.026	0.065	0.099	0.089	1
海2018	0.068	0.032	0.059	0.054	0.009	0.033	0.065	0.102	0.092	1
海2019	0.01	0.003	0.028	0.324	0.045	0.02	0.005	0.151	0.008	1
海2020	0.014	0.061	0.101	0.139	0.01	0.025	0.008	0.024	0.196	1

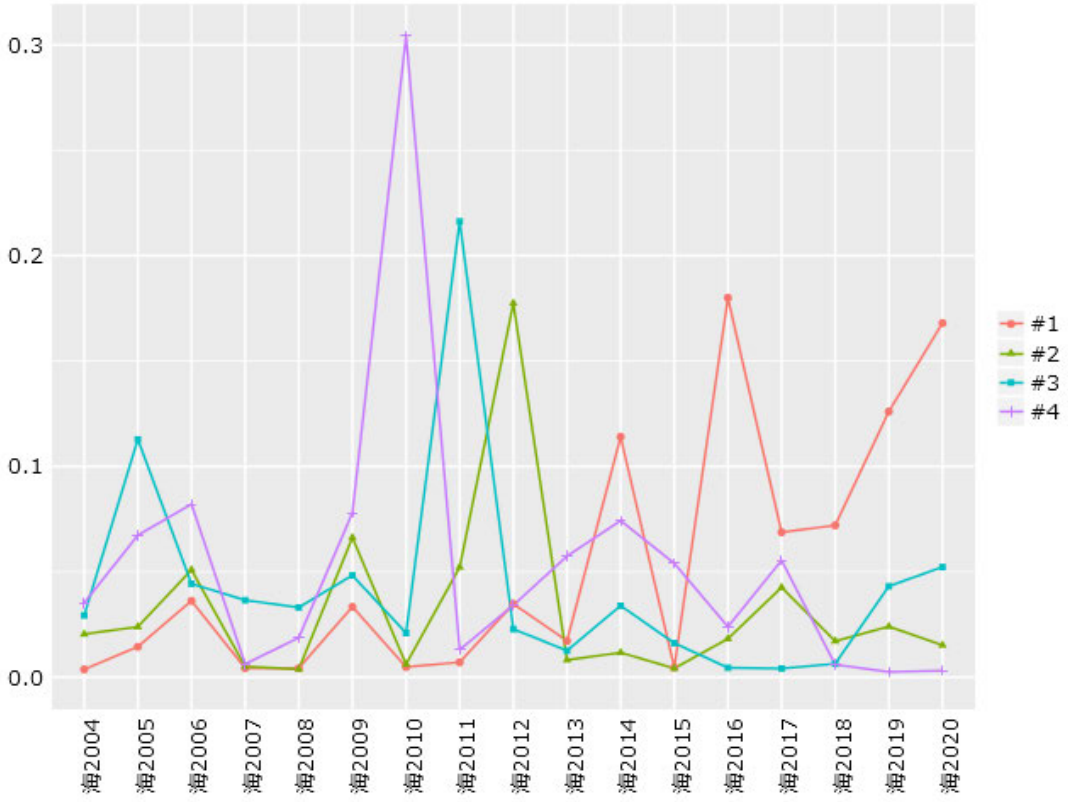


図 4-11　1～4 トピックの比率（18 トピックス、分析対象語数：153）

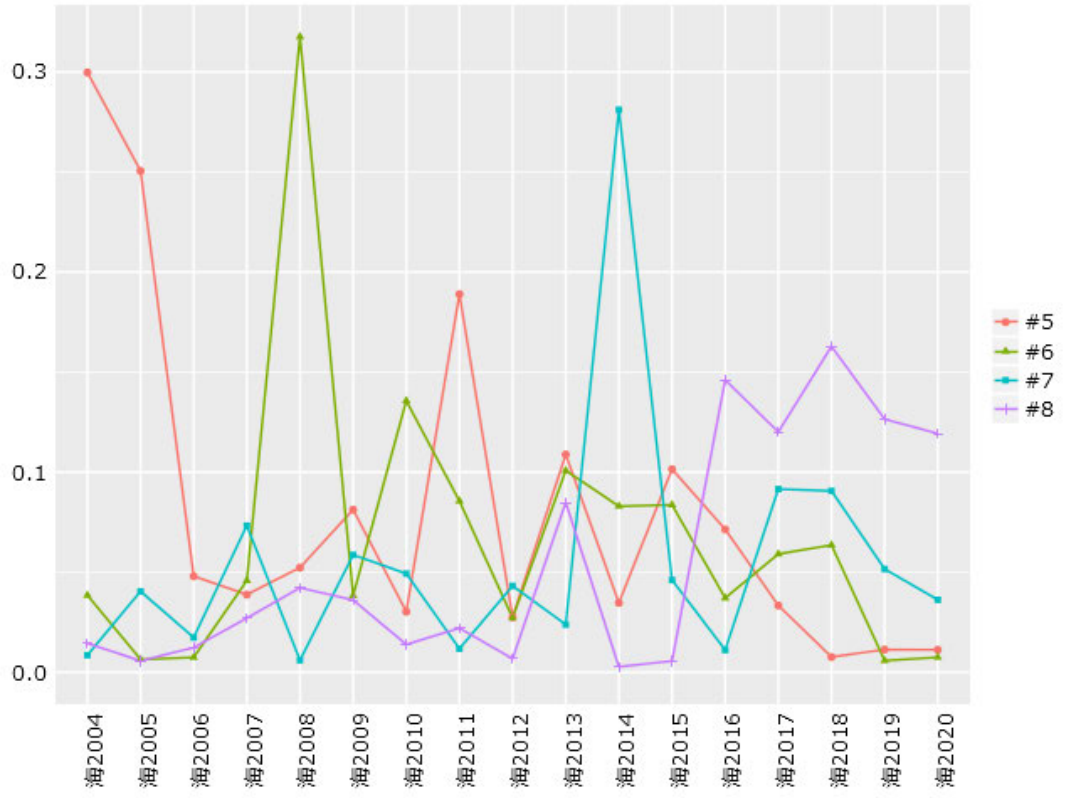


図 4-12　5～8 トピックの比率（18 トピックス、分析対象語数：153）

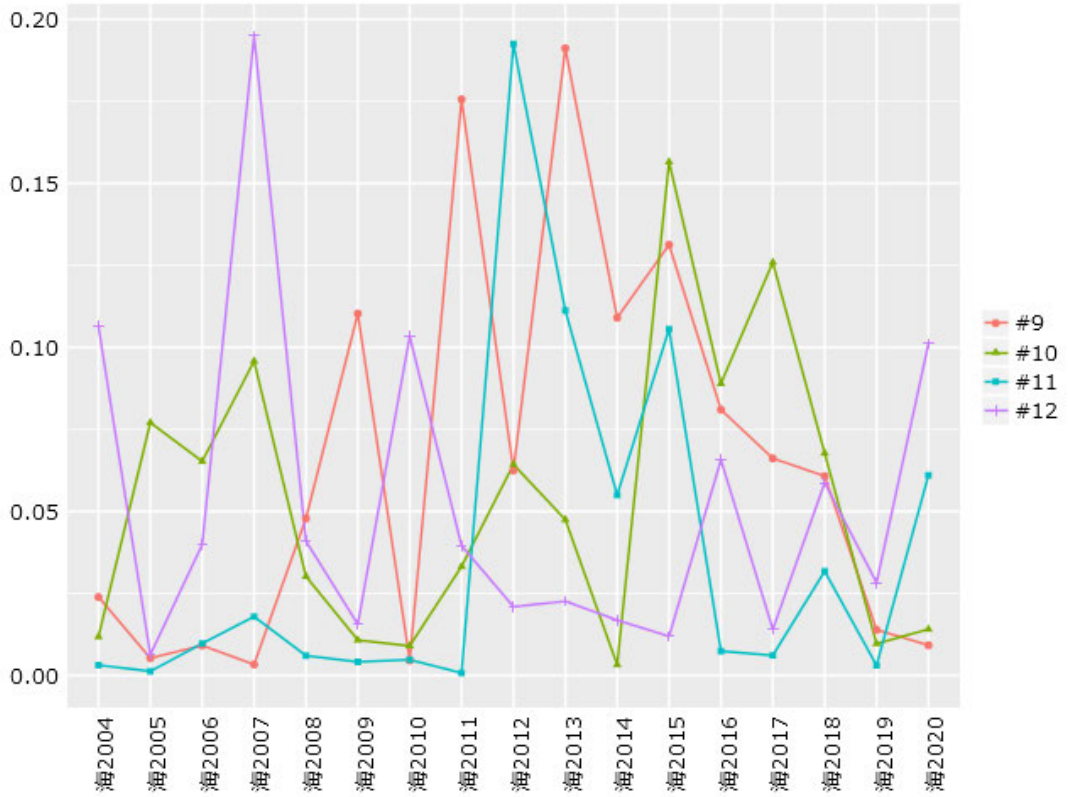


図 4-13 9～12 トピックの比率（18 トピックス、分析対象語数：153）

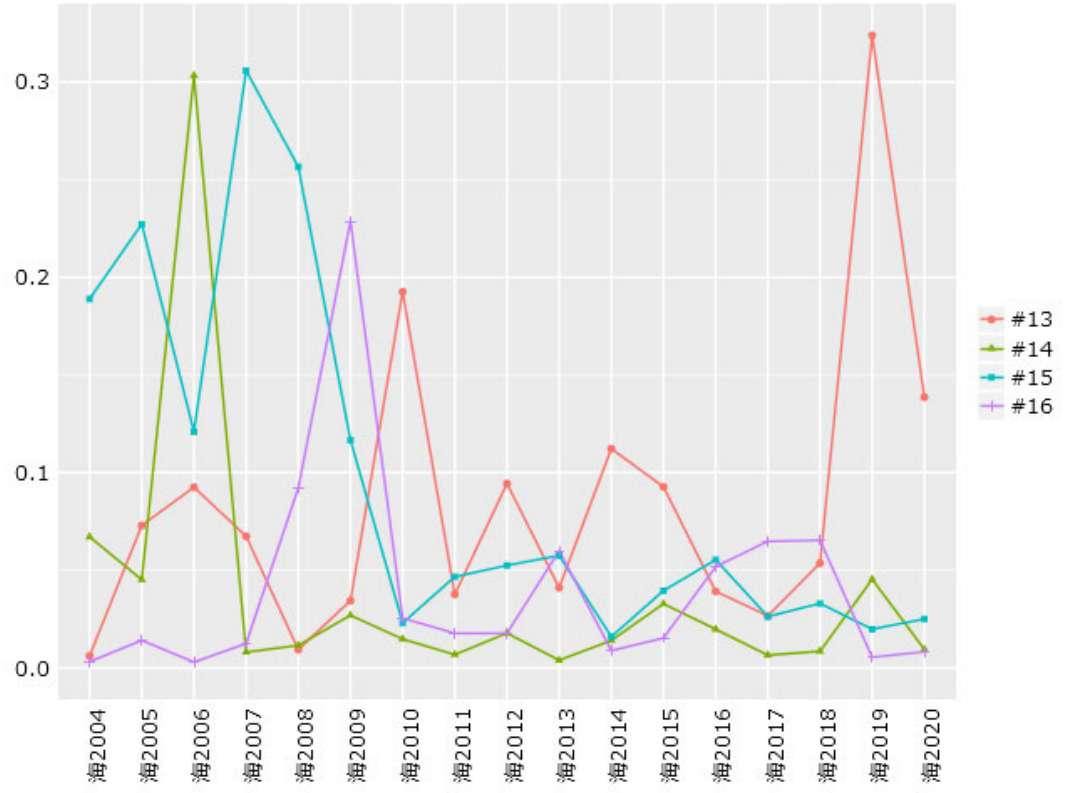


図 4-14 13～16 トピックの比率（18 トピックス、分析対象語数：153）

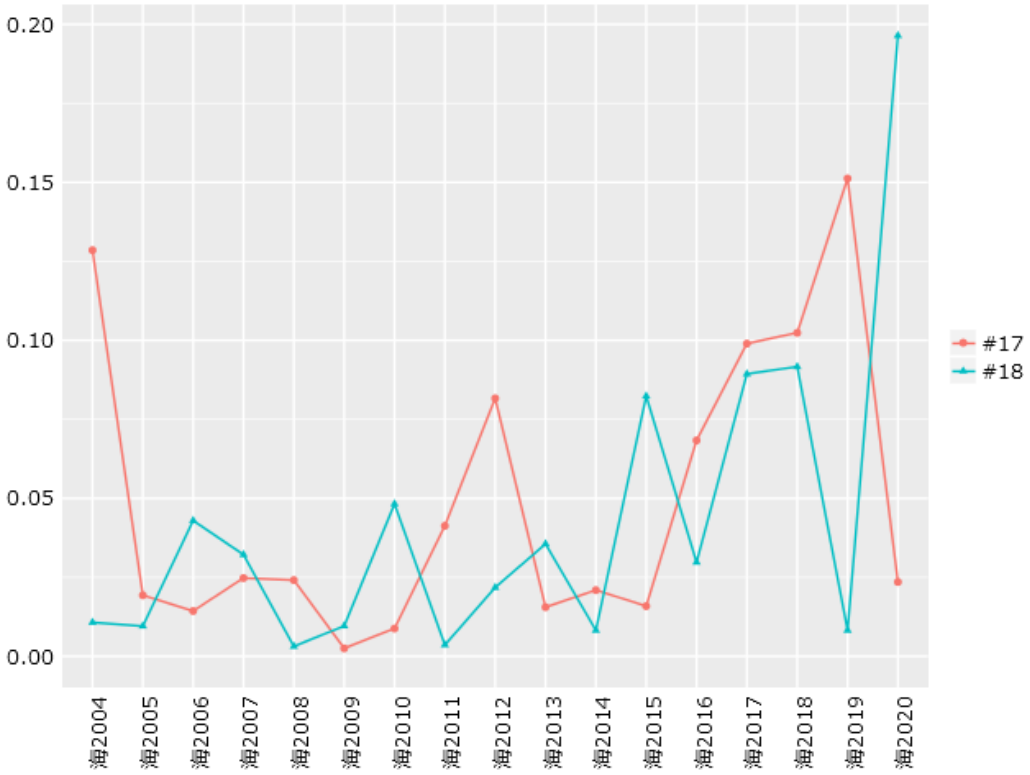


図 4-15 17～18 トピックの比率（18 トピックス、分析対象語数：153）

付録 7「環境白書（2008～2020 年）の分析結果」

文書番号：JRDN-21-026

1. 前処理結果

環境白書･海洋白書･水産白書（2008～2020 年）の分析で設定した強制抽出語（316 語）と 54 語の除外語を設定し、「動詞、感動詞、動詞 B、副詞 B」を除外して前処理を実行した。

Chanse での前処理の結果、総抽出語数：2,239,286、異なり語数：23,447 のうち 16,490 語が分析処理で使用された。抽出語出現数の頻度分布を図 1-1、抽出語リスト（上位 200 語）を図 1 -2 に示す。

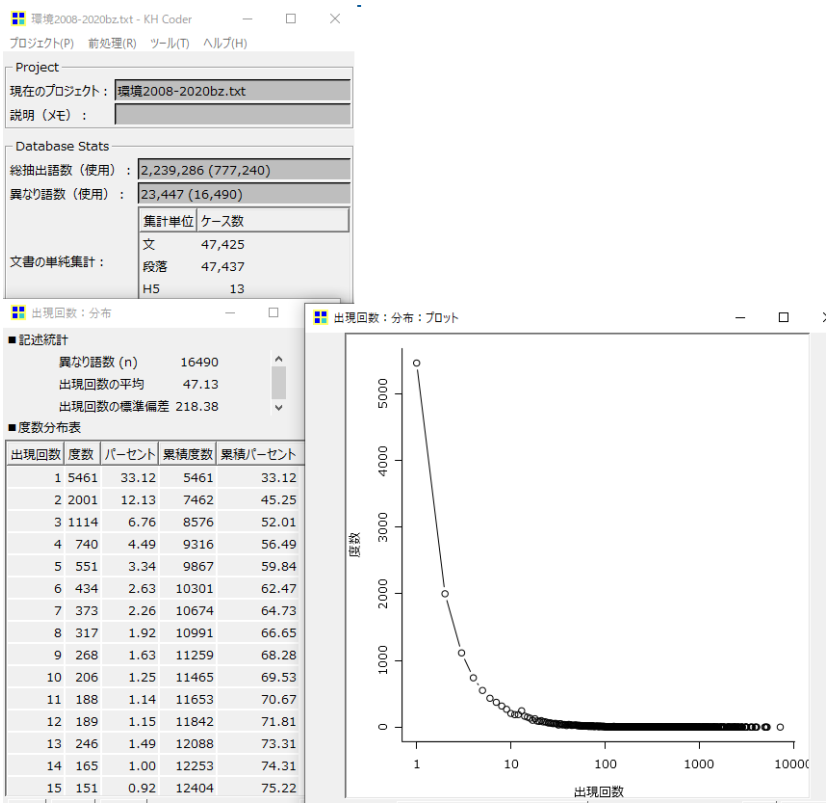


図 1-1 抽出語出現数の頻度分布

#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度
1	実施	サ変名詞	7214	36	基準	名詞	2311				
2	対策	サ変名詞	5234	37	排出量	タグ	2286				
3	処理	サ変名詞	5150	38	削減	サ変名詞	2272				
4	利用	サ変名詞	5044	39	制度	名詞	2243				
5	廃棄物	タグ	4944	40	研究	サ変名詞	2188				
6	社会	名詞	4057	41	検討	サ変名詞	2186	71	都市	名詞	1581
7	資源	名詞	4006	42	リサイクル	サ変名詞	2137	72	指定	サ変名詞	1555
8	保全	サ変名詞	3909	43	温暖化	タグ	2136	73	生態系	タグ	1544
9	計画	サ変名詞	3648	44	発生	サ変名詞	2076	74	導入	サ変名詞	1529
10	調査	サ変名詞	3547	45	規制	サ変名詞	2057	75	減少	サ変名詞	1523
11	技術	名詞	3514	46	事業者	タグ	2052	76	普及	サ変名詞	1513
12	情報	名詞	3173	47	物質	名詞	2048	77	提供	サ変名詞	1501
13	施設	サ変名詞	3146	48	連携	サ変名詞	2034	78	回収	サ変名詞	1467
14	支援	サ変名詞	3140	49	森林	名詞	2028	79	基本	名詞	1460
15	生物多様性	タグ	3120	50	循環	サ変名詞	2004	80	条約	名詞	1444
16	活動	サ変名詞	3095	51	使用	サ変名詞	1963	81	保護	サ変名詞	1442
17	整備	サ変名詞	3053	52	協力	サ変名詞	1950	82	機関	名詞	1440
18	影響	サ変名詞	2863	53	温室効果ガス	タグ	1840	83	重要	形容動詞	1438
19	評価	サ変名詞	2842	54	防止	サ変名詞	1838	84	措置	サ変名詞	1426
20	地球	名詞	2805	55	日本	地名	1832	85	化学物質	タグ	1420
21	管理	サ変名詞	2793	56	汚染	サ変名詞	1818	86	構築	サ変名詞	1419
22	促進	サ変名詞	2698	57	対象	名詞	1815	87	産業	名詞	1418
23	自然	形容動詞	2635	58	発電	サ変名詞	1810	88	生産	サ変名詞	1403
24	開催	サ変名詞	2627	59	策定	サ変名詞	1762	89	高い	形容詞	1377
25	活用	サ変名詞	2550	60	気候変動	タグ	1759	90	環境省	組織名	1375
26	開発	サ変名詞	2547	61	結果	副詞可能	1724	91	地方公共団体	タグ	1343
27	法律	名詞	2541	62	達成	サ変名詞	1723	92	適正	形容動詞	1332
28	経済	名詞	2537	63	システム	名詞	1697	93	国内	名詞	1325
29	状況	名詞	2464	64	増加	サ変名詞	1694	94	原子力	名詞	1310
30	エネルギー	名詞	2408	65	持続可能	タグ	1692	95	分野	名詞	1298
31	目標	名詞	2396	66	国際	名詞	1676	96	強化	サ変名詞	1286
32	必要	形容動詞	2388	67	企業	名詞	1669	97	生活	サ変名詞	1262
33	関係	サ変名詞	2383	68	再生	サ変名詞	1664	98	実現	サ変名詞	1257
34	排出	サ変名詞	2361	69	問題	ナイ形容	1598	99	製品	名詞	1242
35	世界	名詞	2345	70	自動車	名詞	1589	100	目的	名詞	1238

#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度	#	抽出語	品詞/活用	頻度
101	生物	名詞	1237	136	施行	サ変名詞	988				
102	被害	名詞	1237	137	改正	サ変名詞	986				
103	循環型	タグ	1225	138	拡大	サ変名詞	984				
104	課題	名詞	1205	139	国際的	タグ	972				
105	政策	名詞	1201	140	適切	形容動詞	971				
106	健康	形容動詞	1193	141	水質	名詞	958	171	福島	地名	851
107	委員会	タグ	1192	142	把握	サ変名詞	958	172	変化	サ変名詞	850
108	確保	サ変名詞	1192	143	作成	サ変名詞	944	173	特別	形容動詞	848
109	消費	サ変名詞	1190	144	効果	名詞	939	174	再生可能エネルギー	タグ	845
110	参加	サ変名詞	1167	145	関連	サ変名詞	932	175	エコ	名詞	844
111	会議	サ変名詞	1157	146	プロジェクト	名詞	931	176	サービス	サ変名詞	843
112	都道府県	名詞	1157	147	燃料	名詞	923	177	行政	名詞	841
113	対応	サ変名詞	1151	148	人	名詞C	921	178	報告書	タグ	837
114	政府	名詞	1150	149	団体	名詞	920	179	生息	サ変名詞	834
115	設置	サ変名詞	1138	150	土壌	名詞	912	180	回	名詞C	830
116	戦略	名詞	1135	151	公表	サ変名詞	908	181	グリーン	名詞	820
117	向上	サ変名詞	1127	152	開始	サ変名詞	904	182	積極的	タグ	816
118	行動	サ変名詞	1117	153	可能性	タグ	903	183	法	名詞C	802
119	低炭素	タグ	1104	154	場合	副詞可能	901	184	確認	サ変名詞	798
120	市町村	名詞	1097	155	大臣	名詞	897	185	大気	名詞	796
121	モニタリング	名詞	1095	156	設定	サ変名詞	894	186	COP	タグ	793
122	収集	サ変名詞	1084	157	改善	サ変名詞	887	187	報告	サ変名詞	791
123	形成	サ変名詞	1082	158	自然環境	タグ	886	188	人口	名詞	787
124	特定	サ変名詞	1078	159	3R	未知語	881	189	具体的	タグ	785
125	配慮	サ変名詞	1075	160	途上国	タグ	879	190	家庭	名詞	784
126	公害	名詞	1072	161	省エネ	タグ	872	191	成長	サ変名詞	784
127	観測	サ変名詞	1053	162	中心	名詞	872	192	災害	名詞	774
128	貢献	サ変名詞	1032	163	国民	名詞	869	193	負荷	サ変名詞	771
129	アジア	地名	1023	164	建設	サ変名詞	867	194	方針	名詞	769
130	測定	サ変名詞	1023	165	可能	形容動詞	862	195	手法	名詞	767
131	処分	サ変名詞	1022	166	騒音	名詞	856	196	有効	形容動詞	755
132	CO2	未知語	1014	167	一般	名詞	852	197	ネットワーク	名詞	753
133	製造	サ変名詞	1005	168	リスク	名詞	851	198	交通	名詞	752
134	会合	サ変名詞	1004	169	規模	名詞	851	199	主体	名詞	743
135	認定	サ変名詞	999	170	国立公園	タグ	851	200	割合	名詞	739

図 1-2 抽出語リスト（上位 200 語）

2. 共起ネットワーク

KHCoder により自動設定される最小出現頻度：1595、分析対象語数：69 とした場合と、最小出現頻度：900、分析対象語数：154 の結果を示す。

分析対象語数：69 での共起ネットワークの描画条件を以下に、語・語での共起ネットワークを図 2-1、語・年での共起ネットワークを図 2-2 に示す。

また、分析対象語数：154 での共起ネットワークの描画条件を以下に、語・語での共起ネットワークを図 2-3、語・年での共起ネットワークを図 2-4 に示す。

	最小出現頻度	分析対象語数	描画結果
語・語	1595	69	ノード数：56　エッジ数：69
語・年	1595	69	ノード数：34　エッジ数：69

	最小出現頻度	分析対象語数	描画結果
語・語	900	154	ノード数：112　エッジ数：154
語・年	900	154	ノード数：56　エッジ数：154

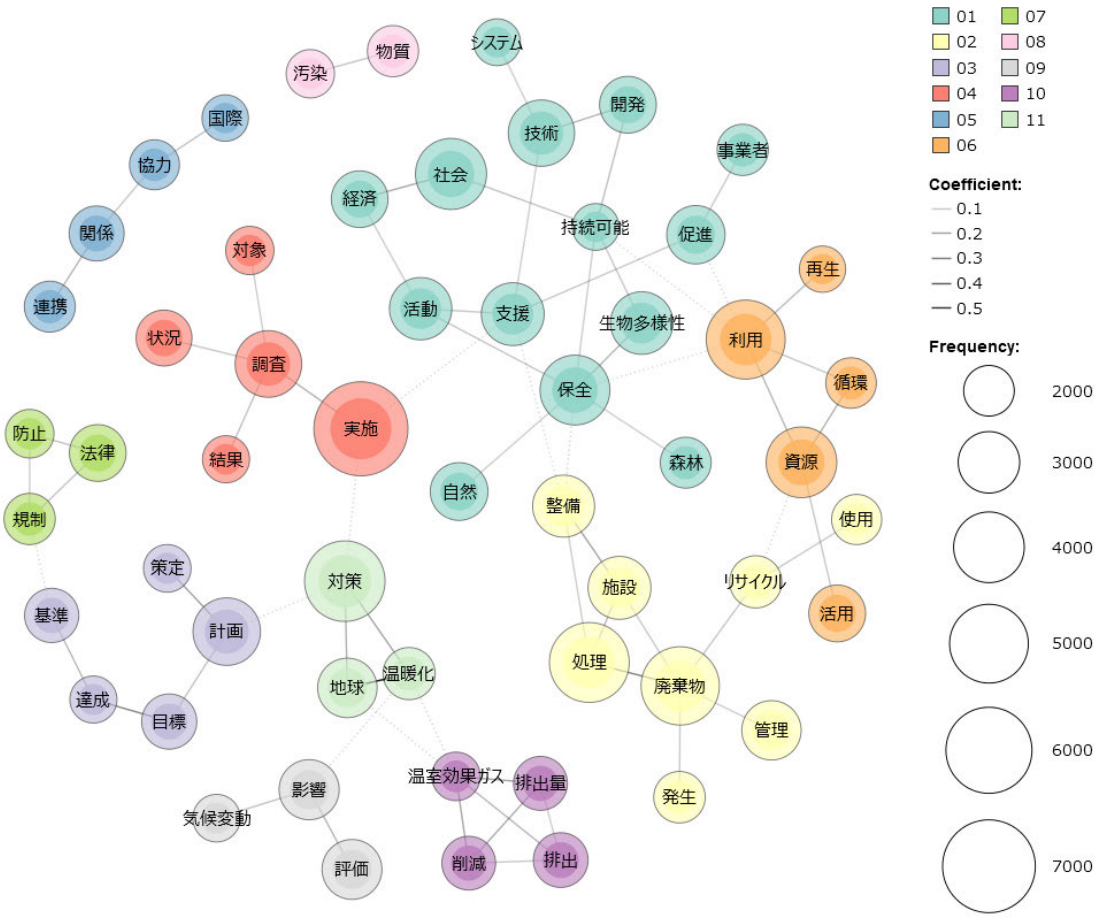


図 2-1　共起ネットワーク（語・語）（分析対象語数：69）

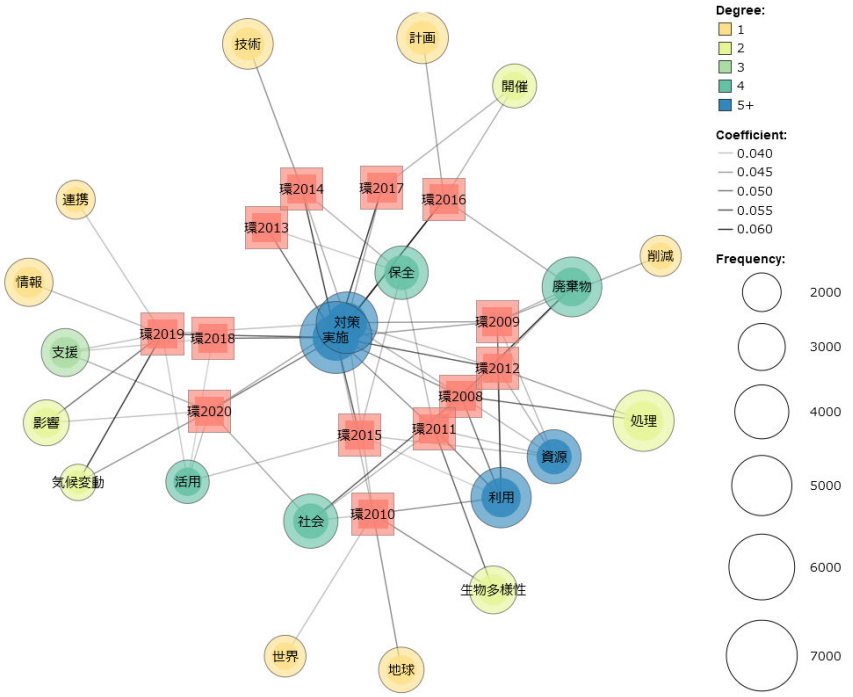


図 2-2 共起ネットワーク（語・年）（分析対象語数：69）

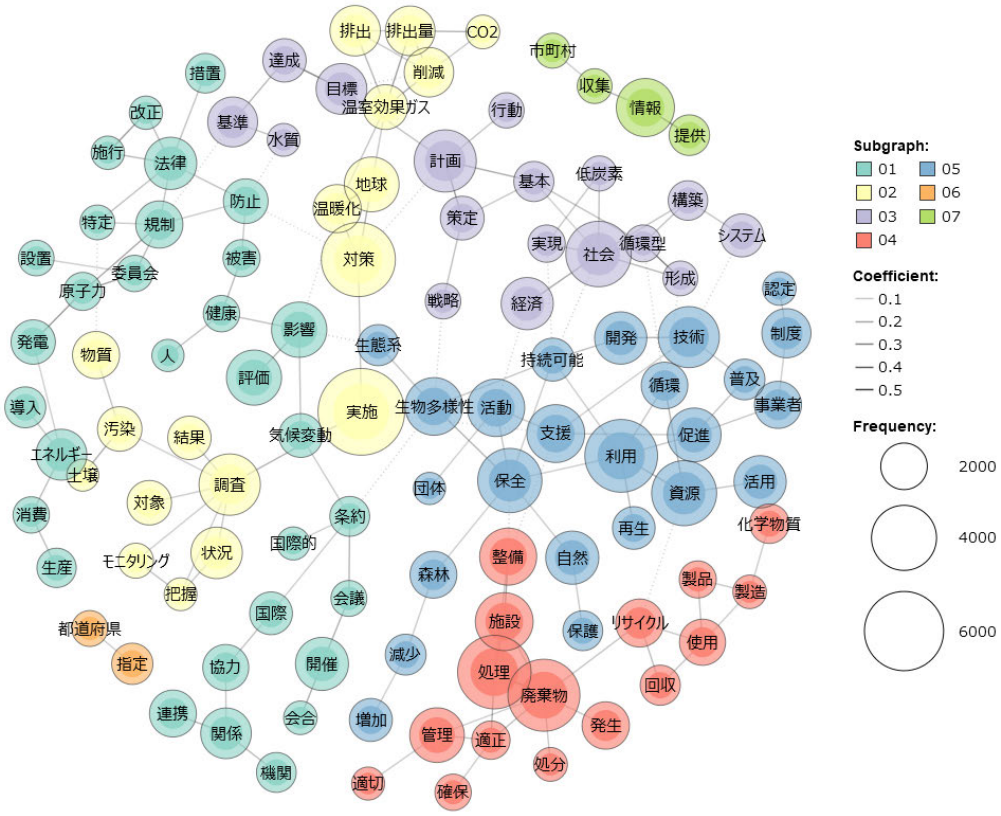


図 2-3 共起ネットワーク（語・語）（分析対象語数：154）

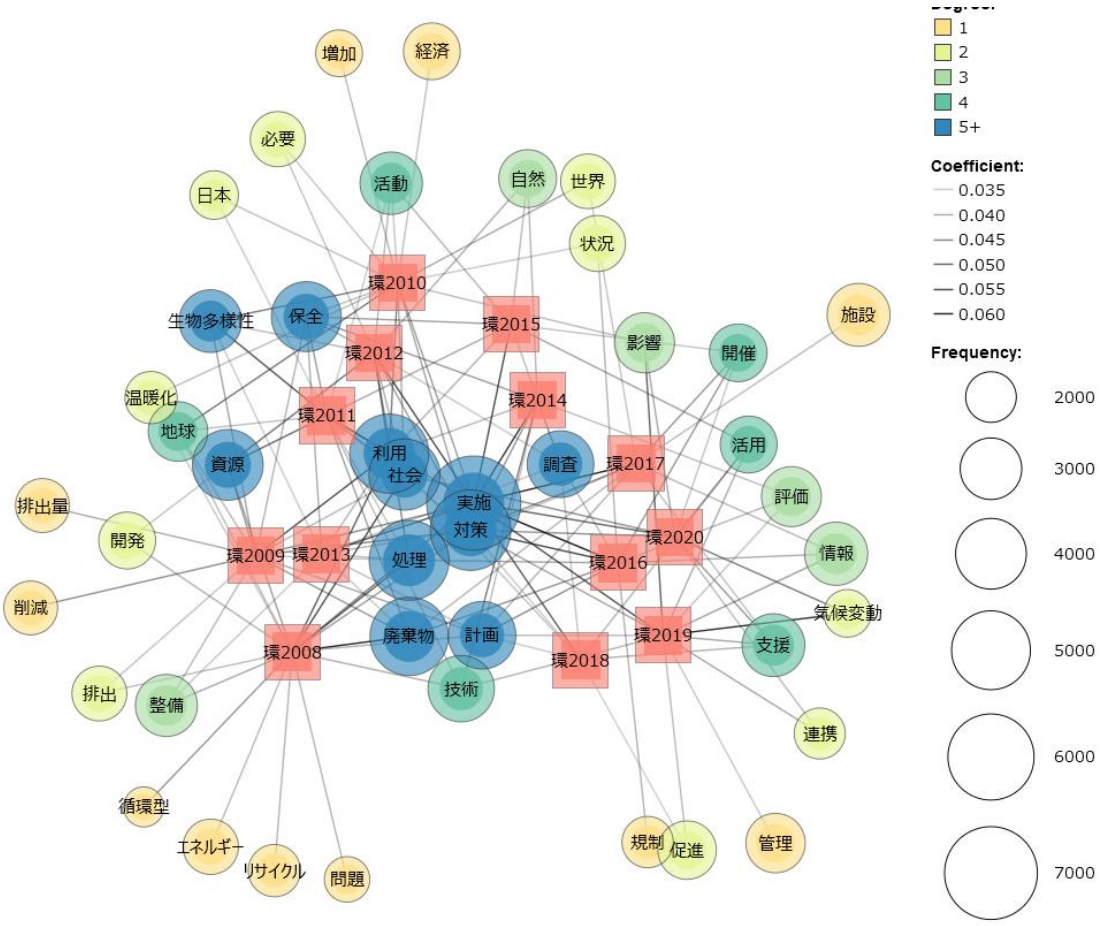


図 2-4 共起ネットワーク（語・年）（分析対象語数：154）

3. 対応分析

KHCoder により自動設定される最小出現頻度：1595、分析対象語数：69 とした場合と、最小出現頻度：900、分析対象語数：154 で対応分析処理を行った。

分析対象語数：69 の累積寄与率は成分 1 と 2 で 54.68%であり、成分 3 と 4 の累積寄与率は約 25%である。その結果を図 3-1,3-2 に示す。

分析対象語数：154 の累積寄与率は成分 1 と 2 で 55.52%であり、成分 3 と 4 の累積寄与率は約 21%である。その結果を図 3-3,3-4 に示す。

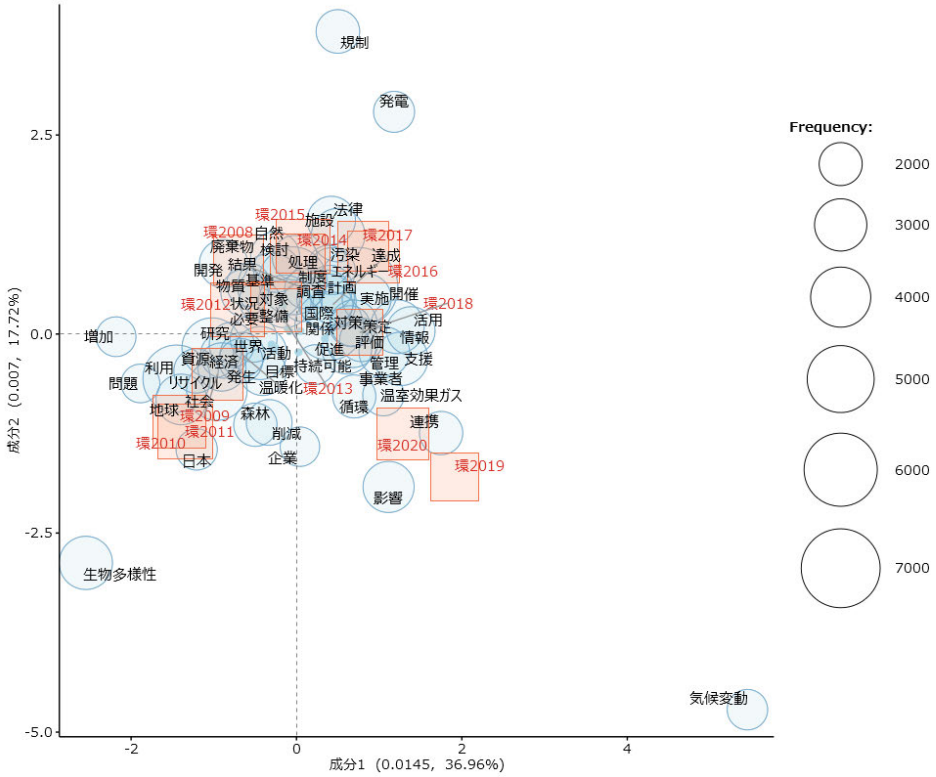


図 3-1 語･年の対応分析結果 （成分 1 と 2）（分析対象語数：69）

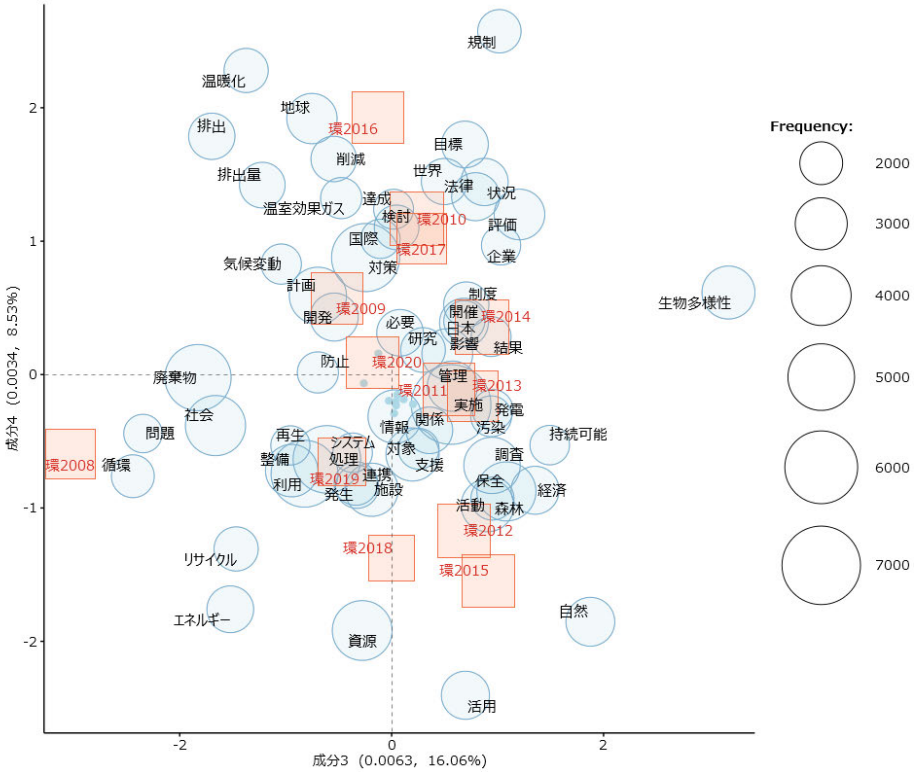


図 3-2 語･年の対応分析結果 （成分 3 と 4）（分析対象語数：69）

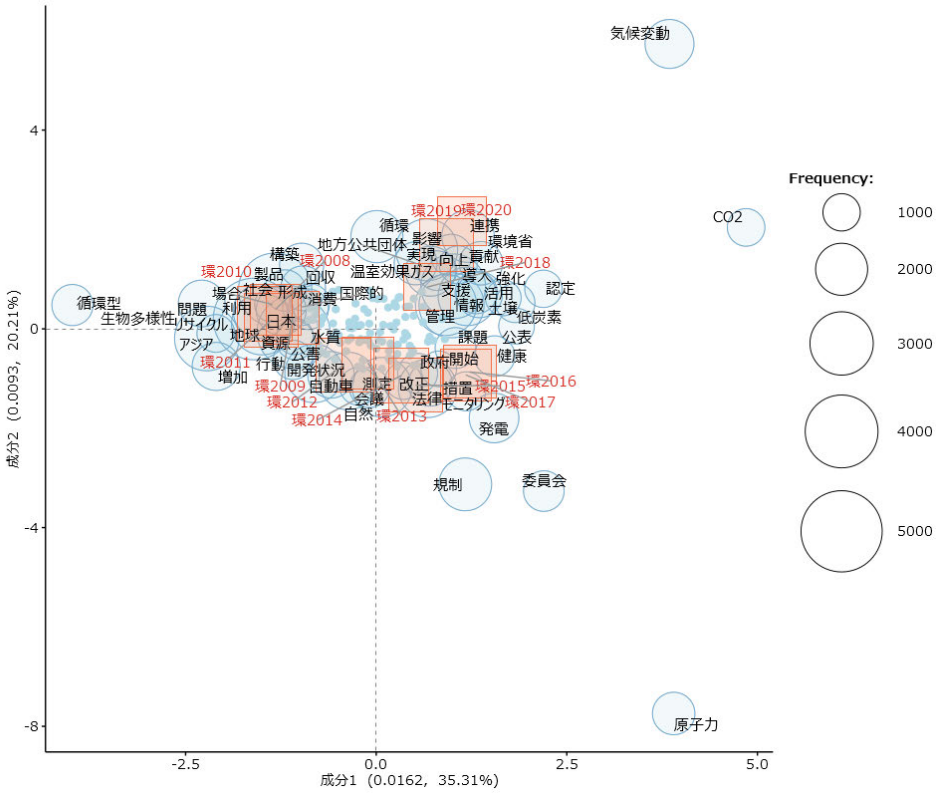


図 3-3 語･年の対応分析結果（成分 1 と 2）（分析対象語数：154）

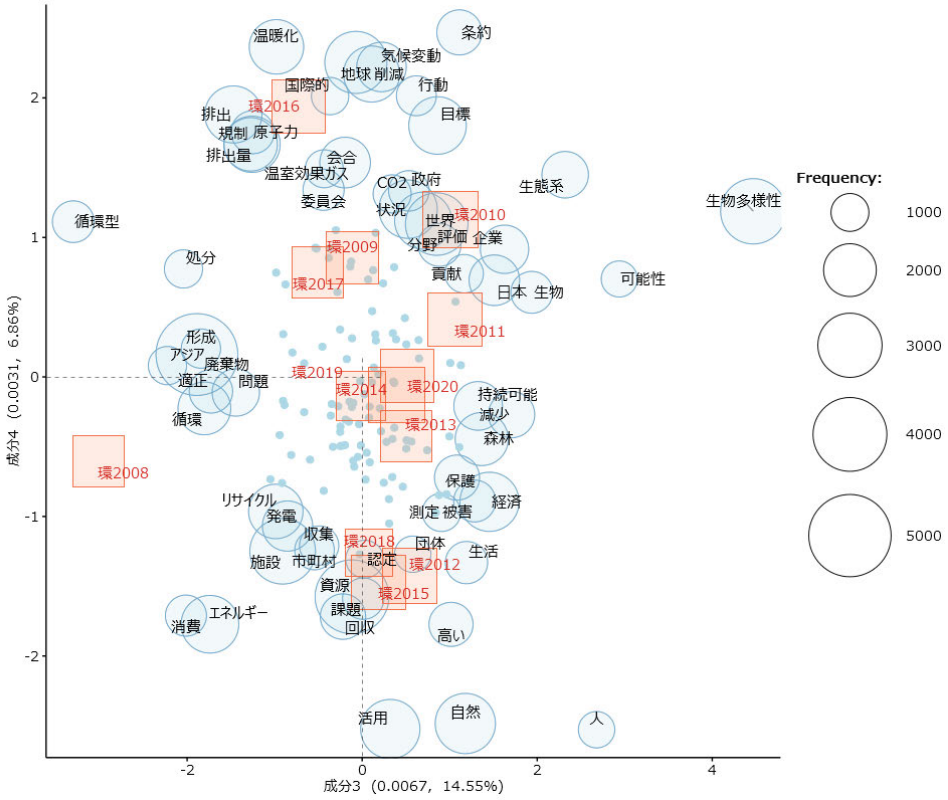


図 3-4 語･年の対応分析結果（成 3 と 4）（分析対象語数：154）

4. LDA 分析

KHCoder により自動設定される最小出現数：1595、分析対象語数：69 で集計単位を H5（年）とした場合と、分析対象語数：154 での LDA 分析を行った。

分析対象語数：69 での LDA トピック数推定結果を図 4-1、分析対象語数：154 での結果を図 4-2 に示す。これらの図から 69 語でのトピック数は 16、154 語では 20 と推察した。

分析対象語数：69、トピック数：16 での LDA 処理結果を表 4-1、そのヒートマップを図 4-3、ヒートマップ樹形図を図 4-4、*トピック比率集計表を表 4-2、トピック比率を図 4-5～8 に示す。

また、分析対象語数：154、トピック数：20 でのその LDA 処理結果を表 4-3、そのヒートマップを図 4-9、ヒートマップ樹形図を図 4-10、*トピック比率集計表を表 4-4、トピック比率を図 4-11～15 に示す。

*は別途 excel 形式で提供。

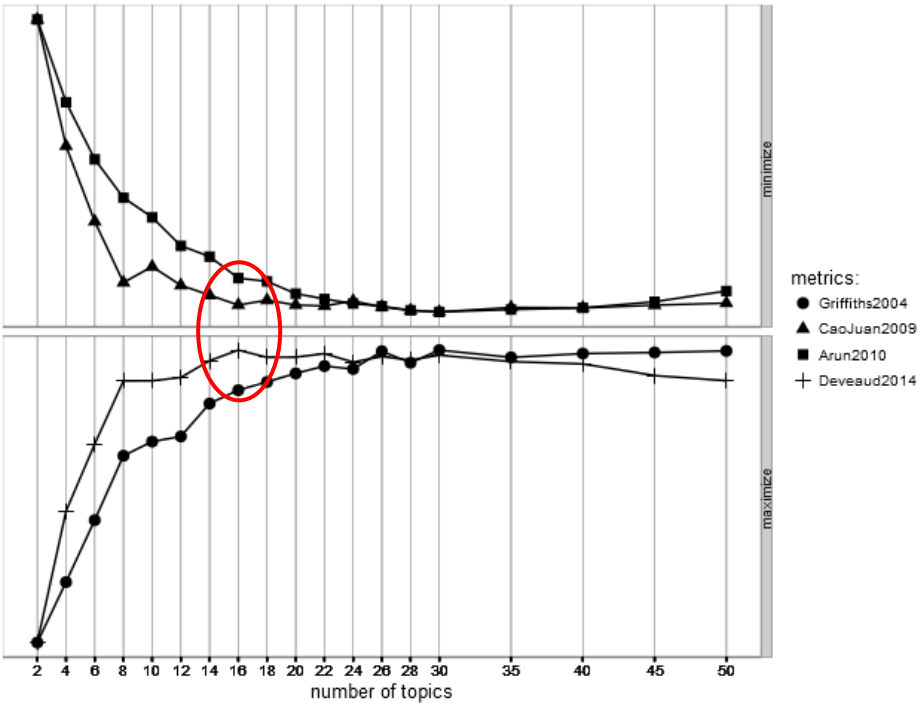


図 4-1 LDAtuning 実行結果（分析対象語数：69）

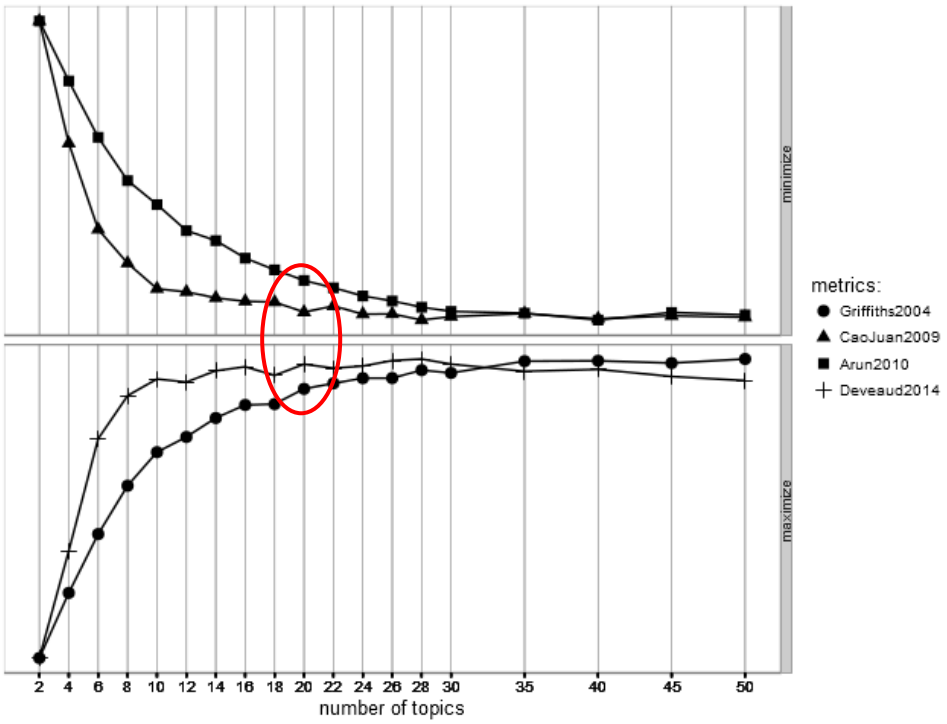


図 4-2 LDAtuning 実行結果（分析対象語数：154）

表 4-1 LDA 処理結果（16 トピックス、分析対象語数：69）

Topics

#1		#2		#3		#4		#5		#6		#7		#8	
削減	0.156	法律	0.143	リサイクル	0.097	廃棄物	0.178	対策	0.202	増加	0.154	技術	0.246	活用	0.110
排出量	0.131	規制	0.142	社会	0.092	地球	0.113	処理	0.122	経済	0.128	基準	0.089	支援	0.099
排出	0.109	発電	0.099	利用	0.092	社会	0.091	必要	0.107	温暖化	0.095	検討	0.079	処理	0.089
対策	0.075	検討	0.077	廃棄物	0.086	実施	0.089	達成	0.102	研究	0.081	防止	0.074	経済	0.085
計画	0.070	開発	0.074	問題	0.082	処理	0.081	目標	0.088	影響	0.056	汚染	0.070	廃棄物	0.080
利用	0.070	状況	0.055	開発	0.071	温暖化	0.061	管理	0.079	世界	0.054	調査	0.066	発電	0.065
温室効果ガス	0.058	計画	0.054	整備	0.065	排出	0.060	実施	0.067	必要	0.053	対象	0.047	関係	0.064
管理	0.047	基準	0.051	協力	0.054	再生	0.053	関係	0.049	利用	0.052	事業者	0.043	施設	0.053
目標	0.042	評価	0.046	処理	0.047	使用	0.050	自然	0.048	保全	0.050	開催	0.043	整備	0.041
整備	0.041	処理	0.032	発生	0.047	システム	0.037	使用	0.019	地球	0.048	関係	0.039	管理	0.040

#9		#10		#11		#12		#13		#14		#15		#16	
情報	0.129	世界	0.119	社会	0.156	気候変動	0.182	実施	0.146	実施	0.198	資源	0.157	生物多様性	0.219
制度	0.096	評価	0.087	法律	0.139	影響	0.124	自然	0.145	開催	0.131	エネルギー	0.155	利用	0.106
計画	0.094	状況	0.085	発生	0.104	連携	0.074	調査	0.082	活動	0.100	利用	0.139	資源	0.087
物質	0.083	国際	0.080	循環	0.073	対策	0.073	対策	0.080	持続可能	0.085	施設	0.097	森林	0.067
策定	0.083	企業	0.075	企業	0.070	情報	0.071	使用	0.070	保全	0.083	循環	0.068	地球	0.058
促進	0.074	支援	0.067	エネルギー	0.065	支援	0.067	評価	0.062	事業者	0.061	森林	0.047	日本	0.045
整備	0.070	施設	0.064	保全	0.057	評価	0.040	再生	0.042	調査	0.055	日本	0.042	影響	0.042
研究	0.069	開発	0.063	必要	0.055	管理	0.040	保全	0.041	協力	0.047	連携	0.034	保全	0.042
調査	0.065	結果	0.054	問題	0.053	促進	0.040	基準	0.038	温室効果ガス	0.025	整備	0.032	目標	0.040
資源	0.045	制度	0.053	増加	0.047	循環	0.038	管理	0.037	防止	0.024	リサイクル	0.032	削減	0.037

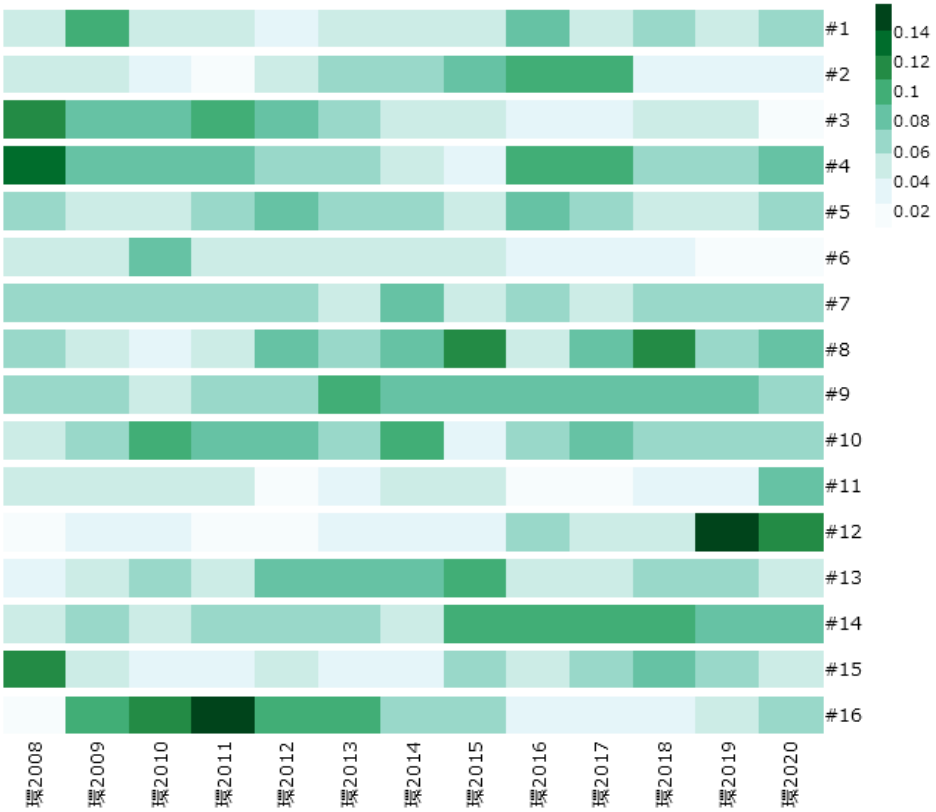


図 4-3　LDA ヒートマップ（16 トピックス、分析対象語数：69）

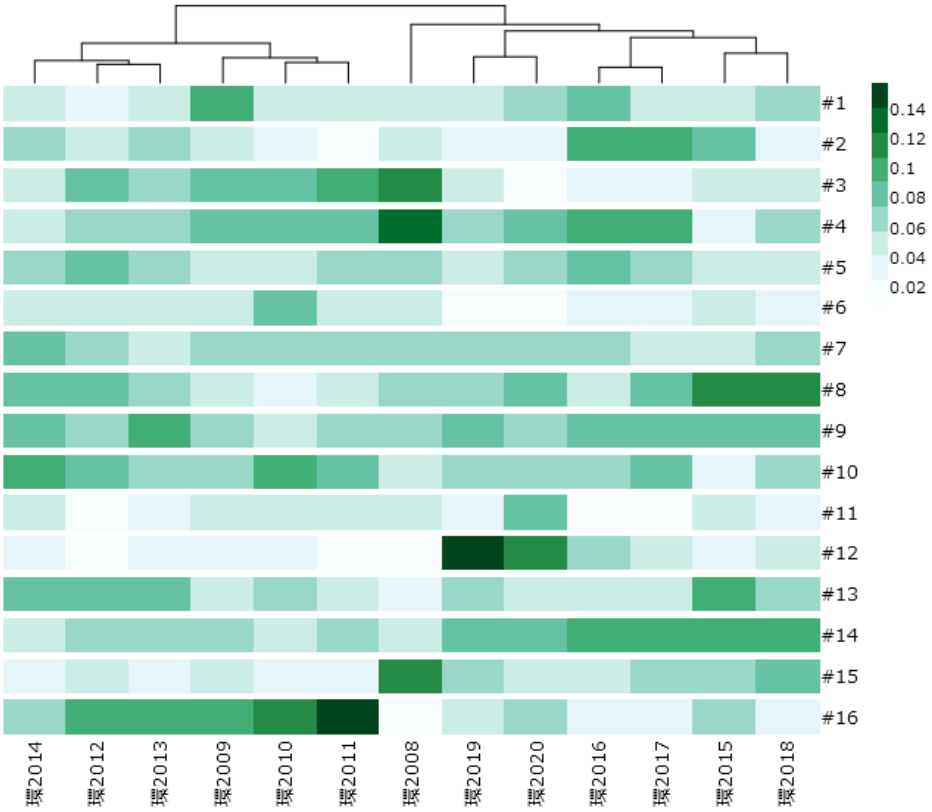


図 4-4　LDA ヒートマップ樹形図（16 トピックス、分析対象語数：69）

表 4-2 トピック比率集計表（16 トピックス、分析対象語数：69）

	#1	#2	#3	#4	#5	#6	#7	#8	
環2008	0.057	0.045	0.122	0.136	0.062	0.054	0.072	0.063	
環2009	0.096	0.043	0.088	0.077	0.055	0.052	0.065	0.048	
環2010	0.055	0.028	0.077	0.084	0.056	0.088	0.071	0.033	
環2011	0.044	0.025	0.095	0.079	0.062	0.047	0.07	0.044	
環2012	0.033	0.043	0.092	0.067	0.077	0.046	0.069	0.078	
環2013	0.045	0.067	0.062	0.074	0.068	0.044	0.054	0.069	
環2014	0.057	0.065	0.057	0.054	0.06	0.043	0.087	0.076	
環2015	0.048	0.077	0.051	0.037	0.042	0.055	0.056	0.108	
環2016	0.077	0.094	0.036	0.106	0.078	0.028	0.06	0.045	
環2017	0.056	0.092	0.03	0.093	0.069	0.037	0.053	0.079	
環2018	0.072	0.028	0.043	0.073	0.053	0.028	0.068	0.109	
環2019	0.052	0.026	0.053	0.074	0.044	0.02	0.072	0.064	
環2020	0.074	0.03	0.015	0.081	0.059	0.014	0.067	0.084	
	#9	#10	#11	#12	#13	#14	#15	#16	ケース数
環2008	0.071	0.045	0.044	0.024	0.037	0.05	0.11	0.009	1
環2009	0.06	0.067	0.049	0.028	0.058	0.062	0.054	0.098	1
環2010	0.049	0.095	0.055	0.039	0.067	0.048	0.032	0.123	1
環2011	0.07	0.083	0.052	0.022	0.053	0.071	0.037	0.145	1
環2012	0.066	0.076	0.024	0.019	0.083	0.072	0.054	0.1	1
環2013	0.096	0.062	0.037	0.032	0.086	0.071	0.036	0.098	1
環2014	0.081	0.092	0.048	0.03	0.091	0.056	0.034	0.07	1
環2015	0.078	0.037	0.053	0.026	0.105	0.092	0.074	0.059	1
環2016	0.083	0.074	0.019	0.073	0.048	0.096	0.049	0.034	1
環2017	0.087	0.084	0.02	0.056	0.043	0.104	0.061	0.036	1
環2018	0.088	0.068	0.03	0.053	0.068	0.098	0.084	0.038	1
環2019	0.076	0.062	0.034	0.158	0.068	0.085	0.059	0.053	1
環2020	0.066	0.069	0.078	0.111	0.052	0.088	0.051	0.061	1

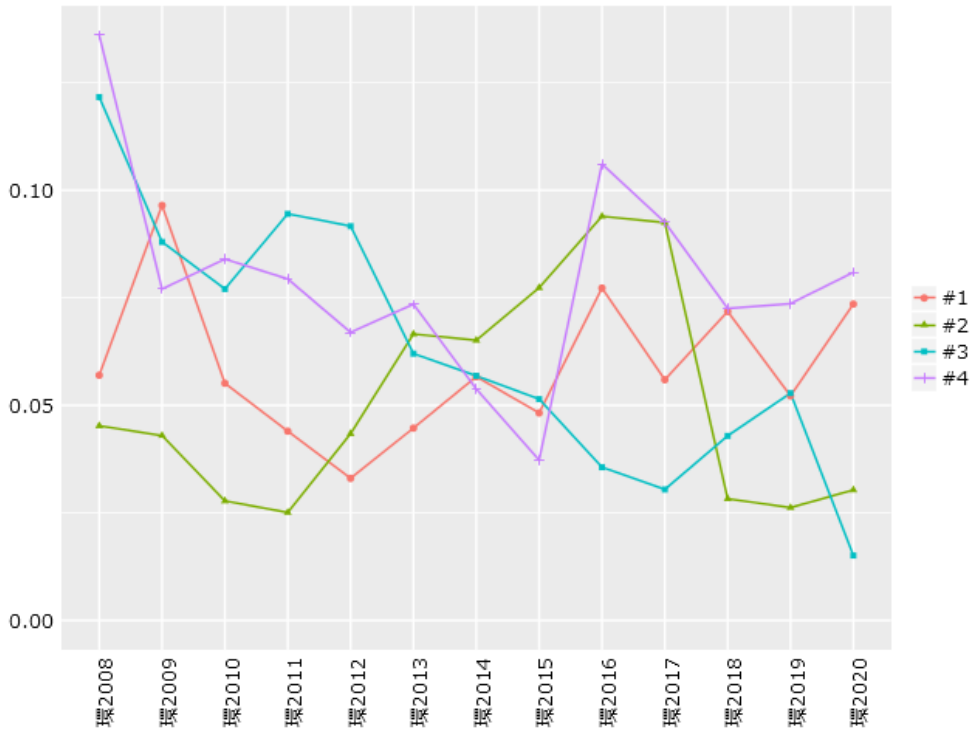


図 4-5 1～4 トピックの比率（16 トピックス、分析対象語数：69）

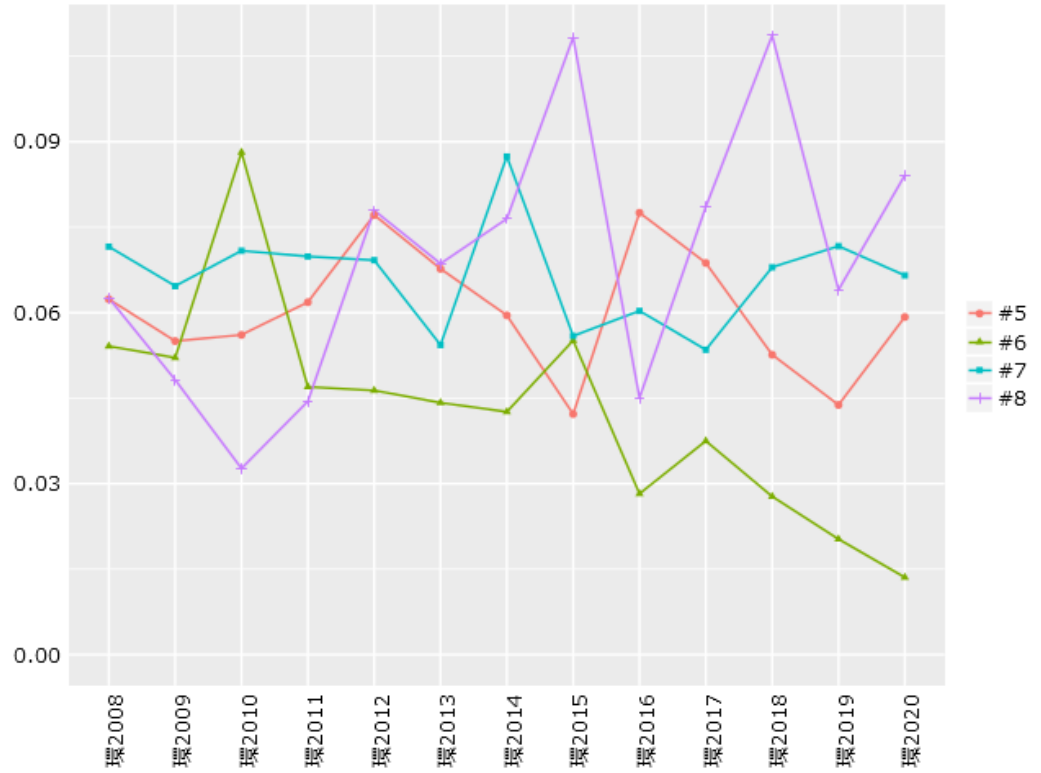


図 4-6 5～8 トピックの比率（16 トピックス、分析対象語数：69）

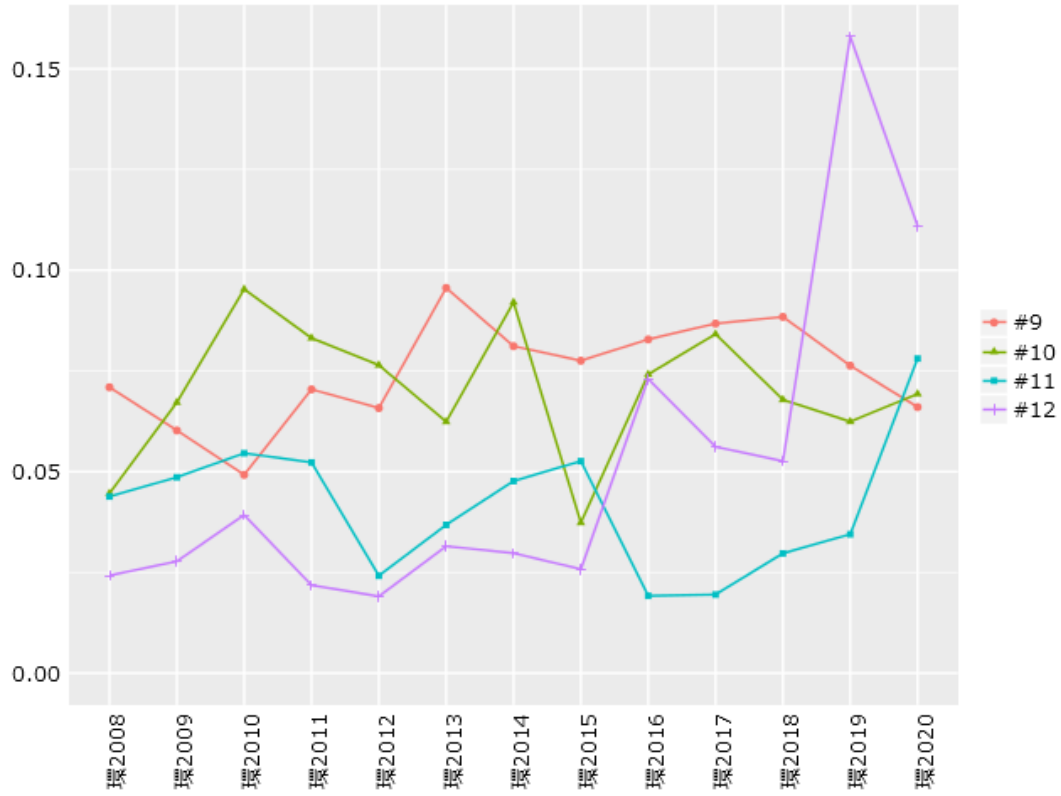


図 4-7 9～12 トピックの比率（16 トピックス、分析対象語数：69）

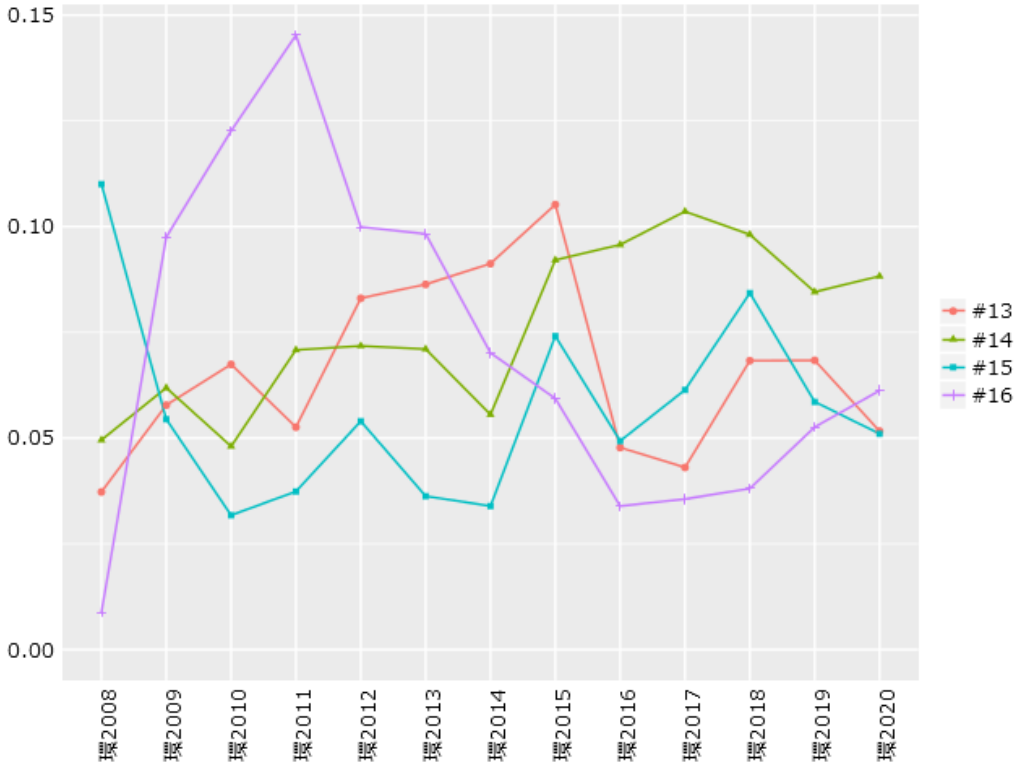


図 4-8 13～16 トピックの比率（16 トピックス、分析対象語数：69）

表 4-3 LDA 処理結果（20 トピックス、分析対象語数：154）

Topics

#1		#2		#3		#4		#5		#6	
自然	0.103	回収	0.090	気候変動	0.214	実施	0.154	CO2	0.089	経済	0.107
実施	0.074	支援	0.085	影響	0.120	処理	0.121	支援	0.088	増加	0.084
施設	0.059	リサイクル	0.072	情報	0.057	資源	0.087	連携	0.067	活動	0.071
情報	0.044	連携	0.070	循環	0.051	重要	0.047	開催	0.039	減少	0.067
影響	0.042	活動	0.068	評価	0.039	開催	0.046	技術	0.039	必要	0.058
支援	0.038	地球	0.065	地方公共団体	0.036	政策	0.040	持続可能	0.036	自動車	0.044
健康	0.038	収集	0.047	計画	0.028	システム	0.035	実施	0.033	保全	0.040
認定	0.031	再生	0.044	社会	0.028	対象	0.034	導入	0.032	結果	0.036
指定	0.030	整備	0.041	構築	0.023	高い	0.033	低炭素	0.032	研究	0.034
対策	0.029	構築	0.038	強化	0.023	協力	0.033	土壌	0.032	高い	0.032

#7		#8		#9		#10		#11		#12		#13	
生物多様性	0.210	処理	0.075	利用	0.087	地球	0.112	社会	0.152	企業	0.090	利用	0.169
社会	0.081	廃棄物	0.069	循環型	0.072	計画	0.090	廃棄物	0.114	世界	0.072	資源	0.125
日本	0.062	制度	0.069	問題	0.062	法律	0.077	エネルギー	0.087	関係	0.058	森林	0.080
生態系	0.040	情報	0.057	技術	0.055	世界	0.071	循環	0.080	目標	0.052	持続可能	0.059
可能性	0.037	事業者	0.054	開発	0.051	目標	0.060	処理	0.073	戦略	0.051	生活	0.045
条約	0.036	確保	0.047	資源	0.045	評価	0.044	消費	0.051	人	0.048	開発	0.039
影響	0.035	管理	0.038	温暖化	0.039	状況	0.041	排出	0.039	削減	0.046	会議	0.037
保全	0.034	保全	0.037	増加	0.037	廃棄物	0.030	日本	0.035	実現	0.046	提供	0.031
発生	0.034	策定	0.036	アジア	0.037	促進	0.029	適正	0.033	政府	0.043	団体	0.028
保護	0.033	適正	0.036	リサイクル	0.032	生産	0.028	処分	0.026	貢献	0.043	保全	0.025

#14		#15		#16		#17		#18		#19		#20	
原子力	0.151	法律	0.132	計画	0.104	都市	0.088	対策	0.214	削減	0.116	技術	0.093
規制	0.110	発生	0.064	整備	0.085	実施	0.072	達成	0.065	実施	0.080	調査	0.091
委員会	0.060	活用	0.062	基準	0.061	活用	0.053	温暖化	0.058	排出量	0.076	評価	0.059
発電	0.055	発電	0.052	廃棄物	0.057	管理	0.051	検討	0.056	排出	0.060	国内	0.042
開発	0.039	エネルギー	0.051	利用	0.056	生態系	0.049	排出	0.055	温室効果ガス	0.047	被害	0.040
措置	0.038	保全	0.041	施設	0.043	化学物質	0.046	必要	0.054	参加	0.045	状況	0.037
法律	0.034	問題	0.041	物質	0.042	配慮	0.041	結果	0.048	条約	0.044	普及	0.036
関係	0.031	検討	0.040	地方公共団体	0.034	規制	0.040	社会	0.044	防止	0.037	開発	0.035
開催	0.028	自然	0.039	低炭素	0.031	調査	0.037	温室効果ガス	0.036	企業	0.033	国際	0.033
検討	0.026	課題	0.036	燃料	0.027	拡大	0.037	排出量	0.034	普及	0.031	制度	0.032

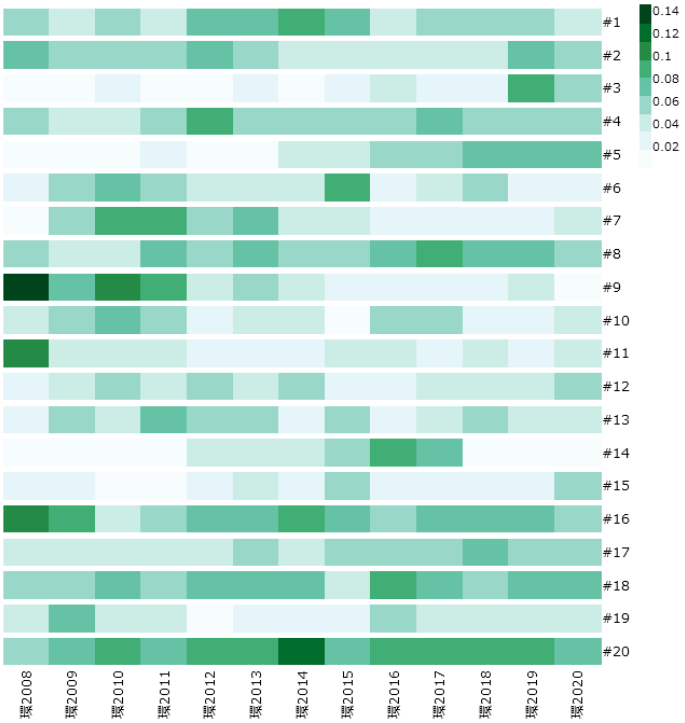


図 4-9　LDA ヒートマップ（20 トピックス、分析対象語数：154）

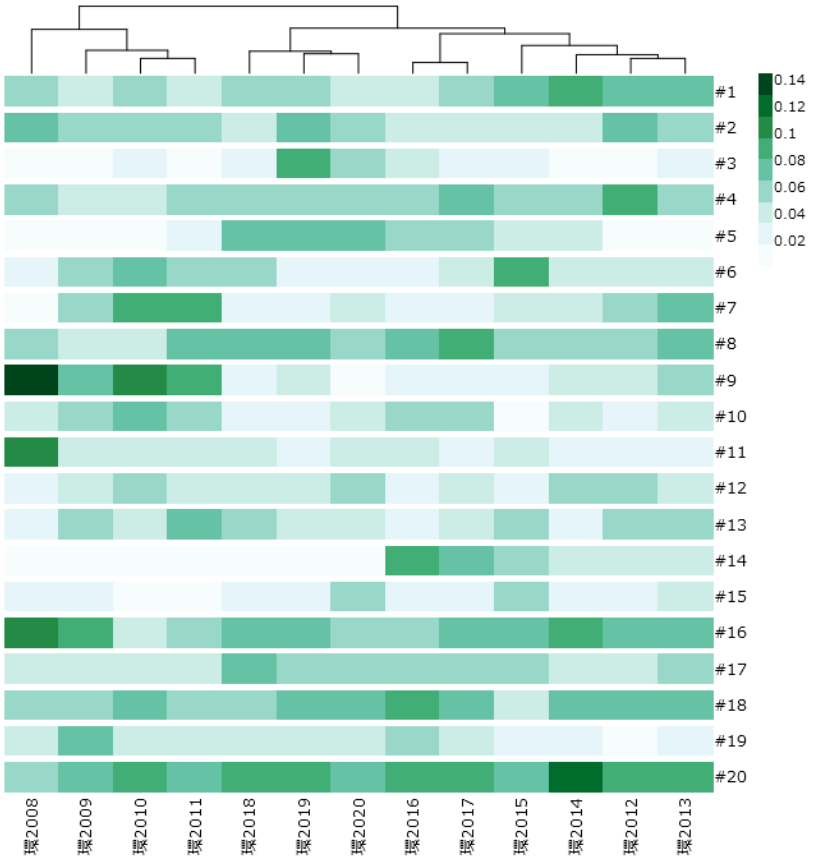


図 4-10　LDA ヒートマップ樹形図（20 トピックス、分析対象語数：154）

表 4-4 トピック比率集計表（20 トピックス、分析対象語数：154）

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	
環2008	0.053	0.073	0.01	0.062	0.01	0.025	0.008	0.064	0.145	0.048	
環2009	0.045	0.063	0.015	0.048	0.011	0.053	0.064	0.047	0.077	0.057	
環2010	0.054	0.054	0.018	0.039	0.009	0.077	0.085	0.048	0.099	0.075	
環2011	0.047	0.057	0.011	0.061	0.019	0.051	0.094	0.067	0.09	0.059	
環2012	0.072	0.078	0.007	0.089	0.008	0.047	0.058	0.061	0.046	0.023	
環2013	0.075	0.059	0.018	0.063	0.018	0.039	0.067	0.069	0.05	0.042	
環2014	0.087	0.046	0.013	0.063	0.039	0.043	0.048	0.055	0.049	0.038	
環2015	0.067	0.043	0.018	0.062	0.044	0.084	0.04	0.058	0.025	0.017	
環2016	0.049	0.037	0.038	0.06	0.061	0.027	0.025	0.075	0.033	0.063	
環2017	0.055	0.037	0.033	0.068	0.063	0.036	0.026	0.083	0.024	0.055	
環2018	0.065	0.049	0.029	0.053	0.078	0.051	0.026	0.074	0.02	0.025	
環2019	0.051	0.066	0.095	0.064	0.075	0.03	0.028	0.077	0.041	0.019	
環2020	0.04	0.057	0.061	0.055	0.081	0.032	0.04	0.059	0.016	0.039	
	#11	#12	#13	#14	#15	#16	#17	#18	#19	#20	ケース数
環2008	0.113	0.029	0.033	0.01	0.027	0.097	0.039	0.065	0.034	0.056	1
環2009	0.045	0.034	0.057	0.013	0.023	0.094	0.042	0.065	0.079	0.067	1
環2010	0.044	0.054	0.046	0.002	0.013	0.047	0.049	0.067	0.036	0.083	1
環2011	0.045	0.049	0.067	0.003	0.014	0.056	0.042	0.051	0.041	0.077	1
環2012	0.031	0.062	0.061	0.037	0.031	0.074	0.036	0.07	0.015	0.093	1
環2013	0.024	0.04	0.05	0.035	0.036	0.079	0.055	0.068	0.027	0.085	1
環2014	0.02	0.052	0.03	0.047	0.031	0.084	0.042	0.066	0.031	0.115	1
環2015	0.04	0.031	0.052	0.063	0.063	0.073	0.063	0.047	0.031	0.08	1
環2016	0.034	0.029	0.02	0.082	0.019	0.065	0.05	0.083	0.061	0.088	1
環2017	0.022	0.034	0.036	0.076	0.025	0.081	0.05	0.073	0.043	0.082	1
環2018	0.037	0.048	0.056	0.005	0.033	0.081	0.07	0.059	0.048	0.093	1
環2019	0.021	0.041	0.037	0.006	0.031	0.071	0.051	0.071	0.04	0.084	1
環2020	0.044	0.058	0.046	0.006	0.062	0.055	0.059	0.081	0.039	0.071	1

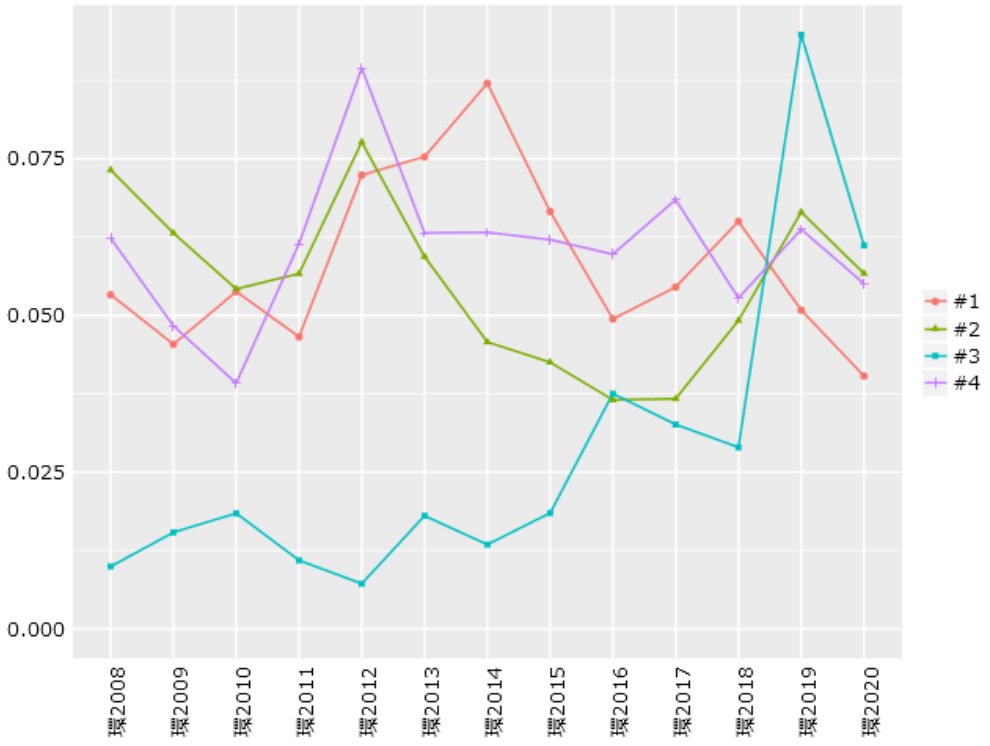


図 4-11 1～4 トピックの比率（20 トピックス、分析対象語数：154）

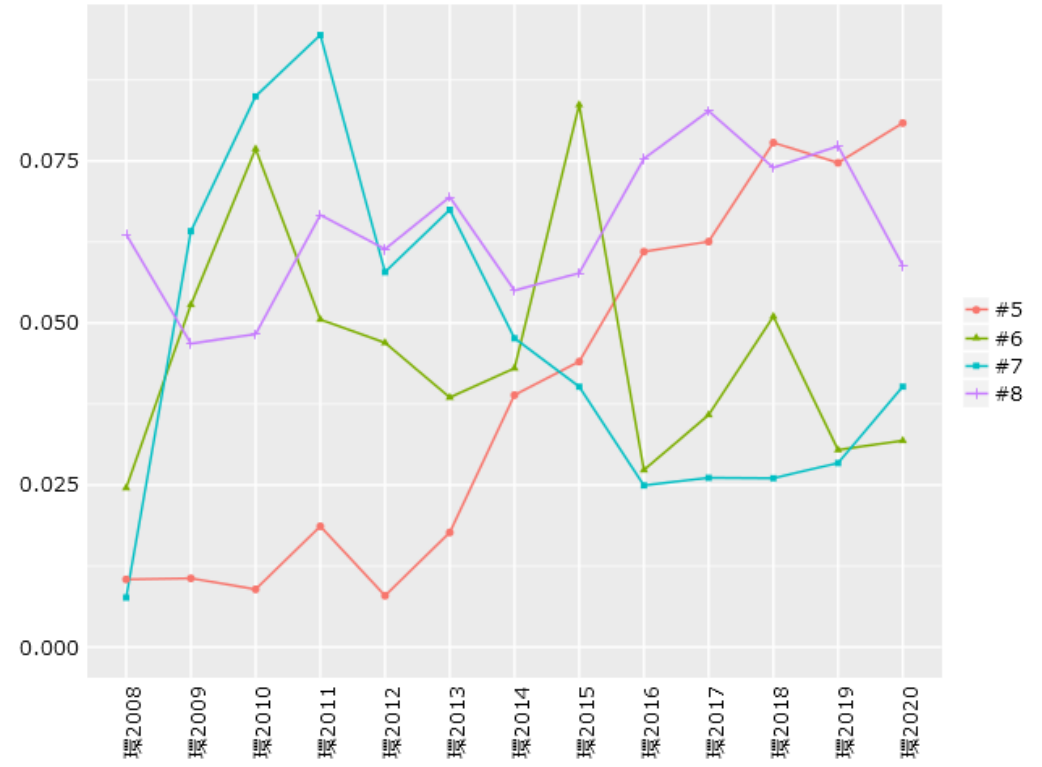


図 4-12 5～8 トピックの比率（20 トピックス、分析対象語数：154）

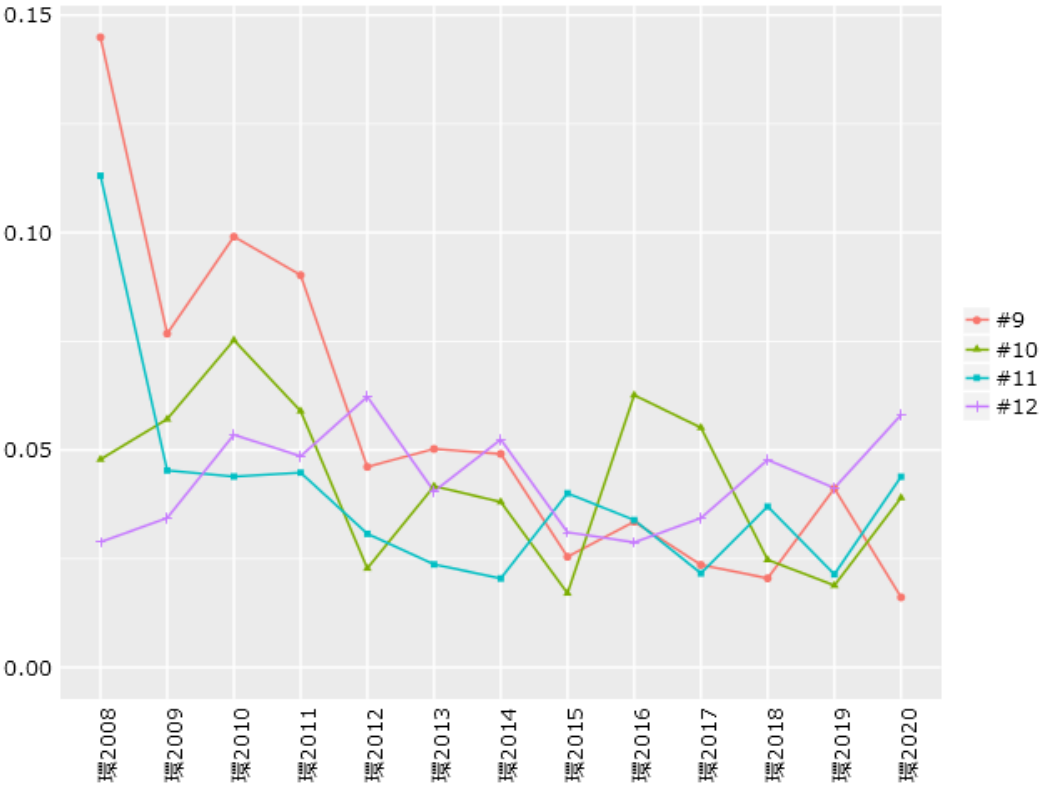


図 4-13 9～12 トピックの比率（20 トピックス、分析対象語数：154）

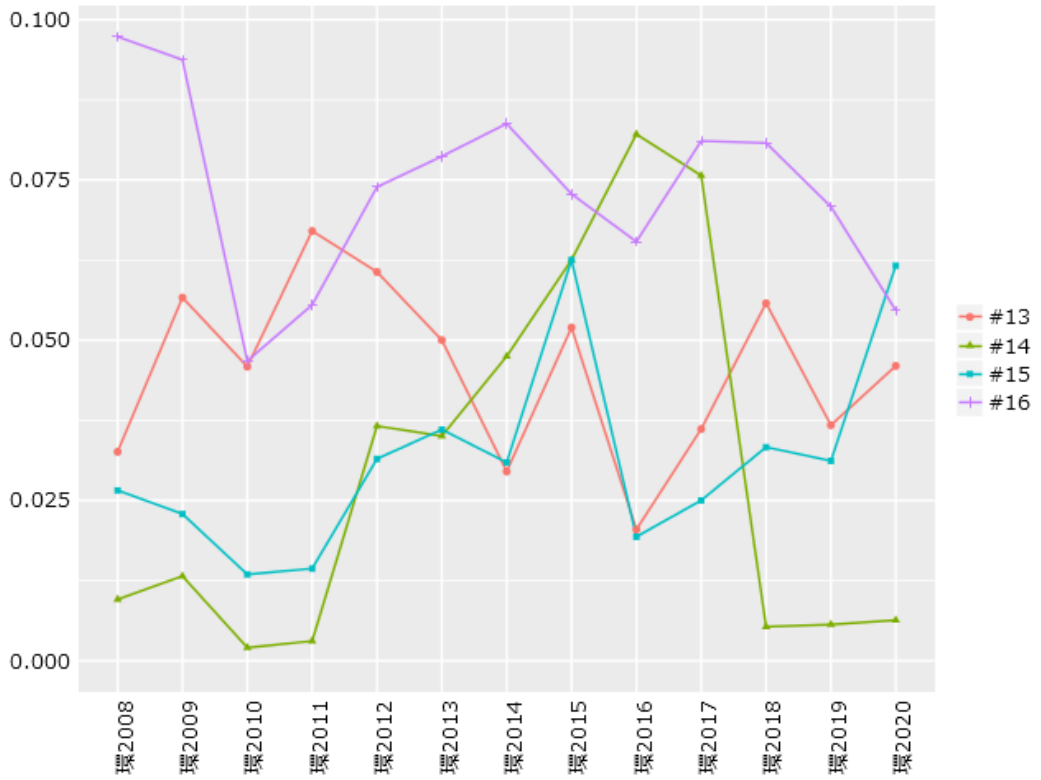


図 4-14　13～16 トピックの比率（20 トピックス、分析対象語数：154）

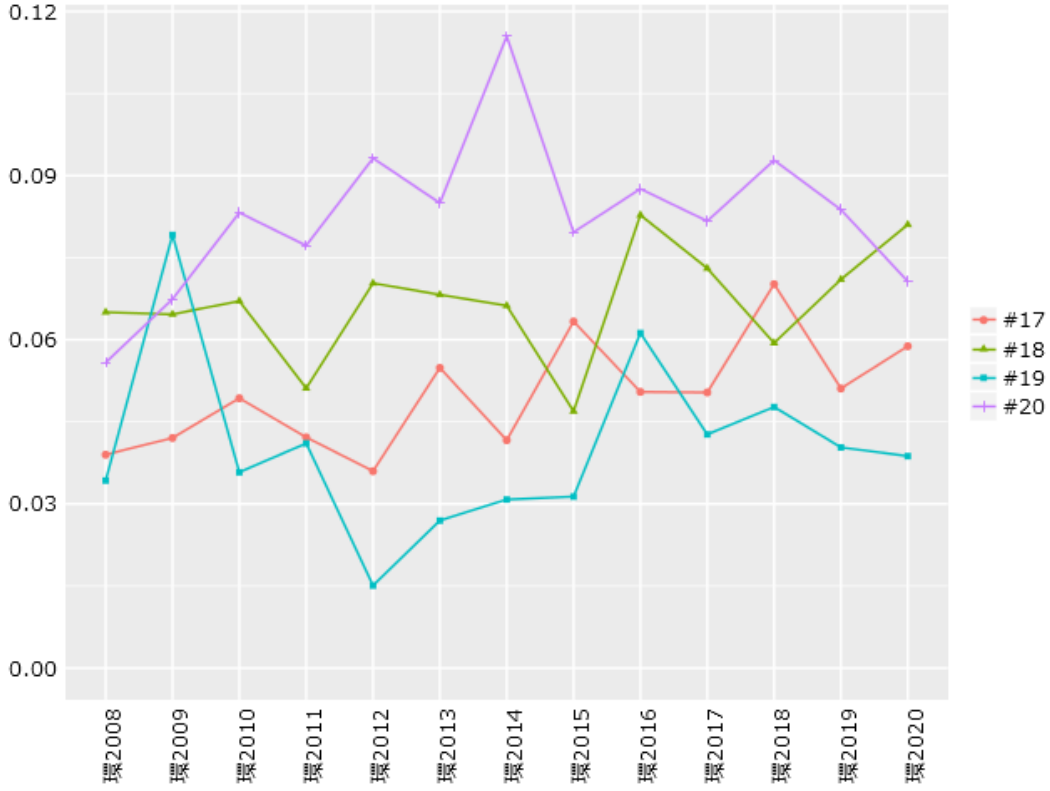


図 4-15　17～20 トピックの比率（20 トピックス、分析対象語数：154）

付録 8「環境･水産･海洋白書（2008～2020 年）の分析結果」

文書番号：JRDN-21-027

1. 前処理結果

環境白書･海洋白書･水産白書（2008～2020 年）の分析で設定した強制抽出語（131 語）と 54 語の除外語を設定し、「動詞、感動詞、動詞 B、副詞 B」を除外して前処理を実行した。

Chanse での前処理の結果、総抽出語数：3791910、異なり語数：34505 のうち 24920 語が分析処理で使用された。抽出語出現数の頻度分布を図 1-1、抽出語リスト（上位 200 語）を図 1-2 に示す。

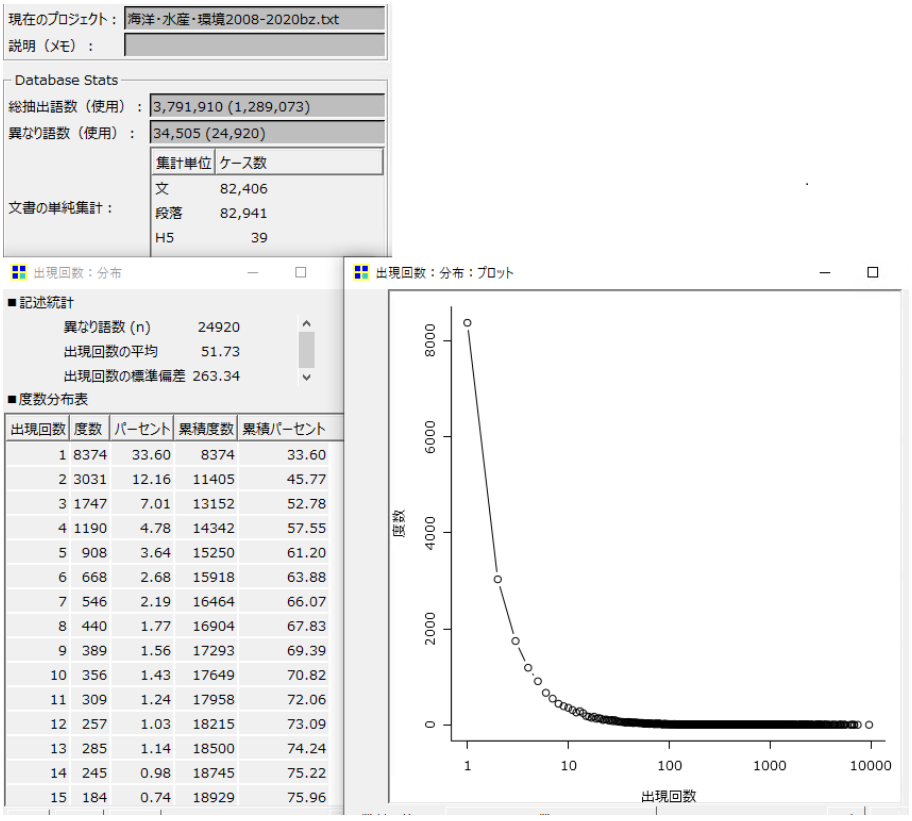


図 1-1 抽出語出現数の頻度分布

#	抽出語	品詞/活用	頻度
1	実施	サ変名詞	9571
2	利用	サ変名詞	7394
3	管理	サ変名詞	6749
4	資源	名詞	6495
5	対策	サ変名詞	6447
6	計画	サ変名詞	6220
7	調査	サ変名詞	5533
8	保全	サ変名詞	5528
9	処理	サ変名詞	5408
10	開発	サ変名詞	5260
11	技術	名詞	5097
12	廃棄物	タグ	5027
13	社会	名詞	4957
14	情報	名詞	4910
15	活動	サ変名詞	4901
16	必要	形容動詞	4637
17	関係	サ変名詞	4449
18	支援	サ変名詞	4355
19	施設	サ変名詞	4233
20	影響	サ変名詞	4185
21	整備	サ変名詞	4097
22	状況	名詞	3878
23	世界	名詞	3826
24	開催	サ変名詞	3700
25	日本	地名	3681
26	生物多様性	タグ	3559
27	評価	サ変名詞	3557
28	研究	サ変名詞	3553
29	基本	名詞	3509
30	促進	サ変名詞	3425
31	活用	サ変名詞	3411
32	地球	名詞	3398
33	目標	名詞	3235
34	検討	サ変名詞	3206
35	重要	形容動詞	3195
36	経済	名詞	3170
37	連携	サ変名詞	3168
38	自然	形容動詞	3136
39	制度	名詞	3121
40	国際	名詞	3086
41	エネルギー	名詞	3075
42	発生	サ変名詞	3065
43	増加	サ変名詞	3019
44	問題	ナイ形容	2959
45	減少	サ変名詞	2955
46	協力	サ変名詞	2950
47	法律	名詞	2927
48	対象	名詞	2899
49	規制	サ変名詞	2849
50	生産	サ変名詞	2828
51	産業	名詞	2821
52	策定	サ変名詞	2663
53	結果	副詞可能	2647
54	海域	名詞	2600
55	削減	サ変名詞	2600
56	政策	名詞	2592
57	基準	名詞	2570
58	機関	名詞	2554
59	温暖化	タグ	2474
60	排出	サ変名詞	2473
61	措置	サ変名詞	2454
62	持続可能	タグ	2400
63	課題	名詞	2383
64	教育	サ変名詞	2370
65	確保	サ変名詞	2348
66	使用	サ変名詞	2335
67	排出量	タグ	2325
68	システム	名詞	2316
69	高い	形容詞	2310
70	導入	サ変名詞	2303
66	使用	サ変名詞	2335
67	排出量	タグ	2325
68	システム	名詞	2316
69	高い	形容詞	2310
70	導入	サ変名詞	2303
71	強化	サ変名詞	2300
72	生態系	タグ	2296
73	事業者	タグ	2288
74	防止	サ変名詞	2255
75	気候変動	タグ	2228
76	水産物	名詞	2177
77	物質	名詞	2175
78	リサイクル	サ変名詞	2174
79	提供	サ変名詞	2166
80	分野	名詞	2150
81	発電	サ変名詞	2144
82	循環	サ変名詞	2123
83	構築	サ変名詞	2105
84	再生	サ変名詞	2086
85	対応	サ変名詞	2084
86	森林	名詞	2079
87	政府	名詞	2065
88	条約	名詞	2059
89	会議	サ変名詞	2056
90	被害	名詞	2049
91	観測	サ変名詞	2047
92	企業	名詞	2040
93	汚染	サ変名詞	2039
94	設置	サ変名詞	2033
95	委員会	タグ	2022
96	達成	サ変名詞	2006
97	保護	サ変名詞	1995
98	水産	名詞	1984
99	目的	名詞	1979
100	消費	サ変名詞	1976
101	向上	サ変名詞	1966
102	漁船	名詞	1956
103	生物	名詞	1939
104	温室効果ガス	タグ	1894
105	国内	名詞	1894
106	普及	サ変名詞	1887
107	実現	サ変名詞	1817
108	漁獲	サ変名詞	1801
109	総合的	タグ	1793
110	参加	サ変名詞	1787
111	都市	名詞	1756
112	拡大	サ変名詞	1755
113	中国	地名	1753
114	指定	サ変名詞	1746
115	生活	サ変名詞	1717
116	可能	形容動詞	1665
117	国際的	タグ	1664
118	中心	名詞	1650
119	回収	サ変名詞	1642
120	自動車	名詞	1635
121	漁業者	タグ	1629
122	変化	サ変名詞	1623
123	行動	サ変名詞	1622
124	多く	副詞可能	1605
125	関連	サ変名詞	1604
126	地方公共団体	タグ	1601
127	大きい	形容詞	1588
128	適切	形容動詞	1582
129	戦略	名詞	1575
130	場合	副詞可能	1567
131	都道府県	名詞	1529
132	形成	サ変名詞	1527
133	多い	形容詞	1510
134	機能	サ変名詞	1502
135	環境省	組織名	1497
136	原子力	名詞	1485
137	会合	サ変名詞	1480
138	適正	形容動詞	1463
139	プロジェクト	名詞	1444
140	化学物質	タグ	1444
141	議論	サ変名詞	1431
142	設定	サ変名詞	1430
143	総合	サ変名詞	1430
144	養殖	サ変名詞	1428
145	開始	サ変名詞	1422
146	期待	サ変名詞	1404
147	健康	形容動詞	1404
148	可能性	タグ	1391
149	把握	サ変名詞	1369
150	効果	名詞	1366
151	貢献	サ変名詞	1366
152	製品	名詞	1363
153	モニタリング	名詞	1347
154	具体的	タグ	1345
155	団体	名詞	1339
156	安全	形容動詞	1338
157	特定	サ変名詞	1338
158	収集	サ変名詞	1331
159	経営	サ変名詞	1320
160	国民	名詞	1312
161	規模	名詞	1307
162	福島	地名	1304
163	作成	サ変名詞	1288
164	人	名詞C	1284
165	昭和	固有名詞	1277
166	体制	名詞	1274
167	積極的	タグ	1272
168	改善	サ変名詞	1269
169	データ	名詞	1267
170	改正	サ変名詞	1261
166	体制	名詞	1274
167	積極的	タグ	1272
168	改善	サ変名詞	1269
169	データ	名詞	1267
170	改正	サ変名詞	1261
171	配慮	サ変名詞	1261
172	循環型	タグ	1239
173	アジア	地名	1233
174	市町村	名詞	1227
175	沿岸域	タグ	1224
176	作業	サ変名詞	1224
177	維持	サ変名詞	1221
178	製造	サ変名詞	1217
179	操業	サ変名詞	1212
180	割合	名詞	1210
181	食品	名詞	1208
182	採択	サ変名詞	1200
183	消費者	タグ	1194
184	認定	サ変名詞	1192
185	供給	サ変名詞	1190
186	大臣	名詞	1186
187	成長	サ変名詞	1182
188	太平洋	地名	1181
189	国連	組織名	1178
190	資源管理	タグ	1177
191	それぞれ	副詞可能	1176
192	施行	サ変名詞	1165
193	確認	サ変名詞	1163
194	船舶	名詞	1163
195	共同	サ変名詞	1162
196	CO2	未知語	1160
197	一般	名詞	1156
198	公表	サ変名詞	1145
199	方針	名詞	1141
200	多様	形容動詞	1139

図 1-2 抽出語リスト（上位 200 語）

2. 共起ネットワーク

KHCoder による自動設定値は、最小出現頻度：2225、分析対象語数：75 であった共起ネットワークの描画条件を以下に、語・語での共起ネットワークを図 2-1、語・年での共起ネットワークを図 2-2 に示す。また、最小出現頻度：1340、分析対象語数：154 での共起ネットワークの描画条件を以下に、語・語での共起ネットワークを図 2-3、語・年での共起ネットワークを図 2-4 に示す。

	最小出現頻度	分析対象語数	描画結果
語・語	2225	75	ノード数：57　エッジ数：75
語・年	2225	75	ノード数：60　エッジ数：75

	最小出現頻度	分析対象語数	描画結果
語・語	1340	154	ノード数：83　エッジ数：75
語・年	1340	154	ノード数：53　エッジ数：75

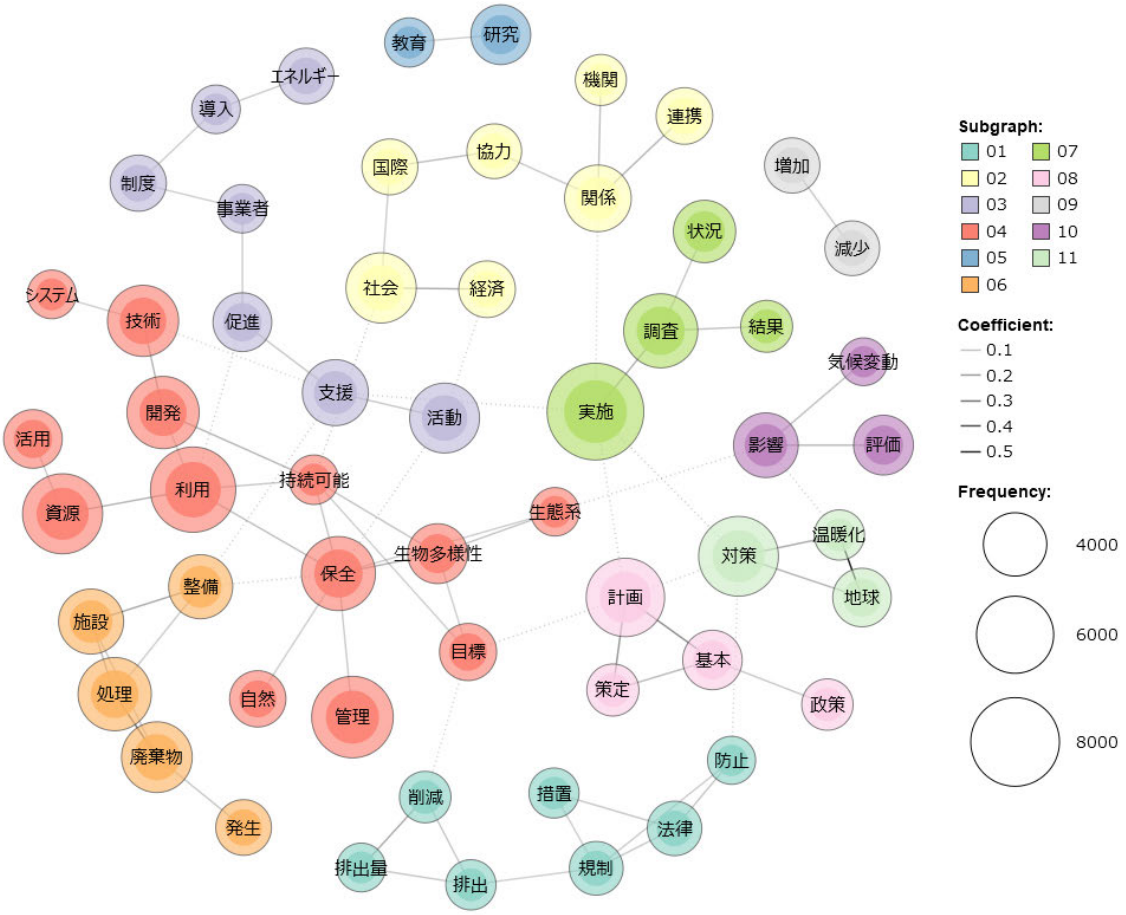


図 2-1　共起ネットワーク（語・語、75 語）

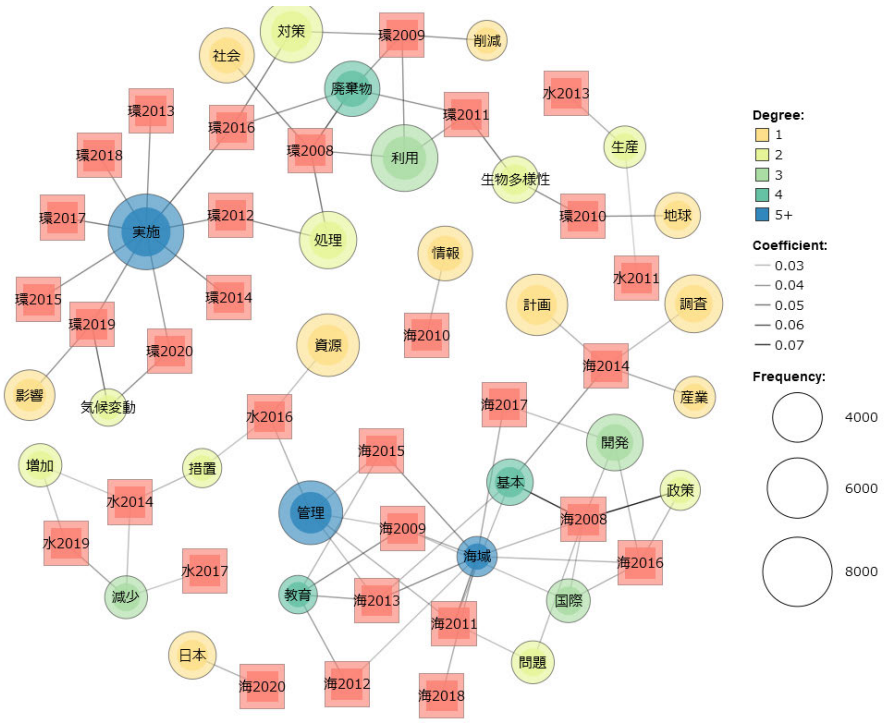


図 2-2 共起ネットワーク（語・年、75 語）

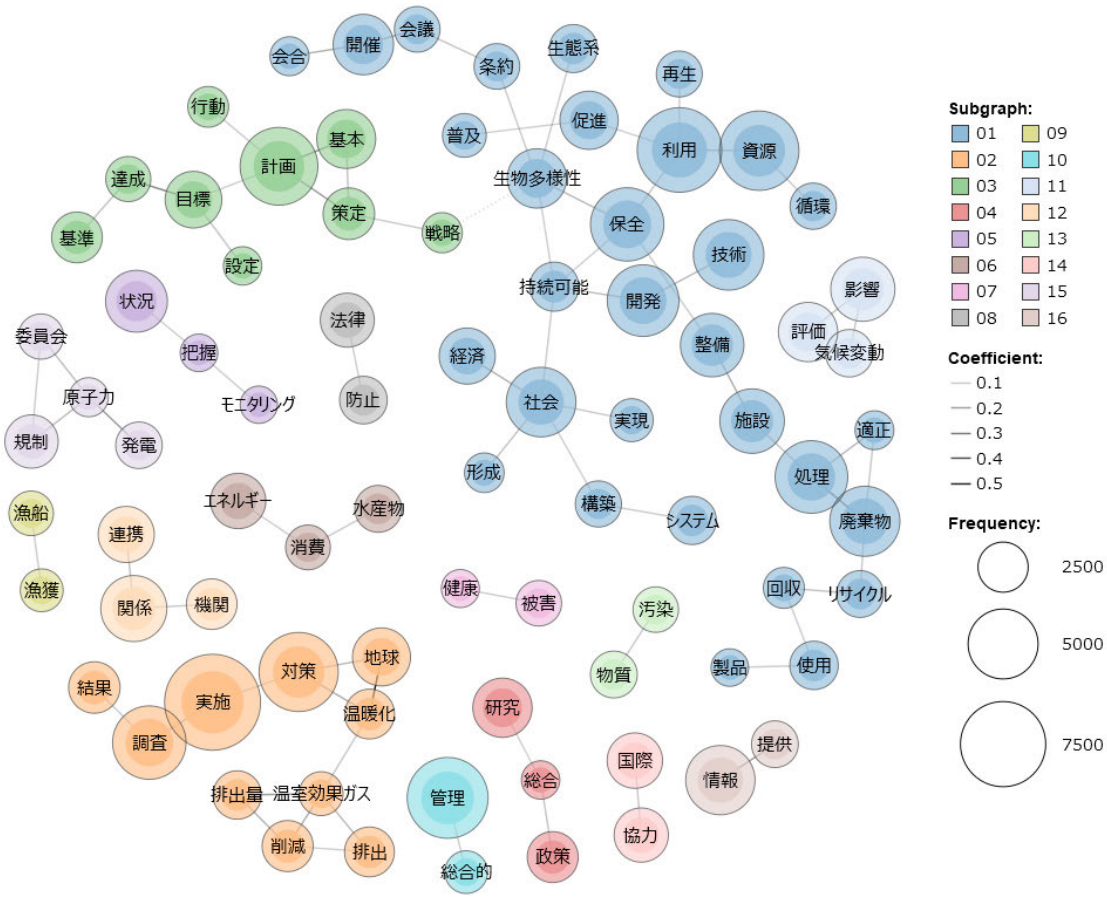


図 2-3 共起ネットワーク（語・語、154 語）

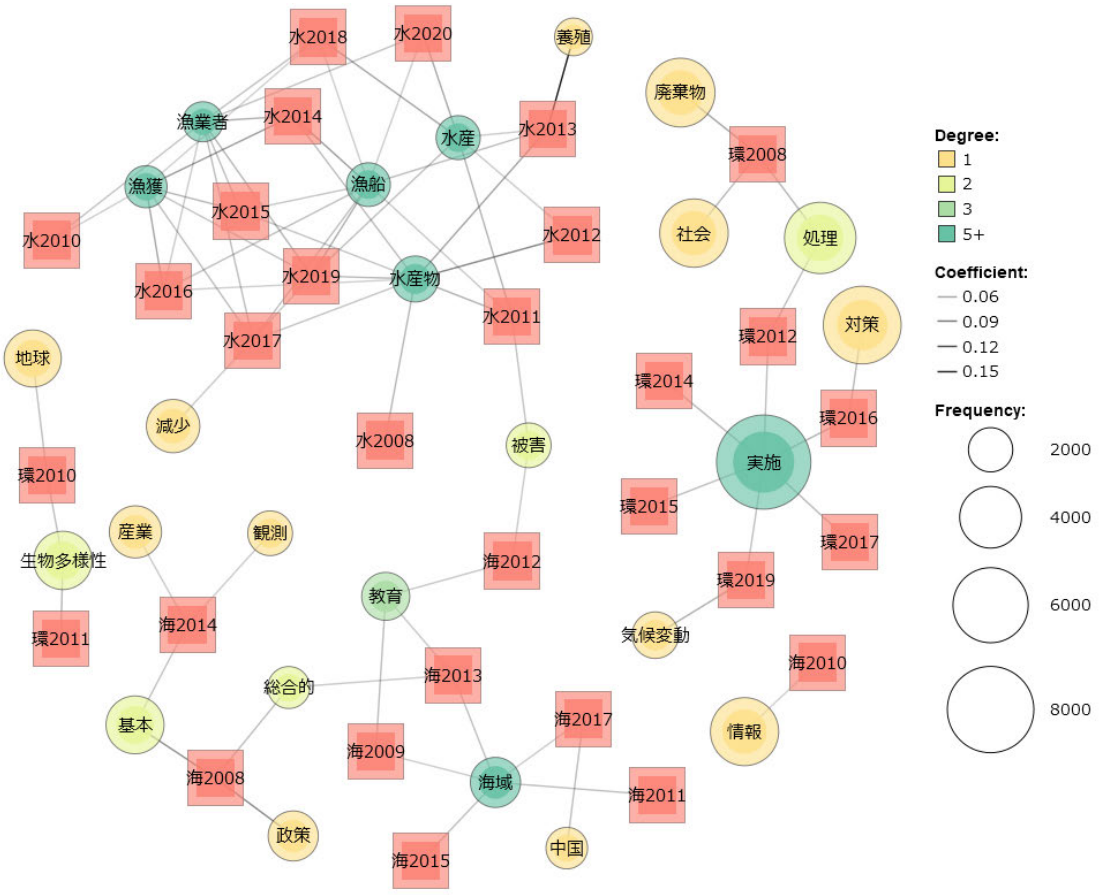


図 2-4 共起ネットワーク（語・年、154 語）

3. 対応分析

KHCoder により自動設定される最小出現数：2225、分析対象語数：75 で対応分析処理を行った。累積寄与率は成分 1 と 2 で 64%、成分 3 と 4 を加えて 74%となった。その結果を図 3-1～2 に示す。更に、分析対象語数：154 でも対応分析処理を行った。その累積寄与率は成分 1 と 2 で 72%となった。その結果を図 3-3 に示す。

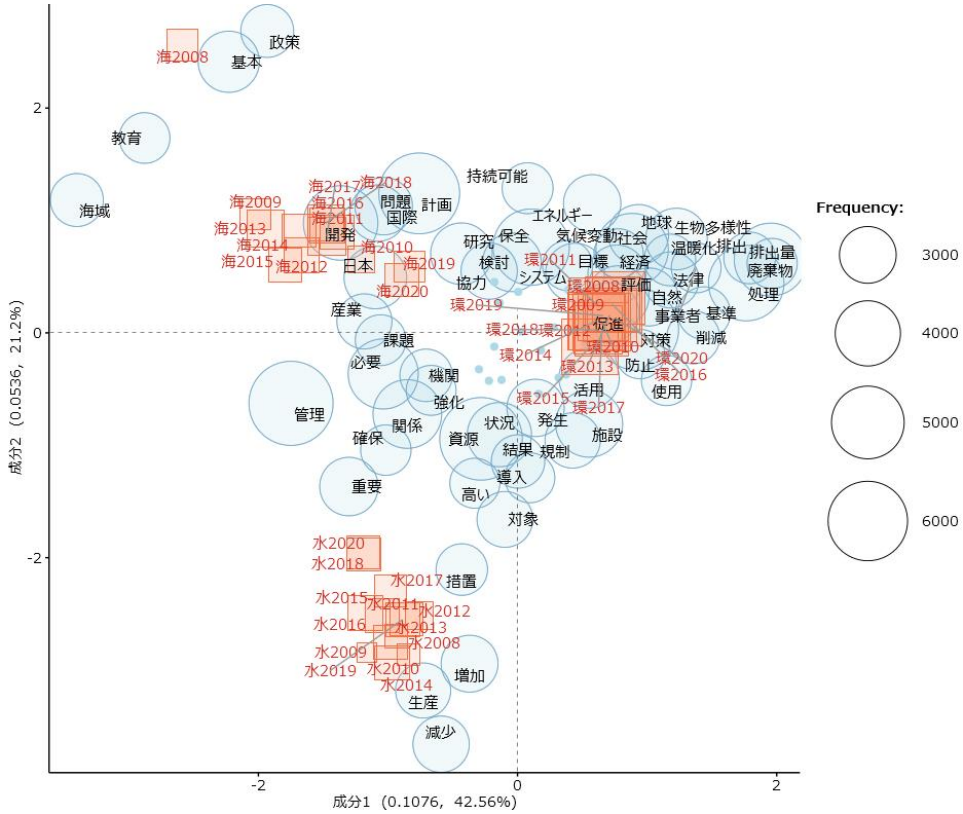


図 3-1 語･年の対応分析結果（成分 1と2、分析対象語数：75）

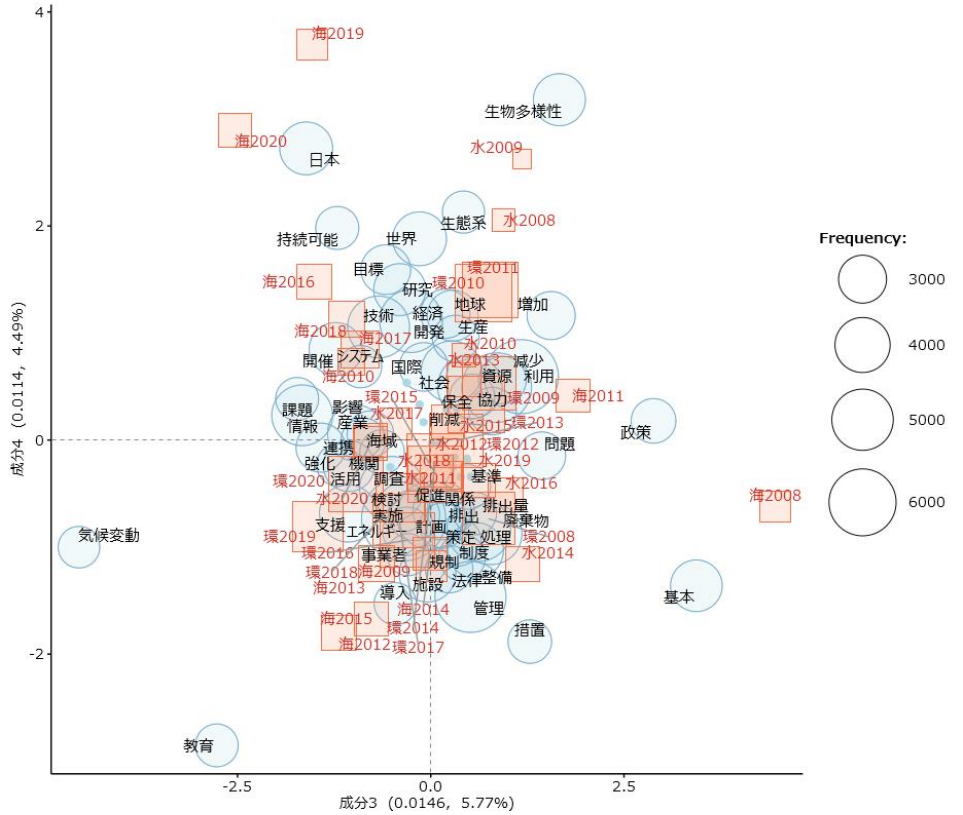


図 3-2 語･年の対応分析結果（成分 3と4、分析対象語数：75）

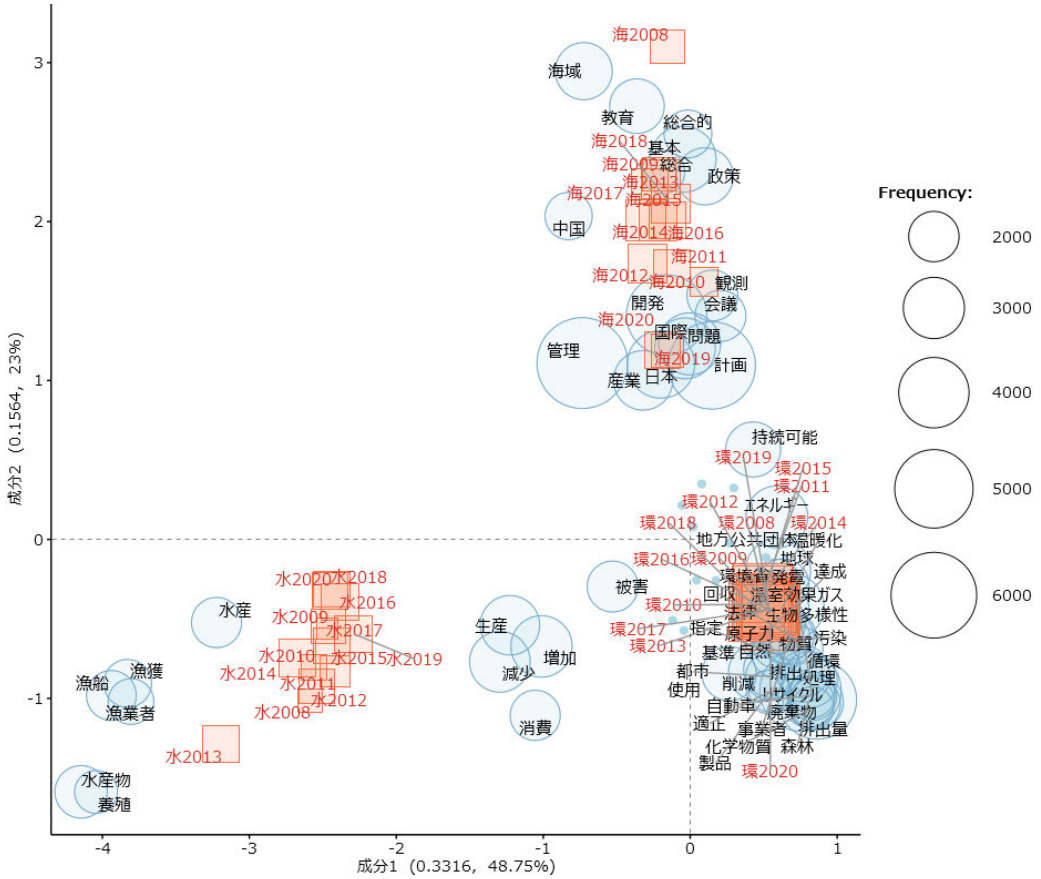


図 3-3 語・年の対応分析結果（成分 1 と 2、分析対象語数：154）

4. LDA 分析

KHCoder により自動設定される最小出現数：2225、分析対象語数：75 で集計単位を H5（年）とした場合と、分析対象語数：154 での LDA 分析を行った。

分析対象語数：75 での LDA トピック数推定結果を図 4-1、分析対象語数：154 での結果を図 4-2 に示す。これらの図から 75 語でのトピック数は 14 或いは 20、154 語では 12～14 或いは 20～22 と推察した。そこでそれぞれ、トピック数 14 と 20 の場合について分析処理を実行した。

分析対象語数	75	75	154	154
トピック数	14	20	14	20
分析結果	表 4- 1	表 4-2	表 4-3	表 4-4
ヒートマップ	図 4-3	図 4-9	図 4-16	図 4-22
ヒートマップ樹形図	図 4-4	図 4-10	図 4-17	図 4-23
トピック比率図	図 4-5～8	図 4-11～15	図 4-18～21	図 4-24～28

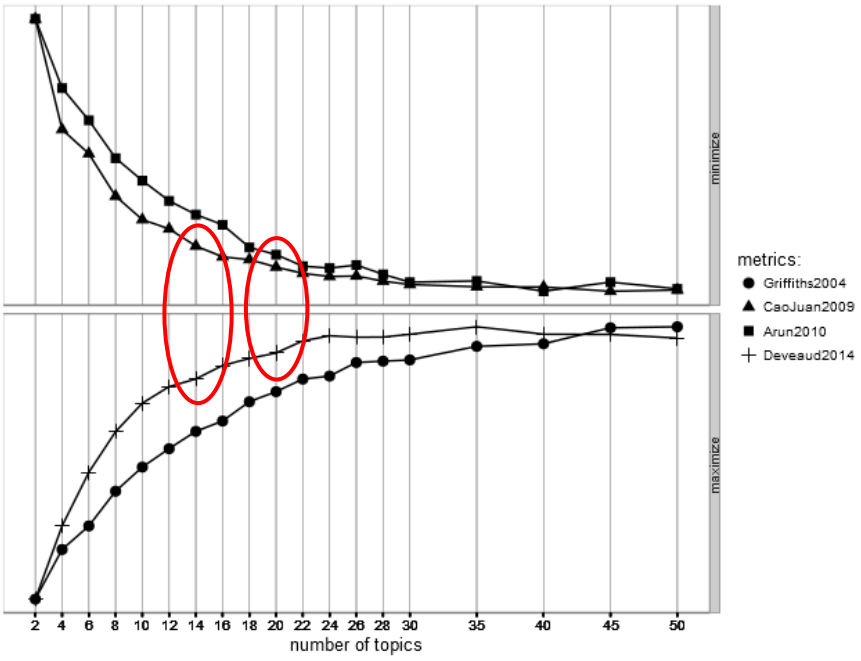


図 4-1 LDAtuning 実行結果（分析対象語数：75）

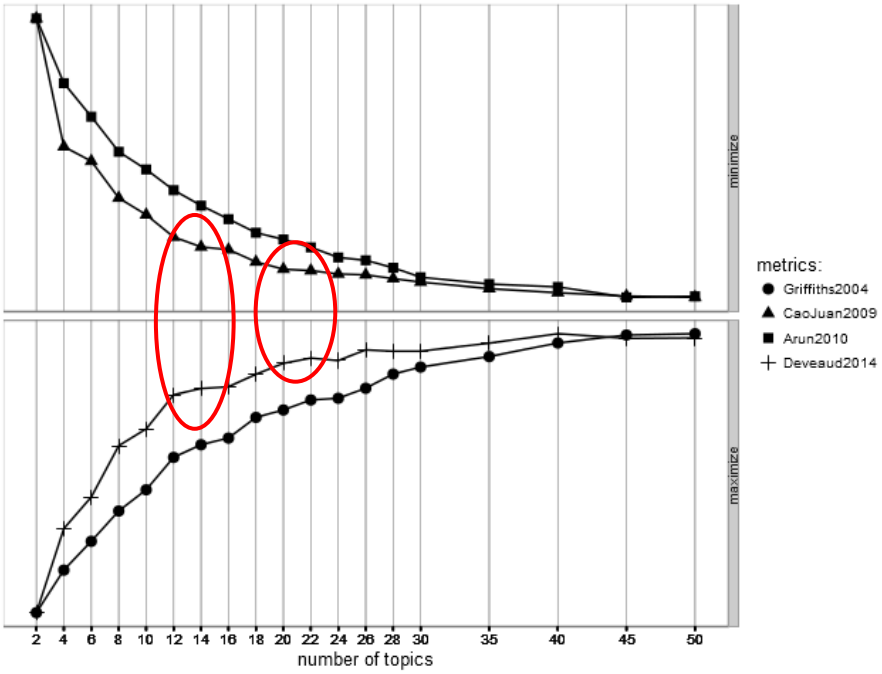


図 4-2 LDAtuning 実行結果（分析対象語数：154）

表 4-1 LDA 処理結果（14 トピックス、分析対象語数：75）

Topics

#1		#2		#3		#4		#5		#6		#7	
自然	0.094	実施	0.129	増加	0.107	調査	0.161	海域	0.148	教育	0.258	施設	0.091
活用	0.078	規制	0.098	減少	0.103	情報	0.160	計画	0.125	管理	0.071	支援	0.086
保全	0.069	制度	0.089	生産	0.084	資源	0.059	管理	0.087	活動	0.069	関係	0.071
施設	0.068	調査	0.080	利用	0.076	必要	0.058	日本	0.058	研究	0.062	発生	0.069
実施	0.056	評価	0.077	資源	0.070	技術	0.058	開発	0.058	産業	0.048	生産	0.064
経済	0.053	検討	0.054	連携	0.036	研究	0.055	産業	0.050	必要	0.039	実施	0.049
支援	0.050	状況	0.053	技術	0.035	開発	0.053	利用	0.050	機関	0.039	対策	0.049
対策	0.046	導入	0.052	高い	0.031	システム	0.043	エネルギー	0.047	検討	0.034	状況	0.046
技術	0.042	措置	0.046	影響	0.030	活動	0.042	保全	0.043	実施	0.034	必要	0.046
開催	0.038	計画	0.043	対象	0.028	重要	0.040	問題	0.040	連携	0.030	世界	0.044

#8		#9		#10		#11		#12		#13		#14	
開発	0.088	生物多様性	0.161	地球	0.112	管理	0.208	廃棄物	0.097	気候変動	0.129	基本	0.185
開催	0.073	保全	0.068	社会	0.093	措置	0.087	処理	0.094	影響	0.077	政策	0.117
持続可能	0.072	利用	0.055	技術	0.073	資源	0.069	整備	0.045	実施	0.072	国際	0.065
国際	0.063	生態系	0.053	温暖化	0.062	関係	0.062	エネルギー	0.045	情報	0.069	計画	0.062
世界	0.050	経済	0.050	世界	0.055	重要	0.056	排出量	0.045	連携	0.050	利用	0.061
目標	0.049	活動	0.049	利用	0.052	必要	0.054	排出	0.043	対策	0.050	管理	0.058
研究	0.043	評価	0.045	対策	0.051	機関	0.050	利用	0.042	活用	0.049	開発	0.054
日本	0.043	目標	0.044	状況	0.046	実施	0.040	対策	0.038	評価	0.046	関係	0.047
技術	0.041	社会	0.036	問題	0.041	対象	0.039	社会	0.035	支援	0.046	問題	0.047
情報	0.041	実施	0.035	開発	0.039	活動	0.037	削減	0.031	計画	0.040	協力	0.046

図 4-3 LDA ヒートマップ（14 トピックス、分析対象語数：75）

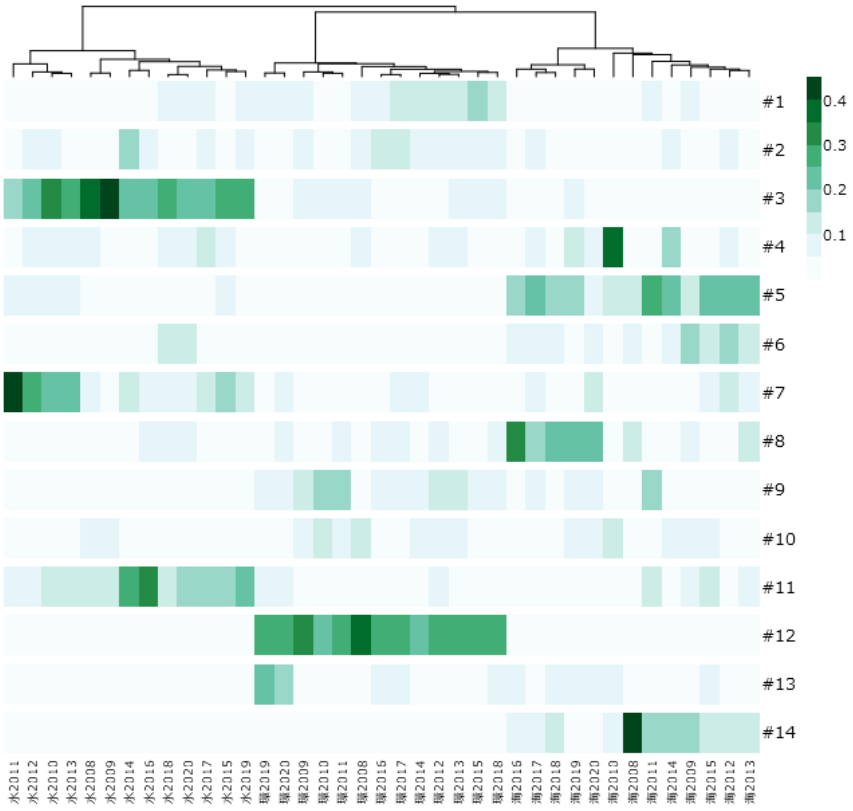


図 4-4　LDA ヒートマップ樹形図（14 トピックス、分析対象語数：75）

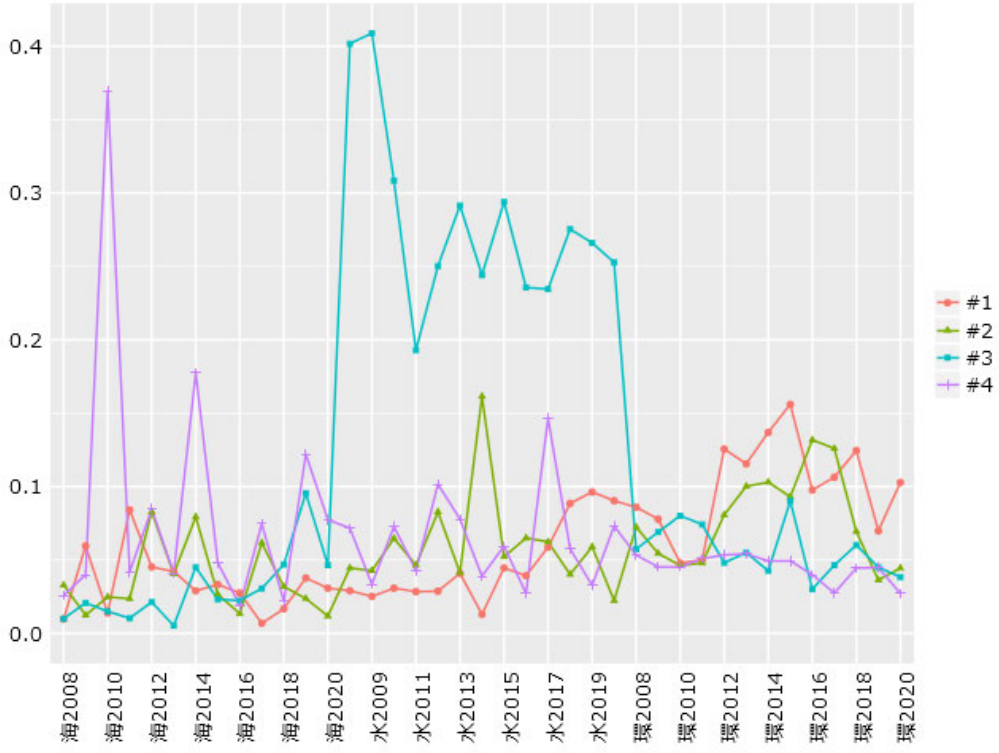


図 4-5　1～4 トピックの比率（14 トピックス、分析対象語数：75）

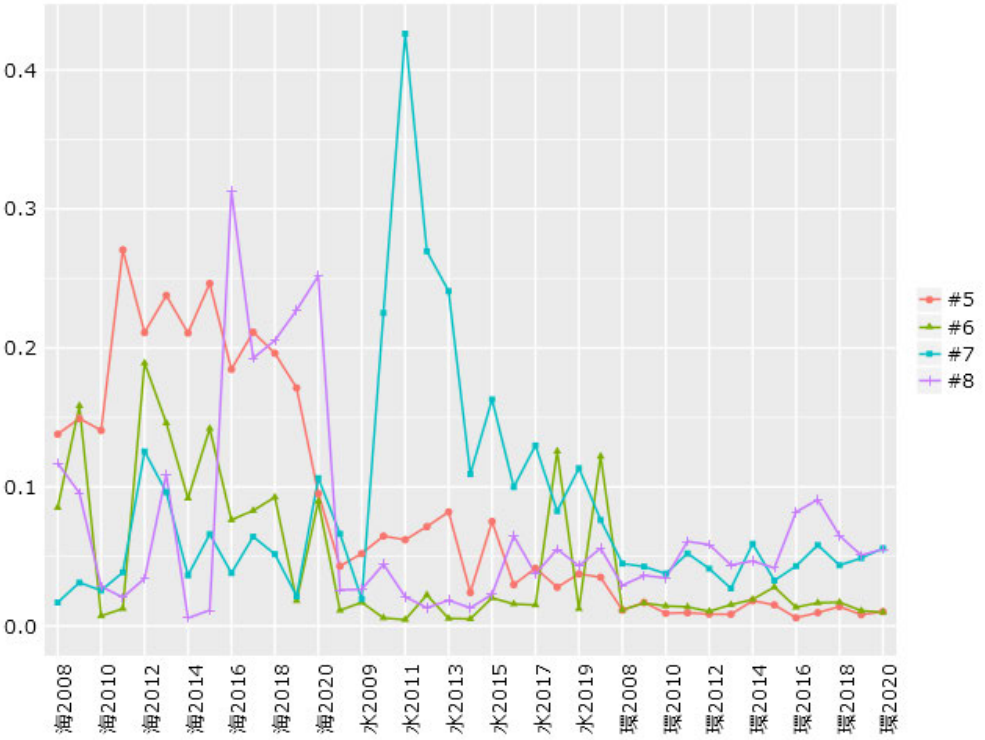


図 4-6 5～7 トピックの比率（14 トピックス、分析対象語数：75）

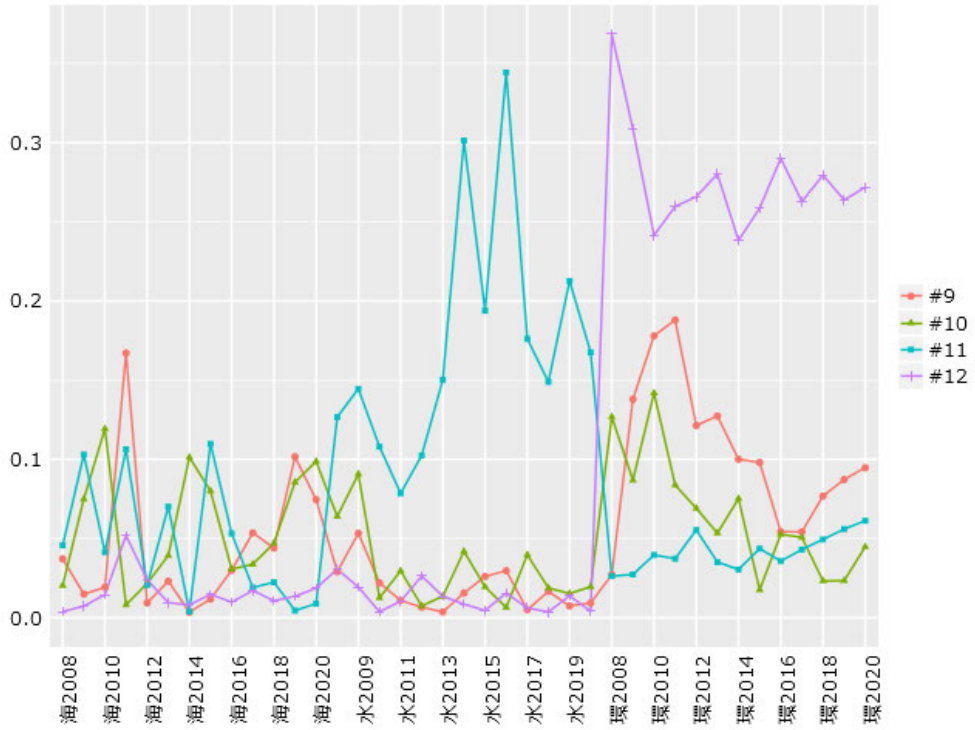


図 4-7 9～12 トピックの比率（14 トピックス、分析対象語数：75）

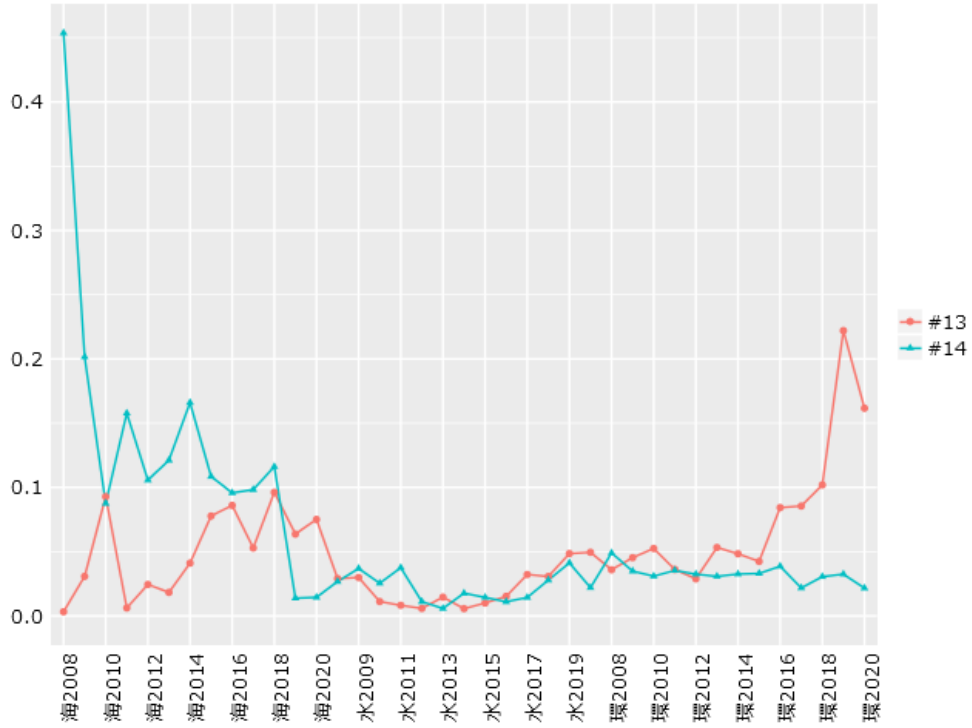


図 4-8 13～14 トピックの比率（14 トピックス、分析対象語数：75）

表 4-2 LDA 処理結果（20 トピックス、分析対象語数：75）

Topics

#1		#2		#3		#4		#5		#6		#7	
自然	0.113	海域	0.169	廃棄物	0.123	管理	0.230	施設	0.128	開発	0.156	エネルギー	0.132
保全	0.097	管理	0.157	処理	0.107	資源	0.152	調査	0.115	持続可能	0.111	利用	0.103
実施	0.073	計画	0.110	対策	0.078	措置	0.070	支援	0.072	開催	0.094	活用	0.068
法律	0.065	問題	0.073	排出	0.057	増加	0.066	発生	0.065	国際	0.071	資源	0.068
施設	0.062	保全	0.055	排出量	0.057	減少	0.059	実施	0.060	実施	0.068	経済	0.053
活用	0.059	重要	0.054	実施	0.046	対象	0.043	計画	0.055	政策	0.066	促進	0.049
処理	0.050	日本	0.052	地球	0.040	機関	0.039	整備	0.055	目標	0.065	制度	0.044
経済	0.049	関係	0.040	削減	0.037	導入	0.039	活動	0.053	資源	0.049	検討	0.043
基準	0.043	活動	0.036	基準	0.036	影響	0.037	関係	0.047	課題	0.041	連携	0.038
活動	0.040	必要	0.032	法律	0.034	結果	0.036	利用	0.039	強化	0.040	社会	0.036

#8		#9		#10		#11		#12		#13		#14	
調査	0.150	教育	0.331	実施	0.144	情報	0.228	生物多様性	0.188	利用	0.103	支援	0.181
計画	0.132	管理	0.067	規制	0.138	調査	0.129	利用	0.079	資源	0.092	対策	0.147
基本	0.075	研究	0.066	評価	0.107	必要	0.082	評価	0.055	社会	0.090	協力	0.081
産業	0.068	産業	0.062	制度	0.089	活動	0.071	保全	0.053	問題	0.080	促進	0.079
開発	0.067	必要	0.051	検討	0.074	技術	0.063	社会	0.050	開発	0.079	関係	0.076
情報	0.054	検討	0.046	技術	0.067	重要	0.058	経済	0.050	協力	0.062	世界	0.075
地球	0.049	活動	0.045	状況	0.059	確保	0.044	活動	0.040	国際	0.054	開催	0.072
法律	0.042	実施	0.042	措置	0.059	課題	0.040	影響	0.039	保全	0.052	機関	0.052
必要	0.036	機関	0.037	結果	0.042	関係	0.039	発生	0.036	温暖化	0.046	状況	0.052
整備	0.035	確保	0.035	導入	0.030	システム	0.033	生態系	0.034	自然	0.041	策定	0.050

#15		#16		#17		#18		#19		#20	
必要	0.125	発生	0.101	気候変動	0.175	日本	0.202	減少	0.134	基本	0.209
重要	0.124	開発	0.077	情報	0.095	研究	0.160	増加	0.130	政策	0.135
生産	0.111	海域	0.075	影響	0.092	技術	0.158	生産	0.093	開発	0.063
活動	0.090	利用	0.075	連携	0.092	世界	0.109	連携	0.080	管理	0.055
関係	0.075	状況	0.066	実施	0.081	生態系	0.063	産業	0.061	計画	0.055
確保	0.074	高い	0.060	評価	0.061	目標	0.058	資源	0.059	利用	0.052
減少	0.063	世界	0.057	活用	0.053	地球	0.054	活用	0.058	国際	0.042
開催	0.044	対策	0.054	社会	0.039	システム	0.044	技術	0.057	関係	0.037
施設	0.043	日本	0.054	技術	0.036	経済	0.034	状況	0.050	必要	0.036
削減	0.039	生産	0.054	保全	0.032	活動	0.028	規制	0.045	策定	0.035

図 4-9　LDA ヒートマップ（20 トピックス、分析対象語数：75）

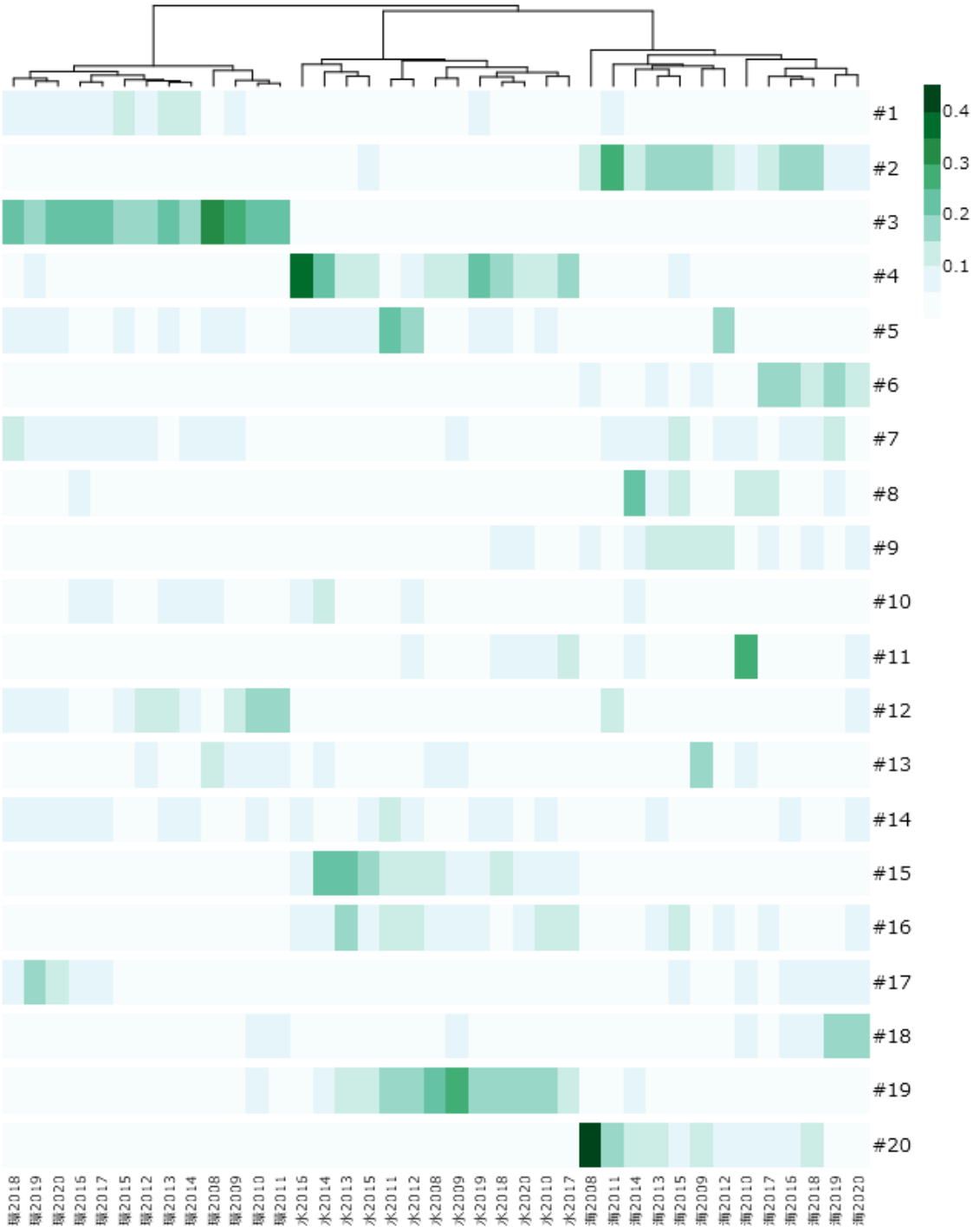


図 4-10　LDA ヒートマップ樹形図（20 トピックス、分析対象語数：75）

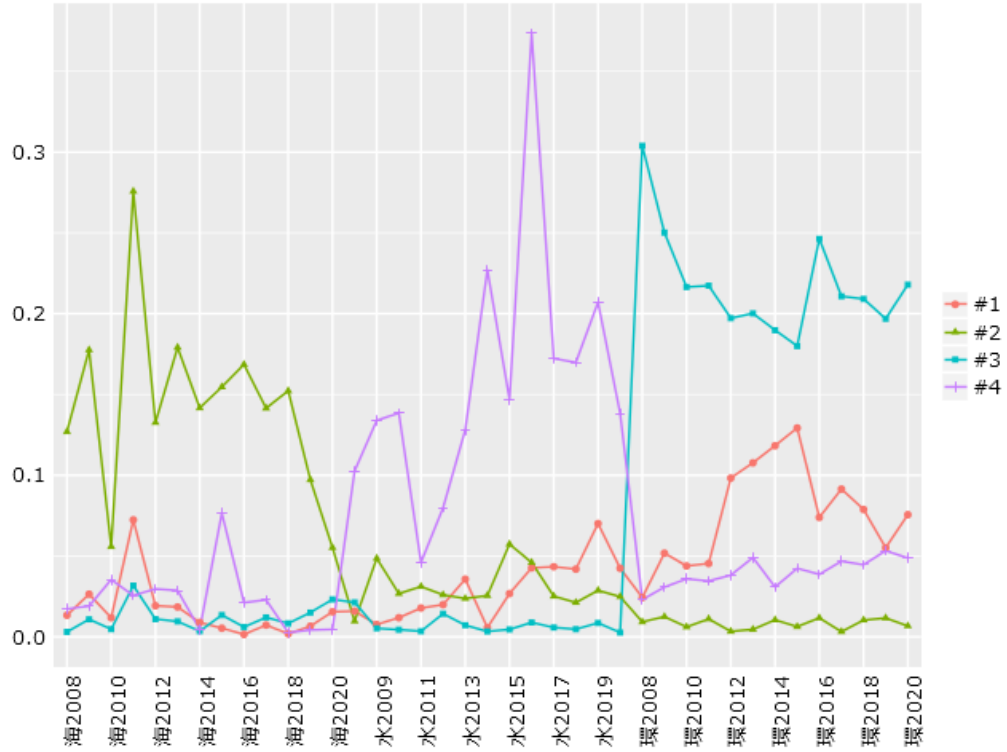


図 4-11　1～4 トピックの比率（20 トピックス、分析対象語数：75）

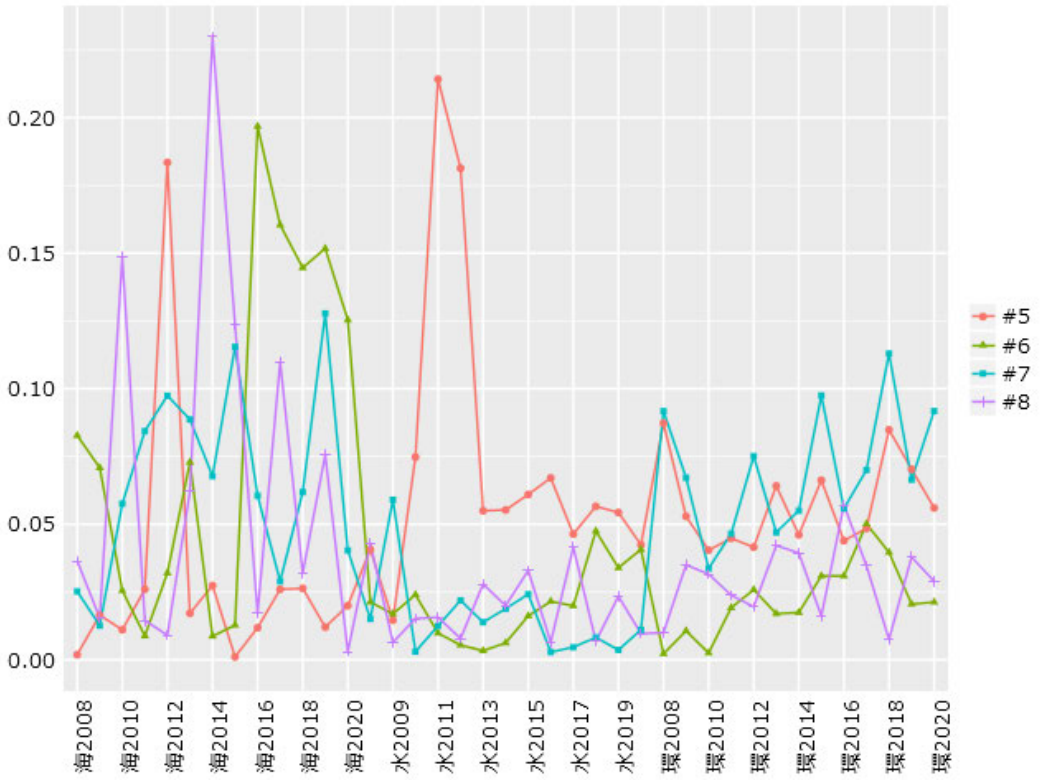


図 4-12　5～8 トピックの比率（20 トピックス、分析対象語数：75）

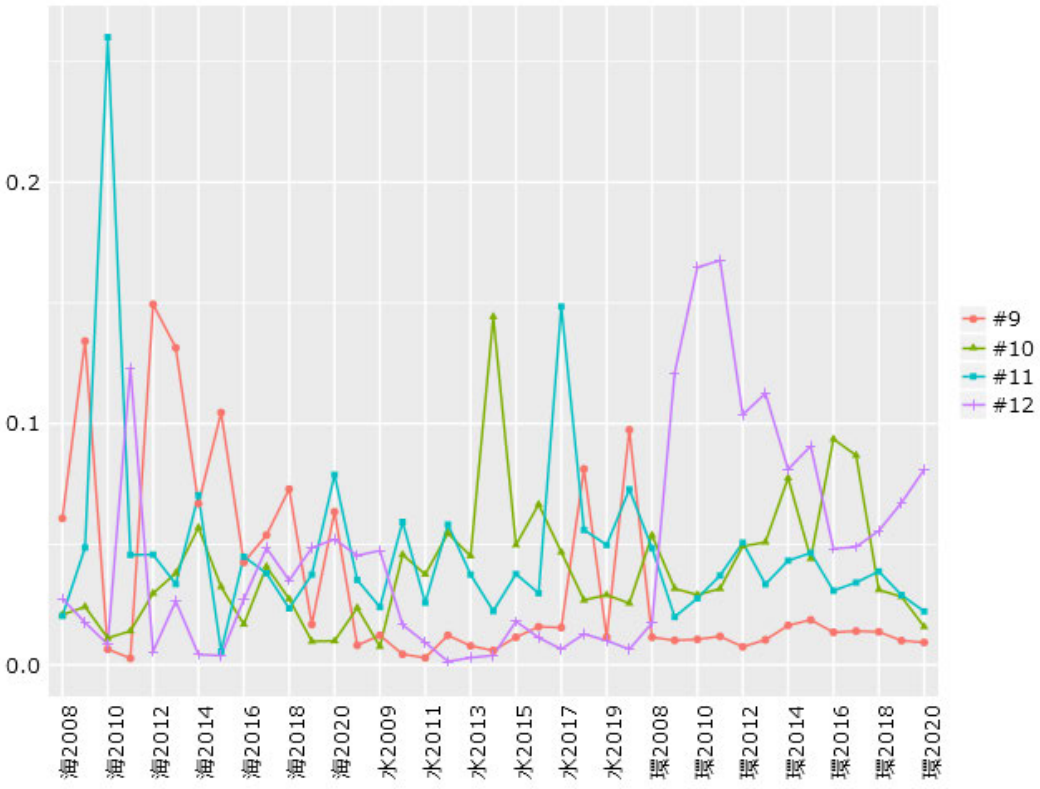


図 4-13 9～12 トピックの比率（20 トピックス、分析対象語数：75）

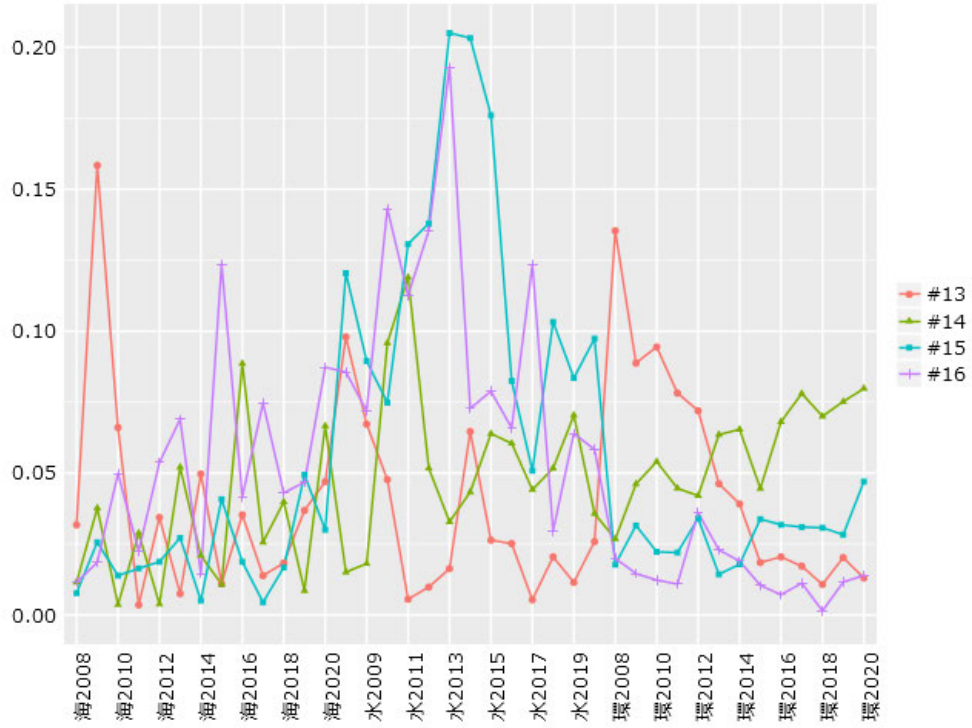


図 4-14 13～16 トピックの比率（20 トピックス、分析対象語数：75）

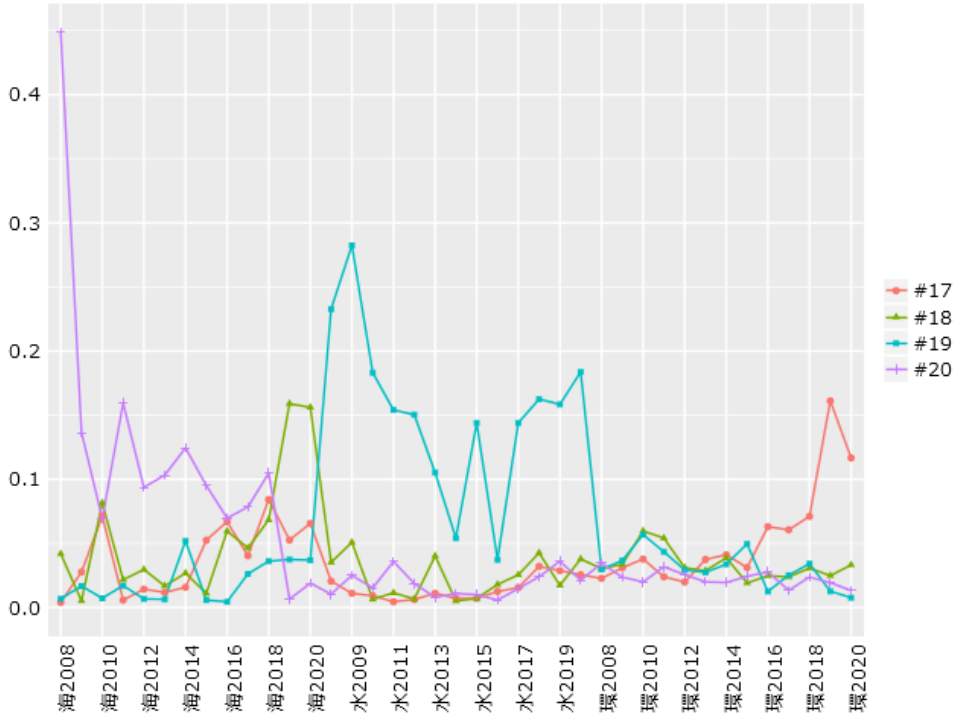


図 4-15 17～20 トピックの比率（20 トピックス、分析対象語数：75）

表 4-3 LDA 処理結果（14 トピックス、分析対象語数：154）

#1		#2		#3		#4		#5		#6		#7	
気候変動	0.090	海域	0.079	生物多様性	0.107	利用	0.064	基本	0.099	情報	0.085	被害	0.093
影響	0.055	計画	0.076	保全	0.059	社会	0.057	管理	0.062	調査	0.082	施設	0.083
実施	0.049	中国	0.067	経済	0.043	対策	0.052	政策	0.059	観測	0.079	関係	0.071
評価	0.042	管理	0.058	生態系	0.035	温暖化	0.043	総合的	0.052	技術	0.045	支援	0.069
対策	0.038	総合	0.035	自然	0.034	資源	0.042	計画	0.050	開発	0.045	調査	0.061
管理	0.027	日本	0.033	評価	0.028	状況	0.039	開発	0.038	計画	0.034	実施	0.046
情報	0.026	開発	0.033	利用	0.027	実施	0.037	利用	0.035	必要	0.030	活動	0.043
連携	0.025	問題	0.025	活動	0.026	制度	0.037	必要	0.035	研究	0.029	資源	0.029
開催	0.025	設置	0.024	生物	0.026	問題	0.036	問題	0.033	システム	0.028	利用	0.029
導入	0.023	利用	0.023	目標	0.024	必要	0.035	海域	0.032	日本	0.028	発生	0.026

#8		#9		#10		#11		#12		#13		#14	
水産物	0.080	廃棄物	0.065	管理	0.128	開発	0.055	原子力	0.077	教育	0.138	実施	0.070
漁船	0.069	処理	0.051	資源	0.097	国際	0.049	実施	0.063	産業	0.060	対策	0.059
漁獲	0.063	計画	0.030	措置	0.048	持続可能	0.046	規制	0.055	研究	0.041	活用	0.048
漁業者	0.056	排出量	0.029	重要	0.040	開催	0.042	発電	0.045	活動	0.040	技術	0.047
水産	0.056	利用	0.029	活動	0.032	会議	0.040	評価	0.041	連携	0.037	施設	0.046
養殖	0.048	エネルギー	0.028	関係	0.031	実施	0.034	制度	0.040	日本	0.035	調査	0.046
減少	0.042	社会	0.027	機関	0.030	目標	0.032	検討	0.035	生産	0.033	支援	0.040
生産	0.038	リサイクル	0.027	導入	0.030	政策	0.031	委員会	0.033	活用	0.032	情報	0.034
増加	0.031	排出	0.027	必要	0.028	世界	0.030	状況	0.031	期待	0.032	策定	0.032
消費	0.028	循環	0.023	制度	0.026	協力	0.028	法律	0.028	課題	0.027	健康	0.024

図 4-16　LDA ヒートマップ（14 トピックス、分析対象語数：154）

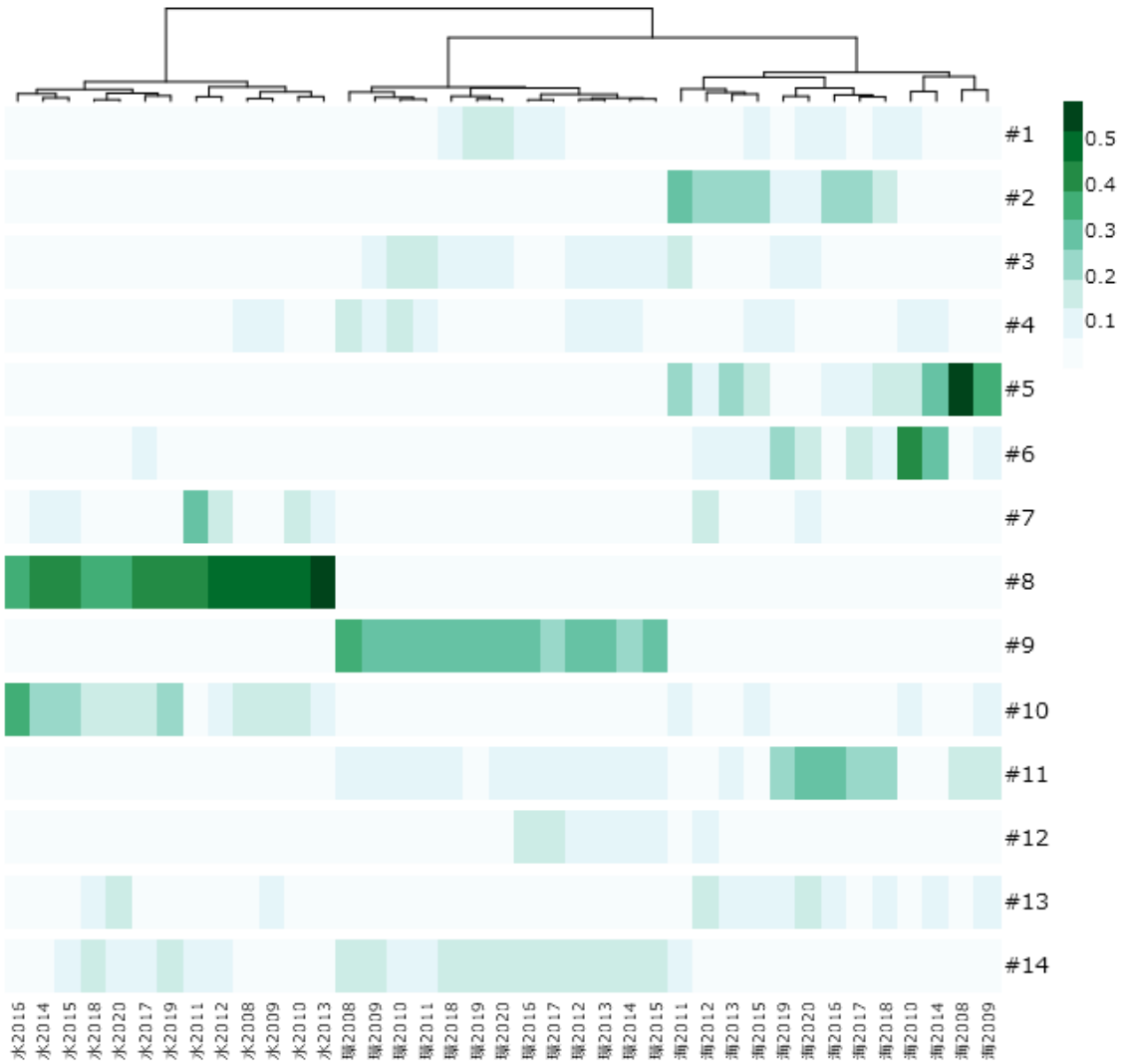


図 4-17　LDA ヒートマップ樹形図（14 トピックス、分析対象語数：154）

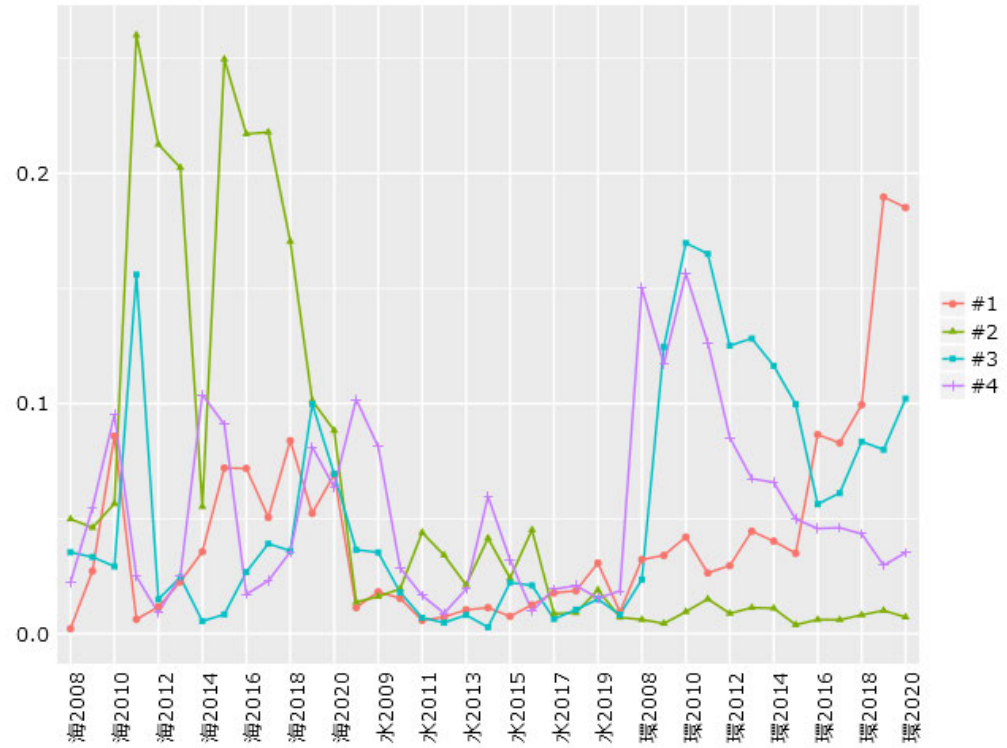


図 4-18 1～4 トピックの比率（14 トピックス、分析対象語数：154）

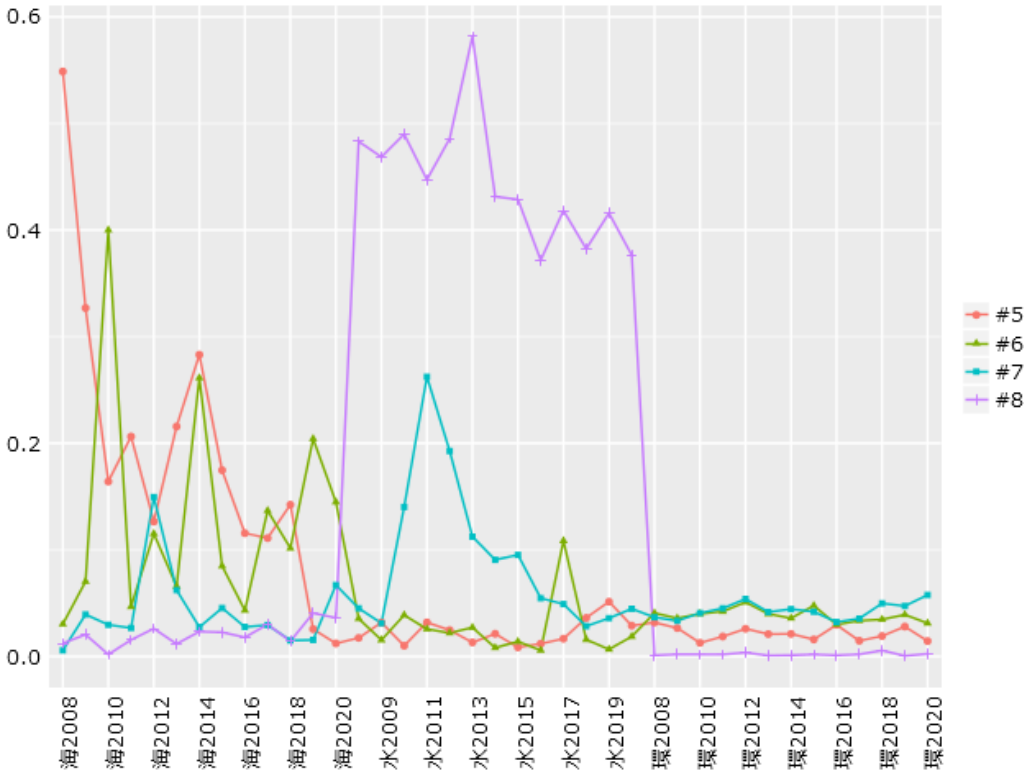


図 4-19 5～8 トピックの比率（14 トピックス、分析対象語数：154）

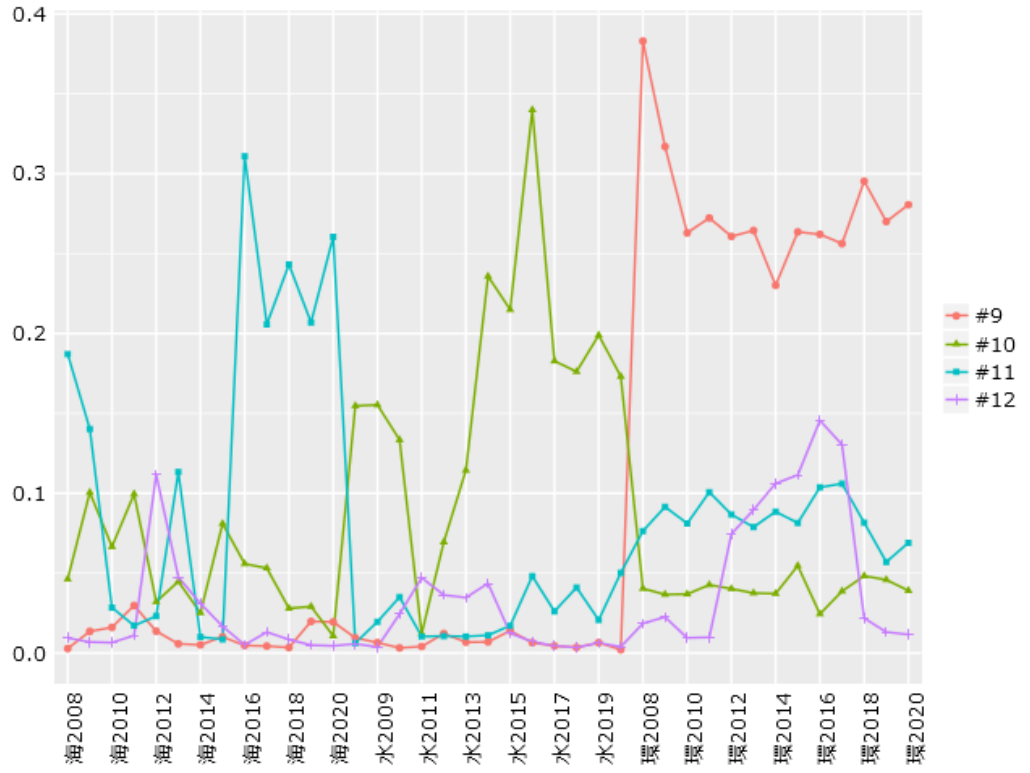


図 4-20 9～12 トピックの比率（14 トピックス、分析対象語数：154）

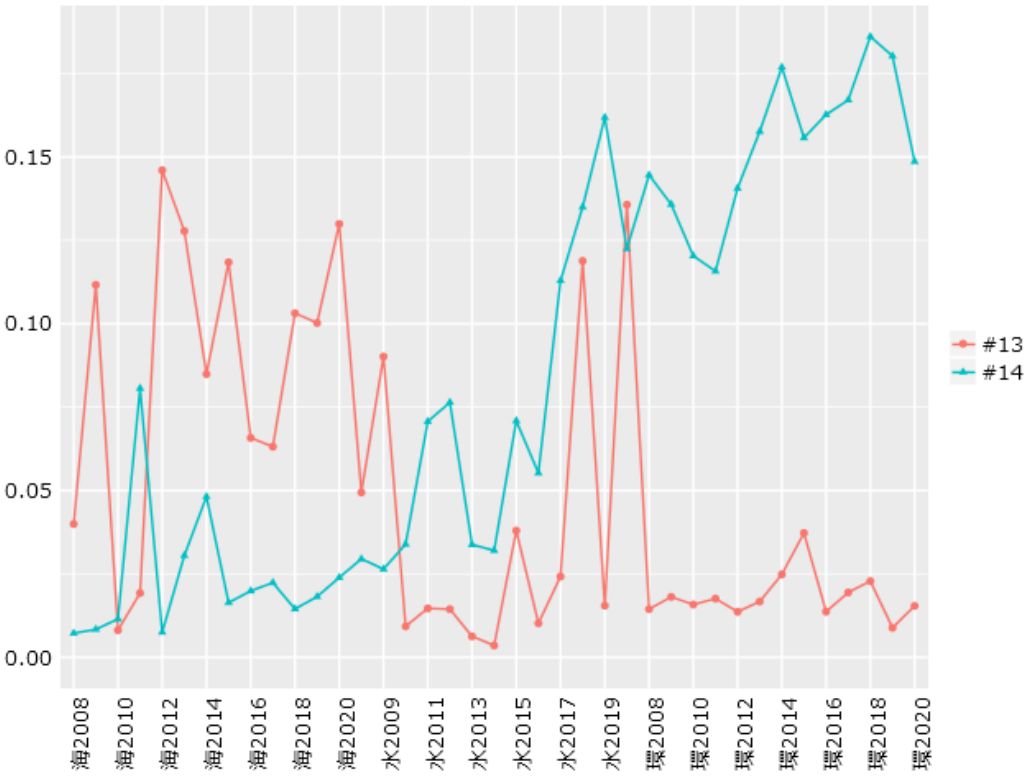


図 4-21 13～14 トピックの比率（14 トピックス、分析対象語数：154）

表 4-4 LDA 処理結果（20 トピックス、分析対象語数：154）

Topics

#1		#2		#3		#4		#5		#6		#7	
開発	0.066	教育	0.265	利用	0.069	日本	0.095	管理	0.123	管理	0.096	基本	0.110
開催	0.061	研究	0.051	処理	0.063	技術	0.085	保全	0.060	海域	0.093	管理	0.072
持続可能	0.059	機関	0.039	社会	0.055	研究	0.070	整備	0.050	計画	0.088	政策	0.060
会議	0.053	目的	0.038	廃棄物	0.054	目標	0.047	生物	0.043	中国	0.061	総合的	0.053
政策	0.044	水産	0.038	資源	0.046	開発	0.044	生態系	0.041	総合	0.041	計画	0.052
国際	0.044	必要	0.037	リサイクル	0.042	期待	0.040	機関	0.039	実施	0.036	問題	0.039
協力	0.037	活動	0.036	施設	0.042	中国	0.037	調査	0.034	開発	0.033	開発	0.038
政府	0.034	連携	0.027	整備	0.032	世界	0.031	適切	0.032	問題	0.030	利用	0.036
参加	0.032	産業	0.026	問題	0.030	分野	0.031	保護	0.031	産業	0.029	海域	0.031
資源	0.030	適切	0.025	回収	0.030	システム	0.028	自然	0.030	検討	0.028	国際	0.030

#8		#9		#10		#11		#12		#13		#14	
減少	0.104	エネルギー	0.070	被害	0.125	漁獲	0.096	産業	0.131	廃棄物	0.044	気候変動	0.126
増加	0.089	活動	0.060	調査	0.093	漁船	0.092	技術	0.079	実施	0.037	影響	0.081
資源	0.069	必要	0.059	施設	0.073	漁業者	0.082	利用	0.061	計画	0.033	実施	0.053
管理	0.054	日本	0.042	多く	0.055	水産	0.058	生産	0.054	処理	0.033	情報	0.038
向上	0.054	促進	0.037	支援	0.049	水産物	0.055	連携	0.052	排出量	0.032	評価	0.031
消費	0.051	重要	0.033	関係	0.049	重要	0.033	活用	0.047	対策	0.031	強化	0.029
拡大	0.040	課題	0.032	状況	0.038	措置	0.032	計画	0.041	排出	0.030	連携	0.029
利用	0.039	経済	0.032	発生	0.035	資源	0.031	実現	0.035	地球	0.025	貢献	0.025
機能	0.038	資源	0.031	対応	0.034	対象	0.029	高い	0.033	削減	0.024	施設	0.024
影響	0.037	発電	0.031	活動	0.033	実施	0.028	可能	0.031	温室効果ガス	0.023	向上	0.022

#15		#16		#17		#18		#19		#20	
調査	0.123	生物多様性	0.143	原子力	0.098	養殖	0.171	対策	0.100	社会	0.056
情報	0.123	保全	0.065	実施	0.080	水産物	0.148	情報	0.078	地球	0.056
観測	0.114	自然	0.040	規制	0.054	水産	0.068	支援	0.069	状況	0.048
計画	0.052	資源	0.038	発電	0.047	生産	0.052	実施	0.058	対策	0.046
必要	0.046	経済	0.035	委員会	0.039	漁船	0.036	世界	0.053	世界	0.046
開発	0.041	活動	0.033	法律	0.038	必要	0.031	状況	0.045	温暖化	0.040
地球	0.031	生態系	0.033	施設	0.037	消費	0.028	関係	0.043	必要	0.040
課題	0.031	利用	0.031	処理	0.036	重要	0.027	活用	0.037	影響	0.035
重要	0.030	企業	0.027	汚染	0.033	大きい	0.023	設置	0.035	利用	0.034
基本	0.029	実施	0.025	結果	0.032	利用	0.023	技術	0.034	検討	0.031

図 4-22　LDA ヒートマップ（20 トピックス、分析対象語数：154）

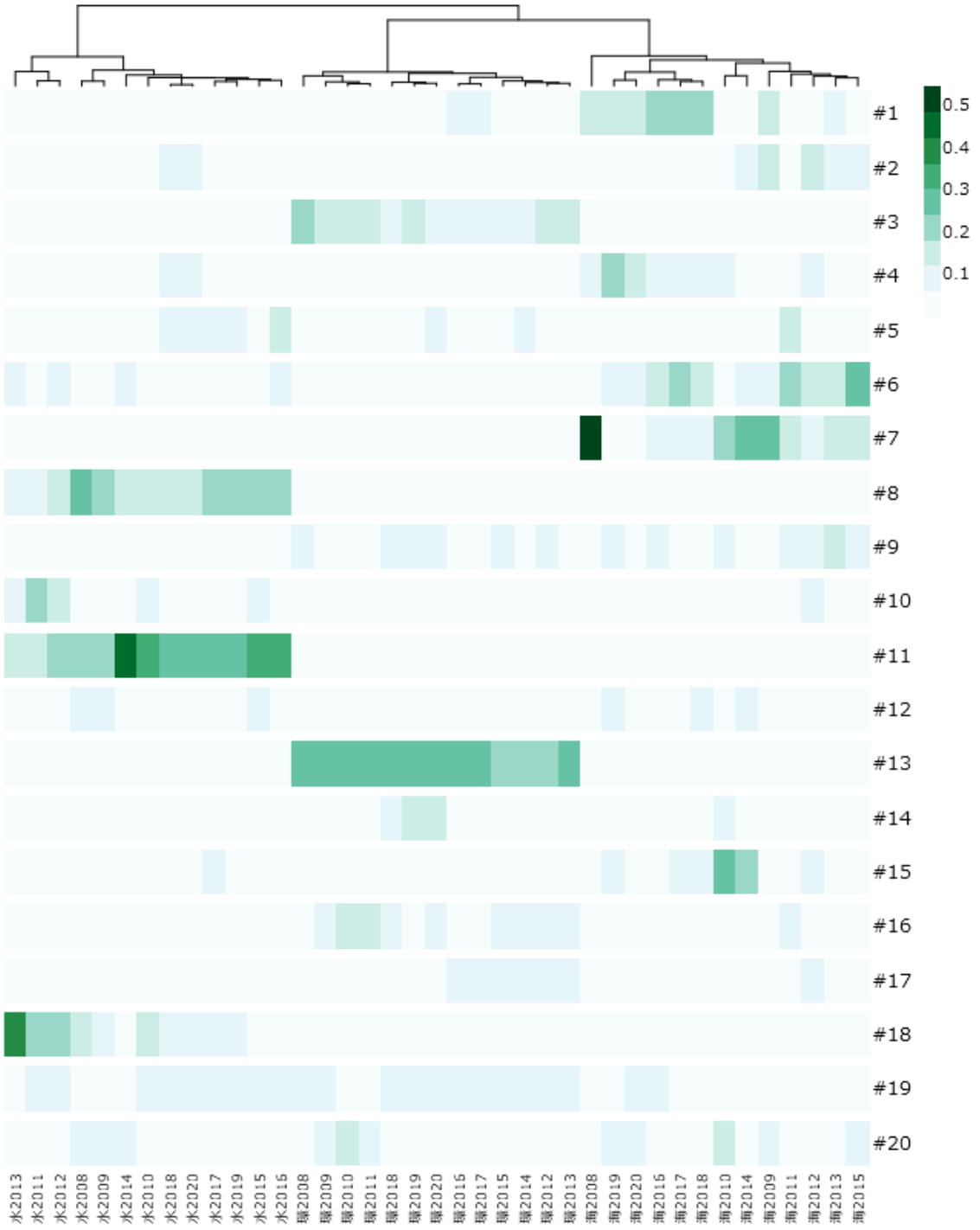


図 4-23　LDA ヒートマップ樹形図（20 トピックス、分析対象語数：154）

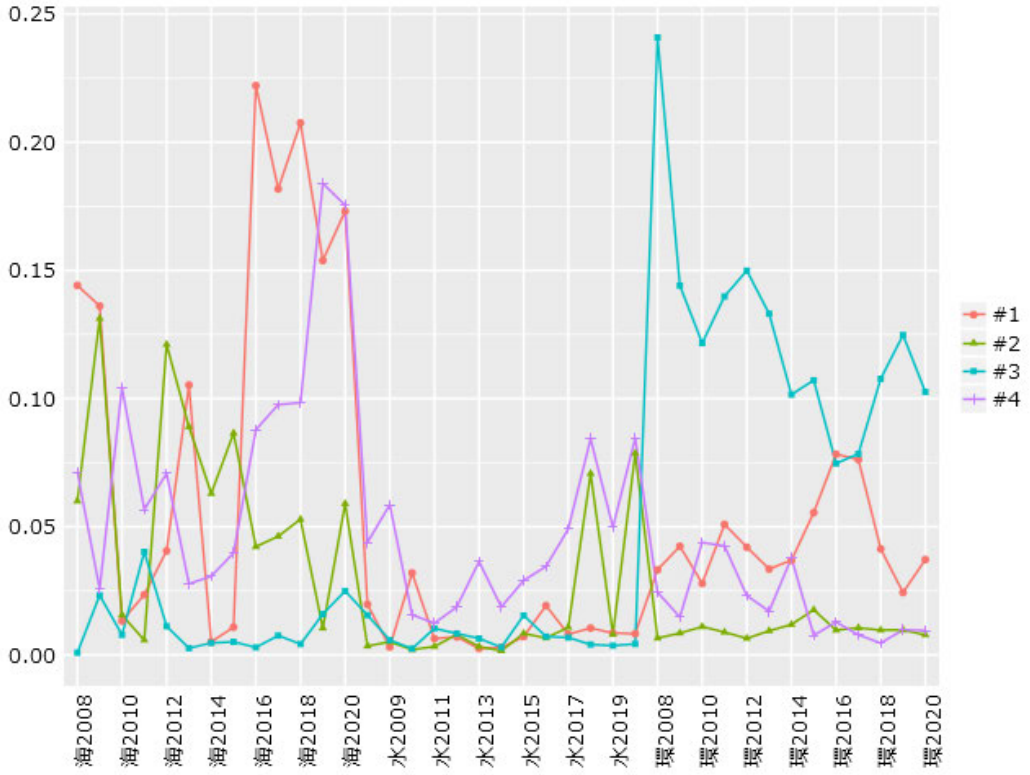


図 4-24 1～4 トピックの比率（20 トピックス、分析対象語数：154）

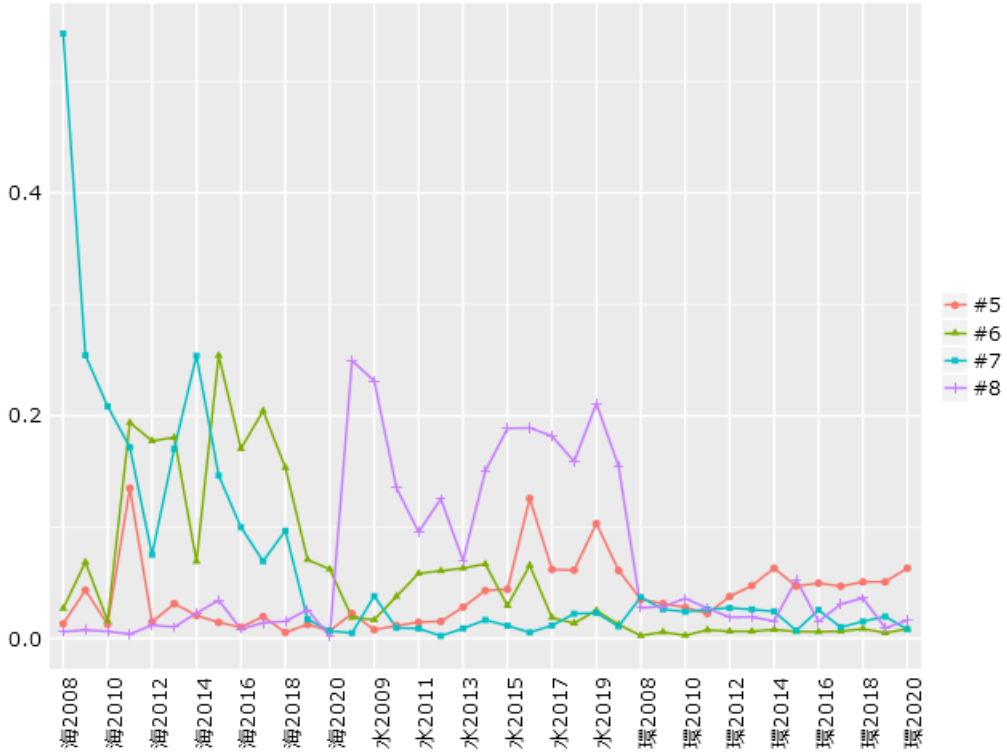


図 4-25 5～8 トピックの比率（20 トピックス、分析対象語数：154）

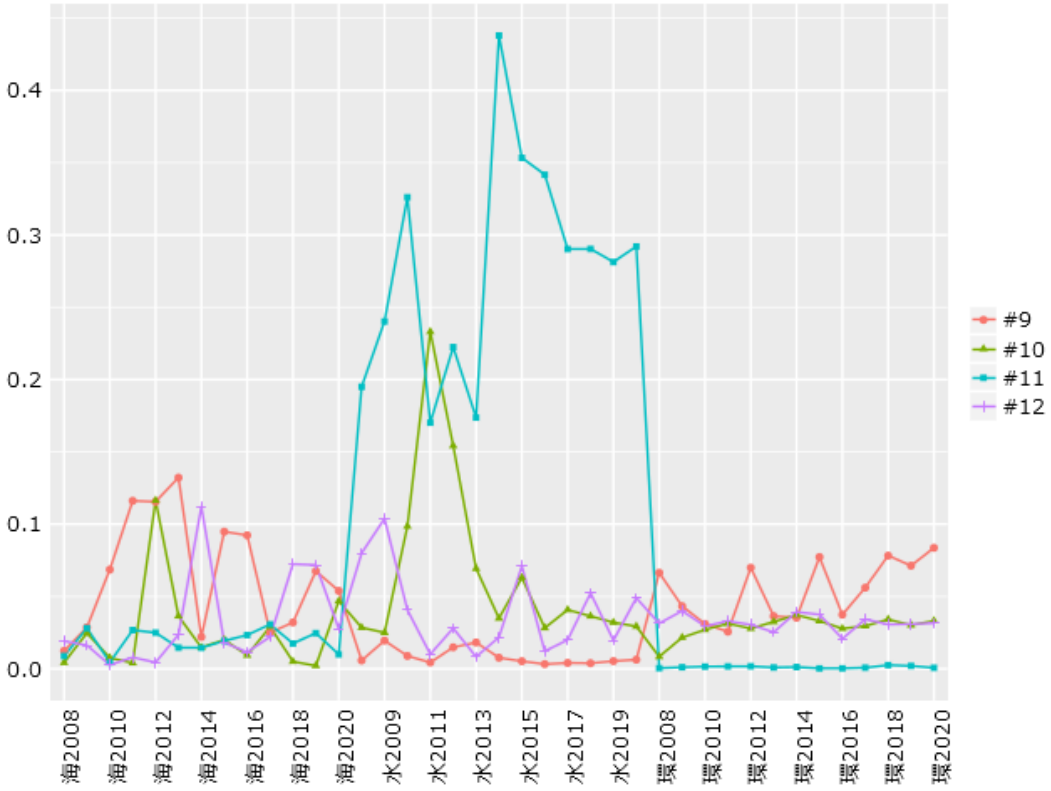


図 4-26 9～12 トピックの比率（20 トピックス、分析対象語数：154）

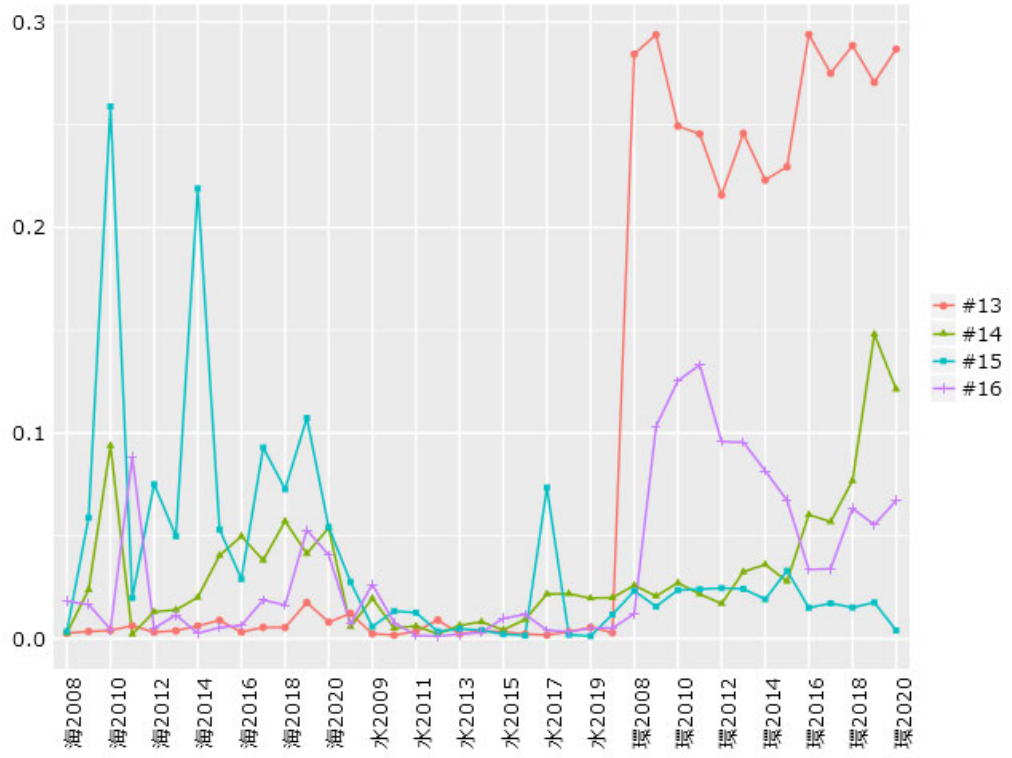


図 4-27 13～16 トピックの比率（20 トピックス、分析対象語数：154）

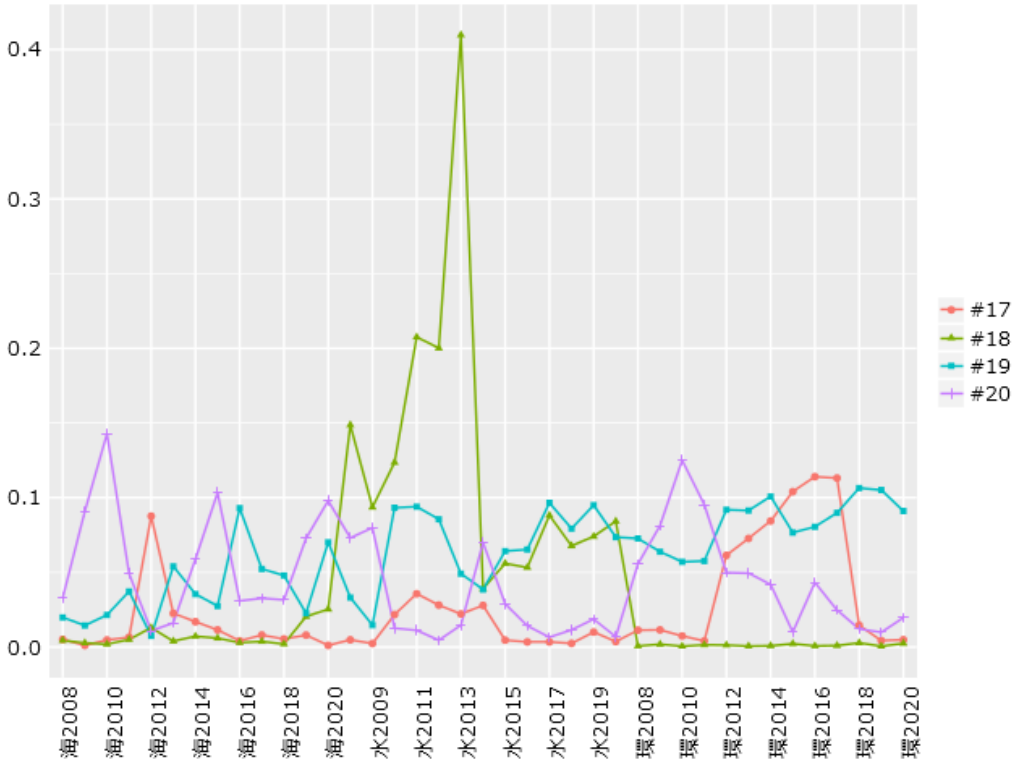


図 4-28 17～20 トピックの比率（20 トピックス、分析対象語数：154）

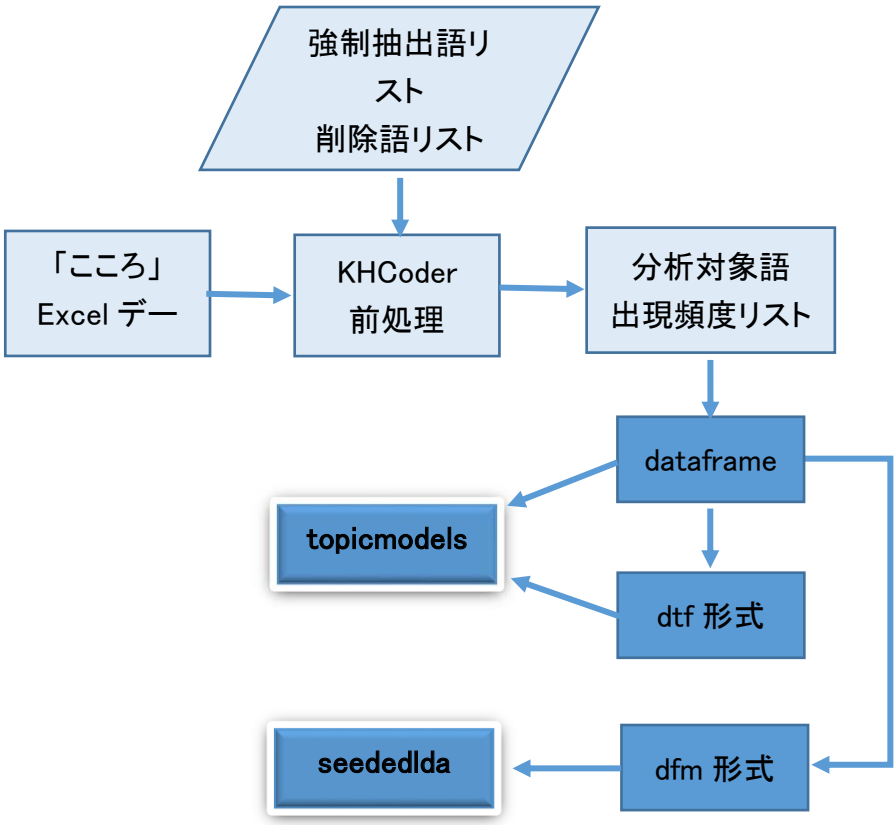
付録 9 「KHCoder での LDA 分析結果と R パッケージでの処理結果の比較」

文書番号：JRDN-21-028B

1. 概要

KH Coder のチュートリアル資料では、夏目漱石の「こころ」を題材とした LDA トピックモデル分析が紹介されている。そこで、「こころ」の Excel データ（C:¥KH Coder3¥tutorial_jp¥kokoro.xls）について、KH Coder での処理結果と R のパッケージでの処理結果との整合性を調査した。

KH Coder では、形態素解析ツール Chasen で処理した結果について、ユーザが設定した「強制抽出語、削除語、抽出対象品詞」での前処理結果が抽出語リスト（単語の出現頻度表）として得られる。この抽出語リストから LDA トピックモデル分析での分析対象語を CSV 形式で出力し、KH Coder での LDA トピック分析で使用している R toopicmodel パッケージの LDA 処理用入力データに変換して R-Studio で処理し、両者の整合性を確認する。更に、R seedelda パッケージの LDA（textmodel_lda）での処理結果とも比較し、教師付き LDA 分析（textmodel_seededlda）を試みる。KH Coder での処理と R-Studio での処理を図 1-1 に示す。



（注） は KH Coder での処理、 は R-Studio での処理をあらわす。

図 1-1 KH Coder での処理と R-Studio での処理

KH Coder では topicmodels の LDA 実行に際して、乱数の初期値を 1234567、フィッテイング手法として Gibbs サンプリング（デフォルト：VEM アルゴリズム）、burnin（開始後の無視する Gibbs イテレーション数、デフォルト:0）を 1000、iter（Gibbs イテレーション数）を 2000（デフォルト）と固定設定しており、分析対象語が同じであれば同じ結果が得られる。乱数の初期値、burnin の値、Gibbs イテレーション数 iter の値を変更すると、トピックモデルの結果は違ってくる。

R-Studio で topicmodels の LDA を実行する場合、入力データはデータフレーム形式の出現頻度行列または単語の出現頻度（TF）を重みとした dtm 形式（DocumentTermMatrix）である。一方、seededlda での入力データは、quanteda パッケージで使用されている dfm 形式である。

KH Coder に付帯されている R は 3.1 版と古く、CRAN からは各種パッケージを取得できない。現時点で広く普及している R 3.6.3 版と 4.1 版であれば CRAN 経由で R-Studio に各種パッケージをインストールできる。

topicmodels の LDA への入力データには、データフレーム形式の出現頻度行列を設定していることが判明したので、R の 3.1 版、3.6.3 版、4.1 版での topicmodels の LDA 実行結果に差異の無いことを確認した。また、KH Coder でのトピック分析結果と R-Studio でデータフレーム形式の出現頻度行列を入力データとした topicmodels の LDA 処理結果が一致することを確認した。しかし、出現頻度行列を tm パッケージで使用されている dtm 形式に変換した場合には、KH Coder でのトピック分析結果とは一致しない。これは dtm 形式への変換時に文書や単語がソートされて格納されるため、結果的に出現頻度行列（行：文章、列：単語）の行交換や列交換が発生し、topicmodels の LDA への入力データである出現頻度行列に違いが生ずることに起因している。

seededlda パッケージの textmodel_lda での処理結果と topicmodels の LDA 処理結果の比較を行った。なお、seededlda パッケージでも Gibbs サンプリングを使用しており、その最大イテレーション数をパラメータとして設定できる。

2. KH Coder での前処理

KH Coder チュートリアル資料（KH Coder_tutorial.pdf）に記載されているように、Excel 形式データ（kokoro.xls）での前処理にあたり強制抽出語（「一人、二人」）を設定する。抽出語リストを図 2-1 に示す。この結果は、解説書の図 A.23 と一致している。

強制抽出語を指定した KH Coder での「こころ（kokoro.xls）」の前処理結果を図 2-2 に示す。文：5064、段落・H5：1215、使用抽出語数：36194、使用異なり語数：5484 となっている。「こころ」は 3 部構成で、「上」：36 節、「中」：18 節、「下」：56 節の全 110 節であり、「こころ（kokoro.xls）」には、「テキスト」の他に「部」、「章」、「章（ラベル）」の列がある。

解説書の図 A.24「トピックモデルの推定結果」は、集計単位：H5（部）、トピック数：20、分析対象語数：71 と設定した処理結果であるが、この条件でのトピック数の推定（LDAtuning）結果を図 2-4 に示す。この図からトピック数は 6 或いは 12 と推定され、トピック数 6 でのトピック推定結果を図 2-5 に示す。

抽出語リスト

Filter Entry

OR検索　部分一致　フィルタ

List

#	抽出語	品詞/活用	頻度
1	先生	名詞	595
2	K	タグ	411
3	奥さん	名詞	388
⊞ 4	思う	動詞	296
5	父	名詞C	269
6	自分	名詞	264
⊞ 7	見る	動詞	225
⊞ 8	聞く	動詞	219
⊟ 9	出る	動詞	185
├	出	連用形	128
├	出る	基本形	33
├	出	未然形	18
└	出よ	未然ウ接続	6
10	人	名詞C	182
11	母	名詞C	170
12	お嬢さん	名詞	168
13	前	副詞可能	164

図 2-1　抽出語リスト

Project

現在のプロジェクト：kokoro.xls [テキスト]

説明（メモ）：

Database Stats

総抽出語数（使用）：106,073 (36,194)

異なり語数（使用）：6,064 (5,484)

文書の単純集計：

集計単位	ケース数
文	5,064
段落	1,215
H5	1,215

図 2-2 「こころ」の前処理結果

1	テキスト	部	章	章(ラベル)
1194	酒は止めたけれども、何もする気にはなりません。仕方がないから書物を読	[3]下_先生と遺書	3_53	下・五十三
1195	同時に私はKの死因を繰り返し繰り返し考えたのです。その当座は頭がた	[3]下_先生と遺書	3_53	下・五十三
1196	「その内妻の母が病気になりました。医者に見せると到底癒らないという診断	[3]下_先生と遺書	3_54	下・五十四
1197	母は死にました。私と妻はたった二人ぎりになりました。妻は私に向って、こ	[3]下_先生と遺書	3_54	下・五十四
1198	母の亡くなった後、私はできるだけ妻を親切に取り扱ってやりました。ただ、	[3]下_先生と遺書	3_54	下・五十四
1199	妻はある時、男の心と女の心とはどうしてもぴたりと一つになれないものだろ	[3]下_先生と遺書	3_54	下・五十四
1200	私の胸にはその時分から時々恐ろしい影が閃きました。初めはそれが偶然	[3]下_先生と遺書	3_54	下・五十四
1201	私はただ人間の罪というものを深く感じたのです。その感じが私をKの墓へ引	[3]下_先生と遺書	3_54	下・五十四
1202	私がそう決心してから今日まで何年になるでしょう。私と妻とは元の通り仲好	[3]下_先生と遺書	3_54	下・五十四
1203	「死んだつもりで生きて行こうと決心した私の心は、時々外界の刺戟で躍り上	[3]下_先生と遺書	3_55	下・五十五
1204	波瀾も曲折もない単調な生活を続けて来た私の内面には、常にこうした苦し	[3]下_先生と遺書	3_55	下・五十五
1205	私は今日に至るまですでに二、三度運命の導いて行く最も楽な方向へ進もう	[3]下_先生と遺書	3_55	下・五十五
1206	同時に私だけがいなくなった後の妻を想像してみるといかにも不憫でした。私	[3]下_先生と遺書	3_55	下・五十五
1207	記憶して下さい。私はこんな風にして生きて来たのです。始めてあなたに鎌	[3]下_先生と遺書	3_55	下・五十五
1208	すると夏の暑い盛りに明治天皇が崩御になりました。その時私は明治の精神	[3]下_先生と遺書	3_55	下・五十五
1209	「私は殉死という言葉をほとんど忘れていました。平生使う必要のない字だか	[3]下_先生と遺書	3_56	下・五十六
1210	それから約一カ月ほど経ちました。御大葬の夜私はいつもの通り書斎に坐	[3]下_先生と遺書	3_56	下・五十六
1211	私は新聞で乃木大将の死ぬ前に書き残して行ったものを読みました。西南	[3]下_先生と遺書	3_56	下・五十六
1212	それから二、三日して、私はとうとう自殺する決心をしたのです。私に乃木さ	[3]下_先生と遺書	3_56	下・五十六
1213	私は妻を残して行きます。私がいなくなっても妻に衣食住の心配がないのは	[3]下_先生と遺書	3_56	下・五十六
1214	私が死のうと決心してから、もう十日以上になりますが、その大部分はあな	[3]下_先生と遺書	3_56	下・五十六
1215	しかし私は今その要求を果たしました。もう何にもする事はありません。この	[3]下_先生と遺書	3_56	下・五十六
1216	私は私の過去を善悪ともに他の参考に供するつもりです。しかし妻だけはた	[3]下_先生と遺書	3_56	下・五十六

（注） この Excel データにより、KH Coder では段落：1215 となっている。

図 2-3 「こころ（kokoro.xls）」データ

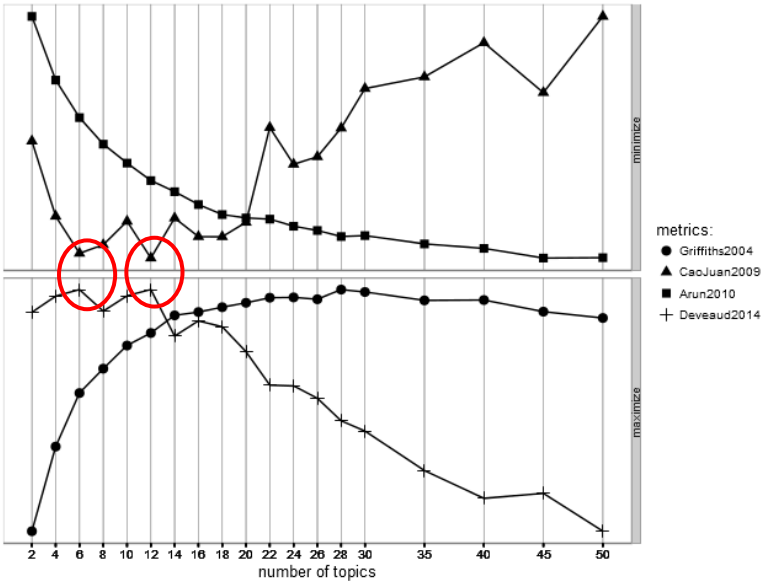


図 2-4 トピック数の推定結果（ldatuning）

Topics

#1		#2		#3		#4		#5		#6	
出る	0.138	先生	0.443	K	0.278	聞く	0.109	父	0.202	思う	0.212
前	0.123	人	0.130	奥さん	0.242	知る	0.090	母	0.113	自分	0.172
今	0.092	来る	0.095	お嬢さん	0.111	顔	0.077	帰る	0.102	考える	0.084
眼	0.090	手紙	0.055	答える	0.061	言葉	0.075	死ぬ	0.065	妻	0.074
立つ	0.073	書く	0.053	室	0.046	見る	0.073	好い	0.062	心	0.074
行く	0.070	聞く	0.041	声	0.041	口	0.064	叔父	0.050	人間	0.047
一人	0.054	急	0.040	見る	0.038	二人	0.061	東京	0.048	意味	0.046
宅	0.051	返事	0.038	坐る	0.036	話	0.055	病気	0.047	男	0.037
手	0.050	外	0.018	取る	0.030	知れる	0.051	出す	0.042	解る	0.031
頭	0.045	今	0.012	女	0.026	少し	0.048	卒業	0.041	態度	0.027

図 2-5 トピック数 6 での推定結果

KH Coder での前処理結果である抽出語リストを CSV ファイルで出力するには、プロジェクト ＞ エクスポート ＞「文書 x 抽出語」表 ＞ CSV ファイル と選択操作し、ポップアップされたパラメータ設定画面で「OK」ボタンを押すと、出力ファイルの設定画面でファイル名を設定後「保存（S）」ボタンを押す。保存された CSV ファイルを図 2-7 に示す。この CSV ファイルには、文書（行）ごとに単語（列）の出現頻度が保存されており、出現頻度がゼロの文書（H5）を除外すれば出現頻度行列となる。この CSV ファイルを R-Studio で読み込む場合には、エンコーディング指定が必要である。

```
read.csv("ファイル名", fileEncoding="UTF-8-BOM")
```

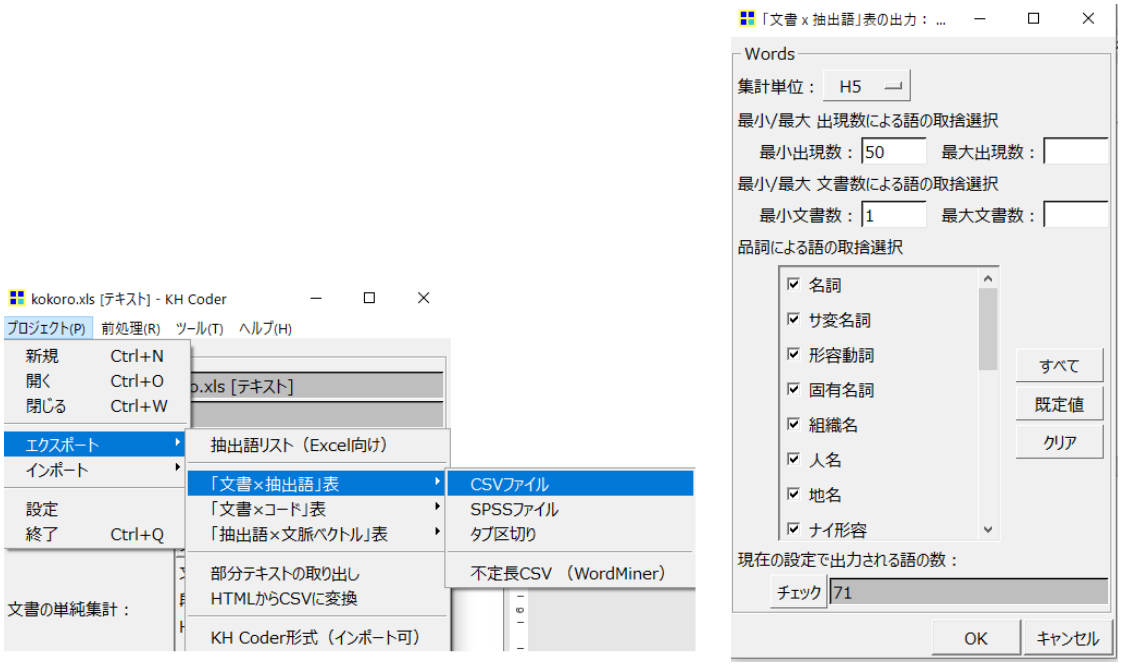


図 2-6 抽出語リストの CSV ファイル出力操作

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	h1	h2	h3	h4	h5	id	length_c	length_w	先生	奥さん	自分	お嬢さん	言葉	手紙	叔父	人間	様子	心持	態度
2	0	0	0	0	1	1	152	95	3	0	0	0	0	0	0	0	0	1	0
3	0	0	0	0	2	2	431	277	1	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	3	3	171	114	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	4	4	141	88	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	5	5	197	137	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	6	6	273	186	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	7	7	205	134	5	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	8	8	383	278	0	0	0	0	0	0	0	0	1	0	0
10	0	0	0	0	9	9	116	75	1	0	1	0	0	0	0	0	0	0	0
11	0	0	0	0	10	10	210	135	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	11	11	117	77	1	0	0	0	0	0	0	0	0	0	0

1207	0	0	0	0	1206	1206	243	171	0	0	0	0	0	0	0	0	0	0	
1208	0	0	0	0	1207	1207	185	126	0	0	0	0	0	0	0	0	0	0	
1209	0	0	0	0	1208	1208	184	123	0	0	1	0	2	0	0	0	0	1	
1210	0	0	0	0	1209	1209	139	97	0	0	0	0	0	0	0	0	0	0	
1211	0	0	0	0	1210	1210	278	183	0	0	0	0	0	0	0	0	0	0	
1212	0	0	0	0	1211	1211	214	136	0	0	0	0	0	0	0	1	0	0	
1213	0	0	0	0	1212	1212	156	111	0	0	0	0	0	0	0	0	0	0	
1214	0	0	0	0	1213	1213	431	272	0	0	2	0	0	0	0	2	0	1	
1215	0	0	0	0	1214	1214	187	127	0	0	0	0	0	1	0	0	0	0	
1216	0	0	0	0	1215	1215	183	117	0	0	0	0	0	0	0	0	0	0	

図 2-7 抽出語リスト CSV ファイルの内容

3. dtm 形式への変換

R-Studio で抽出語リスト CSV ファイルの出現頻度行列部分を読み込み、出現頻度がゼロの文書（H5）を除外すると、出現頻度行列（データフレーム）が得られる。KH Coder では、このデータフレーム形式のデータを topicmodels への入力データとして処理している。

抽出語リスト CSV ファイル（kokoto_H5-71w.csv）を読み込み、topicmodels::LDA で処理

```
dtm <- read.csv("kokoto_H5-71w.csv", fileEncoding="UTF-8-BOM")
dtmx <- dtm[,9:79]
dtmy <- dtmx[rowSums(dtmx) > 0,]

library(topicmodels)
kokoro_lda6 <- topicmodels::LDA(dtmy, k = 6, method = "Gibbs",
   control = list(seed = 1234567, burnin = 1000) ) # KH Coder での処理設定と同じ

terms(kokoro_lda6,10)　# トピック毎の単語群を出力
```

```
> terms(result_lda6,10)
      Topic 1 Topic 2 Topic 3   Topic 4 Topic 5 Topic 6
 [1,] "出る"  "先生"  "K"       "聞く"  "父"    "思う"
 [2,] "前"    "人"    "奥さん"  "知る"  "母"    "自分"
 [3,] "今"    "来る"  "お嬢さん" "顔"   "帰る"  "考える"
 [4,] "眼"    "手紙"  "答える"  "言葉"  "死ぬ"  "妻"
 [5,] "立つ"  "書く"  "室"      "見る"  "好い"  "心"
 [6,] "行く"  "聞く"  "声"      "口"    "叔父"  "人間"
 [7,] "一人"  "急"    "見る"    "二人"  "東京"  "意味"
 [8,] "宅"    "返事"  "坐る"    "話"    "病気"  "男"
 [9,] "手"    "外"    "取る"    "知れる" "出す"  "解る"
[10,] "頭"    "今"    "女"      "少し"  "卒業"  "態度"
```

参考：KH Coder での LDA 処理結果

Topics

#1		#2		#3		#4		#5		#6	
出る	0.138	先生	0.443	K	0.278	聞く	0.109	父	0.202	思う	0.212
前	0.123	人	0.130	奥さん	0.242	知る	0.090	母	0.113	自分	0.172
今	0.092	来る	0.095	お嬢さん	0.111	顔	0.077	帰る	0.102	考える	0.084
眼	0.090	手紙	0.055	答える	0.061	言葉	0.075	死ぬ	0.065	妻	0.074
立つ	0.073	書く	0.053	室	0.046	見る	0.073	好い	0.062	心	0.074
行く	0.070	聞く	0.041	声	0.041	口	0.064	叔父	0.050	人間	0.047
一人	0.054	急	0.040	見る	0.038	二人	0.061	東京	0.048	意味	0.046
宅	0.051	返事	0.038	坐る	0.036	話	0.055	病気	0.047	男	0.037
手	0.050	外	0.018	取る	0.030	知れる	0.051	出す	0.042	解る	0.031
頭	0.045	今	0.012	女	0.026	少し	0.048	卒業	0.041	態度	0.027

tidytext パッケージの cast_dtm により、データフレーム形式データを tm パッケージでの dtm 形式データへ変換する。

出現頻度行列(dtmy)を dtm 形式へ変換

```
library(tidytext)   # tidytext パッケージをインストール
library(dplyr)      # パイプ演算子（%>%）を利用するため dplyr パッケージをインストール
library(reshape2) # melt()単語頻度行列を one-term-per-row 形式データフレームに変換
str(dtmy)
```

```
> str(dtmy)
'data.frame':   1062 obs. of  71 variables:
 $ 先生    : int  3 1 0 0 0 1 5 0 1 0 ...
 $ 奥さん  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 自分    : int  0 0 0 0 0 0 0 0 1 0 ...
 $ お嬢さん: int  0 0 0 0 0 0 0 0 0 0 ...
```

colname(dtmy)

```
> rownames(dtmy)
  [1] "1"    "2"    "3"    "4"    "5"    "6"    "7"    "8"    "9"    "10"   "11"   "12"   "13"   "14"   "15"   "16"
 [17] "17"   "18"   "19"   "20"   "21"   "22"   "23"   "24"   "25"   "26"   "28"   "30"   "31"   "32"   "35"   "36"
 [33] "37"   "38"   "39"   "40"   "43"   "44"   "48"   "52"   "54"   "55"   "56"   "57"   "58"   "60"   "61"   "62"

[961] "1113" "1114" "1115" "1116" "1117" "1118" "1119" "1120" "1122" "1123" "1124" "1125" "1126" "1127" "1128" "1129"
[977] "1130" "1131" "1132" "1133" "1134" "1135" "1136" "1137" "1138" "1139" "1140" "1141" "1142" "1143" "1144" "1145"
[993] "1146" "1147" "1148" "1149" "1150" "1151" "1152" "1153"
[ reached getOption("max.print") -- omitted 62 entries ]
```

rawname(dtmy)

```
> colnames(dtmy)
 [1] "先生"   "奥さん" "自分"   "お嬢さん" "言葉"   "手紙"   "叔父"   "人間"   "様子"   "心持"
[11] "態度"   "話"     "意味"   "病気"     "卒業"   "返事"   "急"     "東京"   "問題"   "前"
[21] "今"     "K"      "二人"   "一人"     "思う"   "見る"   "聞く"   "出る"   "帰る"   "来る"
[31] "考える" "知る"   "行く"   "立つ"     "答える" "死ぬ"   "見える" "知れる" "書く"   "解る"
[41] "出す"   "話す"   "取る"   "坐る"     "付く"   "笑う"   "歩く"   "読む"   "好い"   "悪い"
[51] "少し"   "父"     "人"     "母"       "顔"     "眼"     "心"     "妻"     "口"     "女"
[61] "頭"     "手"     "事"     "宅"       "家"     "室"     "男"     "気"     "兄"     "声"
[71] "外"
```

```
dtmy_mat <- as.matrix(dtmy)   # dtmy（dataframe） を行列に変換
str(dtmy_mat)
```

```
> str(dtmy_mat)
 int [1:1062, 1:71] 3 1 0 0 0 1 5 0 1 0 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:1062] "1" "2" "3" "4" ...
  ..$ : chr [1:71] "先生" "奥さん" "自分" "お嬢さん" ...
```

```
kokoro_df <- melt(dtmy_mat)

# tidytext パッケージの cast_dtm()でデータフレームを DTM 形式に変換
kokoro_dtm <- kokoro_df %>% cast_dtm(document=Var1, term=Var2,
                                     value=value)
str(kokoro_dtm)
```

```
> str(wkokoro_dtm)
List of 6
 $ i       : int [1:75402] 1 2 3 4 5 6 7 8 9 10 ...
 $ j       : int [1:75402] 1 1 1 1 1 1 1 1 1 1 ...
 $ v       : num [1:75402] 3 1 0 0 0 1 5 0 1 0 ...
 $ nrow    : int 1062
 $ ncol    : int 71
 $ dimnames:List of 2
  ..$ Docs : chr [1:1062] "1" "2" "3" "4" ...
  ..$ Terms: chr [1:71] "先生" "奥さん" "自分" "お嬢さん" ...
 - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
 - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
```

（注） dtmy_mat と kokoro_df で、単語の順序が同じ。

4. dfm 形式への変換

R-Studio で抽出語リスト CSV ファイルの出現頻度行列部分を読み込み、出現頻度がゼロの文書（H5）を除外すると、出現頻度行列（データフレーム）が得られる。tidytext パッケージの cast_dfm により、データフレーム形式のデータを quanteda パッケージで使用されている dfm 形式へ変換する。

出現頻度行列(dtmy)を dfm 形式へ変換

```
library(tidytext)   # tidytext パッケージをインストール
library(dplyr)      # パイプ演算子（%>%）を利用するため dplyr パッケージをインストール
library(reshape2) # melt()単語頻度行列を one-term-per-row 形式データフレームに変換


str(dtmy)
```

```
> str(dtmy)
'data.frame':   1062 obs. of  71 variables:
 $ 先生    : int  3 1 0 0 0 1 5 0 1 0 ...
 $ 奥さん  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 自分    : int  0 0 0 0 0 0 0 0 1 0 ...
 $ お嬢さん: int  0 0 0 0 0 0 0 0 0 0 ...
```

```
dtmy_mat <- as.matrix(dtmy)   # dtmy（dataframe） を行列に変換


str(dtmy_mat)
```

```
> str(dtmy_mat)
 int [1:1062, 1:71] 3 1 0 0 0 1 5 0 1 0 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:1062] "1" "2" "3" "4" ...
  ..$ : chr [1:71] "先生" "奥さん" "自分" "お嬢さん" ...
```

```
kokoro_df <- melt(dtmy_mat)


# tidytext パッケージの cast_dfm でデータフレームを DTM 形式に変換
```

```
kokoro_dfm <- kokoro_df %>% cast_dfm(document=Var1, term=Var2,
                                     value=value)
str(kokoro_dfm)
```

```
> str(wkokoro_dfm)
Formal class 'dfm' [package "quanteda"] with 8 slots
  ..@ docvars :'data.frame':    1062 obs. of  3 variables:
  .. ..$ docname_: chr [1:1062] "1" "2" "3" "4" ...
  .. ..$ docid_  : Factor w/ 1062 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9
  .. ..$ segid_  : int [1:1062] 1 1 1 1 1 1 1 1 1 ...
  ..@ meta    :List of 3
  .. ..$ system:List of 5
  .. .. ..$ package-version:Classes 'package_version', 'numeric_version'  hidd
  .. .. .. ..$ : int [1:3] 3 2 1
  .. .. ..$ r-version      :Classes 'R_system_version', 'package_version', 'nu
  .. .. .. ..$ : int [1:3] 3 6 3
  .. .. ..$ system         : Named chr [1:3] "Windows" "x86-64" "Masataka"
  .. .. .. ..- attr(*, "names")= chr [1:3] "sysname" "machine" "user"
  .. .. ..$ directory      : chr "D:/R_workspacex"
  .. .. ..$ created        : Date[1:1], format: "2022-05-03"
  .. ..$ object:List of 9
  .. .. ..$ unit         : chr "documents"
  .. .. ..$ what         : chr "word"
  .. .. ..$ ngram        : int 1
  .. .. ..$ skip         : int 0
  .. .. ..$ concatenator: chr "_"
  .. .. ..$ weight_tf    :List of 3
  .. .. .. ..$ scheme: chr "count"
  .. .. .. ..$ base  : NULL
  .. .. .. ..$ k     : NULL
  .. .. ..$ weight_df    :List of 5
  .. .. .. ..$ scheme    : chr "unary"
  .. .. .. ..$ base      : NULL
  .. .. .. ..$ c         : NULL
  .. .. .. ..$ smoothing: NULL
  .. .. .. ..$ threshold: NULL
  .. .. ..$ smooth       : num 0
  .. .. ..$ summary      :List of 2
  .. .. .. ..$ hash: chr(0)
  .. .. .. ..$ data: NULL
  .. ..$ user   : list()
  ..@ i       : int [1:75402] 0 1 2 3 4 5 6 7 8 9 ...
  ..@ p       : int [1:72] 0 1062 2124 3186 4248 5310 6372 7434 8496 9558 ...
  ..@ Dim     : int [1:2] 1062 71
  ..@ Dimnames:List of 2
  .. ..$ docs     : chr [1:1062] "1" "2" "3" "4" ...
  .. ..$ features: chr [1:71] "先生" "奥さん" "自分" "お嬢さん" ...
  ..@ x       : num [1:75402] 3 1 0 0 0 1 5 0 1 0 ...
  ..@ factors : list()
```

5. topicmodels での処理結果

KH Coder では、単語出現頻度行列データを topicmodels への入力データとして処理しているが、ここでは tm パッケージで使用される dtm 形式データを入力データとし、KH Coder での処理と同じパラメータを設定して処理した結果を比較した。Reshape2 パッケージの melt()関数を利用すれば、単語出現頻度行列データと dtm 形式データでの LDA 処理結果は同じとなった。

出現頻度行列(データフレーム形式：dtmy)で LDA 処理

```
kokoro_lda6 <- topicmodels::LDA(dtmy, k = 6, method = "Gibbs",
  control = list(seed = 1234567, burnin = 1000) ) # KH Coder での処理設定と同じ
terms(kokoro_lda6,10)  # トピック毎の単語群を出力
```

```
> terms(result_lda6,10)
      Topic 1 Topic 2 Topic 3   Topic 4  Topic 5 Topic 6
 [1,] "出る"  "先生"  "K"       "聞く"   "父"    "思う"
 [2,] "前"    "人"    "奥さん"  "知る"   "母"    "自分"
 [3,] "今"    "来る"  "お嬢さん" "顔"     "帰る"  "考える"
 [4,] "眼"    "手紙"  "答える"  "言葉"   "死ぬ"  "妻"
 [5,] "立つ"  "書く"  "室"      "見る"   "好い"  "心"
 [6,] "行く"  "聞く"  "声"      "口"     "叔父"  "人間"
 [7,] "一人"  "急"    "見る"    "二人"   "東京"  "意味"
 [8,] "宅"    "返事"  "坐る"    "話"     "病気"  "男"
 [9,] "手"    "外"    "取る"    "知れる" "出す"  "解る"
[10,] "頭"    "今"    "女"      "少し"   "卒業"  "態度"
```

DTM 形式データ（kokoro_dtm）について LDA 処理

```
kokoro_LDA6 <- LDA(kokoro_dtm, k=6, method="Gibbs",
                   control=list(seed=1234567, burnin=1000) )
terms(kokoro_LDA6, 10)
```

```
      Topic 1 Topic 2 Topic 3   Topic 4  Topic 5 Topic 6
 [1,] "出る"  "先生"  "K"       "聞く"   "父"    "思う"
 [2,] "前"    "人"    "奥さん"  "知る"   "母"    "自分"
 [3,] "今"    "来る"  "お嬢さん" "顔"     "帰る"  "考える"
 [4,] "眼"    "手紙"  "答える"  "言葉"   "死ぬ"  "妻"
 [5,] "立つ"  "書く"  "室"      "見る"   "好い"  "心"
 [6,] "行く"  "聞く"  "声"      "口"     "叔父"  "人間"
 [7,] "一人"  "急"    "見る"    "二人"   "東京"  "意味"
 [8,] "宅"    "返事"  "坐る"    "話"     "病気"  "男"
 [9,] "手"    "外"    "取る"    "知れる" "出す"  "解る"
[10,] "頭"    "今"    "女"      "少し"   "卒業"  "態度"
```

6. seededlda での処理結果

Seededlda パッケージには、教師なし（textmodel_lda）と教師付き（textmodel_seededlda）LDA 処理が含まれている。設定可能なパラメータは Gibbs サンプリング数の他、α値、β値であるが、topicmodels::LDA での burnin に相当するパラメータは無い。乱数の初期値は、set.seed で設定している。

seededlda パッケージの textmodel_lda で処理

```
detach("package:topicmodels", unload = TRUE)
library(seededlda)  # seededlda パッケージをインストール
```

```
set.seed(1234567)
work_LDAs <- textmodel_lda( work_dfm, k = 6, max_iter = 2000,
                    alpha = NULL, beta = NULL, model = NULL,
                    verbose = quanteda_options("verbose") )
terms(work_LDAs)
```

```
      topic1      topic2 topic3 topic4 topic5      topic6
 [1,] "奥さん"    "父"    "先生" "思う" "K"         "出る"
 [2,] "お嬢さん" "母"    "人"    "自分" "自分"      "行く"
 [3,] "顔"        "手紙" "聞く" "妻"    "お嬢さん" "叔父"
 [4,] "見る"      "書く" "出る" "人"    "聞く"      "東京"
 [5,] "前"        "来る" "言葉" "死ぬ" "見る"      "考える"
 [6,] "少し"      "病気" "眼"    "人間" "室"        "帰る"
 [7,] "笑う"      "兄"    "来る" "意味" "眼"        "思う"
 [8,] "話す"      "卒業" "見る" "解る" "坐る"      "家"
 [9,] "口"        "出す" "思う" "心"    "付く"      "今"
[10,] "話"        "見る" "態度" "言葉" "思う"      "宅"
```

topicmodels パッケージの LDA 処理を実行。但し、burnin=1000 パラメータを設定していない。

```
detach("package:seededlda", unload = TRUE)
library(topicmodels)

set.seed(1234567)  # 乱数の初期値を設定
kokoro_LDA6df <- LDA(kokoro_dtm, k=6, method="Gibbs" )
terms(kokoro_LDA6df, 10)
```

```
      Topic 1  Topic 2 Topic 3    Topic 4 Topic 5 Topic 6
 [1,] "出る"   "先生"  "K"         "見る"  "父"    "自分"
 [2,] "今"     "人"    "奥さん"    "顔"    "母"    "思う"
 [3,] "帰る"   "来る"  "お嬢さん" "眼"    "好い"  "妻"
 [4,] "行く"   "書く"  "答える"   "言葉"  "死ぬ"  "知れる"
 [5,] "立つ"   "手紙"  "聞く"     "二人"  "東京"  "事"
 [6,] "考える" "返事"  "話"       "前"    "家"    "人間"
 [7,] "手"     "読む"  "室"       "聞く"  "叔父"  "意味"
 [8,] "一人"   "聞く"  "声"       "思う"  "兄"    "解る"
 [9,] "頭"     "出す"  "坐る"     "口"    "病気"  "男"
[10,] "宅"     "急"    "態度"     "悪い"  "知る"  "女"
```

Seededlda パッケージのサンプルデータ（英語版）で textmodel_seededld が動作することを確認後、dictionary データとして単語（日本語）を設定した場合の動作を確認した。

seededlda モデルで実行

```
dict <- dictionary( list( sensei=c("先生", "手紙", "東京"), friend=c("K", "お嬢さん"),
                          family=c("父", "母", "兄"), private=c("自分", "妻") ) )
set.seed(1234567)
work_sLDA <- textmodel_seededlda( work_dfm, dict, residual=TRUE,
```

```
 max_iter=2000, min_termfreq=10 )
terms(work_sLDA)
```

```
> terms(work_sLDA)
      sensei friend     family private  other
 [1,] "先生" "K"        "父"   "自分"   "奥さん"
 [2,] "手紙" "お嬢さん" "母"   "妻"     "前"
 [3,] "東京" "見る"     "兄"   "思う"   "言葉"
 [4,] "書く" "出る"     "帰る" "死ぬ"   "聞く"
 [5,] "人"   "思う"     "叔父" "人間"   "口"
 [6,] "出る" "室"       "病気" "心"     "顔"
 [7,] "来る" "聞く"     "卒業" "考える" "知れる"
 [8,] "返事" "声"       "思う" "今"     "話す"
 [9,] "聞く" "付く"     "来る" "人"     "今"
[10,] "好い" "歩く"     "見る" "男"     "少し"
```

```
dictj <- dictionary( list( "先生"=c("先生", "手紙", "東京"), "友人"=c("K", "お嬢さん"),
                           "家族"=c("父", "母", "兄"), "自分自身"=c("自分", "妻") ) )
work_sLDAj <- textmodel_seededlda( work_dfm, dictj, residual=TRUE,
                      max_iter=2000,min_termfreq=10 )
terms(work_sLDAj)
```

```
> terms(work_sLDAj)
      先生   友人       家族   自分自身 other
 [1,] "先生" "K"        "父"   "自分"   "奥さん"
 [2,] "手紙" "お嬢さん" "母"   "妻"     "聞く"
 [3,] "東京" "見る"     "兄"   "思う"   "言葉"
 [4,] "書く" "出る"     "帰る" "死ぬ"   "前"
 [5,] "人"   "思う"     "叔父" "人"     "口"
 [6,] "出る" "眼"       "見る" "人間"   "知る"
 [7,] "来る" "室"       "病気" "今"     "見る"
 [8,] "返事" "聞く"     "卒業" "考える" "話"
 [9,] "読む" "顔"       "思う" "男"     "話す"
[10,] "口"   "付く"     "聞く" "心"     "答える"
```

（注） seededlda でのカテゴリ数は、辞書（dictionary）として設定したカテゴリの数に「other」を加えた値となる。

6. Gibbs イテレーション数について

Gibbs サンプリングとは、マルコフ連鎖モンテカルロ法の最も簡単な場合で、潜在変数を分布ではなく、条件付き分布から実際にサンプリングする。原理的には、サンプリングを無限回繰り返せば、真の分布からのサンプルとなる。（参考資料：H24:Introduction to Statistical Topic Models (ism.ac.jp) ISM-2012-TopicModels.ppt）

そこで、topicmodels パッケージの LDA 処理で、乱数初期値を固定し、iter 値による処理結果を比較した。

```
kokoro4k_LDA6c <- LDA(kokoro_dtmc, k=6, method="Gibbs",
        control=list(seed=1234567, burnin=1000, iter=4000) ) # iter=4K
```

iter	トピックの単語群（上位 5 語）

2K	`> terms(kokoro_LDA6c, 10)`
	`      Topic 1  Topic 2    Topic 3  Topic 4 Topic 5  Topic 6`
	`[1,] "人"     "見る"     "思う"   "父"    "K"     "先生"`
	`[2,] "今"     "お嬢さん" "前"     "母"    "自分"  "奥さん"`
	`[3,] "思う"   "顔"       "考える" "帰る"  "聞く"  "聞く"`
	`[4,] "出る"   "女"       "眼"     "言葉"  "妻"    "解る"`
	`[5,] "口"     "話"       "見える" "死ぬ"  "答える" "問題"`

4K	`> terms(kokoro4k_LDA6c, 10)`
	`      Topic 1 Topic 2    Topic 3  Topic 4 Topic 5  Topic 6`
	`[1,] "人"     "見る"     "思う"   "父"    "K"     "先生"`
	`[2,] "今"     "お嬢さん" "自分"   "母"    "聞く"  "奥さん"`
	`[3,] "言葉"   "顔"       "考える" "帰る"  "出る"  "話"`
	`[4,] "知る"   "眼"       "妻"     "死ぬ"  "二人"  "前"`
	`[5,] "口"     "立つ"     "心"     "好い"  "答える" "返事"`

8K	`> terms(kokoro8k_LDA6c, 10)`
	`      Topic 1 Topic 2  Topic 3  Topic 4 Topic 5    Topic 6`
	`[1,] "出る"   "奥さん" "自分"   "父"    "K"       "先生"`
	`[2,] "今"     "見る"   "思う"   "母"    "お嬢さん" "人"`
	`[3,] "二人"   "前"     "考える" "言葉"  "聞く"    "来る"`
	`[4,] "行く"   "顔"     "妻"     "帰る"  "知る"    "見える"`
	`[5,] "好い"   "少し"   "心"     "手紙"  "答える"  "頭"`

15K	`> terms(kokoro15k_LDA6c, 10)`
	`      Topic 1  Topic 2  Topic 3  Topic 4 Topic 5    Topic 6`
	`[1,] "出る"   "奥さん" "思う"   "父"    "K"       "先生"`
	`[2,] "今"     "見る"   "自分"   "母"    "お嬢さん" "人"`
	`[3,] "前"     "顔"     "考える" "帰る"  "眼"      "来る"`
	`[4,] "行く"   "聞く"   "妻"     "死ぬ"  "答える"  "言葉"`
	`[5,] "立つ"   "知る"   "心"     "好い"  "聞く"    "書く"`

30K	`> terms(kokoro30k_LDA6c, 10)`
	`      Topic 1  Topic 2  Topic 3  Topic 4 Topic 5    Topic 6`
	`[1,] "今"     "奥さん" "思う"   "父"    "K"       "先生"`
	`[2,] "聞く"   "前"     "自分"   "母"    "お嬢さん" "来る"`
	`[3,] "出る"   "顔"     "人"     "考える" "見る"    "眼"`
	`[4,] "知る"   "二人"   "妻"     "帰る"  "見える"  "書く"`
	`[5,] "見る"   "言葉"   "心"     "好い"  "女"      "手紙"`

100K	`> terms(kokoro100k_LDA6c, 10)`
	`      Topic 1  Topic 2  Topic 3  Topic 4 Topic 5    Topic 6`
	`[1,] "見る"   "奥さん" "思う"   "父"    "K"       "先生"`
	`[2,] "前"     "二人"   "自分"   "母"    "聞く"    "人"`
	`[3,] "今"     "言葉"   "考える" "帰る"  "お嬢さん" "来る"`
	`[4,] "立つ"   "知れる" "妻"     "出る"  "答える"  "手紙"`
	`[5,] "一人"   "話"     "心"     "死ぬ"  "室"      "書く"`

500K

```
> terms(kokoro500k_LDA6c, 10)
     Topic 1  Topic 2  Topic 3  Topic 4 Topic 5    Topic 6
[1,] "出る"   "奥さん" "思う"   "父"    "K"        "先生"
[2,] "今"     "聞く"   "自分"   "母"    "お嬢さん" "人"
[3,] "前"     "言葉"   "考える" "帰る"  "見る"     "来る"
[4,] "行く"   "口"     "妻"     "死ぬ"  "眼"       "知る"
[5,] "見える" "答える" "心"     "好い"  "顔"       "書く"
```

1M

```
> terms(kokoro1m_LDA6c, 10)
     Topic 1  Topic 2  Topic 3 Topic 4 Topic 5    Topic 6
[1,] "見る"   "奥さん" "自分"  "父"    "K"        "先生"
[2,] "今"     "顔"     "思う"  "母"    "お嬢さん" "来る"
[3,] "考える" "言葉"   "妻"    "帰る"  "二人"     "人"
[4,] "出る"   "知る"   "心"    "死ぬ"  "知れる"   "書く"
[5,] "前"     "聞く"   "人間"  "好い"  "答える"   "手紙"
```

10M

```
> terms(kokoro10m_LDA6c, 10)
     Topic 1  Topic 2  Topic 3  Topic 4 Topic 5    Topic 6
[1,] "自分"   "奥さん" "出る"   "父"    "K"        "先生"
[2,] "思う"   "見る"   "前"     "母"    "聞く"     "人"
[3,] "考える" "顔"     "今"     "帰る"  "お嬢さん" "来る"
[4,] "妻"     "二人"   "行く"   "言葉"  "眼"       "見える"
[5,] "心"     "口"     "知れる" "死ぬ"  "答える"   "書く"
```

100M

```
> terms(kokoro100m_LDA6c, 10)
     Topic 1  Topic 2  Topic 3  Topic 4 Topic 5    Topic 6
[1,] "出る"   "奥さん" "思う"   "父"    "K"        "先生"
[2,] "前"     "見る"   "自分"   "母"    "お嬢さん" "人"
[3,] "二人"   "顔"     "考える" "帰る"  "眼"       "来る"
[4,] "今"     "聞く"   "心"     "好い"  "聞く"     "書く"
[5,] "女"     "少し"   "妻"     "死ぬ"  "知る"     "手紙"
```

200M

```
> terms(kokoro200m_LDA6c, 5)
     Topic 1  Topic 2  Topic 3  Topic 4 Topic 5    Topic 6
[1,] "奥さん" "出る"   "思う"   "父"    "K"        "先生"
[2,] "見る"   "今"     "自分"   "母"    "聞く"     "人"
[3,] "顔"     "立つ"   "考える" "帰る"  "お嬢さん" "来る"
[4,] "前"     "知れる" "妻"     "好い"  "眼"       "知る"
[5,] "少し"   "前"     "心"     "東京"  "二人"     "口"
```

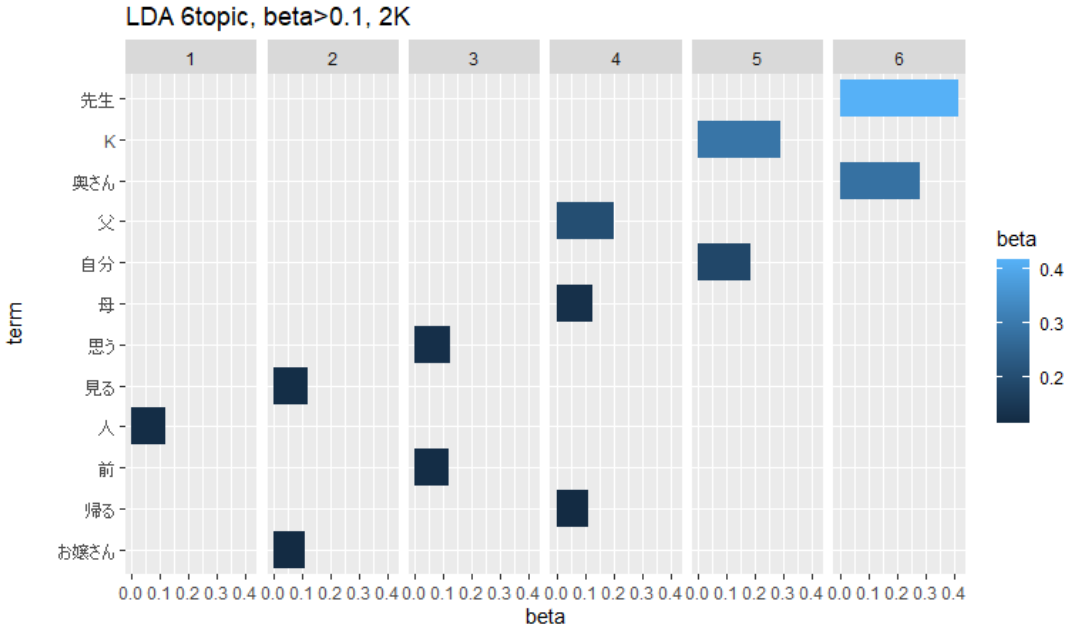
降順β値が 0.1 以上の単語をトピック別に描画した結果をみると、イテレーション 100K 回以降、β値最大の単語として「先生」、「K」、「奥さん」、「父」が 4 トピックに出現する。「自分」と「思う」が 1 トピックに出現し、残り 1 トピックでは出現する単語が固定化されていない。この結果からイテレーション 2K 回（KH Coder での設定値）では、分類されたトピックの単語が安定化する前のように見える。なお、出現頻度行列（データフレーム形式）について処理した場合も、イテレーション 100K 回以降はトピックに出現する単語が固定化されていることが判る。

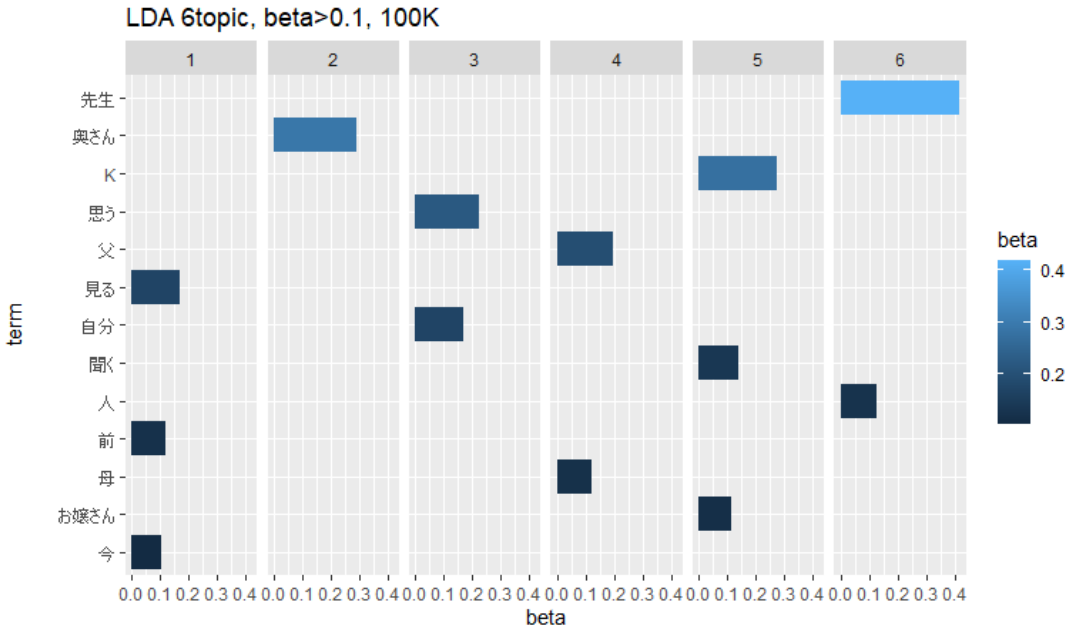
参考資料：H24:Introduction to Statistical Topic Models (ism.ac.jp) ISM-2012-TopicModels.ppt のスライド 111～123 に掲載されている「トピック分布βの学習過程」では、単語数 25 に埋め込まれた 10 トピックを推定する場合で、Gibbs イテレーション 200 回程度で収束することが示されている。これから、イテレーションとしては単語数 x トピック数以上とすれば良いようだが、今回の結果では、分析対象語数 71、トピック数 6 で安定化したトピックが得られたイテレーションは 100 K 回であ

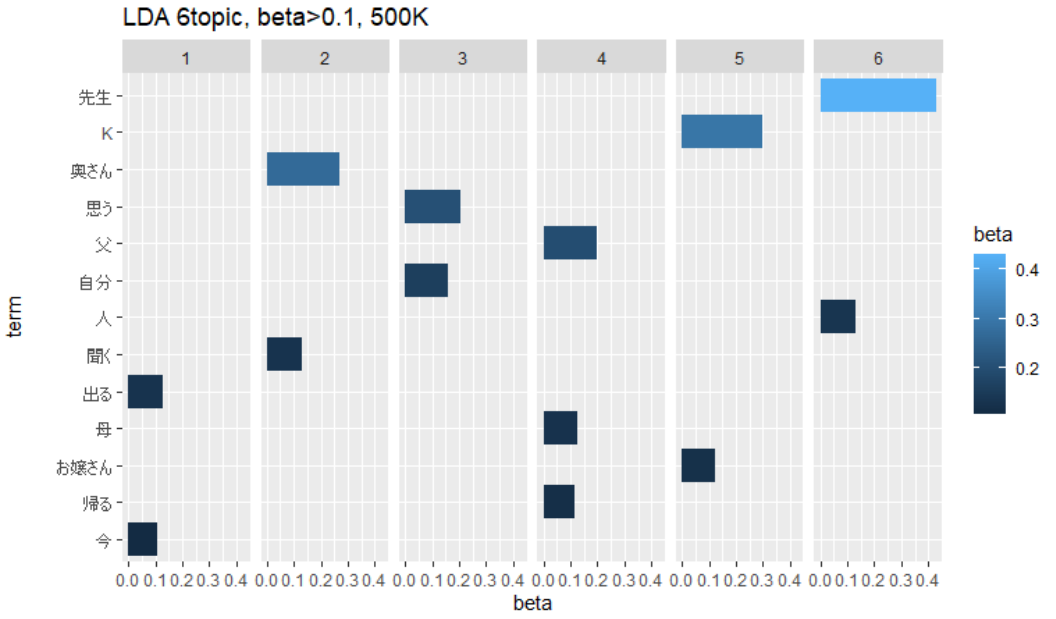
り、2000 回（71x6 より大きい値）よりも遥かに大きい。よって、KH Coder での設定値（2000）では、Gibbs イテレーション数として不十分な可能性がある。

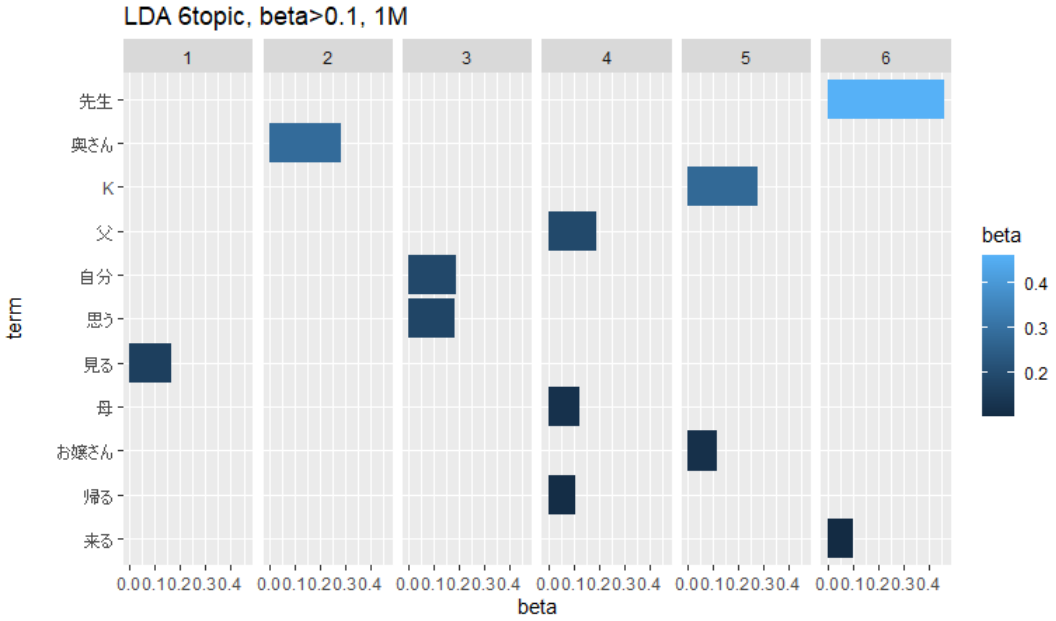
また、seededlda パッケージに準備された textmodel_lda についても、イテレーション数による処理結果の比較も行った。この場合は、2K でも「先生」、「K」、「奥さん」、「父、母」、「自分、妻」が安定的に 5 トピックに出現し、残り 1 トピックでの単語は安定していない。

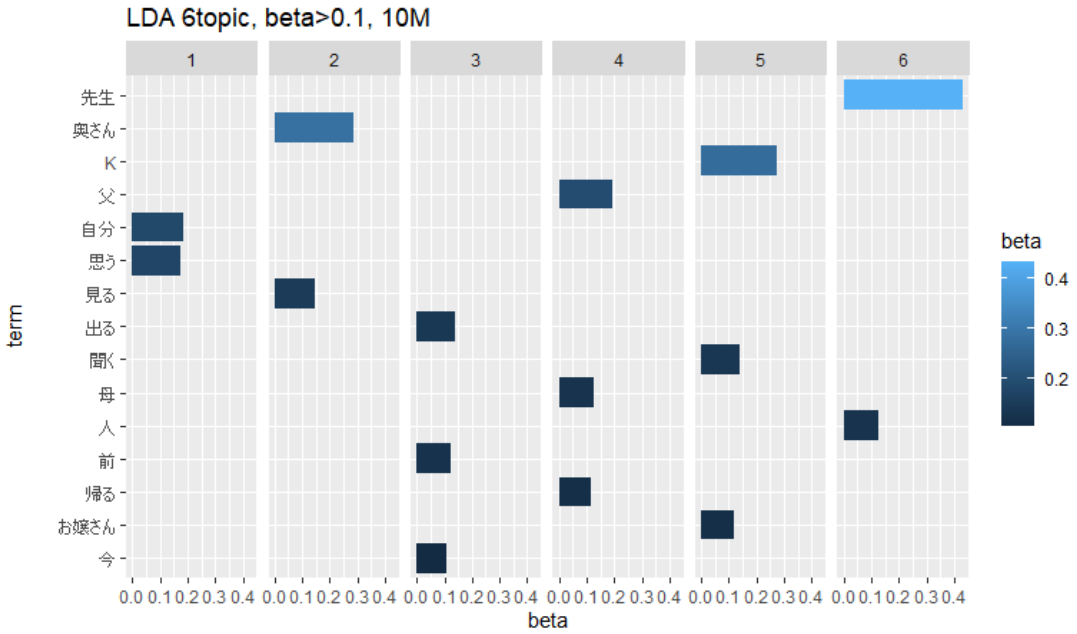
DTM 形式データの LDA 処理結果での降順β値 0.1 以上の単語をトピック別に描画

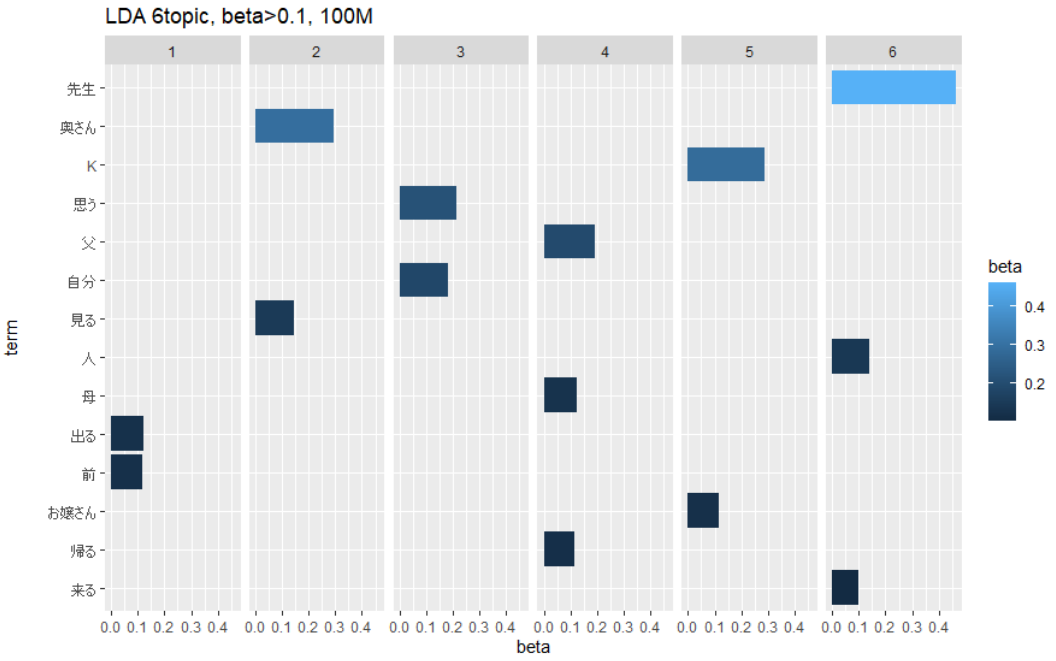


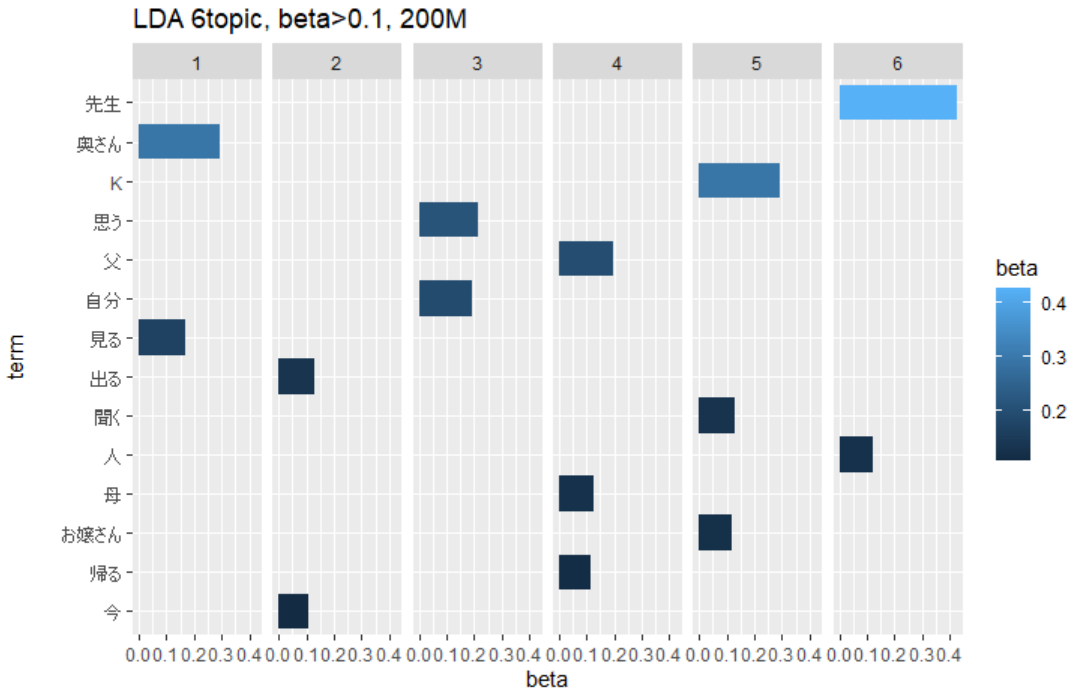



LDA 6topic, beta>0.1, 500K
1
2
3
4
5
6
先生
K
奥さん
思う
父
自分
人
聞く
出る
母
お嬢さん
帰る
今
term
0.0 0.1 0.2 0.3 0.4
beta
beta
0.4
0.3
0.2

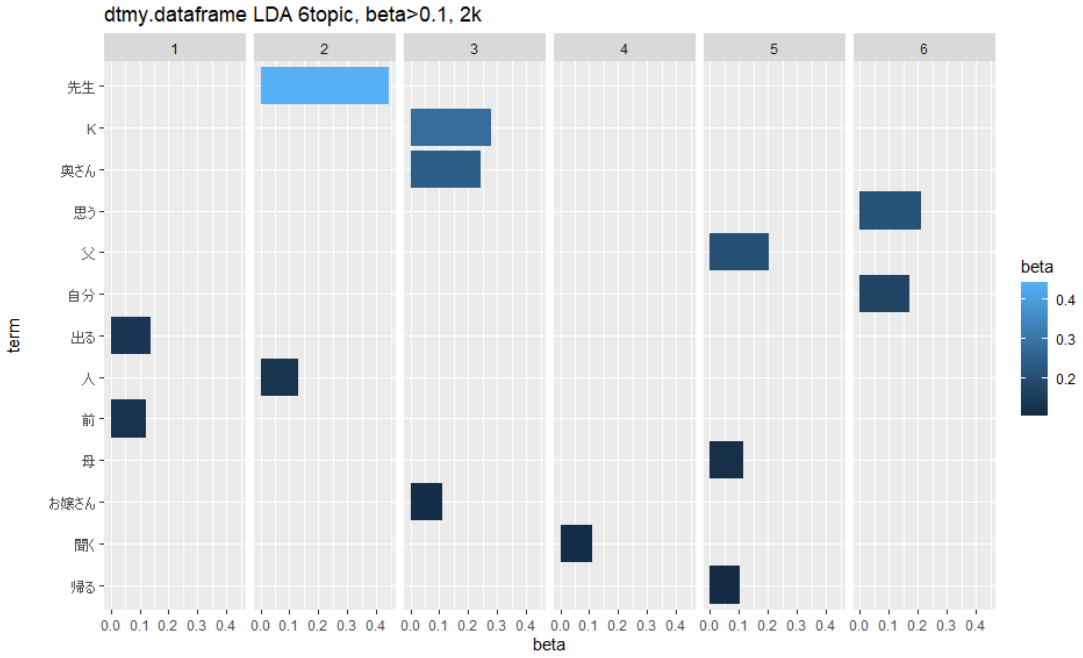

LDA 6topic, beta>0.1, 1M
1
2
3
4
5
6
先生
奥さん
K
父
自分
思う
見る
母
お嬢さん
帰る
来る
term
0.00.10.20.30.4
beta
beta
0.4
0.3
0.2

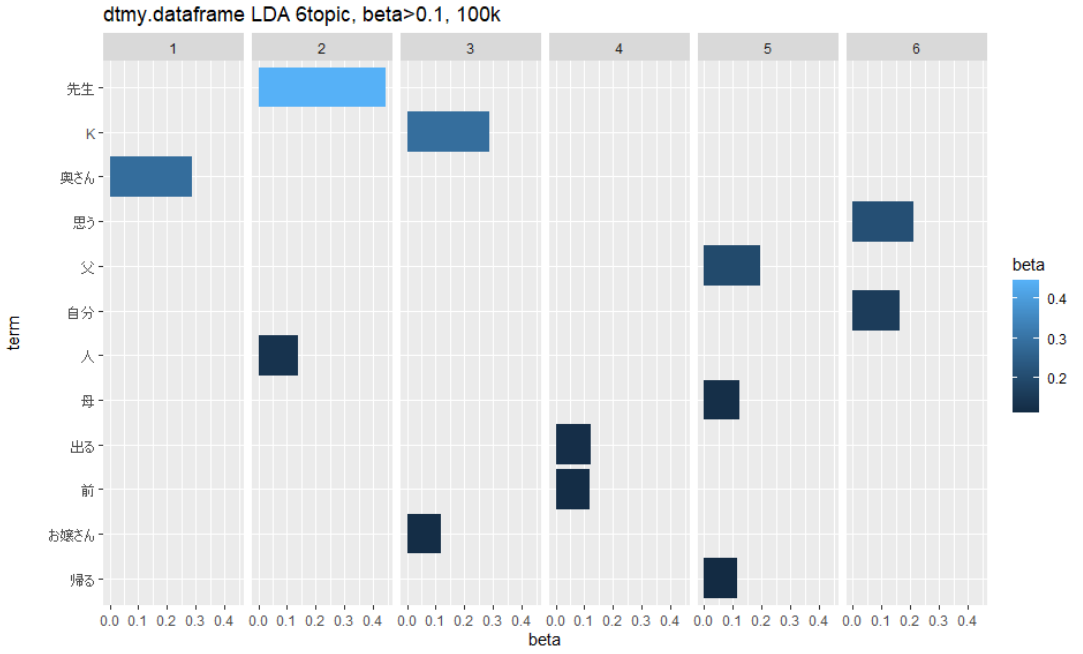

LDA 6topic, beta>0.1, 10M
1
2
3
4
5
6
先生
奥さん
K
父
自分
思う
見る
出る
聞く
母
人
前
帰る
お嬢さん
今
term
0.0 0.1 0.2 0.3 0.4
beta
beta
0.4
0.3
0.2

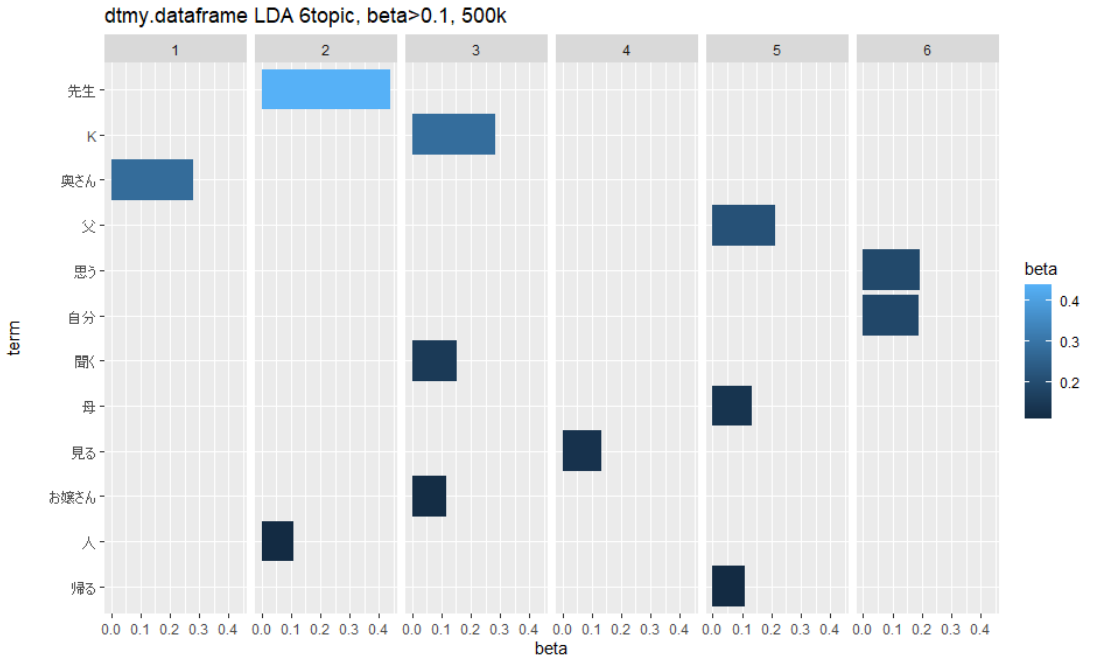

LDA 6topic, beta>0.1, 100M
1
2
3
4
5
6
先生
奥さん
K
思う
父
自分
見る
人
母
出る
前
お嬢さん
帰る
来る
term
0.0 0.1 0.2 0.3 0.4
beta
beta
0.4
0.3
0.2

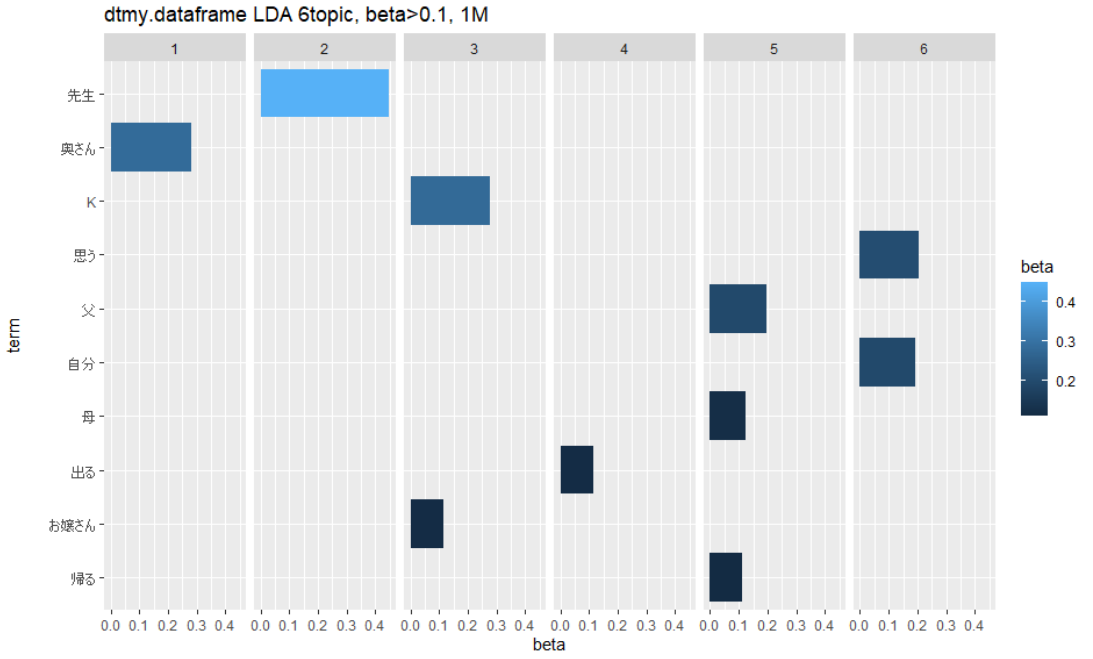

LDA 6topic, beta>0.1, 200M
1
2
3
4
5
6
先生
奥さん
K
思う
父
自分
見る
出る
聞く
人
母
お嬢さん
帰る
今
term
0.00.10.20.30.4
beta
beta
0.4
0.3
0.2

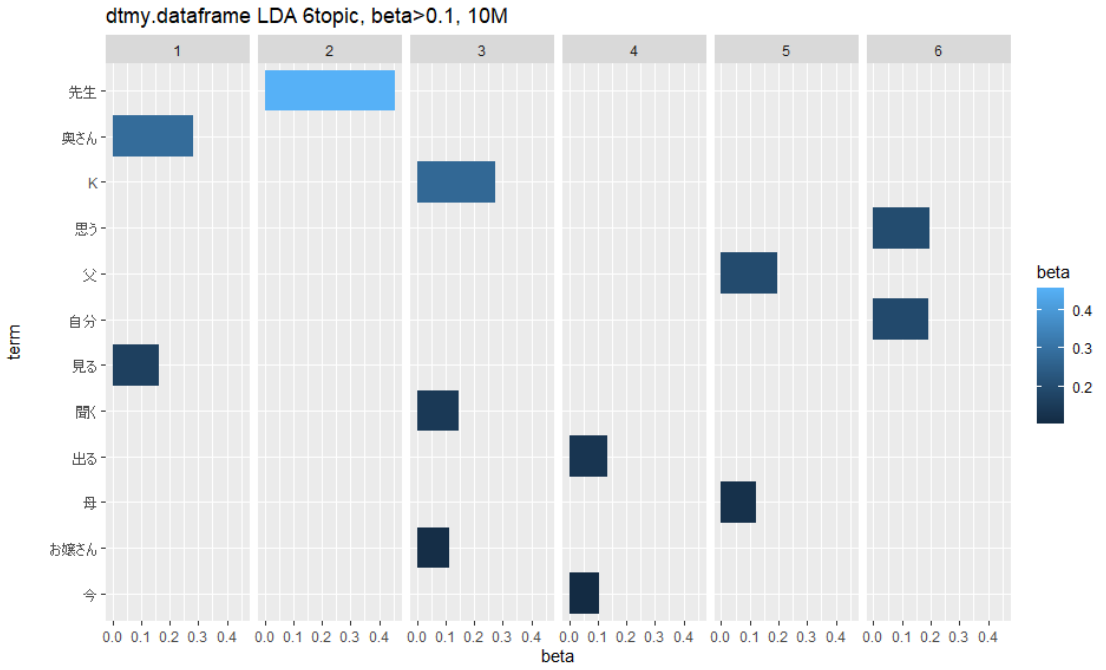
出現頻度行列（データフレーム形式）LDA 処理結果での降順β値 0.1 以上の単語をトピック別に描画





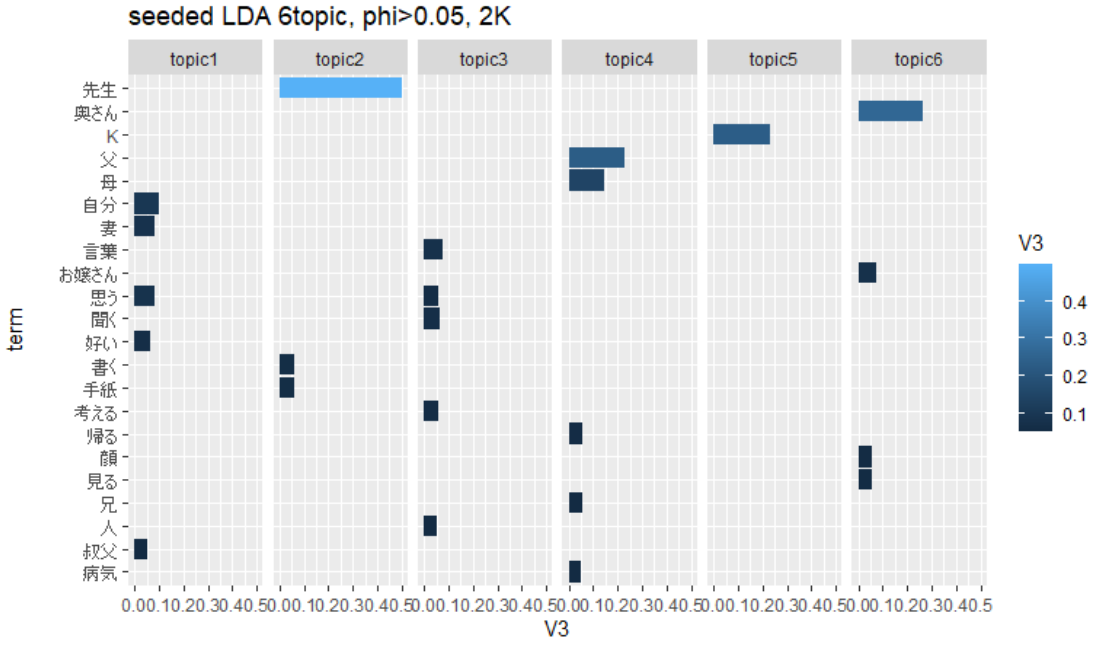

dtmy.dataframe LDA 6topic, beta>0.1, 500k
1
2
3
4
5
6
先生
K
奥さん
父
思う
自分
聞く
母
見る
お嬢さん
人
帰る
term
0.0 0.1 0.2 0.3 0.4
beta
beta
0.4
0.3
0.2

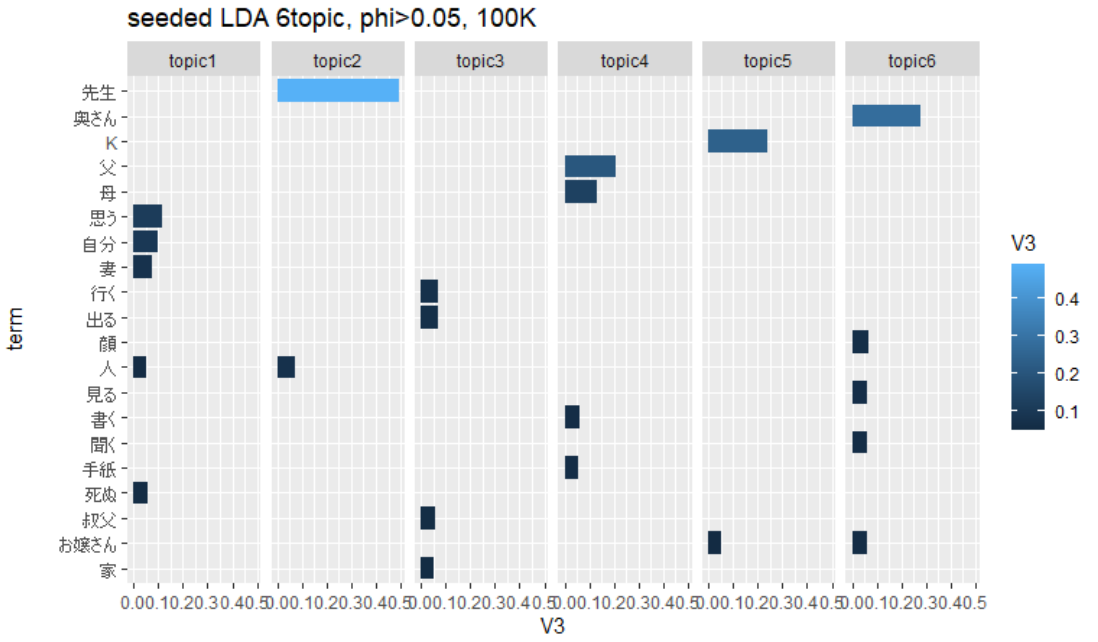

dtmy.dataframe LDA 6topic, beta>0.1, 1M
1
2
3
4
5
6
先生
奥さん
K
思う
父
自分
母
出る
お嬢さん
帰る
term
0.0 0.1 0.2 0.3 0.4
beta
beta
0.4
0.3
0.2

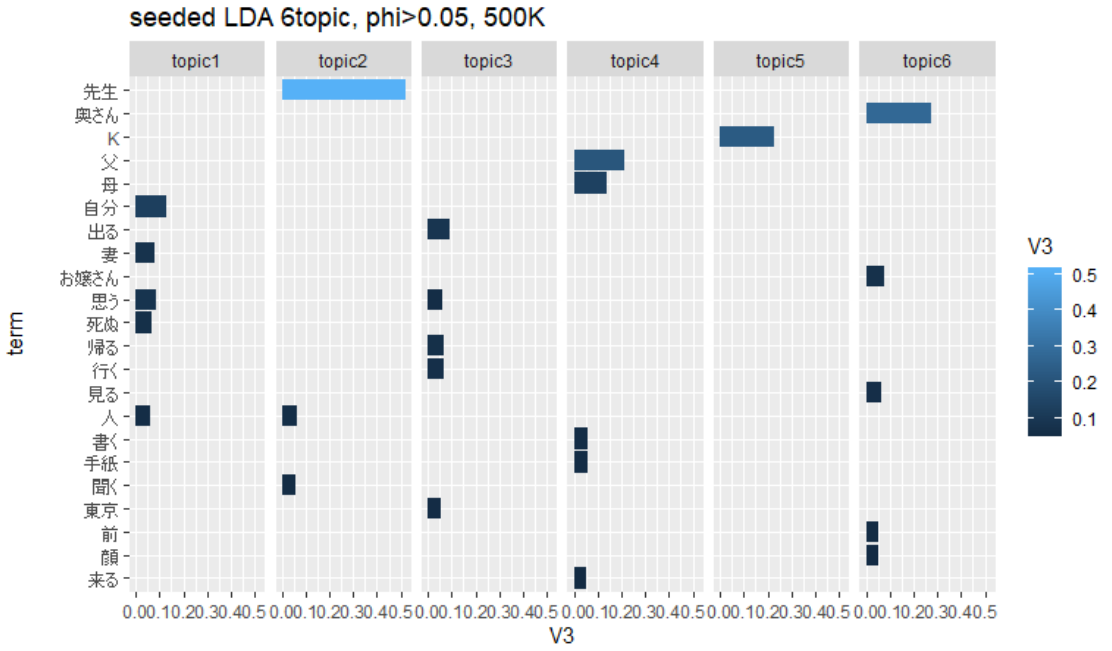


textmodels::LDA で出現頻度行列について 100k ステップ処理し、結果を可視化
dtmy_100 k _LDA6 <- LDA(dtmy, k=6, method="Gibbs", control=list(seed=1234567, burnin=1000, iter=100000)) terms(dtmy_100 k _LDA6, 5)　# Top 5 の単語リスト（ベータ値降順） # ベータ値が 0.1 以上の単語について可視化 dtmy_10m_LDA6 %>% tidy() %>% group_by(topic) %>% top_n(5, beta) %>% ungroup() %>% filter(beta > 0.1) %>% mutate(term=reorder(term, beta)) %>% arrange(topic, -beta) %>% ggplot(aes(term, beta, fill = beta)) + geom_bar(stat = "identity") + facet_wrap(~topic, ncol=6) + coord_flip() + ggtitle("dtmy.dataframe LDA 6topic, beta>0.1, 100K")

DTM 形式データの seededlda（LDA）処理結果での降順ファイ値 0.05 以上の単語をトピック別に描画







seededlda::textmodel_LDA での処理と結果の可視化

```
> Convert DTM to dataframe
work_df   <- tidy(kokoro_dtm)
work_dfc <- tidy(kokoro_dtmc)

> str(work_df)
tibble [6,106 x 3] (S3: tbl_df/tbl/data.frame)
 $ document: chr [1:6106] "1" "2" "6" "7" ...
 $ term      : chr [1:6106] "先生" "先生" "先生" "先生" ...
 $ count     : num [1:6106] 3 1 1 5 1 1 7 3 4 4 ...
> str(work_dfc)
tibble [6,106 x 3] (S3: tbl_df/tbl/data.frame)
 $ document: chr [1:6106] "1010" "1011" "1012" "1013" ...
 $ term      : chr [1:6106] "K" "K" "K" "K" ...
 $ count     : num [1:6106] 3 3 3 2 4 4 2 1 3 3 ...

# tidytext の cast_dfm でデータフレームを変換
work_dfm   <- work_df   %>%
  cast_dfm(document=document, term=term, value=count)
work_dfmc <- work_dfc %>%
  cast_dfm(document=document, term=term, value=count)
```

```
set.seed(1234567)   #  乱数の初期値を指定
work_LDAs <- textmodel_lda( work_dfm, k = 6, max_iter = 2000,
                      alpha = NULL, beta = NULL, model = NULL,
                      verbose = quanteda_options("verbose") )
> terms(work_LDAs)
       topic1 topic2    topic3    topic4 topic5       topic6
 [1,] "自分" "先生"    "言葉"    "父"    "K"           "奥さん"
 [2,] "思う" "書く"    "聞く"    "母"    "眼"          "お嬢さん"
 [3,] "妻"    "手紙"    "思う"    "帰る" "帰る"       "顔"
 [4,] "好い" "思う"    "考える" "兄"    "室"          "見る"
 [5,] "叔父" "来る"    "人"       "病気" "出る"       "前"
 [6,] "人間" "出る"    "知れる" "卒業" "お嬢さん" "二人"
 [7,] "出る" "返事"    "事"       "東京" "見る"       "聞く"
 [8,] "人"    "人"       "意味"    "知る" "自分"       "女"
 [9,] "死ぬ" "見える" "口"       "見る" "声"          "話"
[10,] "心"    "読む"    "自分"    "来る" "聞く"       "立つ"

> library(tidytext)
> library(tidyverse)
work_LDAs_phi <- work_LDAs$phi

#   Phi matrix を結果から抽出
work_LDAs_phi_mat <- as.matrix(work_LDAs$phi)

#   単語リストを verb に、トピック番号を docs に格納
verb <- colnames(work_LDAs_phi_mat)
docs <- rownames(work_LDAs_phi_mat)

# one-term-per-row data frame
out.file <- "phi_work.csv"
# 既存ファイルを消す
if (file.exists(out.file))
  file.remove(out.file)

#   行列の要素データを csv に書き出し
```

```
for ( i in 1:nrow(work_LDAs_phi_mat) ) {
  for ( j in 1:ncol(work_LDAs_phi_mat) ) {

    if( work_LDAs_phi_mat[i,j] > 0 ) {
      buff.txt   <- paste(docs[i], verb[j], work_LDAs_phi_mat[i,j], sep=",")
      write( buff.txt, file=out.file, append=T )
    }
  }
}

# 読み込む際、文字データを character にする。（データフレーム形式）
phi_df <- read.csv(out.file, header=F,   stringsAsFactors=F)

> str(phi_df)
> 'data.frame':  426 obs. of   3 variables:
>   $ V1: chr   "topic1" "topic1" "topic1" "topic1" ...
>   $ V2: chr   "先生" "心持" "書く" "人" ...
>   $ V3: num   7.65e-05 2.91e-02 7.65e-05 4.83e-02 7.65e-05 ...

phi_df %>% group_by(V1) %>% top_n(5, V3) %>% ungroup() %>%
               filter( V3 > 0.05 ) %>%
               mutate( term=reorder(V2, V3)) %>% arrange(V1, -V3) %>%
               ggplot(aes(term, V3, fill = V3)) + geom_bar(stat = "identity") +
               facet_wrap(~V1, ncol=6) + coord_flip() +
               ggtitle("seeded LDA 6topic, phi>0.05, 2K")
```

付録 10　環境･水産･海洋白書の LDA 分析での Gibbs サンプリング数について

文書番号：JRDN-21-029

（注）パワーポイントドキュメントの提出資料を編集。

LDA Gibbsサンプリング数の妥当性

- KHCoderの前処理結果をtopicmodelsのLDAでGibbsサンプリング数を設定して処理
- 分析対象語数：KHCoder自動設定値と推奨最大値近傍
- 各トピックの出現確率上位4単語の可視化結果を比較。

	語数	トピック数	Gibbsサンプリング数					
			2K	10K	50K	100K	500K	1M
KHCoder自動設定値	75	16	図1	図2	図3	図4		
推奨最大値近傍	154	16	図5	図6	図7	図8	図9	図10

- 75語16トピックスの場合（KH Coder自動設定）
  - 上位の単語は、2Kステップで10Kステップ以降と概ね同じ。
- 154語16トピックスの場合
  - 10Kステップ以降では「養殖」が上位に出現するが、2Kステップには無い。
  - 50Kステップと100Kステップの結果は概ね同じ単語が上位に出現。

■ 抽出語リスト（出現頻度）

#	抽出語	品詞/活用	頻度
1	実施	サ変名詞	9571
2	利用	サ変名詞	7394
3	管理	サ変名詞	6749
4	資源	名詞	6495
5	対策	サ変名詞	6447
6	計画	サ変名詞	6220
7	調査	サ変名詞	5533
8	保全	サ変名詞	5528
9	処理	サ変名詞	5408
10	開発	サ変名詞	5260
11	技術	名詞	5097
12	廃棄物	タグ	5027
13	社会	名詞	4957
14	情報	名詞	4910
15	活動	サ変名詞	4901
16	必要	形容動詞	4637
17	関係	サ変名詞	4449
18	支援	サ変名詞	4355
19	施設	サ変名詞	4233
20	影響	サ変名詞	4185
21	整備	サ変名詞	4097
22	状況	名詞	3878
23	世界	名詞	3826
24	開催	サ変名詞	3700
25	日本	地名	3681
26	生物多様性	タグ	3559
27	評価	サ変名詞	3557
28	研究	サ変名詞	3553
29	基本	名詞	3509
30	促進	サ変名詞	3425
31	活用	サ変名詞	3411
32	地球	名詞	3398
33	目標	名詞	3235
34	検討	サ変名詞	3206
35	重要	形容動詞	3195
36	経済	名詞	3170
37	連携	サ変名詞	3168
38	自然	形容動詞	3136
39	制度	名詞	3121
40	国際	名詞	3086
41	エネルギー	名詞	3075
42	発生	サ変名詞	3065
43	増加	サ変名詞	3019
44	問題	ナイ形容	2959
45	減少	サ変名詞	2955
46	協力	サ変名詞	2950
47	法律	名詞	2927
48	対象	名詞	2899
49	規制	サ変名詞	2849
50	生産	サ変名詞	2828
51	産業	名詞	2821
52	策定	サ変名詞	2663
53	結果	副詞可能	2647
54	海域	名詞	2600
55	削減	サ変名詞	2600
56	政策	名詞	2592
57	基準	名詞	2570
58	機関	名詞	2554
59	温暖化	タグ	2474
60	排出	サ変名詞	2473
61	措置	サ変名詞	2454
62	持続可能	タグ	2400
63	課題	名詞	2383
64	教育	サ変名詞	2370
65	確保	サ変名詞	2348
66	使用	サ変名詞	2335
67	排出量	タグ	2325
68	システム	名詞	2316
69	高い	形容詞	2310
70	導入	サ変名詞	2303
66	使用	サ変名詞	2335
67	排出量	タグ	2325
68	システム	名詞	2316
69	高い	形容詞	2310
70	導入	サ変名詞	2303
71	強化	サ変名詞	2300
72	生態系	タグ	2296
73	事業者	タグ	2288
74	防止	サ変名詞	2255
75	気候変動	タグ	2228
76	水産物	名詞	2177
77	物質	名詞	2175
78	リサイクル	サ変名詞	2174
79	提供	サ変名詞	2166
80	分野	名詞	2150
81	発電	サ変名詞	2144
82	循環	サ変名詞	2123
83	構築	サ変名詞	2105
84	再生	サ変名詞	2086
85	対応	サ変名詞	2084
86	森林	名詞	2079
87	政府	名詞	2065
88	条約	名詞	2059
89	会議	サ変名詞	2056
90	被害	名詞	2049
91	観測	サ変名詞	2047
92	企業	名詞	2040
93	汚染	サ変名詞	2039
94	設置	サ変名詞	2033
95	委員会	タグ	2022
96	達成	サ変名詞	2006
97	保護	サ変名詞	1995
98	水産	名詞	1984
99	目的	名詞	1979
100	消費	サ変名詞	1976
101	向上	サ変名詞	1966
102	漁船	名詞	1956
103	生物	名詞	1939
104	温室効果ガス	タグ	1894
105	国内	名詞	1894
106	普及	サ変名詞	1887
107	実現	サ変名詞	1817
108	漁獲	サ変名詞	1801
109	総合的	タグ	1793
110	参加	サ変名詞	1787
111	都市	名詞	1756
112	拡大	サ変名詞	1755
113	中国	地名	1753
114	指定	サ変名詞	1746
115	生活	サ変名詞	1717
116	可能	形容動詞	1665
117	国際的	タグ	1664
118	中心	名詞	1650
119	回収	サ変名詞	1642
120	自動車	名詞	1635
121	漁業者	タグ	1629
122	変化	サ変名詞	1623
123	行動	サ変名詞	1622
124	多く	副詞可能	1605
125	関連	サ変名詞	1604
126	地方公共団体	タグ	1601
127	大きい	形容詞	1588
128	適切	形容動詞	1582
129	戦略	名詞	1575
130	場合	副詞可能	1567
131	都道府県	名詞	1529
132	形成	サ変名詞	1527
133	多い	形容詞	1510
134	機能	サ変名詞	1502
135	環境省	組織名	1497
136	原子力	名詞	1485
137	会合	サ変名詞	1480
138	適正	形容動詞	1463
139	プロジェクト	名詞	1444
140	化学物質	タグ	1444
141	議論	サ変名詞	1431
142	設定	サ変名詞	1430
143	総合	サ変名詞	1430
144	養殖	サ変名詞	1428
145	開始	サ変名詞	1422
146	期待	サ変名詞	1404
147	健康	形容動詞	1404
148	可能性	タグ	1391
149	把握	サ変名詞	1369
150	効果	名詞	1366
151	貢献	サ変名詞	1366
152	製品	名詞	1363
153	モニタリング	名詞	1347
154	具体的	タグ	1345
155	団体	名詞	1339
156	安全	形容動詞	1338
157	特定	サ変名詞	1338
158	収集	サ変名詞	1331
159	経営	サ変名詞	1320
160	国民	名詞	1312
161	規模	名詞	1307
162	福島	地名	1304
163	作成	サ変名詞	1288
164	人	名詞C	1284
165	昭和	固有名詞	1277
166	体制	名詞	1274
167	積極的	タグ	1272
168	改善	サ変名詞	1269
169	データ	名詞	1267
170	改正	サ変名詞	1261
166	体制	名詞	1274
167	積極的	タグ	1272
168	改善	サ変名詞	1269
169	データ	名詞	1267
170	改正	サ変名詞	1261
171	配慮	サ変名詞	1261
172	循環型	タグ	1239
173	アジア	地名	1233
174	市町村	名詞	1227
175	沿岸域	タグ	1224
176	作業	サ変名詞	1224
177	維持	サ変名詞	1221
178	製造	サ変名詞	1217
179	操業	サ変名詞	1212
180	割合	名詞	1210
181	食品	名詞	1208
182	採択	サ変名詞	1200
183	消費者	タグ	1194
184	認定	サ変名詞	1192
185	供給	サ変名詞	1190
186	大臣	名詞	1186
187	成長	サ変名詞	1182
188	太平洋	地名	1181
189	国連	組織名	1178
190	資源管理	タグ	1177
191	それぞれ	副詞可能	1176
192	施行	サ変名詞	1165
193	確認	サ変名詞	1163
194	船舶	名詞	1163
195	共同	サ変名詞	1162
196	CO2	未知語	1160
197	一般	名詞	1156
198	公表	サ変名詞	1145
199	方針	名詞	1141
200	多様	形容動詞	1139

トピック特徴語の比較結果

（75語：100K回、154語：1000K回）

75語にあって154語には無い単語		154語にあって75語には無い単語	
減少	2955	養殖	1428
制度	3121	被害	2049
増加	3019	原子力	1485
社会	4975	水産物	2177
導入	2303	漁獲	1801
法律	2927	漁船	1956
地球	3398	観測	2047
重要	3195	漁業者	1629
活動	4901	中国	1753
整備	4097	発電	2144
		活用	3411
		水産	1984
		排出	2473
		連携	3168

数字は出現頻度。なお、75語の出現頻度は2228、154語では1345。
太字は75語の出現頻度より大きな場合。

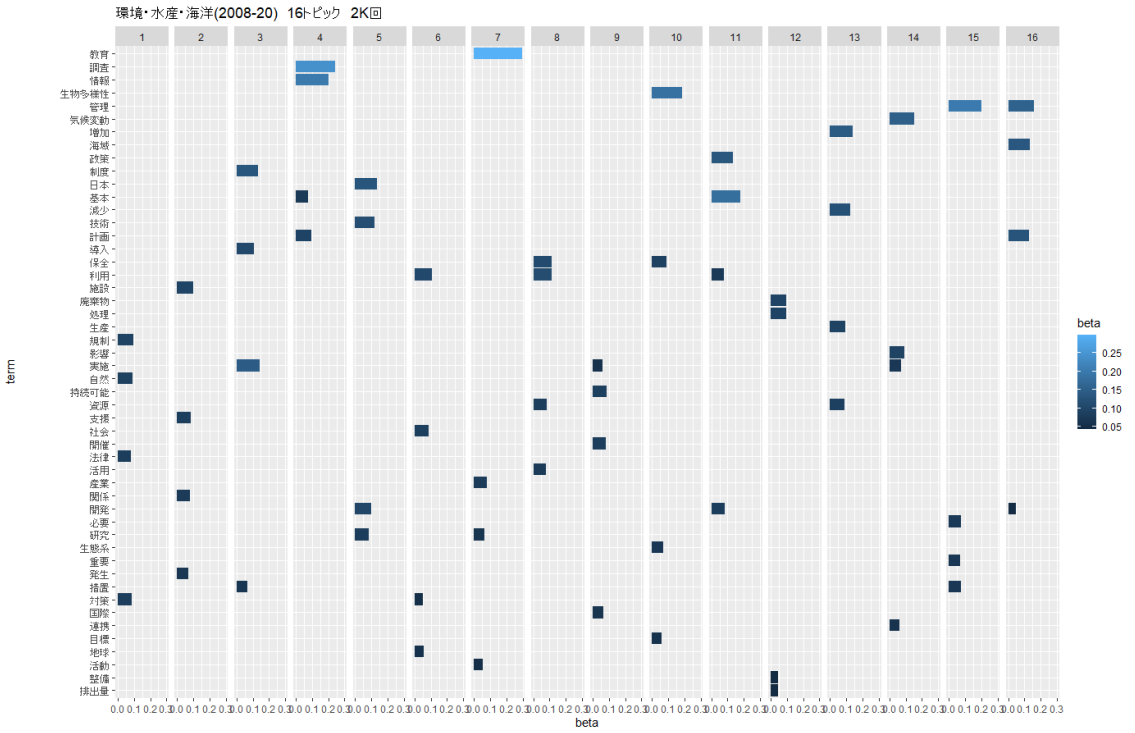


図１　75 語 2K 回

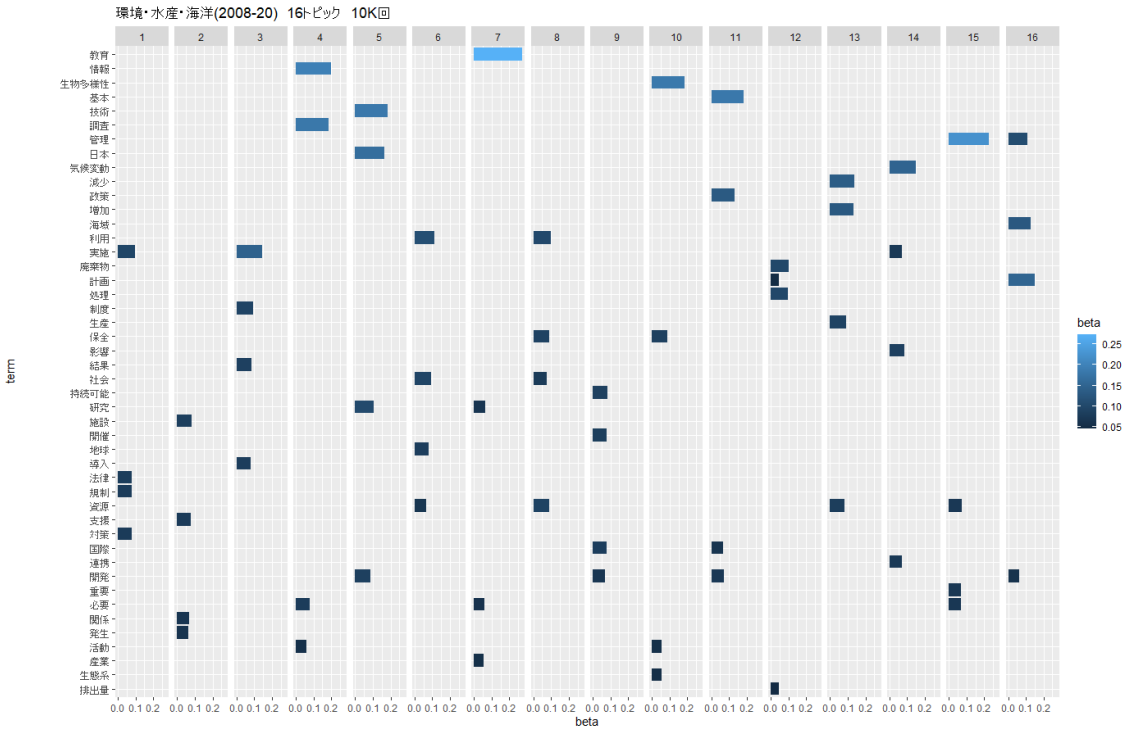


図２　75 語 10K 回

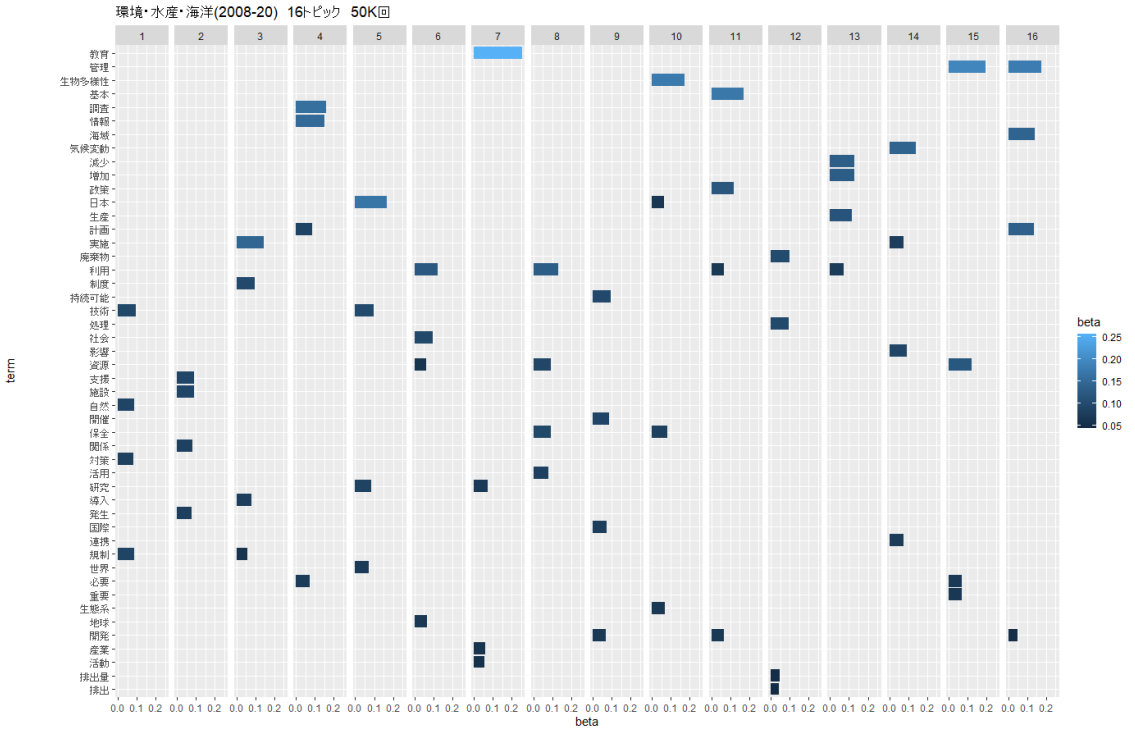


図３　75 語 50K 回

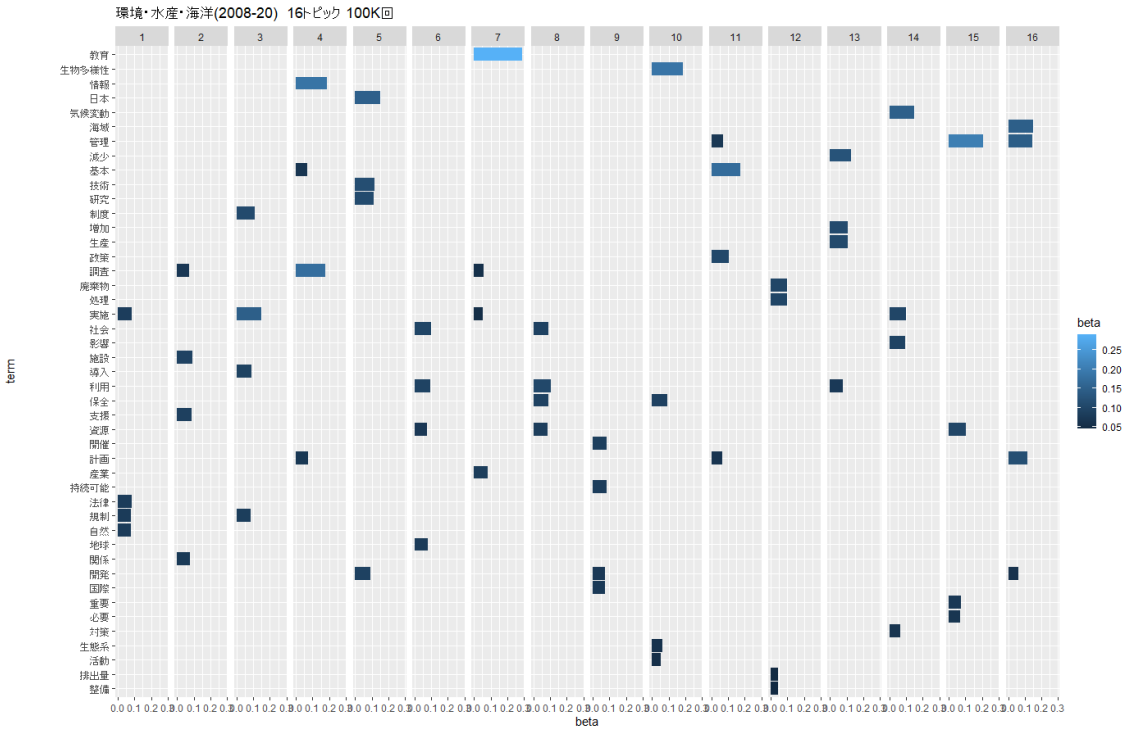


図４　75 語　100K 回

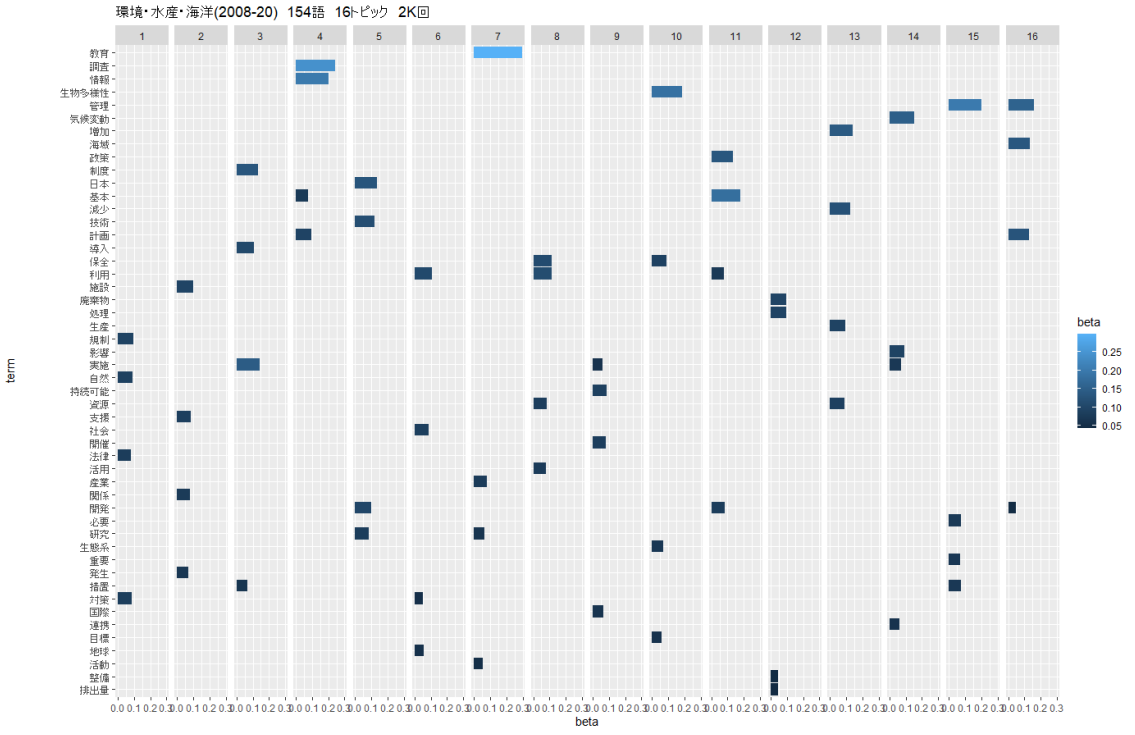


図５　154 語　2K 回

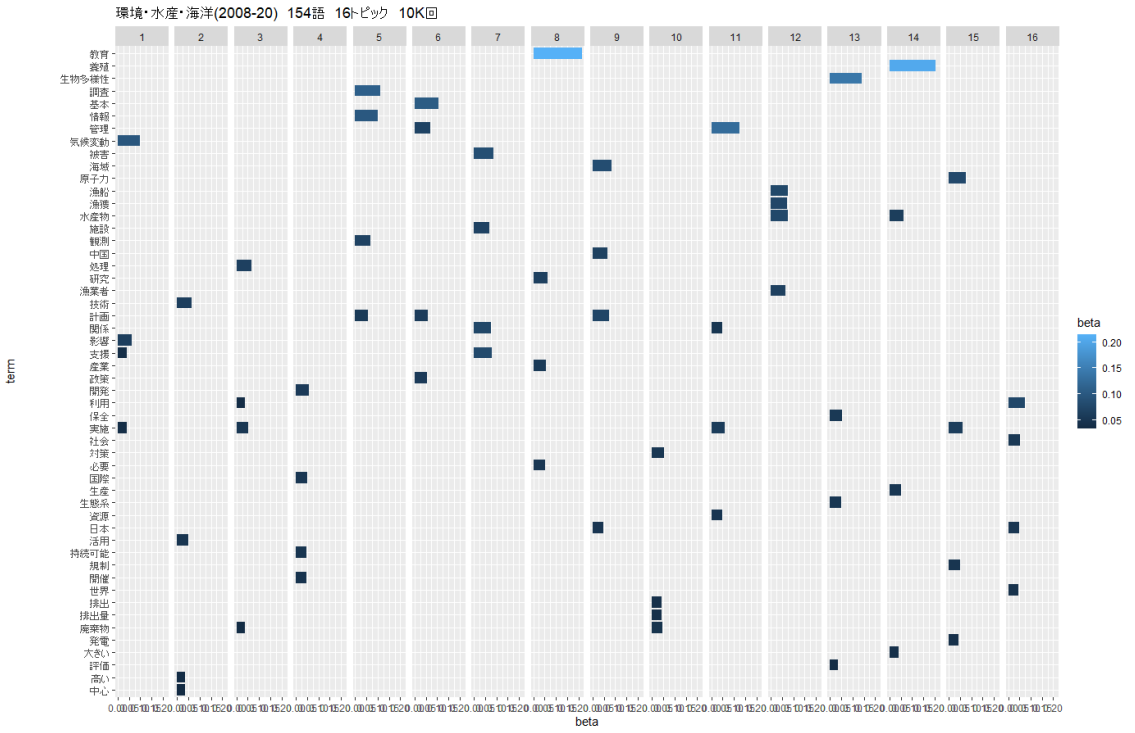


図 6　154 語　10K 回

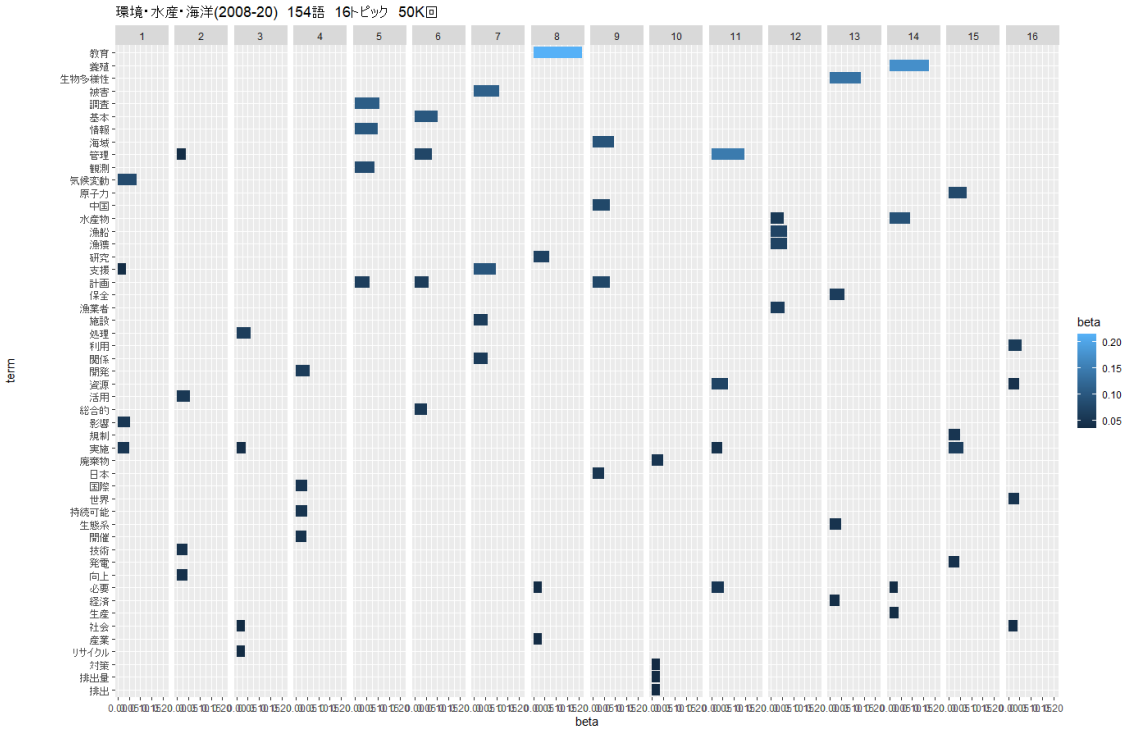


図 7　154 語　50K 回

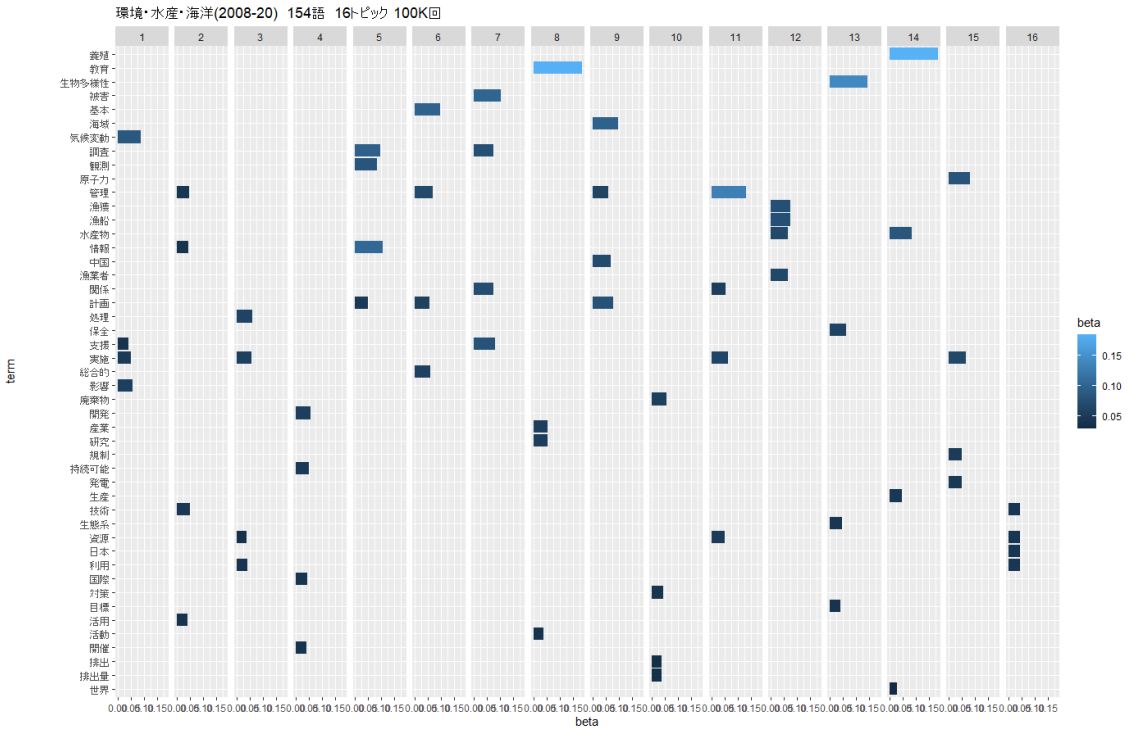


図８　154 語　100K 回

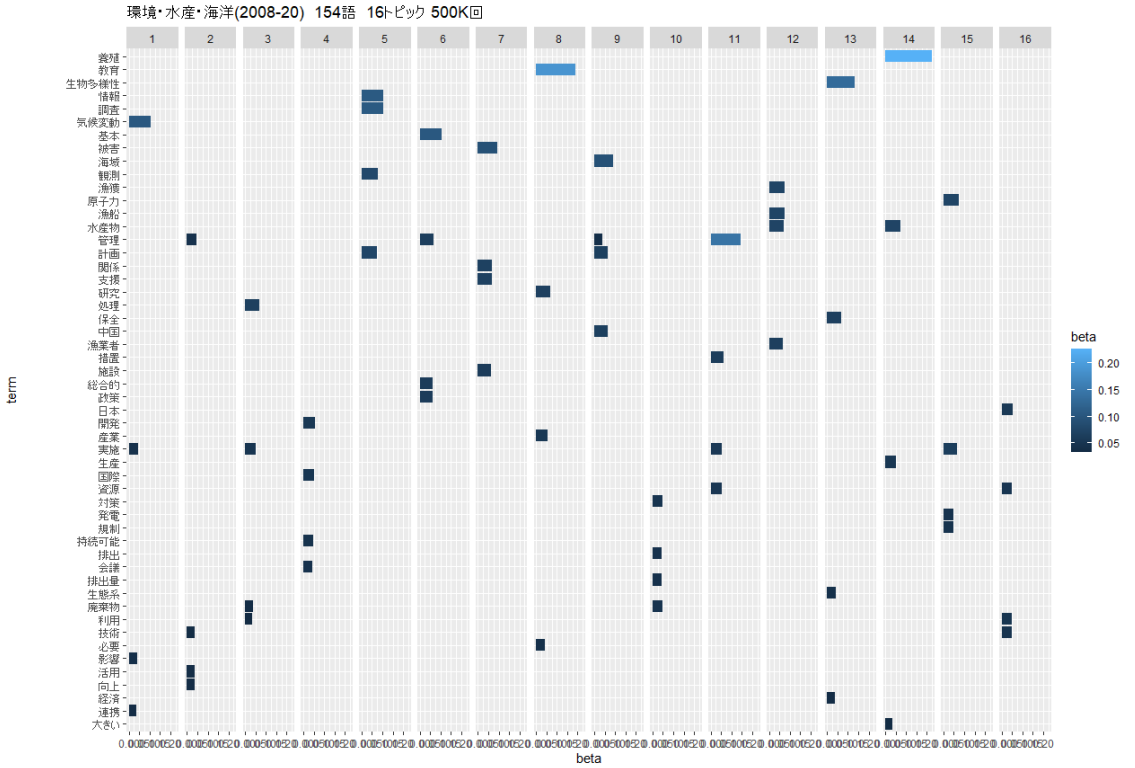


図９　154 語　500K 回

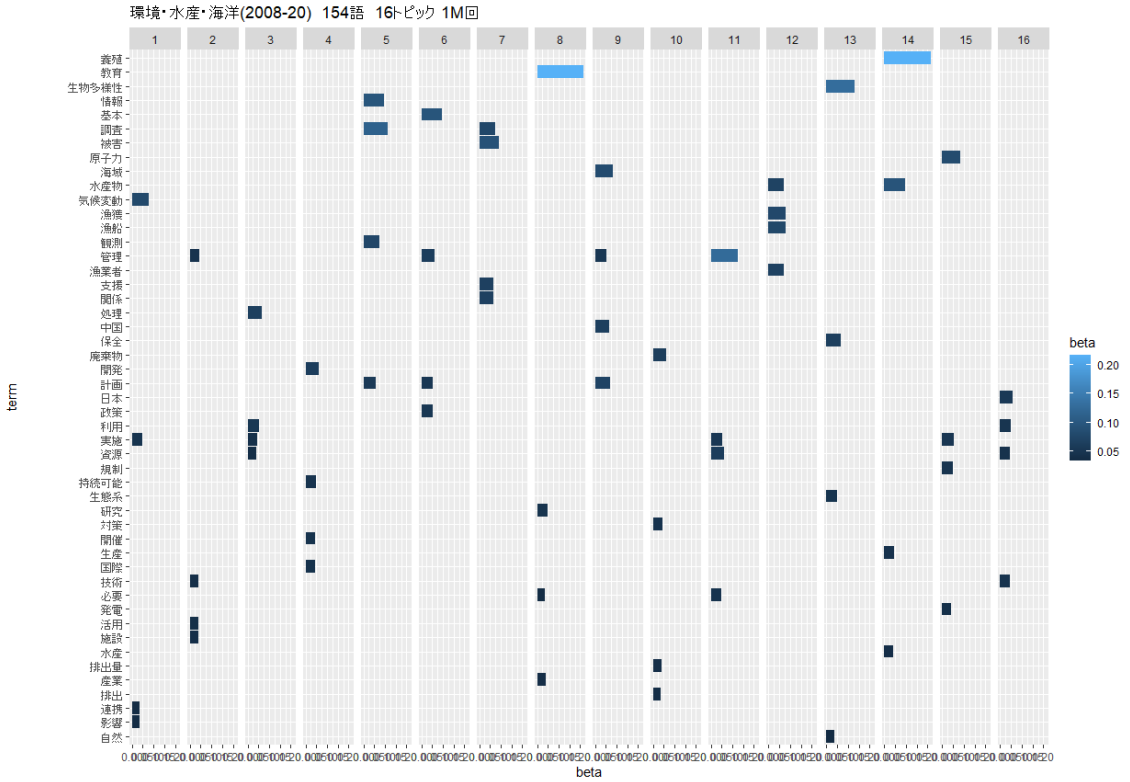


図 10　154 語 1000K 回

抽出語で高い割合のトピックの特徴語

```
> terms(環水海_lda_100k,5)
     Topic 1 Topic 2 Topic 3 Topic 4 Topic 5 Topic 6  Topic 7 Topic 8 Topic 9    Topic 10
[1,] "実施"  "施設"  "実施"  "情報"  "日本"  "社会"   "教育"  "利用"  "開催"  "生物多様性"
[2,] "法律"  "支援"  "制度"  "調査"  "技術"  "利用"   "産業"  "保全"  "持続可能" "保全"
[3,] "自然"  "関係"  "導入"  "計画"  "研究"  "地球"   "調査"  "社会"  "国際"  "生態系"
[4,] "規制"  "調査"  "規制"  "基本"  "開発"  "資源"   "実施"  "資源"  "開発"  "活動"
[5,] "対策"  "発生"  "措置"  "産業"  "世界"  "温暖化" "支援"  "自然"  "政策"  "評価"
     Topic 11 Topic 12 Topic 13 Topic 14   Topic 15 Topic 16
[1,] "基本"   "廃棄物"  "減少"   "気候変動" "管理"    "海域"
[2,] "政策"   "処理"   "増加"   "実施"   "資源"    "管理"
[3,] "管理"   "排出量"  "生産"   "影響"   "重要"    "計画"
[4,] "計画"   "整備"   "利用"   "対策"   "必要"    "開発"
[5,] "開発"   "排出"   "資源"   "評価"   "対象"    "日本"

> terms(環水海2_lda_100k,5)
     Topic 1   Topic 2 Topic 3    Topic 4    Topic 5 Topic 6  Topic 7 Topic 8 Topic 9
[1,] "気候変動" "技術"  "処理"    "開発"    "情報"  "基本"   "被害"  "教育"  "海域"
[2,] "影響"   "管理"  "実施"    "持続可能" "調査"  "管理"   "支援"  "産業"  "計画"
[3,] "実施"   "情報"  "利用"    "国際"    "観測"  "総合的" "調査"  "研究"  "中国"
[4,] "支援"   "活用"  "資源"    "開催"    "計画"  "計画"   "関係"  "活動"  "管理"
[5,] "活用"   "向上"  "リサイクル" "政策"    "産業"  "政策"   "施設"  "連携"  "日本"
     Topic 10 Topic 11 Topic 12 Topic 13   Topic 14 Topic 15 Topic 16
[1,] "廃棄物"  "管理"   "漁獲"   "生物多様性" "養殖"   "原子力"  "技術"
[2,] "対策"   "実施"   "漁船"   "保全"    "水産物"  "実施"   "利用"
[3,] "排出"   "関係"   "水産物"  "生態系"   "生産"   "規制"   "日本"
[4,] "排出量"  "資源"   "漁業者"  "目標"    "世界"   "発電"   "資源"
[5,] "社会"   "必要"   "水産"   "経済"    "必要"   "委員会"  "世界"
```

付録 11 「「こころ」 で KHCoder と R-Studio の LDA 処理結果を比較」

文書番号：JRDN-21-035

1. KHCoder で前処理

入力データ： kokoro_df2.csv 全 110 章（h5 タグ）に分割した文章。

強制抽出語：一人、二人、K

前処理結果：

Project

現在のプロジェクト：	kokoro_df2.csv [text]
説明（メモ）：	

Database Stats

総抽出語数（使用）：	105,912 (36,132)
異なり語数（使用）：	6,021 (5,451)

文書の単純集計：

集計単位	ケース数
文	4,656
段落	110
H5	110

List

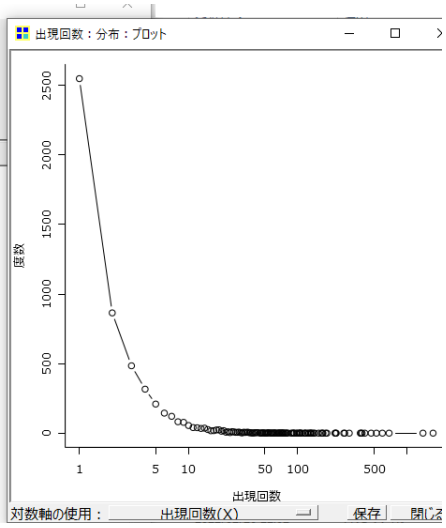
#	抽出語	品詞/活用	頻度
1	先生	名詞	595
2	K	タグ	411
3	奥さん	名詞	388
4	思う	動詞	296
5	父	名詞C	269
6	自分	名詞	264
7	見る	動詞	225
8	聞く	動詞	219
9	出る	動詞	185
10	人	名詞C	180
11	母	名詞C	170
12	お嬢さん	名詞	168
13	前	副詞可能	164
14	帰る	動詞	155
15	今	副詞可能	139
16	顔	名詞C	135
17	来る	動詞	131
18	考える	動詞	130
19	言葉	名詞	126
20	眼	名詞C	123
21	知る	動詞	118
22	二人	タグ	115

■記述統計

異なり語数 (n)	5451
出現回数の平均	6.63
出現回数の標準偏差	38.29

■度数分布表

出現回数	度数	パーセント	累積度数	累積パーセント
1	2548	46.74	2548	46.74
2	865	15.87	3413	62.61
3	485	8.90	3898	71.51
4	317	5.82	4215	77.33
5	210	3.85	4425	81.18
6	145	2.66	4570	83.84
7	122	2.24	4692	86.08
8	83	1.52	4775	87.60
9	79	1.45	4854	89.05
10	56	1.03	4910	90.08
11	41	0.75	4951	90.83
12	40	0.73	4991	91.56
13	35	0.64	5026	92.20
14	37	0.68	5063	92.88
15	27	0.50	5090	93.38



2. KHCoder でのトピック推定

KHCoder による自動設定パラメーは、 最小頻度：50 語数：71 となった。

KHCoder のトピック数の推定（LDAtuning）結果を図 2-1 に示す。この結果からトピック数を 10 として KHCoder で LDA 処理結果を図 2-2 に示す。この処理結果のトピック 2 とトピック 7 に注目し、トピック比率（1 章～110 章）の出力結果を図 2-3、2-4 に示す。

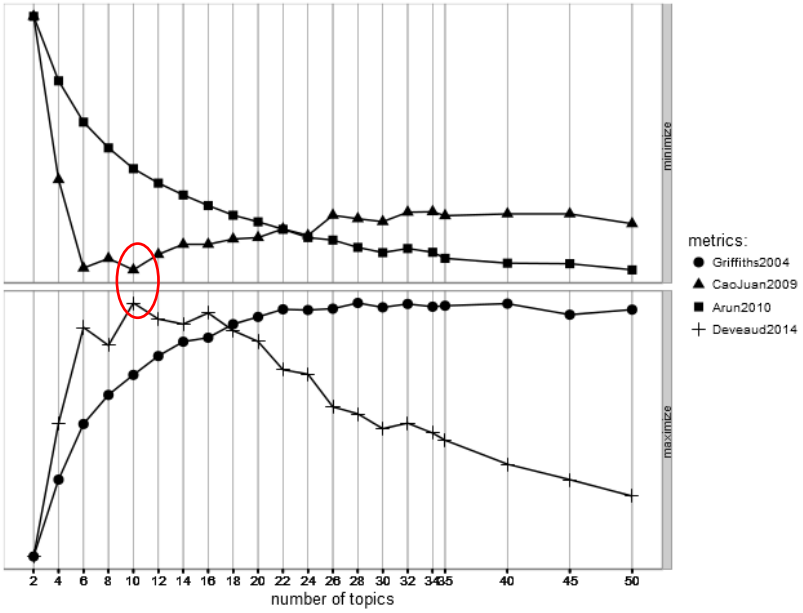


図 2-1 トピック数の推定（LDAtuning）結果

トピックの推定結果

Info

集計単位：h5　トピック数：10　異なり語数：71

Topics

#1		#2		#3		#4		#5		#6		#7		#8		#9		#10	
奥さん	0.459	K	0.486	聞く	0.168	自分	0.285	今	0.141	見る	0.297	父	0.271	帰る	0.190	思う	0.135	先生	0.671
お嬢さん	0.178	室	0.064	言葉	0.146	思う	0.184	人	0.134	眼	0.149	母	0.172	来る	0.131	考える	0.109	二人	0.063
女	0.090	付く	0.060	前	0.112	妻	0.123	聞く	0.111	顔	0.101	手紙	0.075	出る	0.113	叔父	0.089	人	0.051
男	0.037	二人	0.051	口	0.105	人間	0.069	好い	0.099	前	0.093	書く	0.065	一人	0.095	出る	0.080	思う	0.034
知れる	0.032	答える	0.050	話	0.104	死ぬ	0.069	知る	0.087	頭	0.087	兄	0.063	行く	0.083	事	0.078	解る	0.030
坐る	0.032	見える	0.050	意味	0.083	心	0.054	悪い	0.063	手	0.071	病気	0.057	立つ	0.079	心	0.067	答える	0.027
笑う	0.030	知る	0.043	行く	0.056	気	0.046	問題	0.061	心持	0.055	卒業	0.056	声	0.058	家	0.067	見える	0.018
今	0.022	歩く	0.037	急	0.043	取る	0.028	少し	0.056	出る	0.047	出す	0.044	宅	0.036	解る	0.052	立つ	0.017
二人	0.020	知れる	0.020	見える	0.037	知れる	0.022	様子	0.050	少し	0.043	東京	0.037	人	0.030	顔	0.052	様子	0.015
宅	0.019	態度	0.018	宅	0.024	手	0.022	考える	0.039	言葉	0.013	読む	0.032	答える	0.029	外	0.049	人間	0.014

図 2-2 LDA 処理結果（トピック数 10）

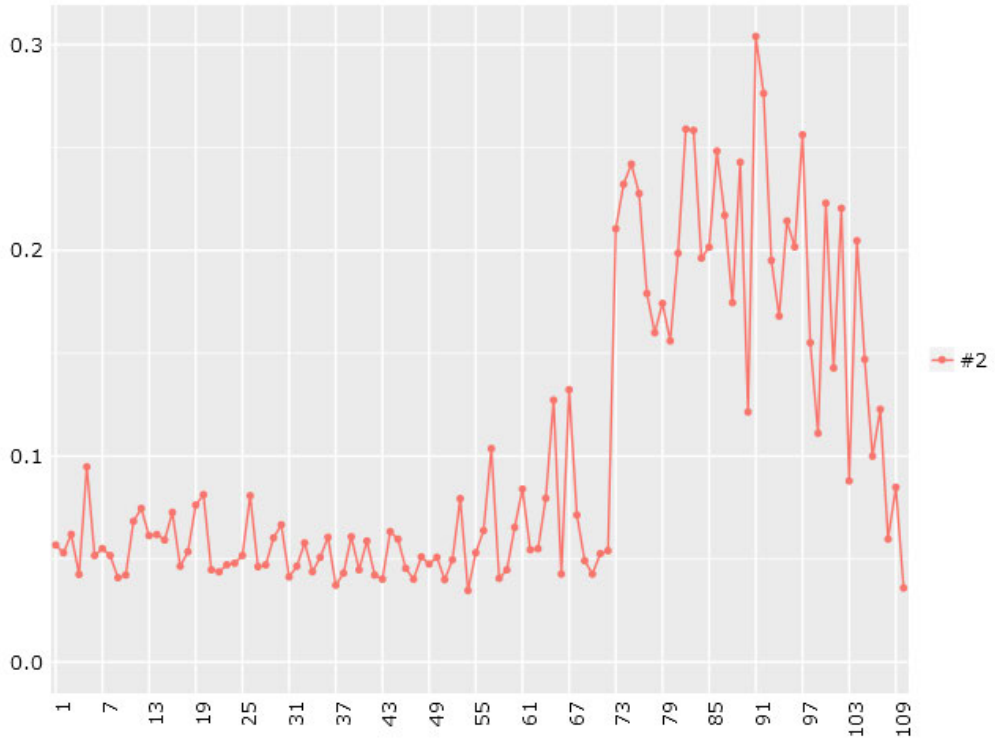


図 2-3 トピック 2 のトピック比率

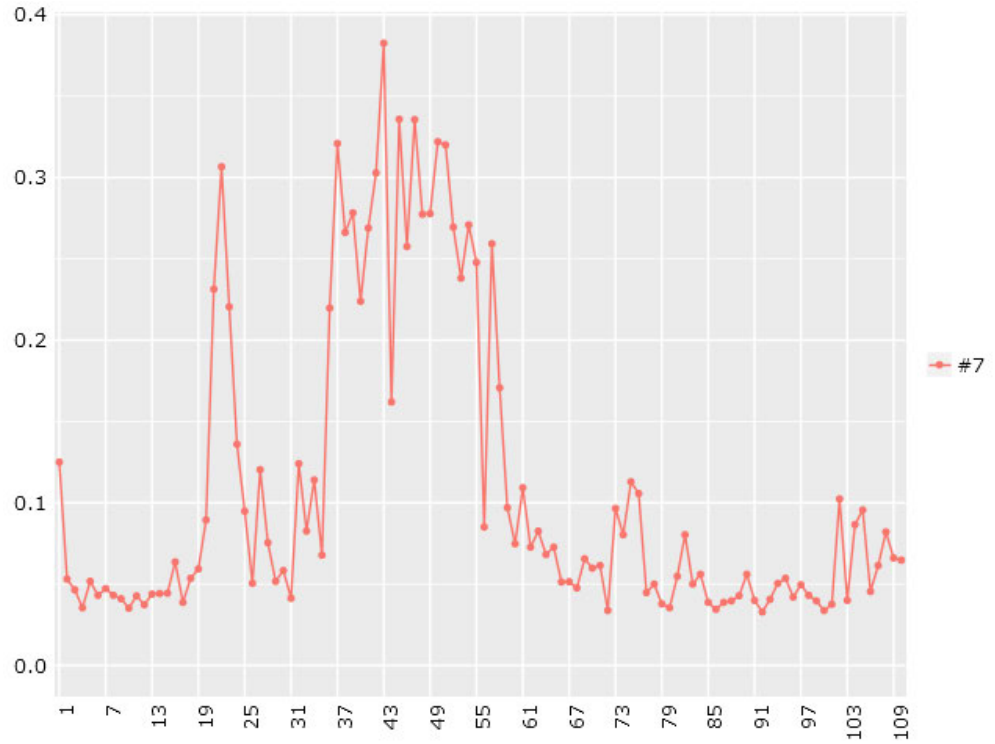


図 2-4 トピック 7 のトピック比率

	A	B	C	D	E	F	G	H	I	J	K	L
1		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	ケース数
2	1	0.068	0.057	0.068	0.114	0.091	0.068	0.125	0.205	0.057	0.148	1
3	2	0.062	0.053	0.044	0.053	0.071	0.168	0.053	0.159	0.106	0.23	1
4	3	0.039	0.062	0.093	0.047	0.078	0.116	0.047	0.093	0.062	0.364	1
5	4	0.057	0.043	0.078	0.113	0.078	0.121	0.035	0.142	0.085	0.248	1
6	5	0.043	0.095	0.112	0.078	0.052	0.069	0.052	0.164	0.078	0.259	1
7	6	0.043	0.052	0.112	0.103	0.103	0.103	0.043	0.069	0.052	0.319	1
8	7	0.047	0.055	0.142	0.118	0.102	0.055	0.047	0.102	0.087	0.244	1
9	8	0.216	0.052	0.06	0.069	0.103	0.069	0.043	0.121	0.078	0.19	1
10	9	0.107	0.041	0.09	0.066	0.139	0.057	0.041	0.139	0.066	0.254	1
11	10	0.092	0.042	0.12	0.085	0.092	0.063	0.035	0.169	0.049	0.254	1
12	11	0.137	0.068	0.094	0.043	0.128	0.085	0.043	0.068	0.12	0.214	1
13	12	0.149	0.075	0.067	0.06	0.164	0.082	0.037	0.052	0.097	0.216	1
100	99	0.214	0.111	0.167	0.103	0.063	0.087	0.04	0.056	0.079	0.079	1
101	100	0.142	0.223	0.068	0.068	0.101	0.101	0.034	0.081	0.142	0.041	1
102	101	0.195	0.143	0.158	0.09	0.098	0.075	0.038	0.083	0.083	0.038	1
103	102	0.094	0.22	0.079	0.071	0.087	0.165	0.102	0.071	0.055	0.055	1
104	103	0.24	0.088	0.096	0.064	0.072	0.248	0.04	0.048	0.048	0.056	1
105	104	0.213	0.205	0.047	0.087	0.055	0.063	0.087	0.094	0.102	0.047	1
106	105	0.14	0.147	0.059	0.213	0.051	0.088	0.096	0.081	0.066	0.059	1
107	106	0.073	0.1	0.082	0.3	0.045	0.082	0.045	0.091	0.127	0.055	1
108	107	0.044	0.123	0.079	0.342	0.053	0.07	0.061	0.105	0.079	0.044	1

図 2-4　トピック比率

3. R-Studio での topicmodels パッケージの LDA 処理

KHCoder のエクスポート機能により、前処理結果（抽出語 71 語）を csv 形式で保存する。その csv ファイルの内容を図 3-1 に示す。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	h1	h2	h3	h4	h5	id	length_c	length_w	先生	奥さん	自分	お嬢さん	言葉	手紙
2	0	0	0	0	1	1	1365	897	5	0	0	0	0	0
3	0	0	0	0	2	2	1379	927	14	0	1	0	0	0
4	0	0	0	0	3	3	1497	995	30	0	0	0	1	0
5	0	0	0	0	4	4	1451	957	25	3	3	0	2	0
6	0	0	0	0	5	5	1345	909	20	0	0	0	2	0
7	0	0	0	0	6	6	1456	930	24	0	3	0	0	0
106	0	0	0	0	105	105	1510	1002	0	3	3	7	0	1
107	0	0	0	0	106	106	1462	948	0	0	6	0	1	0
108	0	0	0	0	107	107	1507	986	0	0	11	0	1	0
109	0	0	0	0	108	108	1549	1036	0	0	12	0	0	0
110	0	0	0	0	109	109	1487	971	0	0	2	0	0	0
111	0	0	0	0	110	110	1772	1166	0	0	3	0	2	1

kokoro_df2-m50-71w

図 3-1　エクスポートされた csv ファイルの内容

Topicmodels::LDA の処理コード （KHcoder での処理と同じパラメータを設定）

```
dtm <- read.csv("D:/R コード/こころ/kokoto_tutorial_H5-71w.csv",
fileEncoding="UTF-8-BOM")
dtmx <- dtm[,9:79]
dtmy <- dtmx[rowSums(dtmx) > 0,]
kokoro_lda <- topicmodels::LDA(dtmy, k = 10,
   method = "Gibbs", control = list(seed = 1234567,burnin = 1000))
```

```
> terms(kokoro_lda, 10)
      Topic 1   Topic 2  Topic 3  Topic 4  Topic 5  Topic 6 Topic 7 Topic 8  Topic 9  Topic 10
 [1,] "奥さん"  "K"      "聞く"   "自分"   "今"     "見る"  "父"    "帰る"   "思う"   "先生"
 [2,] "お嬢さん" "室"     "言葉"   "思う"   "人"     "眼"    "母"    "来る"   "考える" "二人"
 [3,] "女"      "付く"   "豆"     "妻"     "聞く"   "顔"    "手紙"  "出る"   "叔父"   "人"
 [4,] "男"      "二人"   "口"     "人間"   "好い"   "前"    "書く"  "一人"   "出る"   "思う"
 [5,] "知れる"  "答える" "話"     "死ぬ"   "知る"   "頭"    "兄"    "行く"   "事"     "解る"
 [6,] "坐る"    "見える" "意味"   "心"     "悪い"   "手"    "病気"  "立つ"   "心"     "答える"
 [7,] "笑う"    "知る"   "行く"   "気"     "問題"   "心持"  "卒業"  "声"     "家"     "見える"
 [8,] "今"      "歩く"   "急"     "取る"   "少し"   "出る"  "出す"  "宅"     "解る"   "立つ"
 [9,] "二人"    "知れる" "見える" "知れる" "様子"   "少し"  "東京"  "人"     "顔"     "様子"
[10,] "宅"      "態度"   "宅"     "手"     "考える" "言葉"  "読む"  "答える" "外"     "人間"
>
```

トピック 2 の描画コード

```
> library(tidyverse)
> library(tidytext)
gamma2 <- kokoro_lda %>% tidy("gamma") %>%
                group_by(document) %>% filter( topic==2 )
p2 <- ggplot(gamma2, aes(x = as.numeric(document), y = gamma)) +
     geom_line(aes(y   = gamma)) +
     geom_point(aes(y = gamma)) +
     ylab("γ: トピックの出現確率") + xlab("章") +
     scale_x_continuous(breaks = seq(0, 110, by =5)) +
     ggtitle("トピック 2  γ:トピックの出現確率   Gibbs:2K")
plot(p2)
```

Gamma 行列のトピック 2 の値

```
> gamma2$gamma[1:10]
 [1] 0.05681818 0.05309735 0.06201550 0.04255319 0.09482759
     0.05172414 0.05511811 0.05172414 0.04098361 0.04225352
> gamma2$gamma[101:110]
 [1] 0.14285714 0.22047244 0.08800000 0.20472441 0.14705882
     0.10000000 0.12280702 0.05970149 0.08490566 0.03597122
```

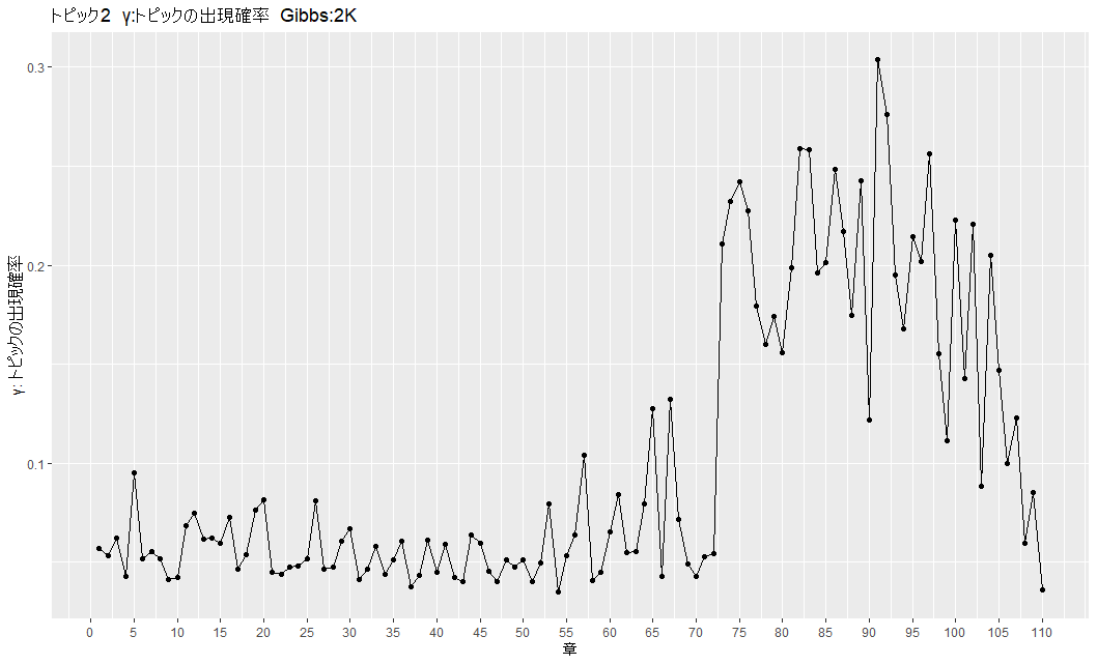


図 3-1 トピック 2 のトピック出現確率

4. Gibbs サンプリング数 500K, 1M での LDA 処理結果

KHCoder での Gibbs サンプリング数：2K と、500K や 1M 回での Topicmodels::LDA での処理結果を比較する。

Topicmodels::LDA で Gibbs サンプリング数を 500K とした処理コード

```
dtm <- read.csv("D:/R コード/こころ/kokoto_tutorial_H5-71w.csv",
fileEncoding="UTF-8-BOM")
dtmx <- dtm[,9:79]
dtmy <- dtmx[rowSums(dtmx) > 0,]
kokoro_lda_500K <- topicmodels::LDA(dtmy, k = 10,
          method = "Gibbs", control = list(seed=1234567,iter=500000))
```

```
> terms(kokoro_lda_500k, 10)
      Topic 1    Topic 2    Topic 3 Topic 4  Topic 5 Topic 6  Topic 7 Topic 8 Topic 9  Topic 10
 [1,] "奥さん"   "K"        "帰る"  "自分"   "人"    "顔"     "父"    "聞く"  "今"     "先生"
 [2,] "女"       "お嬢さん" "出る"  "思う"   "聞く"  "見る"   "母"    "言葉"  "知る"   "人"
 [3,] "立つ"     "室"       "来る"  "妻"     "話"    "眼"     "手紙"  "口"    "少し"   "二人"
 [4,] "考える"   "答える"   "宅"    "心"     "様子"  "頭"     "書く"  "見る"  "考える" "答える"
 [5,] "前"       "二人"     "笑う"  "人間"   "男"    "前"     "兄"    "意味"  "叔父"   "解る"
 [6,] "態度"     "見える"   "行く"  "気"     "急"    "坐る"   "病気"  "好い"  "家"     "前"
 [7,] "お嬢さん" "自分"     "立つ"  "死ぬ"   "一人"  "声"     "卒業"  "問題"  "思う"   "見える"
 [8,] "話す"     "歩く"     "手"    "知れる" "来る"  "知れる" "出す"  "出る"  "事"     "返事"
 [9,] "思う"     "心持"     "歩く"  "外"     "自分"  "手"     "読む"  "悪い"  "東京"   "来る"
[10,] "行く"     "取る"     "心"    "意味"   "心持"  "事"     "好い"  "前"    "見る"   "悪い"
```

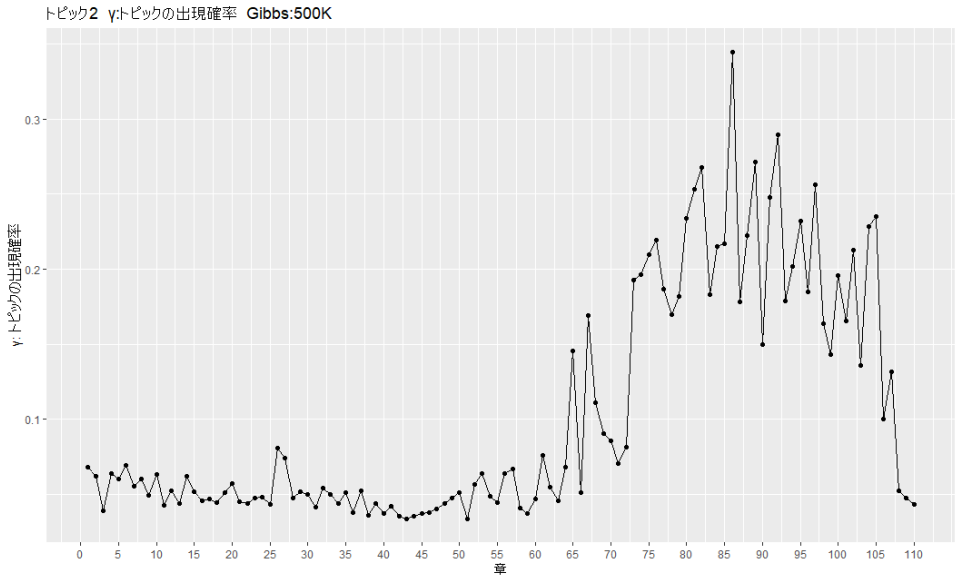


図 4-1 トピック 2 トピック出現確率（textmodels::LDA 500K 回）

Topicmodels::LDA で 1M 回の処理結果を以下に示す。

```
> terms(kokoro_lda_1m, 10)
      Topic 1   Topic 2   Topic 3 Topic 4 Topic 5 Topic 6 Topic 7 Topic 8 Topic 9 Topic 10
 [1,] "奥さん"  "K"       "顔"    "自分"  "二人"  "聞く"  "父"    "帰る"  "今"    "先生"
 [2,] "女"      "お嬢さん" "見る"  "思う"  "考える" "人"    "母"    "来る"  "出る"  "見える"
 [3,] "自分"    "室"      "眼"    "妻"    "話"    "言葉"  "手紙"  "行く"  "叔父"  "人間"
 [4,] "話す"    "見える"  "思う"  "死ぬ"  "頭"    "口"    "書く"  "立つ"  "家"    "前"
 [5,] "聞く"    "答える"  "前"    "心"    "事"    "急"    "兄"    "出る"  "思う"  "少し"
 [6,] "前"      "坐る"    "手"    "気"    "知れる" "解る"  "病気"  "見る"  "東京"  "解る"
 [7,] "見る"    "思う"    "声"    "人間"  "歩く"  "知る"  "読む"  "一人"  "好い"  "答える"
 [8,] "帰る"    "付く"    "知る"  "意味"  "心持"  "問題"  "卒業"  "宅"    "心"    "悪い"
 [9,] "お嬢さん" "取る"   "悪い"  "卒業"  "態度"  "意味"  "返事"  "笑う"  "眼"    "返事"
[10,] "今"      "心持"    "少し"  "歩く"  "人"    "前"    "東京"  "様子"  "外"    "行く"
```

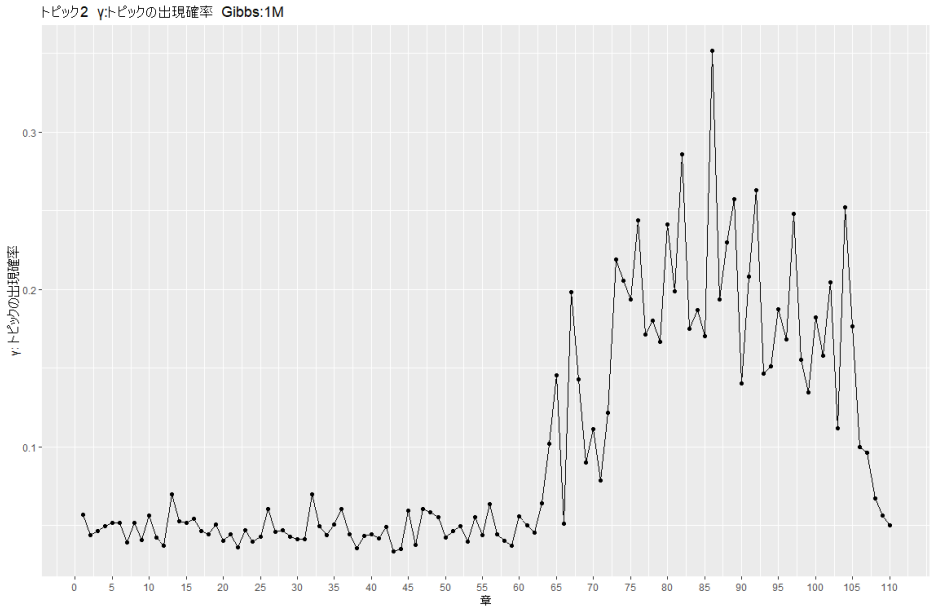


図 4-2 トピック 2 トピック出現確率（textmodels::LDA 1M 回）

トピックの出現確率（γ）描画結果は、Gibbs サンプリング数 2K、500K、1M 回で同じような傾向に見えるが、重ねて描画すると差異が明らかになる。トピック 2 について、KHCoder での 2K 回（黒）をベースに、500K 回（青）、1M 回（赤）の描画結果を比べると、2K 回では 500K 回や 1M 回と顕著な差異（青矢印部分）があるが、500K 回と 1M 回では差異が少ない。

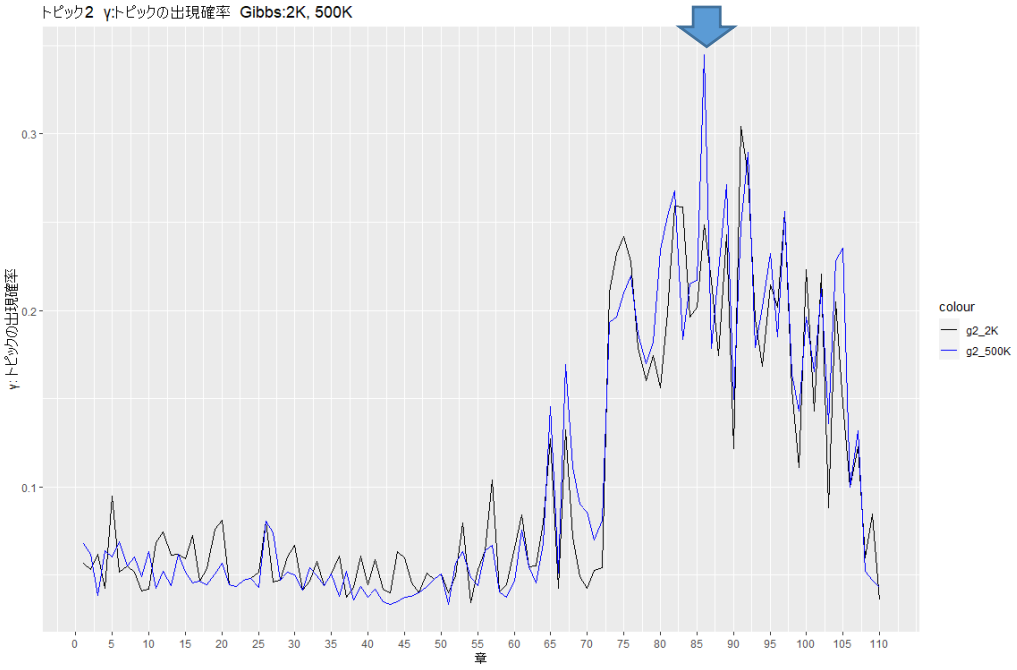


図 4-3 トピック 2 トピック出現確率の比較（textmodels::LDA 2K 回と 500K 回）

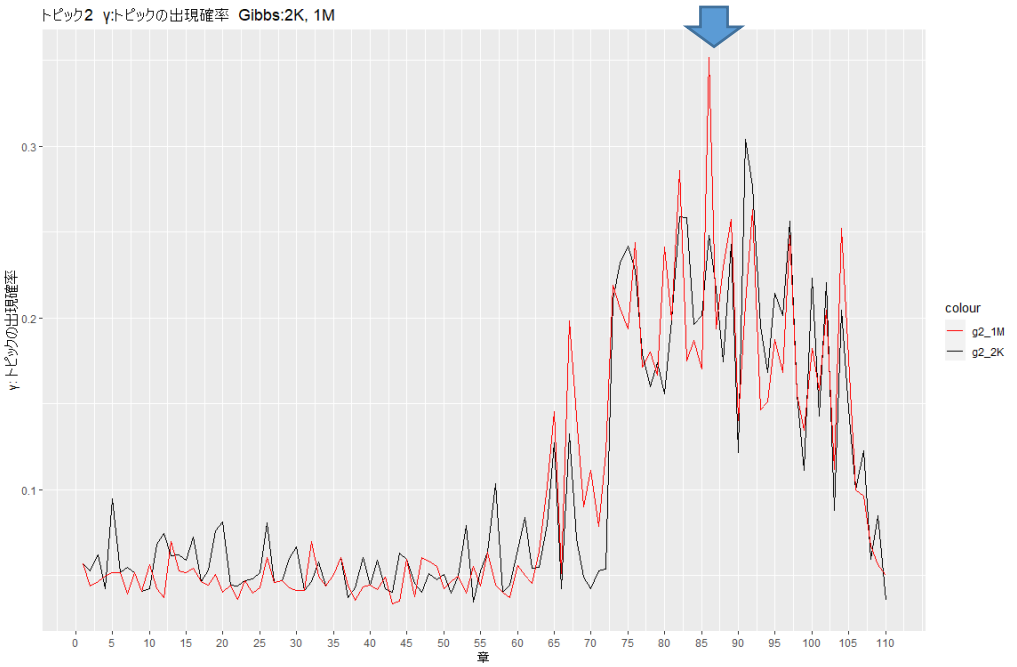


図 4-4 トピック 2 トピック出現確率の比較（textmodels::LDA 2K 回と 1M 回）

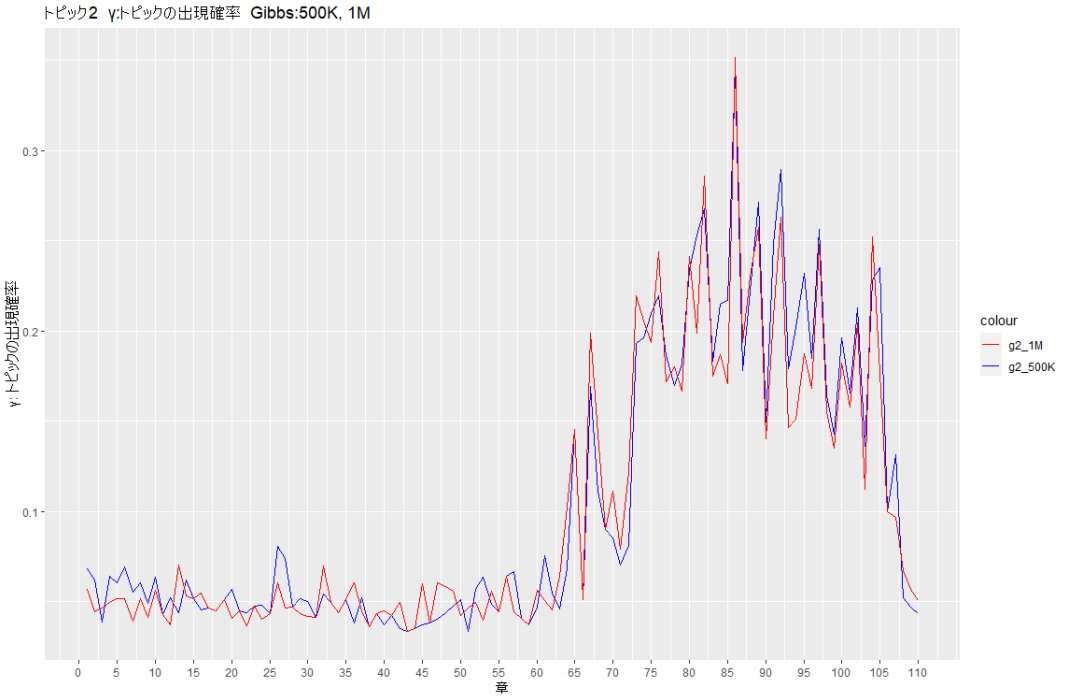


図 4-5 トピック 2 トピック出現確率の比較（textmodels::LDA 500K 回と 1M 回）

トピック 7 についても同様な比較を行うと、2K 回（黒）、500K 回（青）、1M 回（赤）の描画結果は概ね類似している。

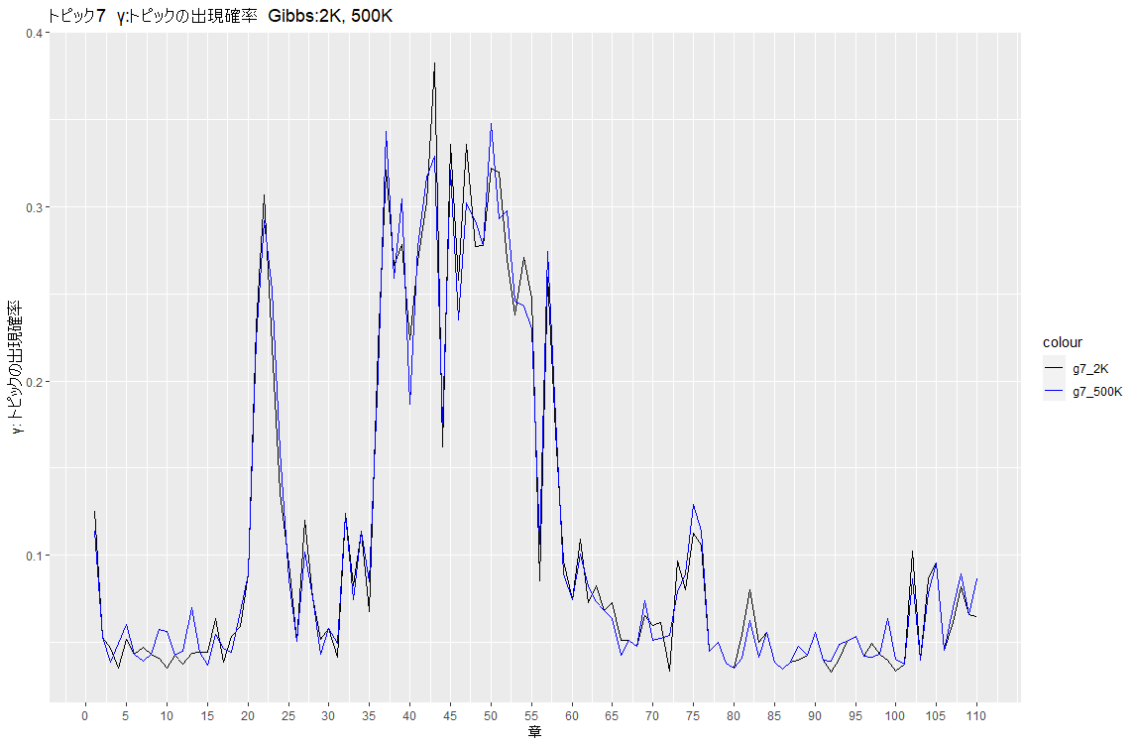


図 4-6 トピック 7 トピック出現確率の比較（textmodels::LDA 2K 回と 500K 回）

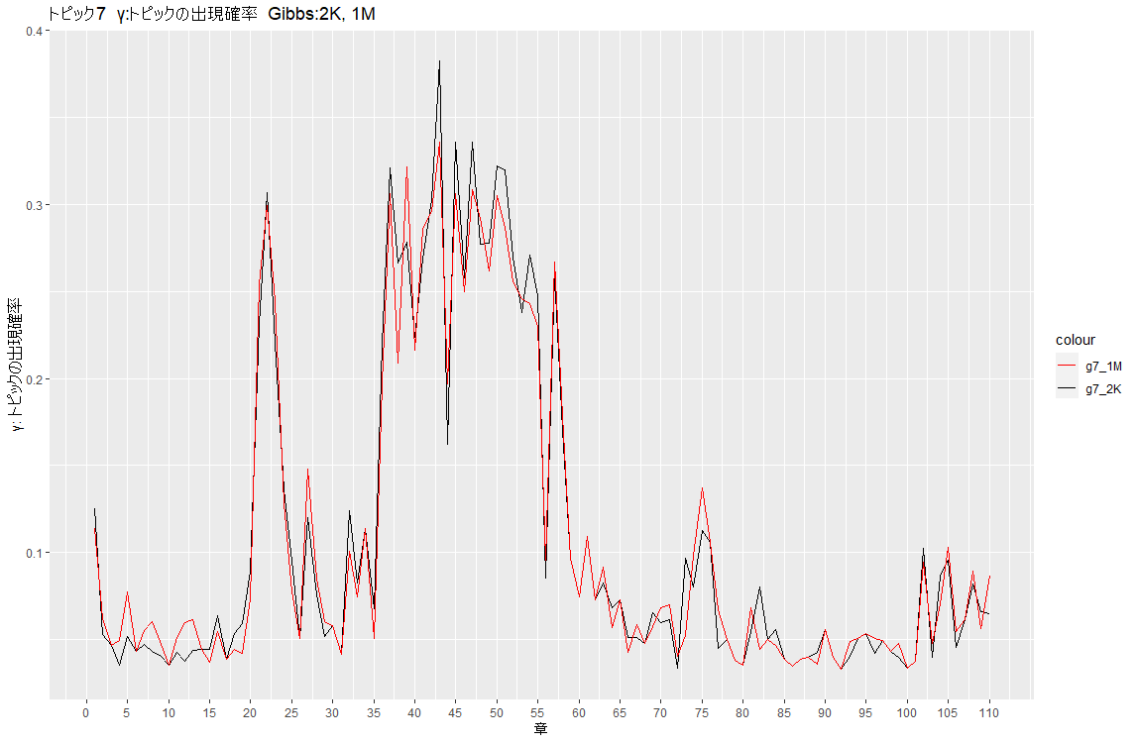


図 4-7 トピック 7 トピック出現確率の比較（textmodels::LDA 2K 回と 1M 回）

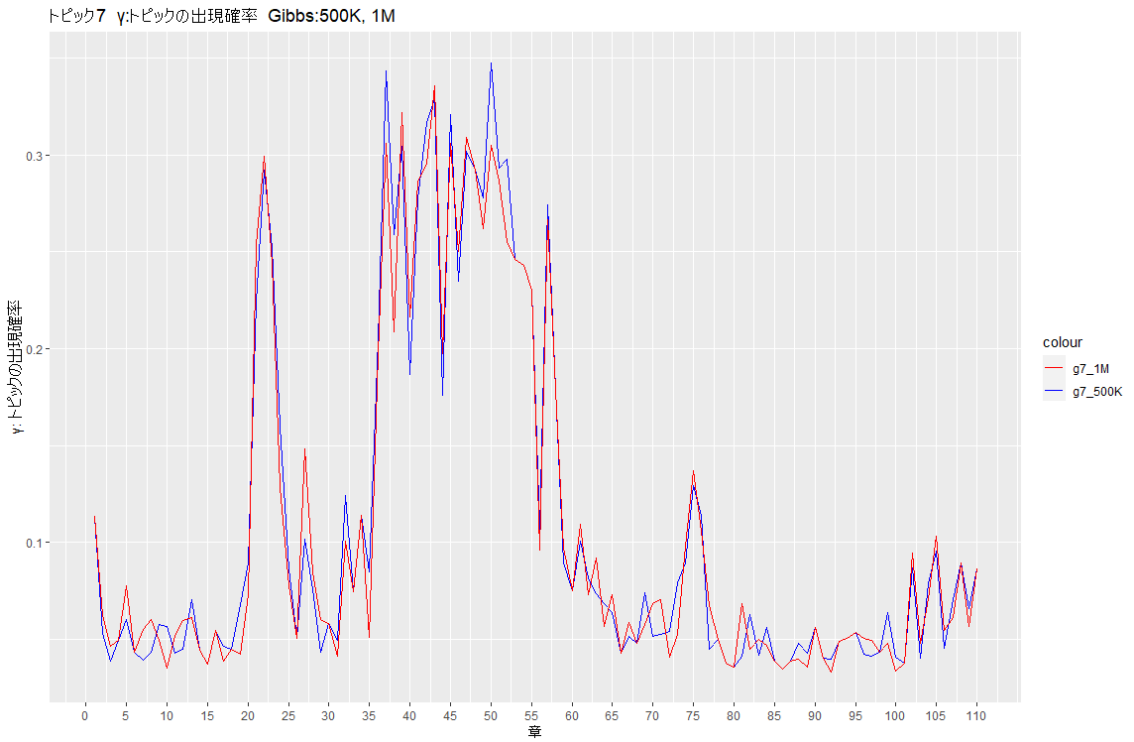


図 4-8 トピック 7 トピック出現確率の比較（textmodels::LDA 500K 回と 1M 回）

5. topicmodels と seededLDA パッケージの LDA 処理結果

こころDTMデータをdtm形式（topicmodelsパッケージLDA用入力）とdfm形式（seededldaパッケージ textmodel_lda 用入力）に変換し、同じ処理パラメータ（Gibbs サンプリング数、乱数初期値）で実行し、両者の結果を比較する。

DTM データを tidyverse パッケージの cast_dtm で dtm 形式に変換し、topicmodels パッケージの LDA で処理（乱数初期化、Gibbs：2000 回）

> library(reshape2)
kokoro_df <- melt(as.matrix(dtmy)) # melt()で変換
wkokoro_dtm <- work_df %>%
cast_dtm(document=Var1, term=Var2, value=value)
kokoro_LDAd <- topicmodels::LDA(kokoro_dtm,
k = 10, method = "Gibbs", control = list(seed = 1234567, iter = 2000))

```
> topicmodels::terms(kokoro_LDAd, 10)
      Topic 1  Topic 2    Topic 3 Topic 4  Topic 5 Topic 6  Topic 7  Topic 8  Topic 9 Topic 10
 [1,] "奥さん" "K"       "自分"  "聞く"   "出る"  "人"     "今"     "見る"   "父"    "先生"
 [2,] "女"     "お嬢さん" "思う"  "言葉"   "帰る"  "知る"   "思う"   "前"     "母"    "答える"
 [3,] "見る"   "室"       "妻"    "考える" "来る"  "見える" "叔父"   "眼"     "書く"  "人"
 [4,] "顔"     "答える"   "死ぬ"  "口"     "立つ"  "自分"   "家"     "顔"     "手紙"  "卒業"
 [5,] "急"     "声"       "心"    "話"     "宅"    "解る"   "事"     "手"     "兄"    "解る"
 [6,] "態度"   "付く"     "人間"  "意味"   "笑う"  "心持"   "東京"   "問題"   "出す"  "人間"
 [7,] "二人"   "坐る"     "一人"  "様子"   "行く"  "頭"     "知れる" "悪い"   "読む"  "外"
 [8,] "少し"   "取る"     "気"    "返事"   "歩く"  "好い"   "考える" "少し"   "病気"  "手"
 [9,] "眼"     "心持"     "外"    "気"     "見る"  "二人"   "頭"     "知れる" "卒業"  "少し"
[10,] "話す"   "聞く"     "帰る"  "二人"   "二人"  "話す"   "心"     "聞く"   "東京"  "一人"
```

DTM データを tidyverse パッケージの cast_dfm で dfm 形式に変換し、seededlda パッケージの textmodel_lda で処理（乱数初期化、Gibbs：2000 回）
kokoro_dfm <- kokoro_df %>% cast_dfm(document=Var1, term=Var2, value=value)

```
> str(kokoro_dfm)
Formal class 'dfm' [package "quanteda"] with 8 slots
  ..@ docvars :'data.frame':    110 obs. of  3 variables:
  .. ..$ docname_: chr [1:110] "1" "2" "3" "4" ...
  .. ..$ docid_  : Factor w/ 110 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
  .. ..$ segid_  : int [1:110] 1 1 1 1 1 1 1 1 1 1 ...
  ..@ meta    :List of 3
  .. ..$ system:List of 5
  .. .. ..$ package-version:Classes 'package_version', 'numeric_version'  hidden list of 1
  .. .. .. ..$ : int [1:3] 3 2 0
```

```
  .. .. .. ..$ data: NULL
  .. ..$ user   : list()
  ..@ i       : int [1:3555] 0 1 2 3 4 5 6 7 8 9 ...
  ..@ p       : int [1:72] 0 53 94 122 162 197 246 345 414 483 ...
  ..@ Dim     : int [1:2] 110 71
  ..@ Dimnames:List of 2
  .. ..$ docs    : chr [1:110] "1" "2" "3" "4" ...
  .. ..$ features: chr [1:71] "先生" "心持" "病気" "急" ...
  ..@ x       : num [1:3555] 5 14 30 25 20 24 19 14 19 22 ...
  ..@ factors : list()
```

```
set.seed(1234567)
kokoro_LDAs <- textmodel_lda( kokoro_dfm, k = 10, max_iter = 2000,
                    alpha = NULL, beta = NULL, model = NULL,
                    verbose = quanteda_options("verbose") )
```

```
> seededlda::terms(kokoro_LDAs, 10)
      topic1 topic2 topic3   topic4   topic5 topic6 topic7     topic8     topic9 topic10
 [1,] "見る" "眼"   "聞く"   "先生"   "妻"   "手紙" "奥さん"   "K"        "父"   "叔父"
 [2,] "行く" "見る" "言葉"   "人"     "自分" "書く" "お嬢さん" "自分"     "母"   "人"
 [3,] "帰る" "帰る" "思う"   "見える" "思う" "人"   "女"       "二人"     "兄"   "思う"
 [4,] "前"   "顔"   "知れる" "答える" "死ぬ" "来る" "自分"     "出る"     "病気" "家"
 [5,] "出る" "頭"   "事"     "解る"   "人間" "読む" "思う"     "お嬢さん" "死ぬ" "東京"
 [6,] "卒業" "声"   "問題"   "人間"   "心"   "返事" "出る"     "答える"   "東京" "考える"
 [7,] "聞く" "室"   "前"     "二人"   "外"   "自分" "男"       "知る"     "好い" "自分"
 [8,] "笑う" "来る" "話"     "知る"   "行く" "気"   "考える"   "考える"   "聞く" "解る"
 [9,] "顔"   "坐る" "口"     "言葉"   "一人" "今"   "今"       "見える"   "口"   "知る"
[10,] "思う" "手"   "話す"   "態度"   "意味" "出す" "好い"     "立つ"     "知る" "心"
>
```

以下で、dtm 形式と dfm 形式に格納されている DTM 行列は、同じ形式であることを確認した。

dtm 形式データと dfm 形式データの内容確認

```
> kokoro_dtm$v
  [1]  5 14 30 25 20 24 19 14 19 22 15 16 12 14 22  3 17  9 11  5 11 1
 [43]  5  3  6  1  7  2  8  1  4  7  1  1  1  3  1  1  1  1  1  2  1
 [85]  2  2  4  1  1  1  1  1  1  2  2  6  3  1  5  4  1  4  1  1  1
[127]  2  1  3  2  1  2  1  1  2  1  1  1  1  1  1  1  1  2  1  1  2
[169]  1  3  1  1  1  1  1  3  1  5  4  2  4  4  1  1  1  3  2  2  7
[211]  3  1  2  2  2  1  1  2  1  1  1  1  1  1  1  2  1  1  1  3  1
[253]  1  1  1  1  4  5  3  4  2  4  3  5  3  2  2  2  2  1  3  2  2
[295]  5  6  3  3  3  2  2  4  3  2  4  2  5  7  2  8  4  3  2  4  4
[337]  3  3  2  3  3  3  5  3  6  5  2  3  4  1  2  6  1  1  1  3  3
[379]  1  1  2  1  3  5  1  3  1  1  1  2  2  2  2  5  5  1  2  3  6
```

```
> kokoro_dfm@x
  [1]  5 14 30 25 20 24 19 14 19 22 15 16 12 14 22  3 17  9 11  5 11
 [43]  5  3  6  1  7  2  8  1  4  7  1  1  1  3  1  1  1  1  1  2  1
 [85]  2  2  4  1  1  1  1  1  1  2  2  6  3  1  5  4  1  4  1  1  1
[127]  2  1  3  2  1  2  1  1  2  1  1  1  1  1  1  1  1  2  1  1  2
[169]  1  3  1  1  1  1  1  3  1  5  4  2  4  4  1  1  1  3  2  2  7
[211]  3  1  2  2  2  1  1  2  1  1  1  1  1  1  1  2  1  1  1  3  1
[253]  1  1  1  1  4  5  3  4  2  4  3  5  3  2  2  2  2  1  3  2  2
[295]  5  6  3  3  3  2  2  4  3  2  4  2  5  7  2  8  4  3  2  4  4
[337]  3  3  2  3  3  3  5  3  6  5  2  3  4  1  2  6  1  1  1  3  3
[379]  1  1  2  1  3  5  1  3  1  1  1  2  2  2  2  5  5  1  2  3  6
```

「K」に関するトピックと「父・母」に関するトピックについて、topicmodels パッケージ LDA の処理結果（赤）と seededlda パッケージ textmodel_lda の処理結果（青）でトピック出現確率を比較した。その結果、出現確率のパターンは似ているが、数値の差異が顕著であった。

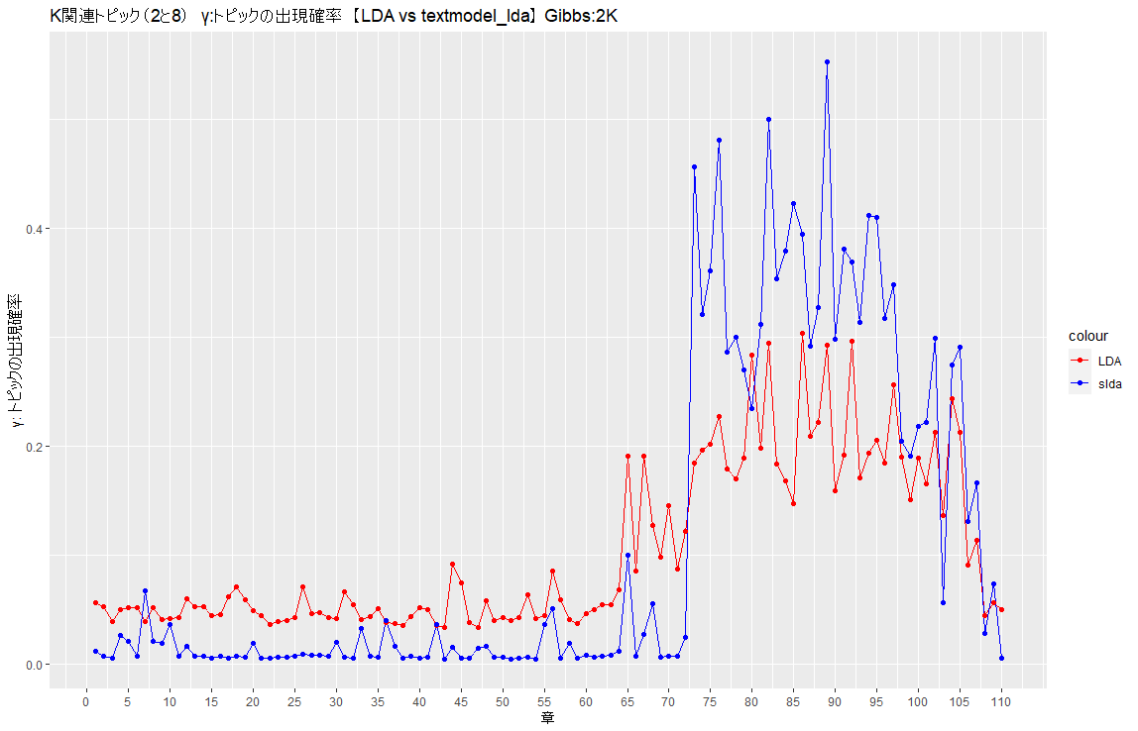


図 5-1 「K」関連トピックのトピック出現確率（textmodels と seededlda）

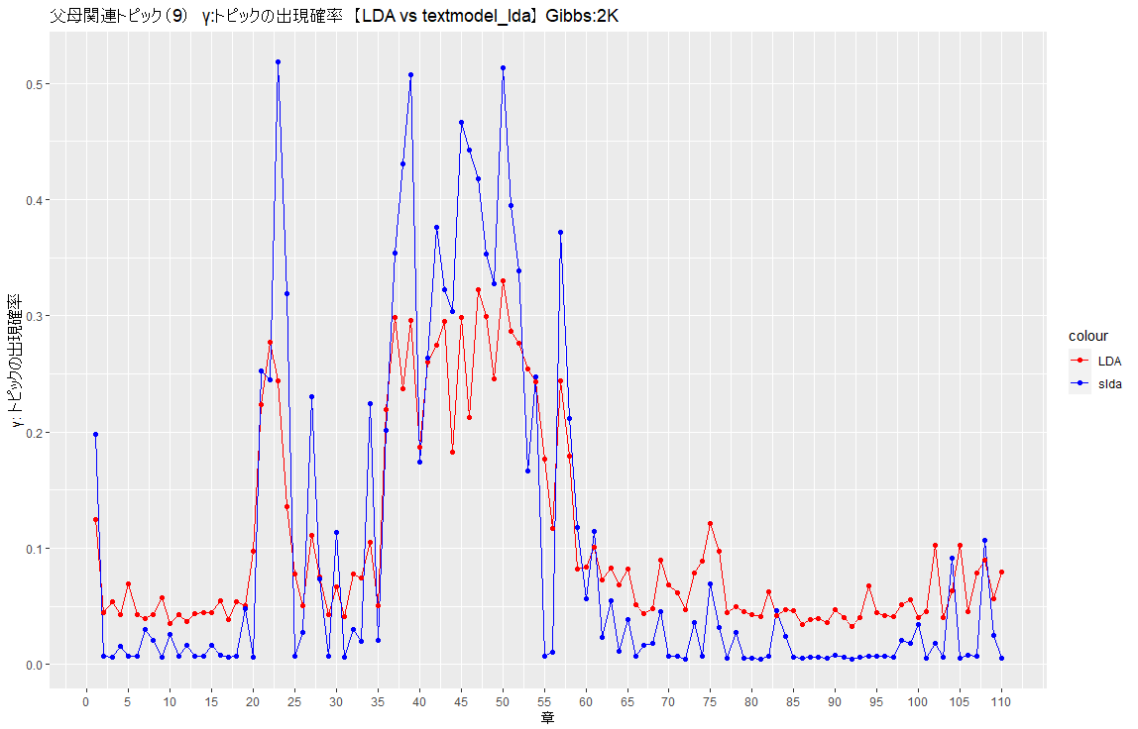


図 5-2 「父・母」関連トピックのトピック出現確率（textmodels と seededlda）

topicmodels パッケージ LDA と seededlda パッケージへの入力データ（DTM）が等価であっても抽出されたトピックや出現確率が異なるので seededlda 開発者に質問したところ、「topicmodels も seededlda も同じ C++コードを利用しており、処理結果の違いは使用する乱数が異なることが原因」との回答を得た。

実際、topicmodels パッケージでは randomMT()関数で疑似一様乱数を生成しているが、seededlda では C++の標準関数 default_random_engine()を使用していた。

topicmodels パッケージ cocus.c

// This is the ``Mersenne Twister'' random number generator MT19937, which
// generates pseudorandom integers uniformly distributed in 0..(2^32 - 1)
// starting from any odd seed in 0..(2^32 - 1). This version is a recode
// by Shawn Cokus (Cokus@math.washington.edu) on March 8, 1998 of a version by
// Takuji Nishimura (who had suggestions from Topher Cooper and Marc Rieffel in
// July-August 1997).
// 以下省略
 	
uint32 randomMT(void)
{

```
    uint32 y;

    if(--left < 0)
        return(reloadMT());

    y  = *next++;
    y ^= (y >> 11);
    y ^= (y <<  7) & 0x9D2C5680U;
    y ^= (y << 15) & 0xEFC60000U;
    y ^= (y >> 18);
    return(y);
}
```

seededlda パッケージ lda.h

```
// LDA model
class LDA {
    public:
        // --- model parameters and variables ---
        int M; // dataset size (i.e., number of docs)
        int V; // vocabulary size
        int K; // number of topics
        double alpha, beta; // LDA hyperparameters
        int niters; // number of Gibbs sampling iterations
        int liter; // the iteration at which the model was saved
        int random; // seed for random number generation
        bool verbose; // print progress messages

        arma::sp_mat data; // transposed document-feature matrix
        arma::vec p; // temp variable for sampling
        Texts z; // topic assignments for words, size M x doc.size()
        arma::umat nw; // cwt[i][j]: number of instances of word/term i assigned to topic j, size V x K
        arma::umat nd; // na[i][j]: number of words in document i assigned to topic j, size M x K
        arma::urowvec nwsum; // nwsum[j]: total number of words assigned to topic j, size K
        arma::ucolvec ndsum; // nasum[i]: total number of words in document i, size M
        arma::mat theta; // theta: document-topic distributions, size M x K
        arma::mat phi; // phi: topic-word distributions, size K x V
```

```
        // prediction with fitted model
        arma::umat nw_ft;
        arma::urowvec nwsum_ft;

        // random number generators
        std::default_random_engine generator;
        std::uniform_real_distribution<double> random_prob;
        std::uniform_int_distribution<int> random_topic;

        // ---------------------------------------
        LDA() {
            set_default_values();
        }
//    以下省略
```

Gibbs サンプリング数を多くすれば、抽出トピックが安定化することが判明しているので、10 トピックで 2K、100K、1M 回での結果を比較する。

●β値の大きな単語を含むトピックで比較

1） 2000 回

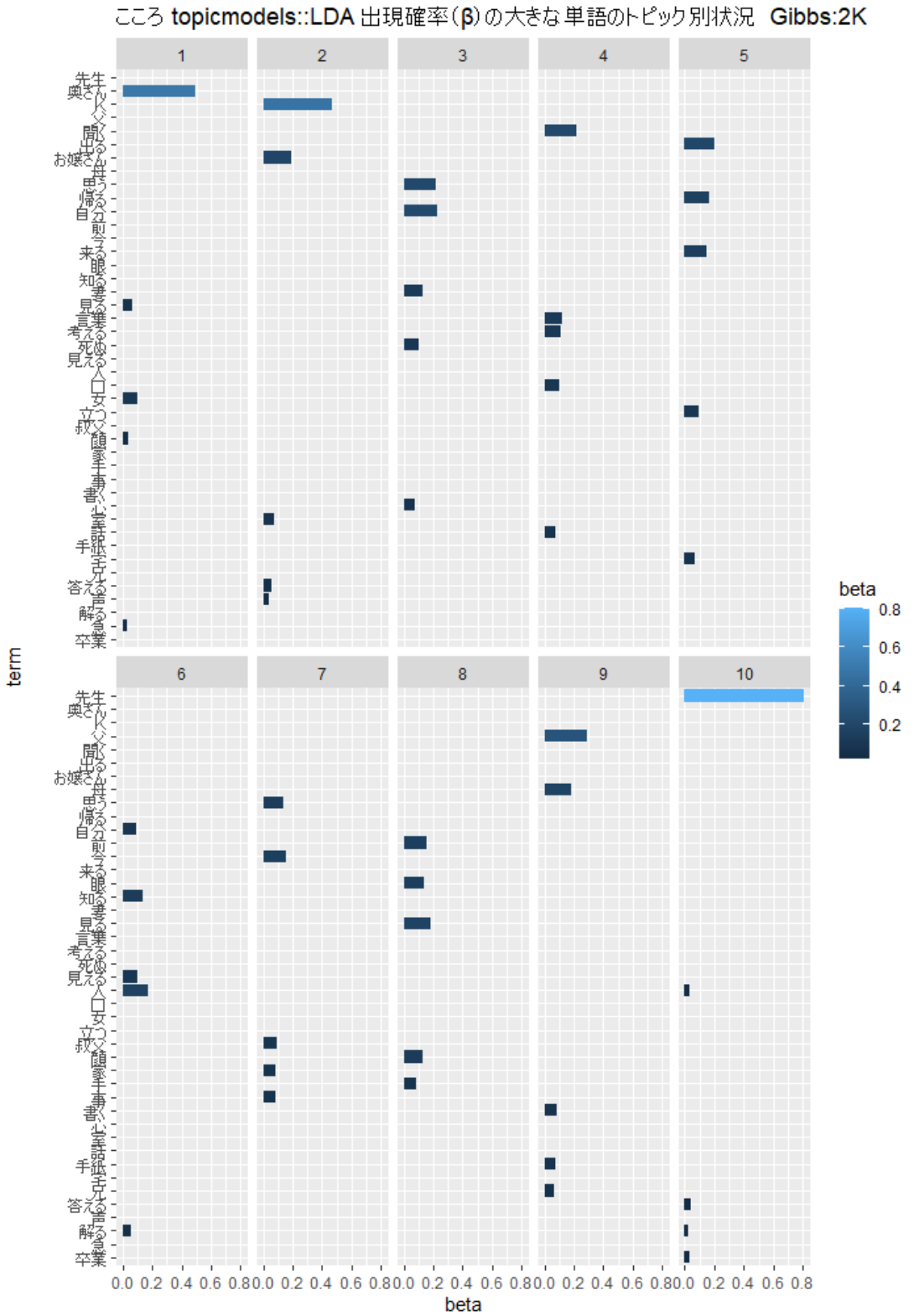


図 5-3 各トピックでの高出現確率 4 語（topicmodels、2000 回）

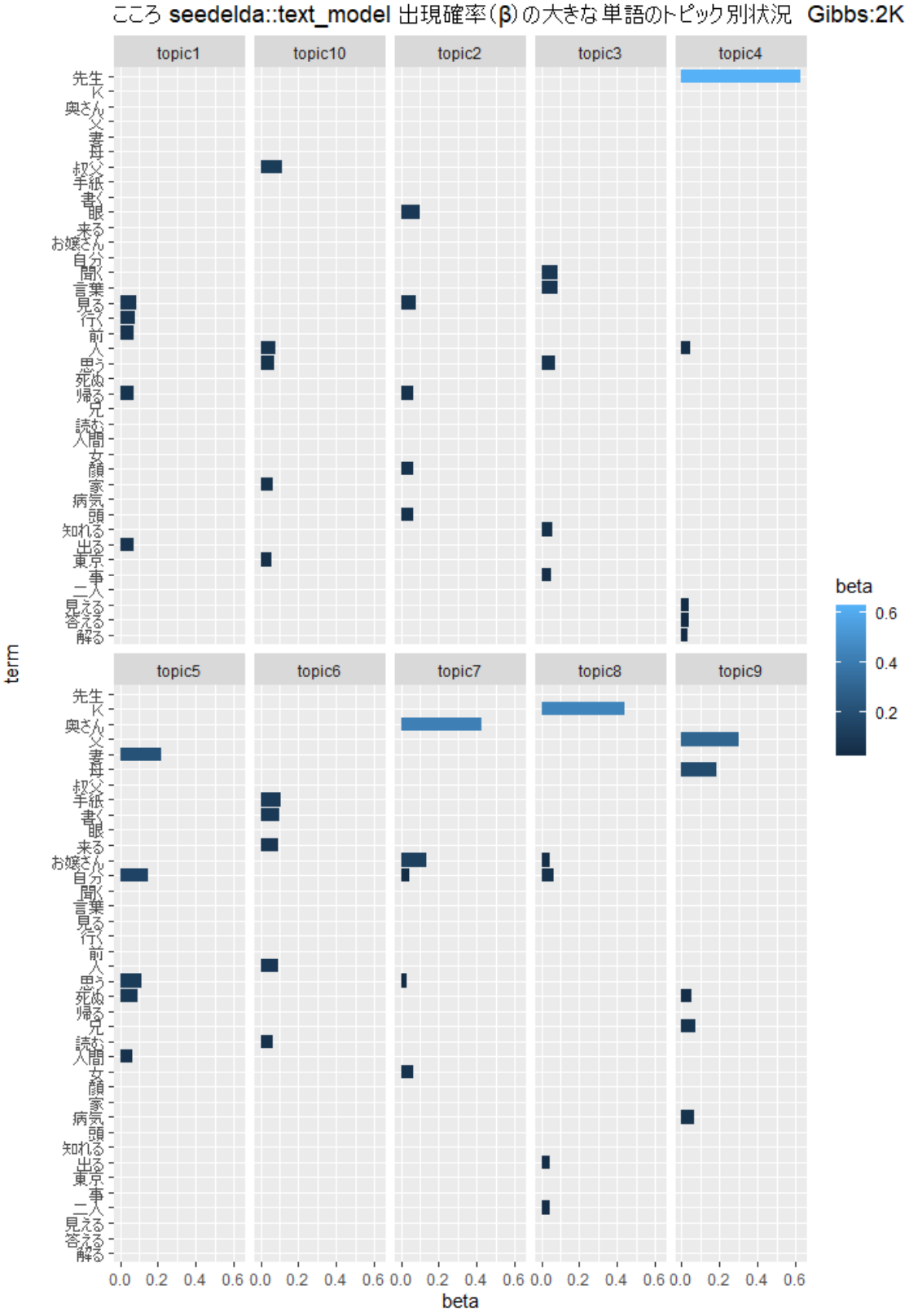


図 5-4 各トピックでの高出現確率 4 語（seededlda、2000 回）

2) 100,000 回

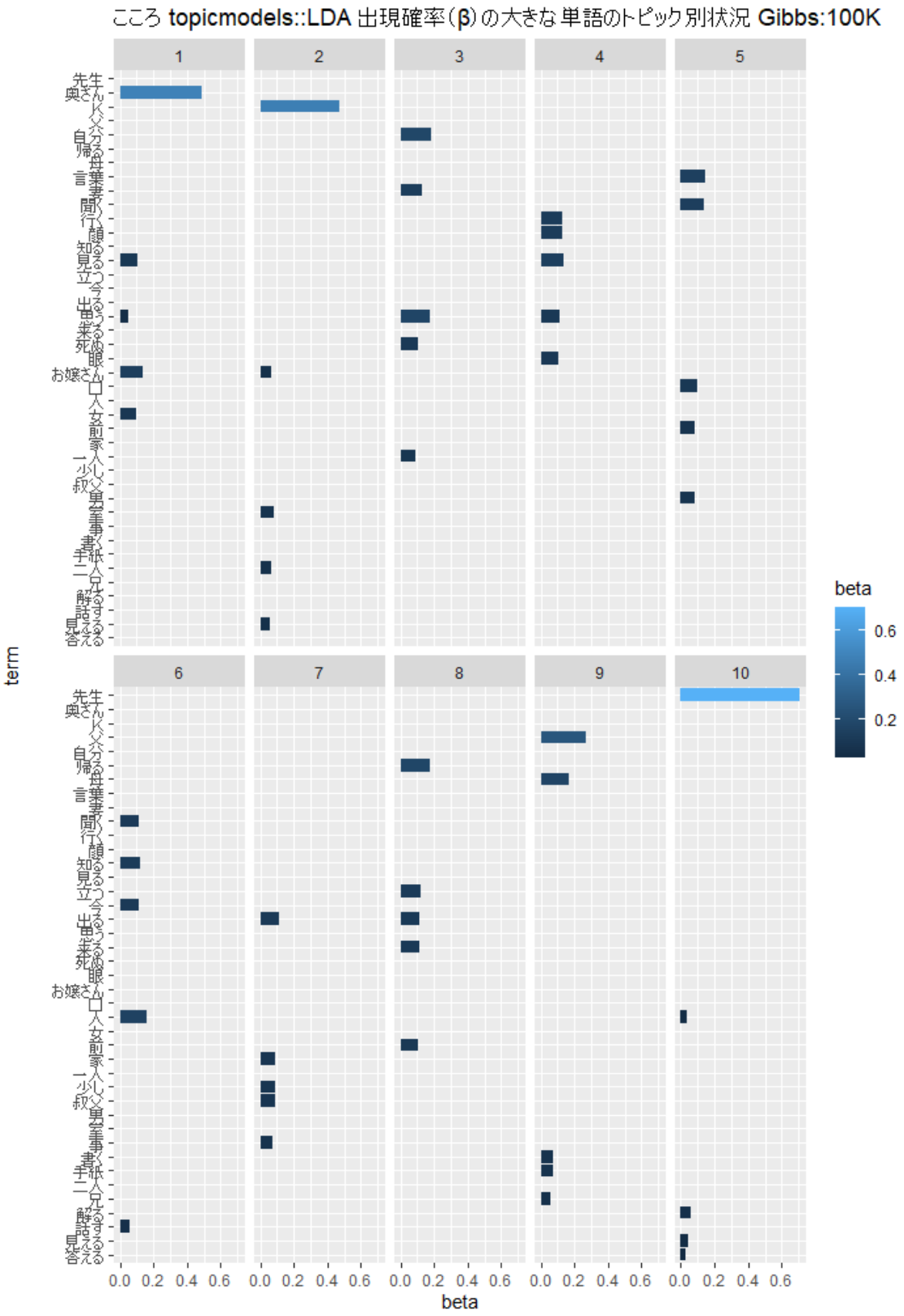


図 5-5 各トピックでの高出現確率 4 語（topicmodels、100,000 回）

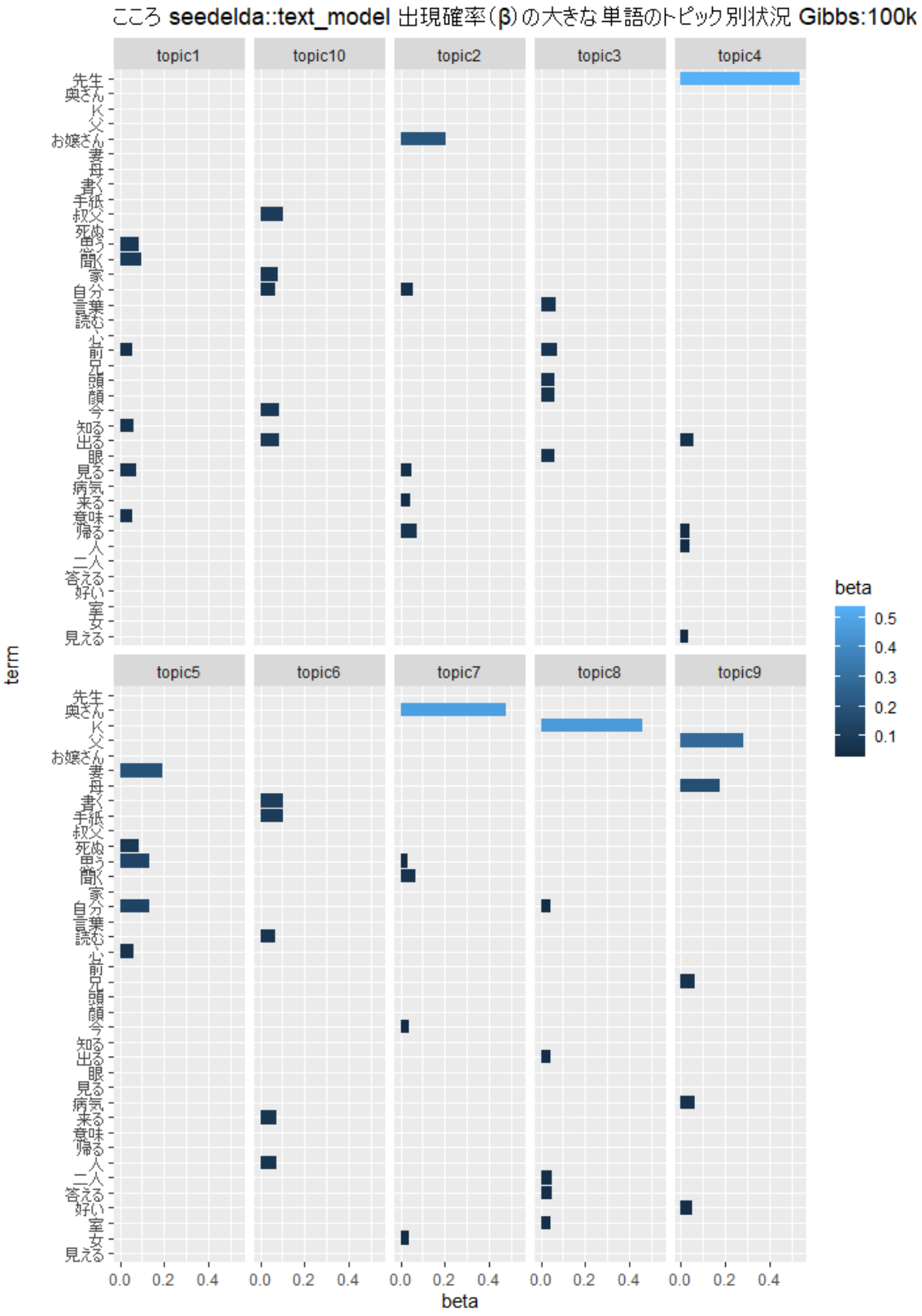


図 5-6 各トピックでの高出現確率 4 語（sseededlda、100,000 回）

3) 1,000,000 回

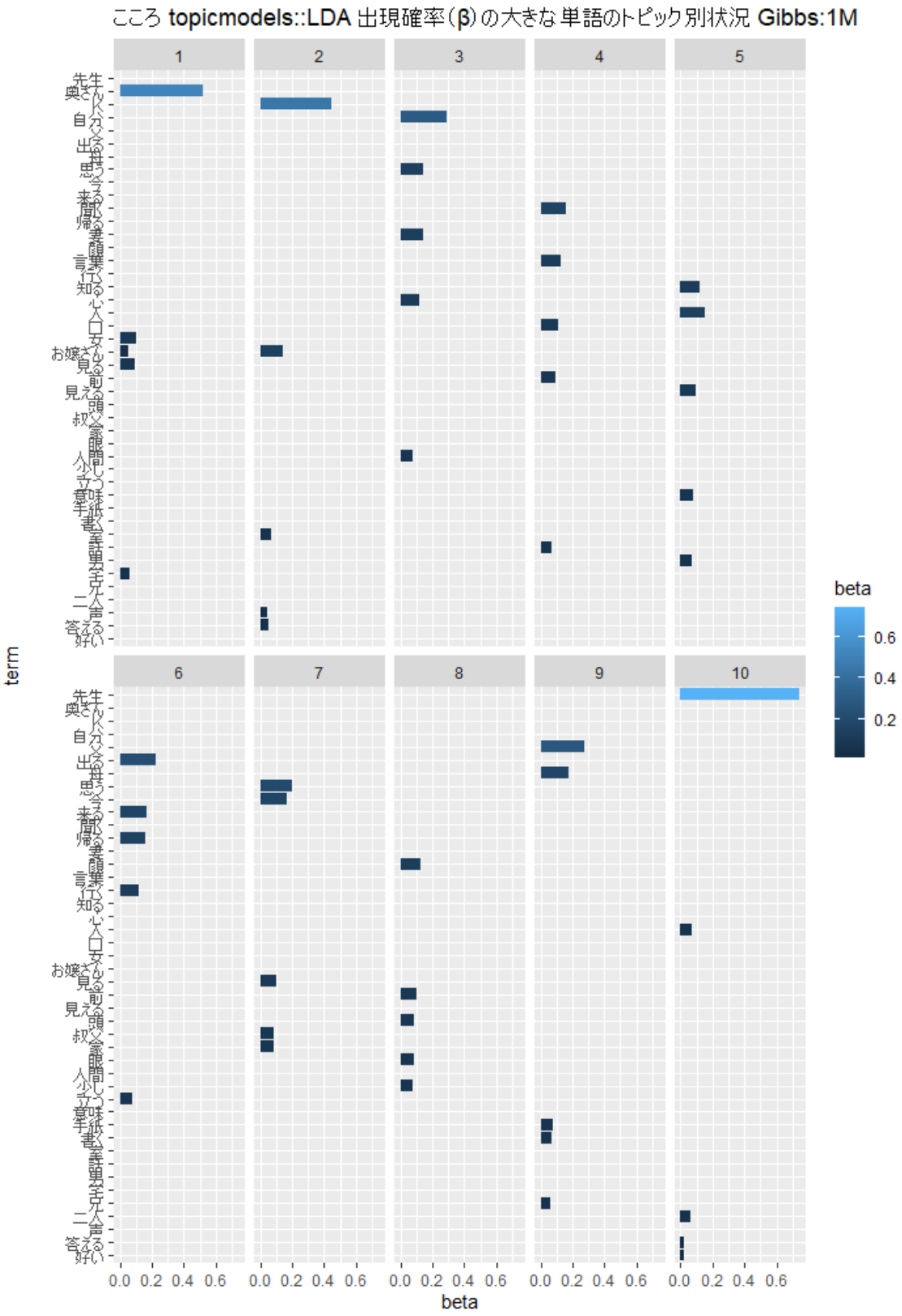


図 5-7 各トピックでの高出現確率 4 語（topicmodels、1,000,000 回）

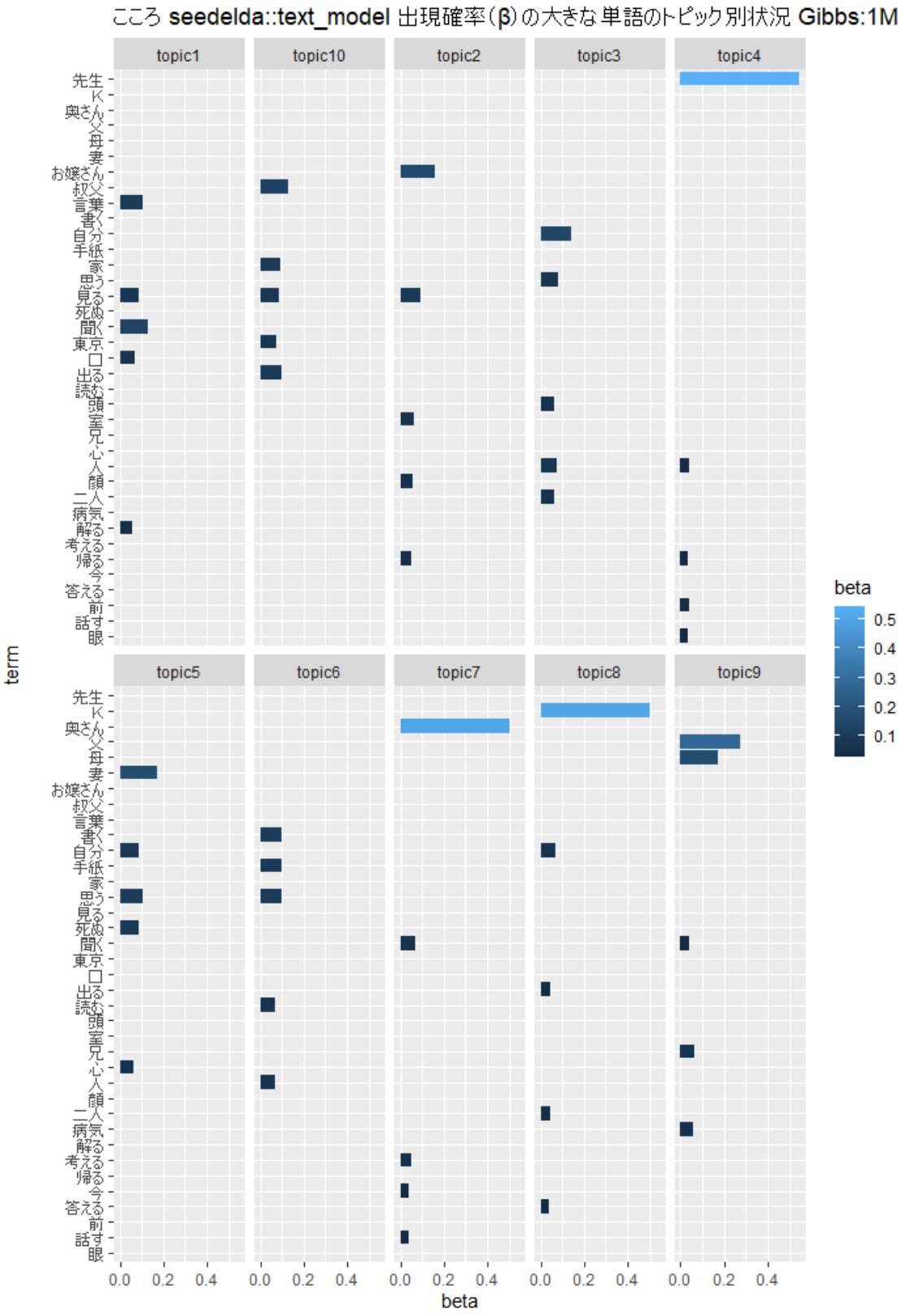


図 5-8 各トピックでの高出現確率 4 語（seedelda、1,000,000 回）

●各文書でのトピック出現確率

Gibbs サンプリング数 2K、100K、1M 回で共通的な高出現確率単語とトピック番号を以下に示す。

出現単語	先生	奥さん	K	父、母	自分、思う、妻
topicmodels	10	1	2	9	3
seededlda	4	7	8	9	5

既に 2K ステップで比較した「K」と「父・母」に加えて、「先生」、「奥さん」、「自分・妻・思う」についても 100K と 1M で対応するトピック間で比較する。

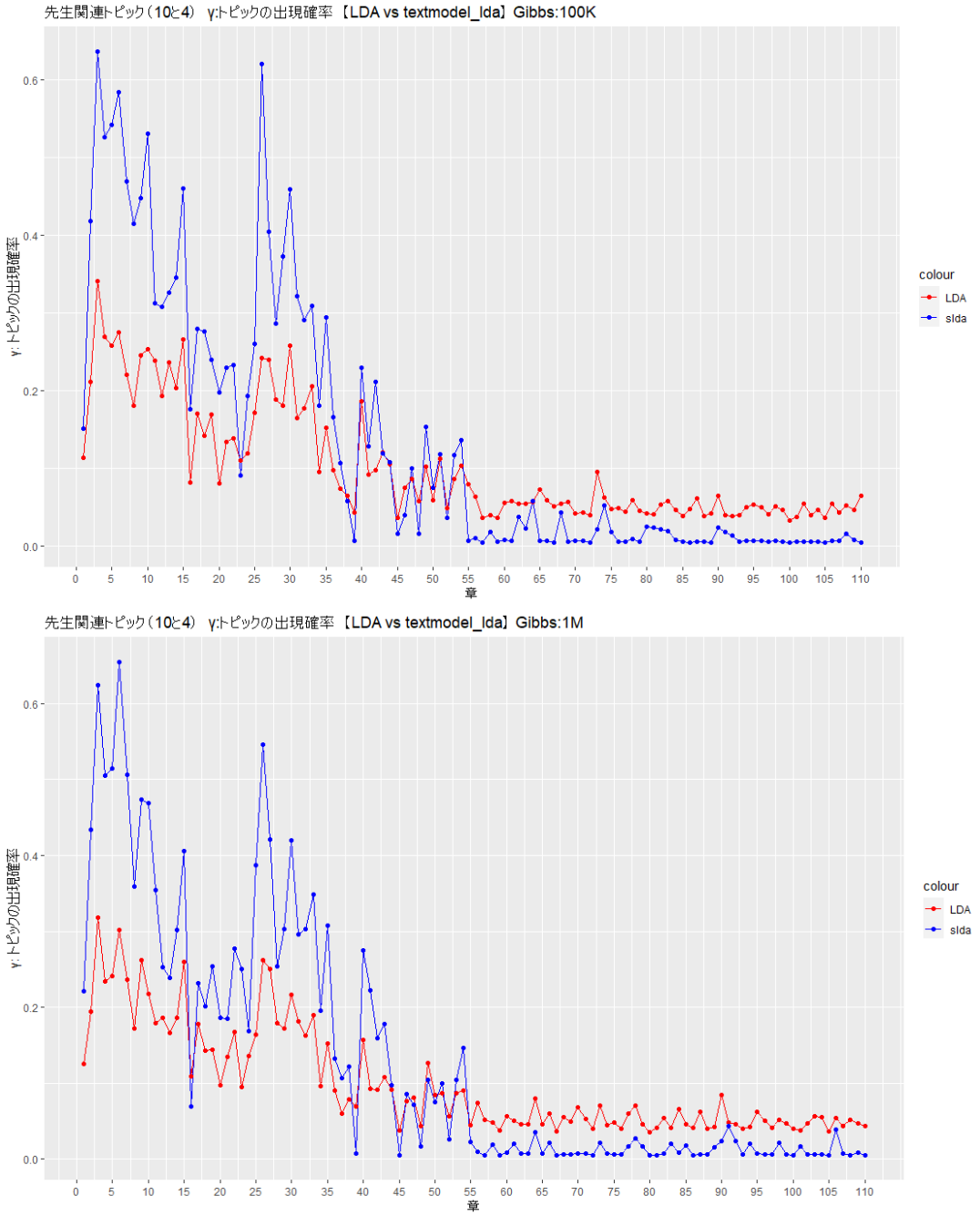
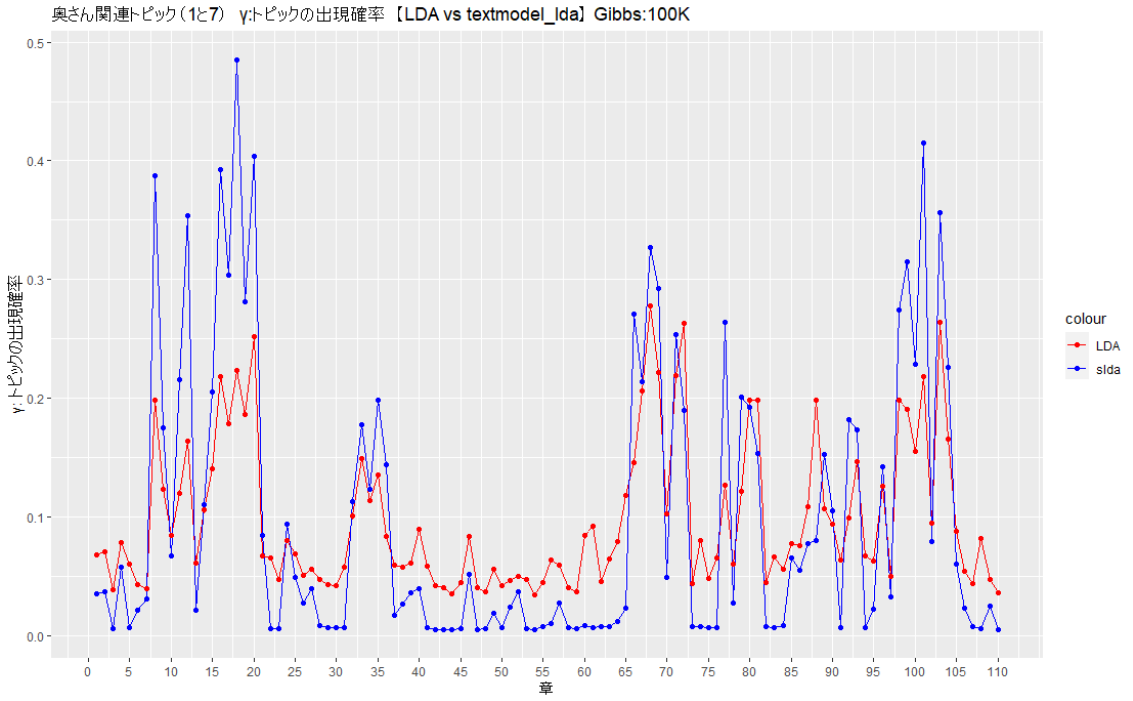


図 5-9 「先生」関連トピックのトピック出現確率（textmodels と seededlda、100K と 1M 回）



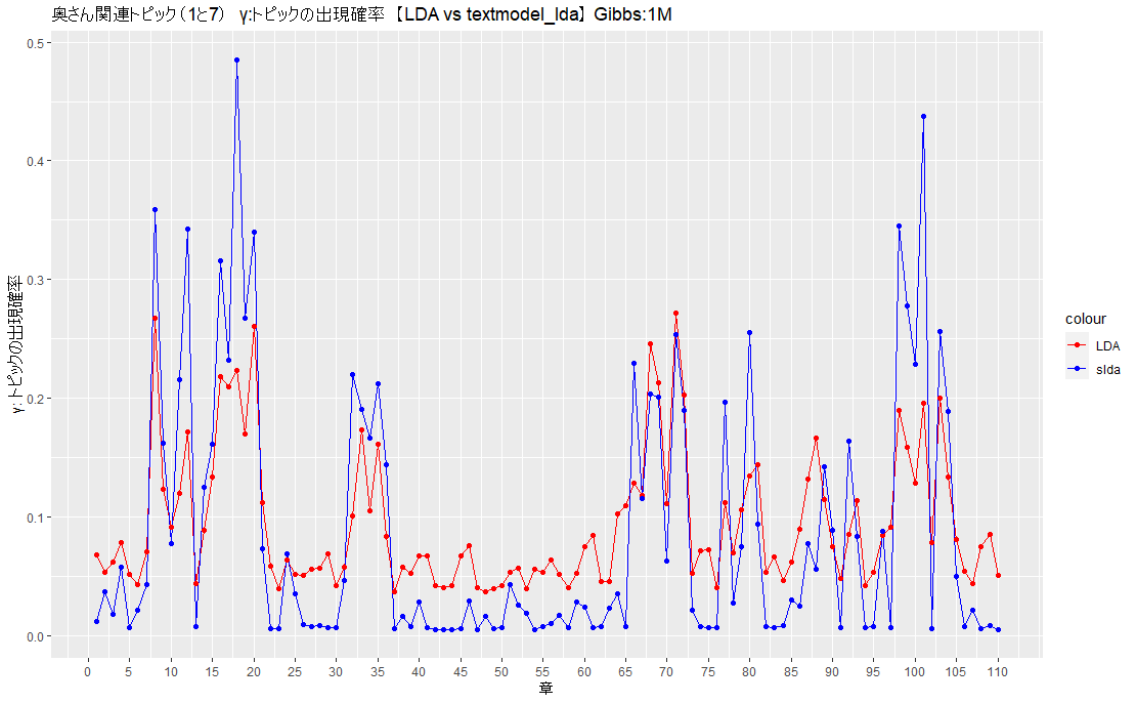


図 5-10 「奥さん」関連トピックのトピック出現確率（textmodels と seededlda、100K と 1M 回）

図 5-11 「K」関連トピックのトピック出現確率（textmodels と seededlda、100K と 1M 回）

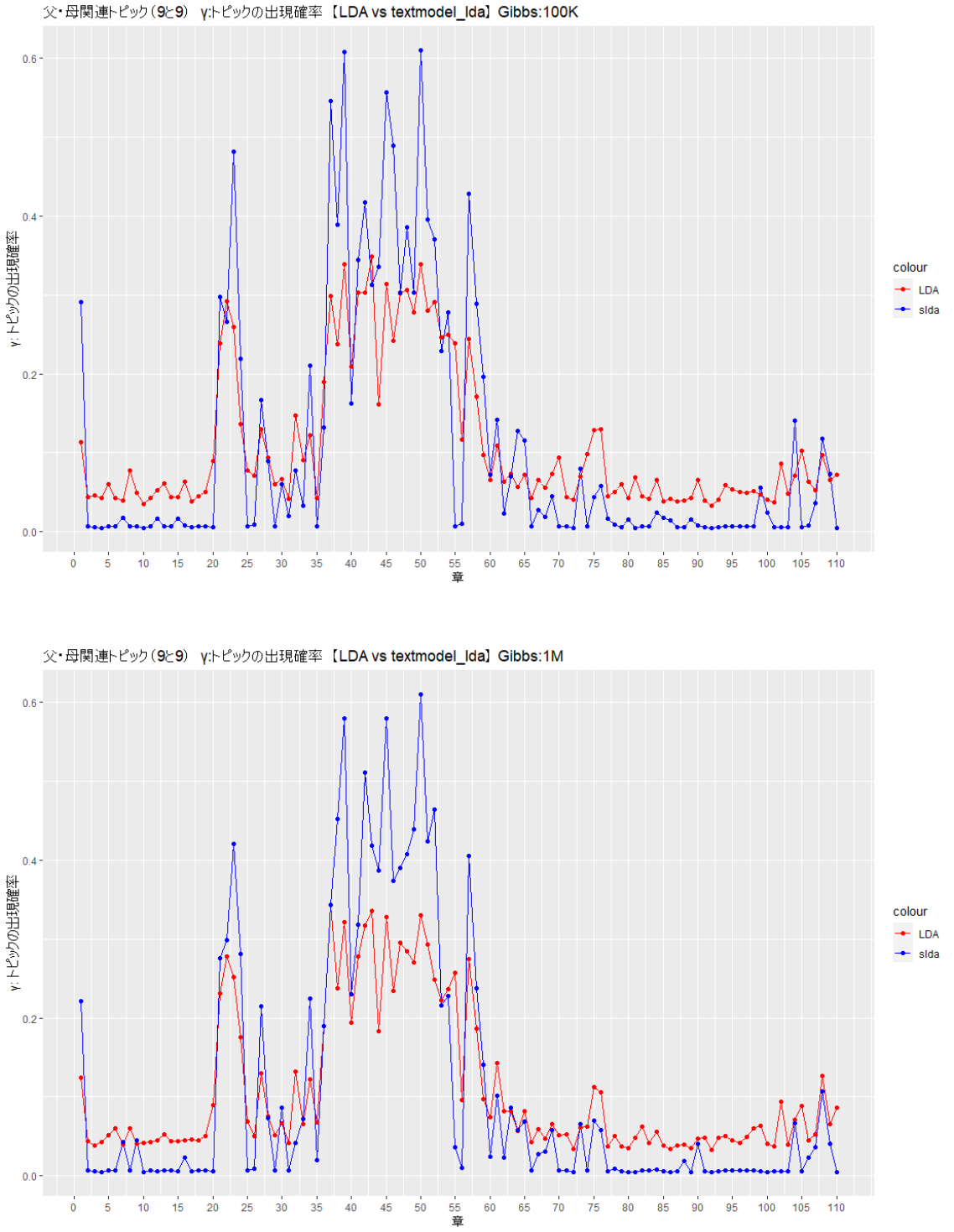


図 5-12 「父・母」関連トピックのトピック出現確率（textmodels と seededlda、100K と 1M 回）

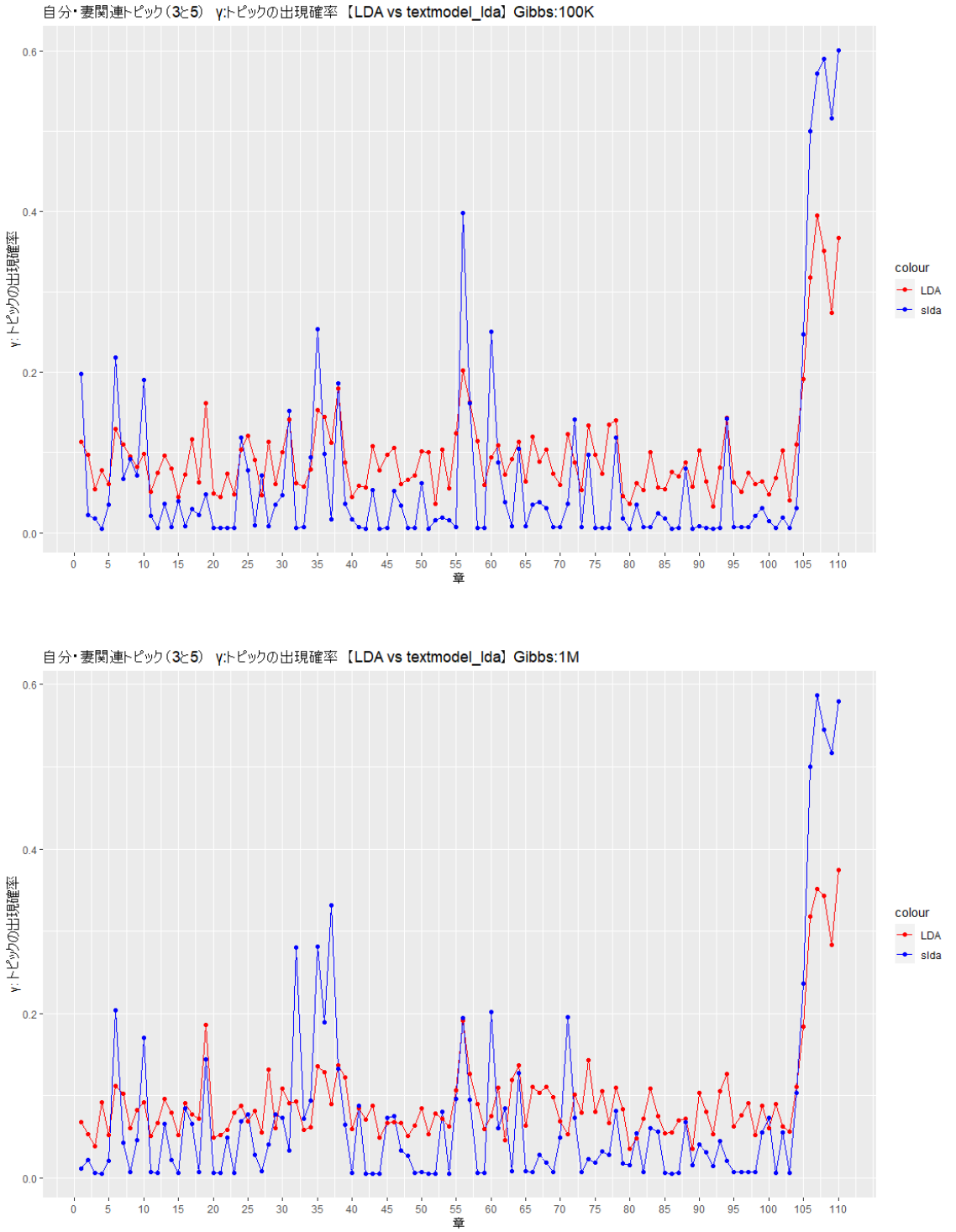


図 5-13 「自分・妻」関連トピックのトピック出現確率
（textmodels と seededlda、100K と 1M 回）

これらの図から、seededlda パッケージの textmodel_lda では、トピックが低出現確率の場合はゼロ付近であり LDA に比べ極端に小さい、一方トピックの出現確率が高い場合には、LDA の 1.6 倍以上と

なっており、γの計算式に違いがあると思われる。

例えば、上記 textmodel_lda で 1M ステップの「自分・妻」のファイ値（Φ）を exp(Φ-1.2)-0.25 と変換すると、雰囲気が似てくる。よって、seededlda でのファイ値の算出方法を更に調べる必要がある。

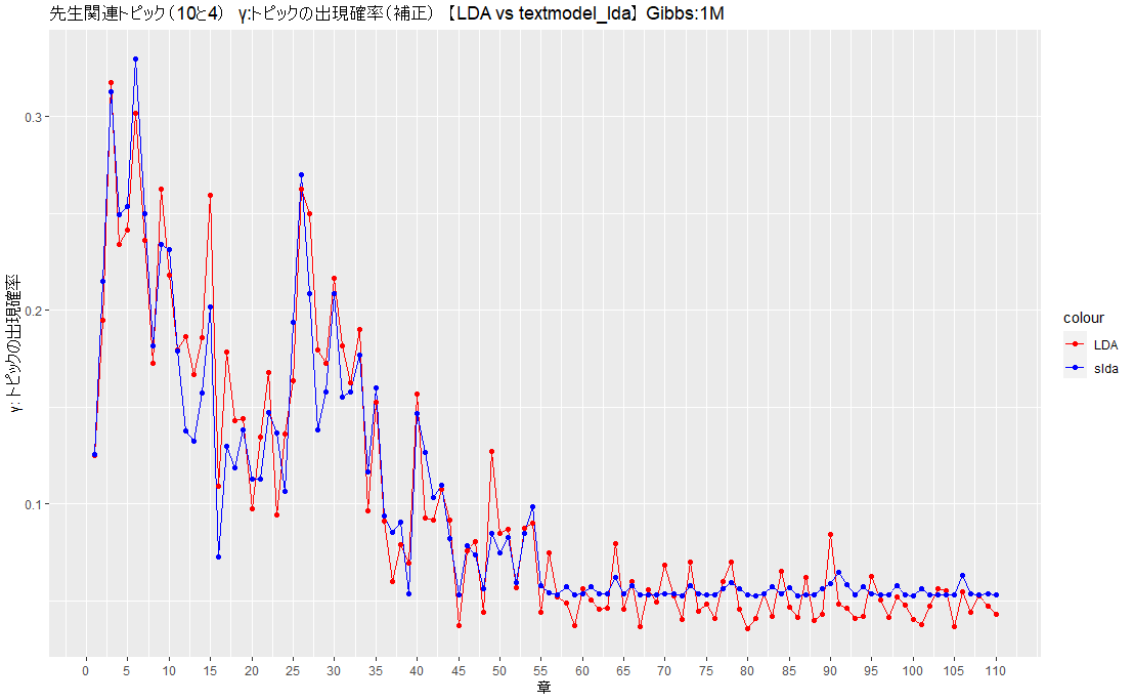


図 5-14 「自分・妻」関連トピックのトピック出現確率を補正した場合
（textmodels と seededlda、100K と 1M 回）

付録 12 環境・水産・海洋白書（2008～2020）教師付き LDA 分析処理結果

（その 1）

文書番号：JRDN-21-033

（注）パワーポイントドキュメントの提出資料を編集。

単語出現頻度と処理パラメータ(5類辞書)

トピックの単語	環境白書	海洋白書	水産白書	環境・海洋・水産白書	海洋基本計画
	2008～2020	2004～2020	2007～2020	2008～2020	3期
気候変動	1759	447	77	2228	54
生物多様性	3120	422	51	3559	60
温暖化 温室効果ガス	2136 1840	350 45	71 15	2474 1894	37 13
水産物	56	102	2111	2177	31
エネルギー 再生可能エネルギー	2408 845	700 189	29 22	3075 1056	92 35

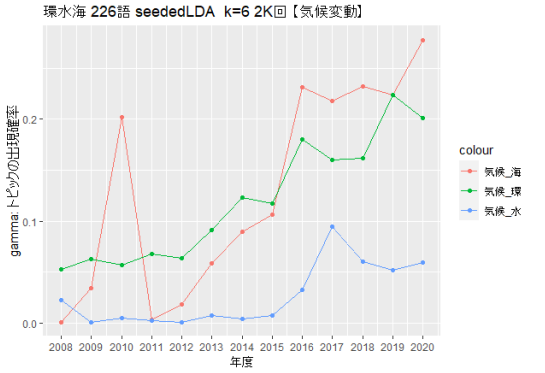
```
dict <- dictionary( list( "気候変動"=c("気候変動"),          "生物多様性"=c("生物多様性"),
                          "温暖化"=c("温暖化","温室効果ガス"), "水産物"=c("水産物"),
                          "エネルギー"=c("エネルギー","再生可能エネルギー") ) )
```

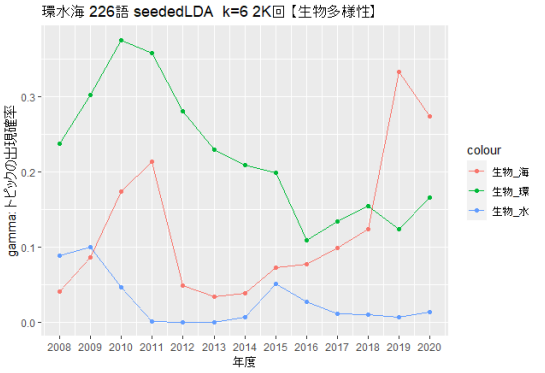
項　目	環境白書	海洋白書	水産白書	環境・海洋・水産白書	海洋基本計画
	2008～2020	2004～2020	2007～2020	2008～2020	3期
最小出現頻度	56	45	15	1056	13
分析対象語数	1995	1501	2212	**226**	566
LDA tuning結果 推定トピック数	18	14	14, 18	20	12
LDA Gibbsサンプリング数		200K以上		50K以上	
教師付きLDAトピック数	6	6	6	6	6

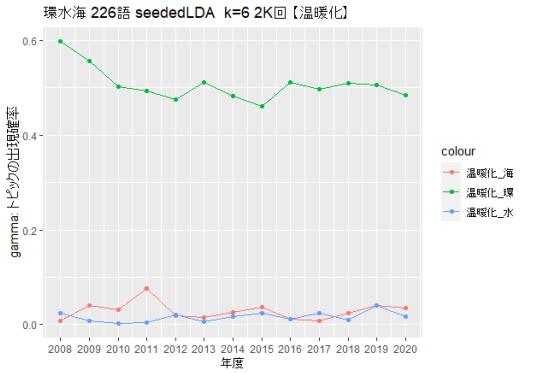
関心の高いキーワードとして154語の中、1）気候変動　2）生物多様性　3）温暖化、温室効果ガス　4）水産物　5）エネルギー　と指定されたが、
環境・海洋・水産白書では、「再生可能エネルギー」の出現頻度が結構高いので、分析対象語数を226とした場合の分析を行った。

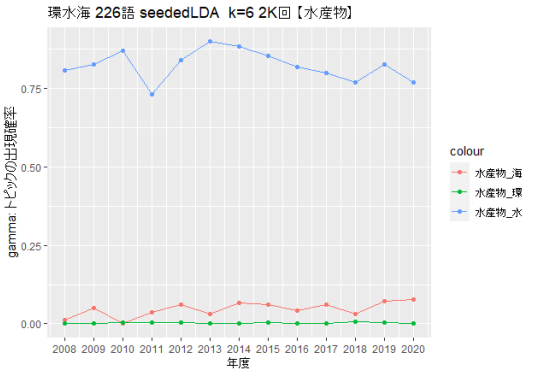
■ 226 語　2K 回

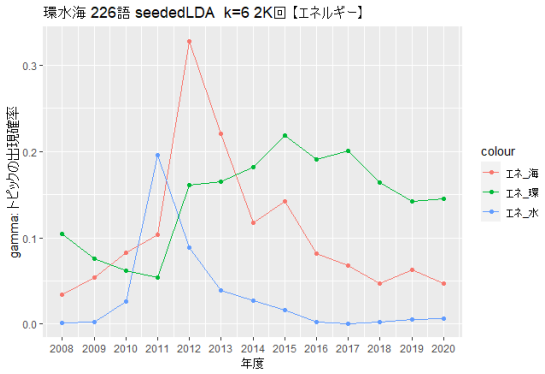
```
> seededlda::terms(環水海226_seededLDA,10)
      気候変動   生物多様性   温暖化     水産物     エネルギー           other
 [1,] "気候変動" "生物多様性" "温暖化"   "水産物"   "エネルギー"         "管理"
 [2,] "情報"     "利用"       "温室効果ガス" "漁船" "再生可能エネルギー" "計画"
 [3,] "技術"     "保全"       "実施"     "漁獲"     "施設"             "基本"
 [4,] "開催"     "活動"       "廃棄物"   "水産"     "発電"             "開発"
 [5,] "実施"     "世界"       "処理"     "漁業者"   "支援"             "海域"
 [6,] "持続可能" "社会"       "対策"     "養殖"     "調査"             "教育"
 [7,] "CO2"      "資源"       "計画"     "管理"     "活動"             "政策"
 [8,] "連携"     "経済"       "利用"     "減少"     "原子力"           "必要"
 [9,] "影響"     "技術"       "整備"     "資源"     "被害"             "利用"
[10,] "対策"     "日本"       "社会"     "操業"     "実施"             "沿岸域"
>
```

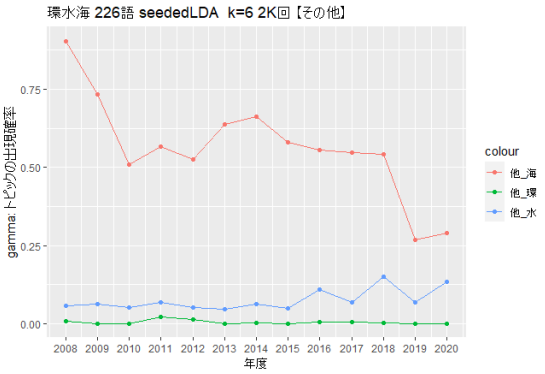






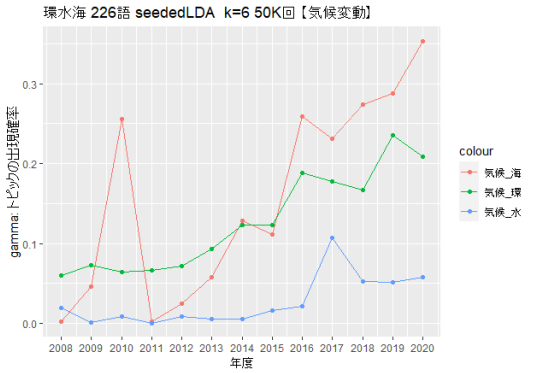


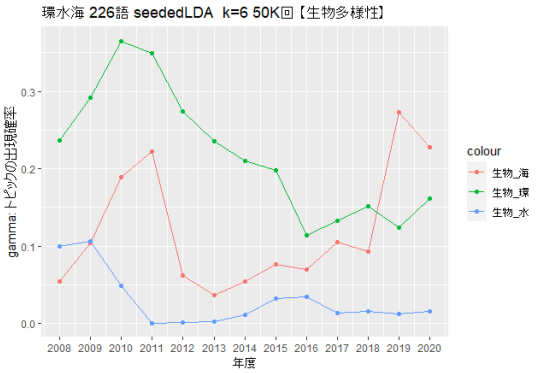


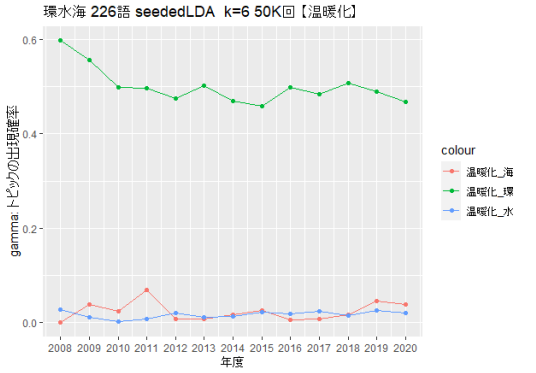


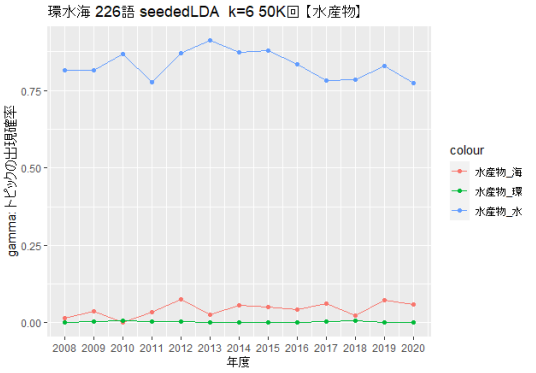
■ 226 語 50K 回

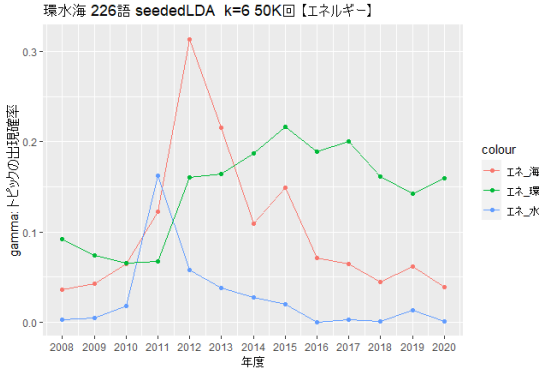
```
> seededlda::terms(環水海226_seededLDA_50k,10)
      気候変動   生物多様性   温暖化       水産物   エネルギー           other
 [1,] "気候変動" "生物多様性" "温暖化"   "水産物" "エネルギー"       "管理"
 [2,] "情報"     "利用"       "温室効果ガス" "漁船" "再生可能エネルギー" "計画"
 [3,] "開催"     "世界"       "実施"     "漁獲"   "施設"           "基本"
 [4,] "実施"     "経済"       "廃棄物"   "水産"   "調査"           "開発"
 [5,] "技術"     "活動"       "処理"     "漁業者" "発電"           "海域"
 [6,] "影響"     "資源"       "対策"     "管理"   "実施"           "教育"
 [7,] "持続可能" "目標"       "計画"     "養殖"   "原子力"         "政策"
 [8,] "目標"     "保全"       "利用"     "資源"   "支援"           "利用"
 [9,] "対策"     "開発"       "整備"     "減少"   "被害"           "沿岸域"
[10,] "調査"     "社会"       "社会"     "操業"   "活動"           "必要"
```

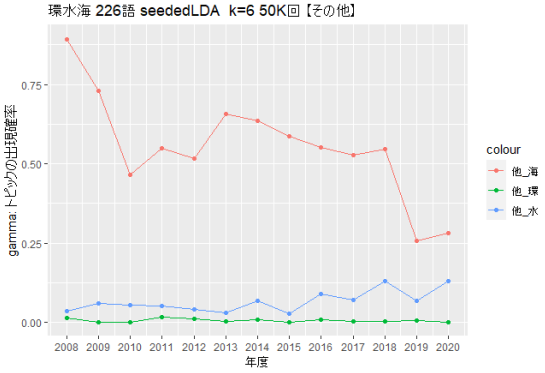






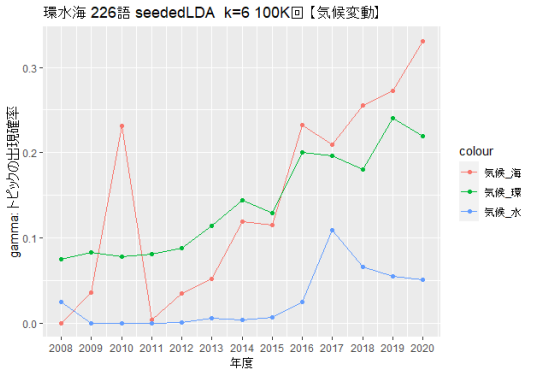


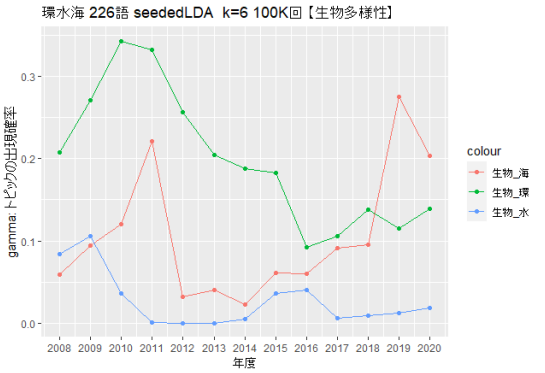


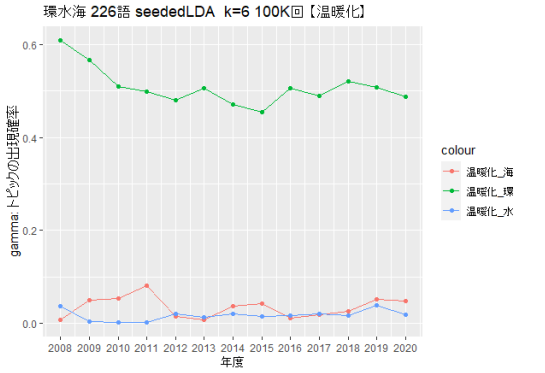


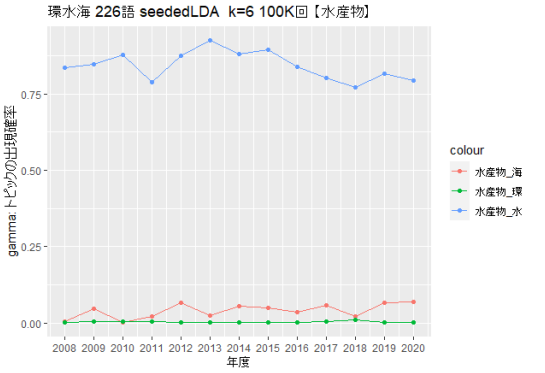
■ 226語 100K回

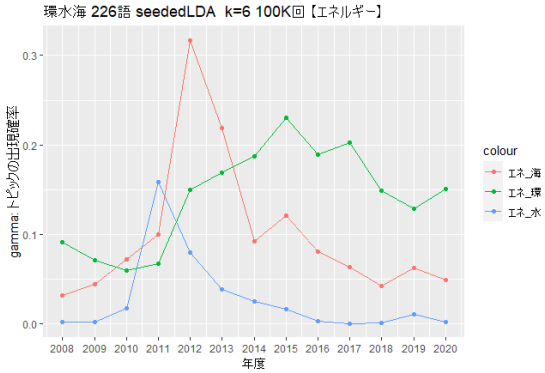
```
> seededlda::terms(環水海226_seededLDA_100k,10)↓
      気候変動   生物多様性   温暖化       水産物     エネルギー       other
 [1,] "気候変動" "生物多様性" "温暖化"   "水産物"   "エネルギー"     "管理"
 [2,] "情報"     "利用"       "温室効果ガス" "漁船"   "再生可能エネルギー" "計画"
 [3,] "技術"     "保全"       "廃棄物"   "漁獲"     "施設"           "開発"
 [4,] "開催"     "活動"       "処理"     "水産"     "支援"           "基本"
 [5,] "実施"     "経済"       "実施"     "漁業者"   "実施"           "海域"
 [6,] "影響"     "世界"       "対策"     "管理"     "発電"           "教育"
 [7,] "目標"     "社会"       "計画"     "養殖"     "原子力"         "政策"
 [8,] "連携"     "資源"       "利用"     "減少"     "活動"           "必要"
 [9,] "世界"     "研究"       "整備"     "資源"     "調査"           "利用"
[10,] "持続可能" "技術"       "資源"     "操業"     "被害"           "沿岸域"
```

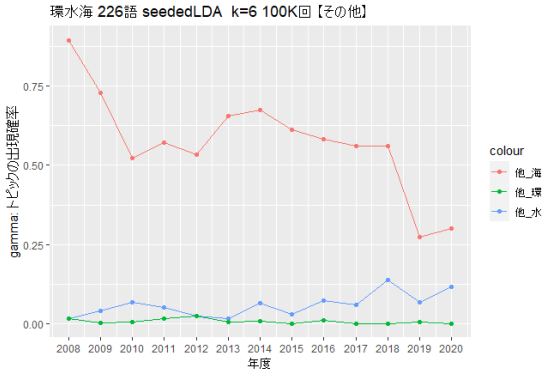






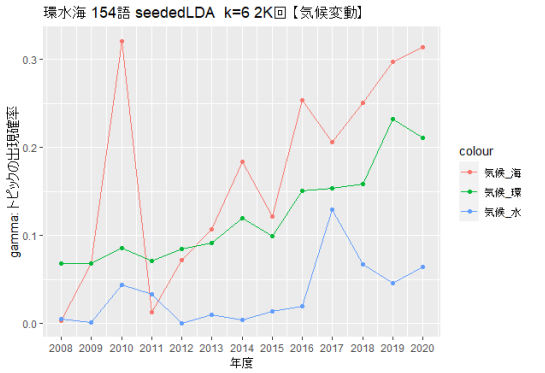


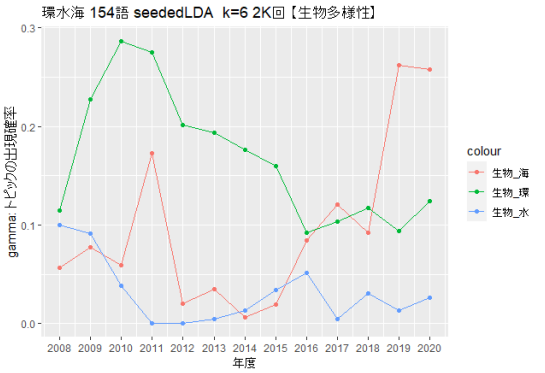


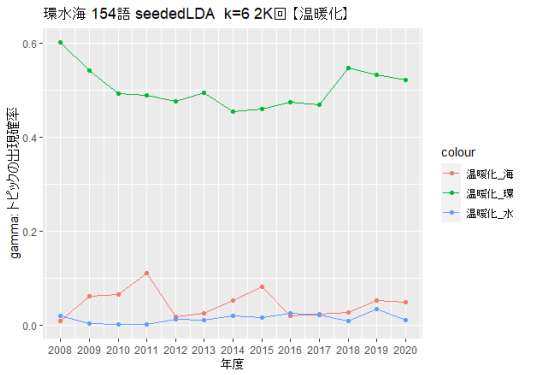


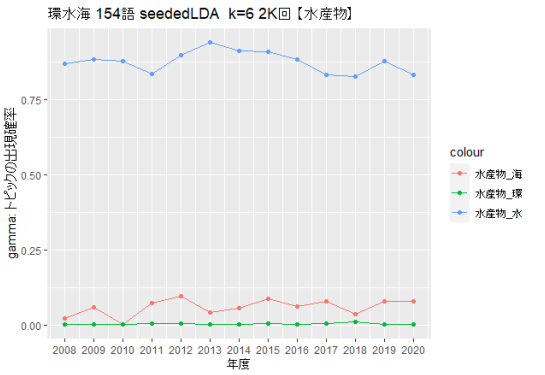
■ 154 語 2K 回 （KH Coder での Gibbs サンプリング数）

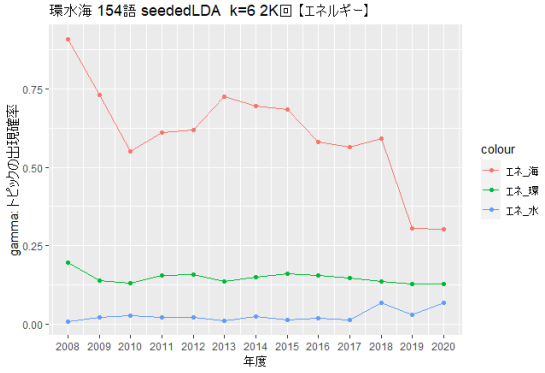
```
> seededlda::terms(環水海2_seededLDA,10)↓
      気候変動   生物多様性   温暖化     水産物     エネルギー   other      ↓
 [1,] "気候変動" "生物多様性" "温暖化"   "水産物"   "エネルギー" "原子力"   ↓
 [2,] "情報"     "利用"       "温室効果ガス" "漁船"   "管理"       "実施"     ↓
 [3,] "実施"     "保全"       "廃棄物"   "漁獲"     "計画"       "発電"     ↓
 [4,] "技術"     "世界"       "処理"     "水産"     "開発"       "施設"     ↓
 [5,] "開催"     "研究"       "実施"     "管理"     "基本"       "規制"     ↓
 [6,] "影響"     "経済"       "対策"     "漁業者"   "政策"       "被害"     ↓
 [7,] "連携"     "社会"       "利用"     "養殖"     "教育"       "委員会"   ↓
 [8,] "調査"     "生態系"     "整備"     "資源"     "必要"       "支援"     ↓
 [9,] "支援"     "目標"       "社会"     "減少"     "海域"       "調査"     ↓
[10,] "観測"     "活動"       "資源"     "生産"     "利用"       "汚染"     ↓
```

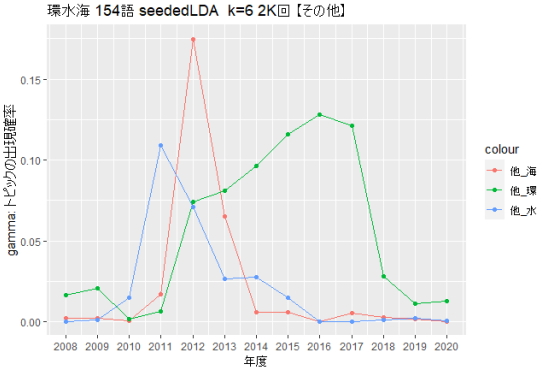






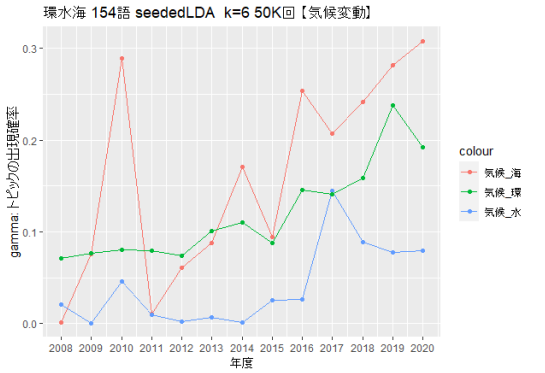


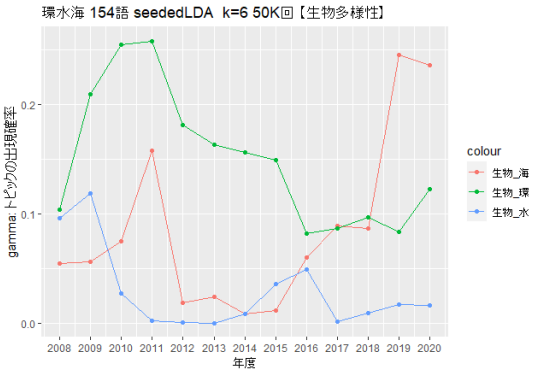


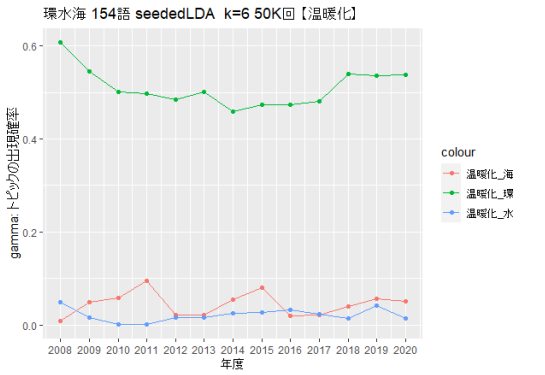


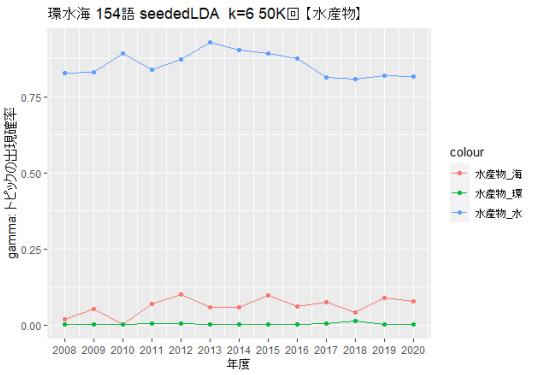
■ 154 語 50K 回

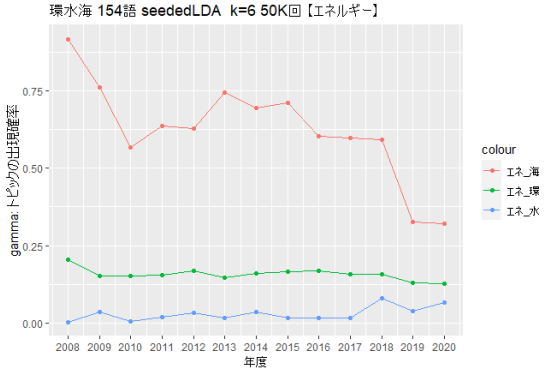
```
> seededlda::terms(環水海2_seededLDA_50k,10)↓
      気候変動   生物多様性   温暖化     水産物     エネルギー   other      ↓
 [1,] "気候変動" "生物多様性" "温暖化" "水産物" "エネルギー" "原子力" ↓
 [2,] "情報"     "利用"       "温室効果ガス" "漁船" "管理"     "実施"     ↓
 [3,] "実施"     "保全"       "廃棄物"   "漁獲"     "計画"     "発電"     ↓
 [4,] "技術"     "世界"       "実施"     "水産"     "開発"     "施設"     ↓
 [5,] "調査"     "目標"       "処理"     "漁業者"   "基本"     "規制"     ↓
 [6,] "影響"     "生態系"     "対策"     "管理"     "必要"     "被害"     ↓
 [7,] "開催"     "技術"       "社会"     "養殖"     "政策"     "調査"     ↓
 [8,] "連携"     "活動"       "利用"     "資源"     "海域"     "委員会"   ↓
 [9,] "目標"     "資源"       "整備"     "減少"     "教育"     "支援"     ↓
[10,] "活用"     "研究"       "排出"     "生産"     "利用"     "汚染"     ↓
```

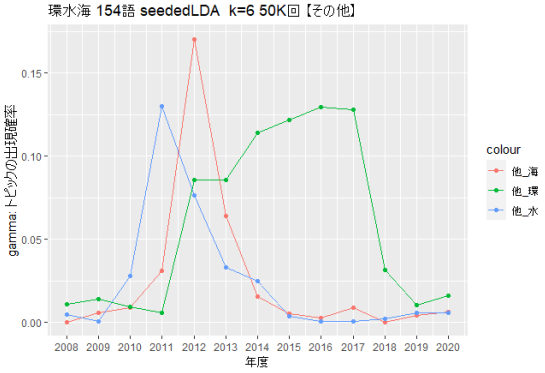






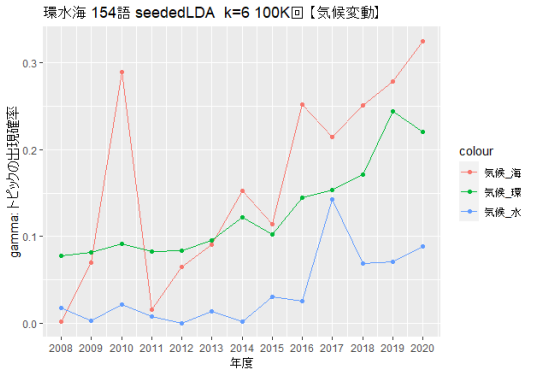


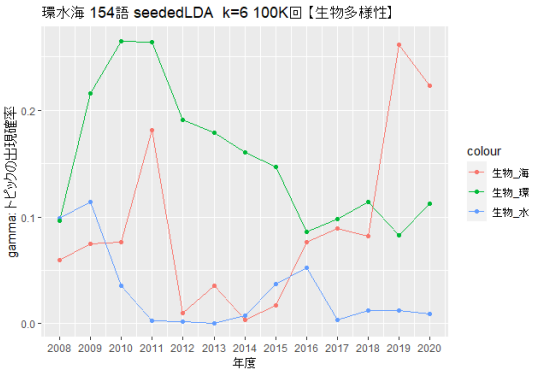


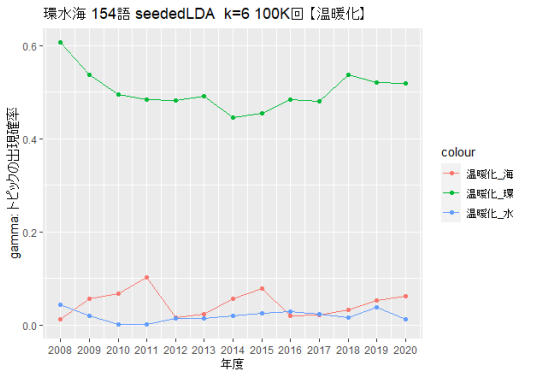


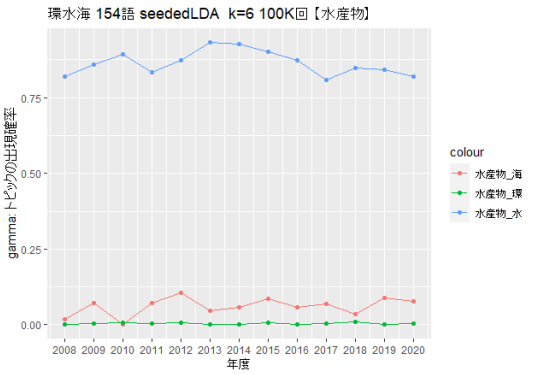
■ 154 語 100K 回

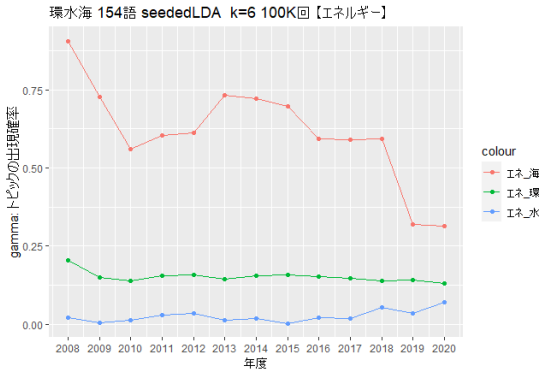
```
> seededlda::terms(環水海2_seededLDA_100k,10)↓
      気候変動     生物多様性   温暖化       水産物     エネルギー   other
 [1,] "気候変動"   "生物多様性" "温暖化"     "水産物"   "エネルギー" "原子力"
 [2,] "情報"       "保全"       "温室効果ガス" "漁船"   "管理"       "実施"
 [3,] "実施"       "利用"       "廃棄物"     "漁獲"     "計画"       "発電"
 [4,] "技術"       "資源"       "処理"       "水産"     "開発"       "規制"
 [5,] "開催"       "世界"       "実施"       "管理"     "基本"       "施設"
 [6,] "影響"       "活動"       "対策"       "漁業者"   "政策"       "被害"
 [7,] "連携"       "生態系"     "利用"       "養殖"     "海域"       "調査"
 [8,] "調査"       "研究"       "社会"       "資源"     "必要"       "委員会"
 [9,] "支援"       "目標"       "整備"       "減少"     "教育"       "検討"
[10,] "活用"       "社会"       "排出"       "生産"     "利用"       "汚染"
```

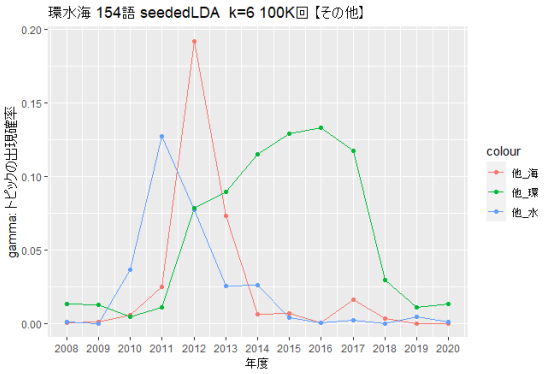

環水海 154語 seededLDA k=6 100K回【気候変動】
gamma:トピックの出現確率
年度
colour
気候_海
気候_環
気候_水


環水海 154語 seededLDA k=6 100K回【生物多様性】
gamma:トピックの出現確率
年度
colour
生物_海
生物_環
生物_水


環水海 154語 seededLDA k=6 100K回【温暖化】
gamma:トピックの出現確率
年度
colour
温暖化_海
温暖化_環
温暖化_水


環水海 154語 seededLDA k=6 100K回【水産物】
gamma:トピックの出現確率
年度
colour
水産物_海
水産物_環
水産物_水


環水海 154語 seededLDA k=6 100K回【エネルギー】
gamma:トピックの出現確率
年度
colour
エネ_海
エネ_環
エネ_水


環水海 154語 seededLDA k=6 100K回【その他】
gamma:トピックの出現確率
年度
colour
他_海
他_環
他_水

付録 13 環境･水産･海洋白書（2008～2020）教師付き LDA 分析処理結果（その２）

文書番号：JRDN-21-034

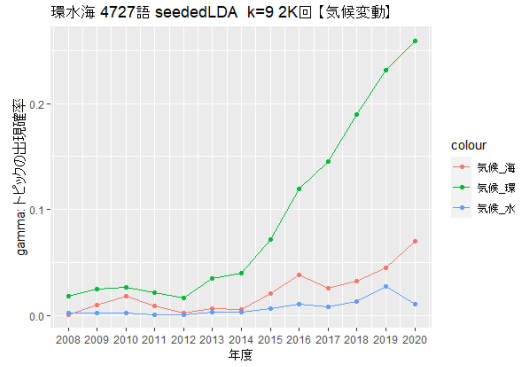
（注）パワーポイントドキュメントの提出資料を編集。

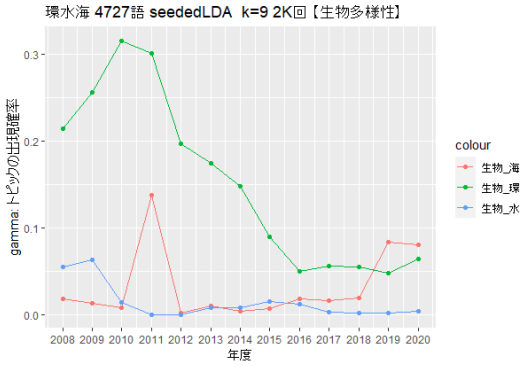
トピック名	トピックの単語	環境白書	海洋白書	水産白書	環境･海洋･水産白書	海洋基本計画
		2008～2020	2004～2020	2007～2020	2008～2020	3期
気候変動	気候変動	1759	447	77	2228	54
プラスチック	海洋プラスチック	12	13	0	**25**	0
	プラスチックごみ	139	42	22	203	0
	プラスチック	578	121	28	726	**1**
	マイクロプラスチック	61	31	15	107	6
生物多様性	生物多様性	3120	422	51	3559	60
水産資源	水産資源	102	165	571	786	59
	漁業資源	19	74	152	227	3
	水産物	56	102	2111	2177	31
地球温暖化	温暖化	2136	350	71	2474	37
	二酸化炭素	727	182	35	887	17
	温室効果ガス	1840	45	15	1894	13
	CO2	1014	144	**2**	1160	**1**
	低炭素	1104	**7**	0	1111	2
地震災害	地震	184	594	102	833	66
	津波	162	723	184	961	112
	防災	329	326	43	658	40
	復興	473	286	330	1085	8
	被災	237	121	95	450	4
	災害	774	234	102	1082	63
	大震災	347	199	201	744	19
再生エネルギー	再生可能エネルギー	845	189	22	1056	35
	風力発電	177	324	9	507	31
	風車	14	129	0	143	3
北極	北極	46	766	7	776	84
	北極圏	**10**	73	0	82	8

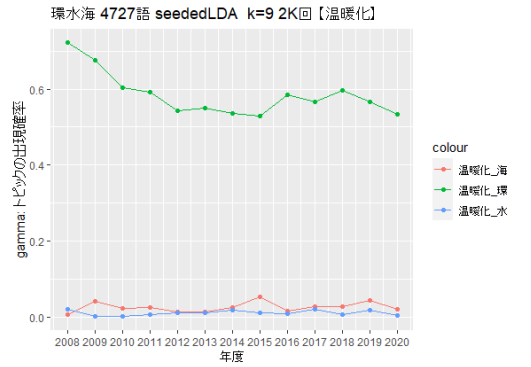
「北極圏」が多数あるので、追加した。

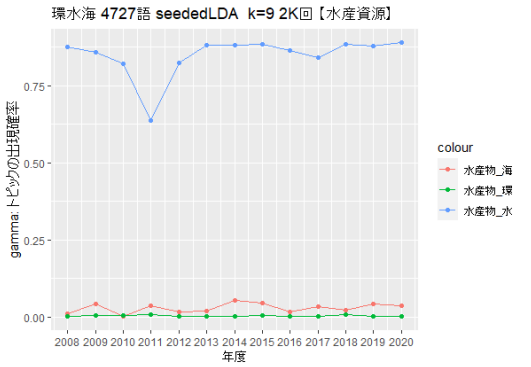
■ 4727 語 8 類辞書 2K 回 （KH Coder での Gibbs サンプリング数）

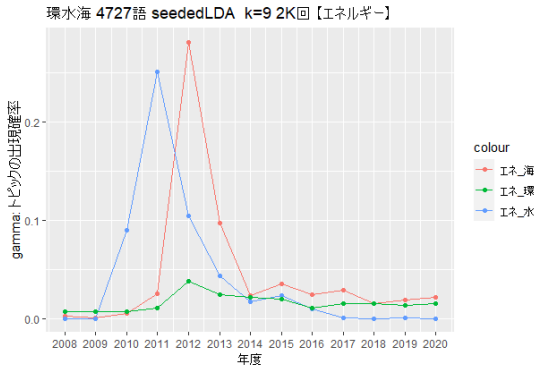
```
> seededlda::terms(環水海4727_seededLDA,10)↓
      気候変動     プラスチック     生物多様性   水産資源   地球温暖化   震災     再可エネルギー       北極域   other
 [1,] "気候変動"   "プラスチック"   "生物多様性" "水産物" "温暖化"     "復興"   "再生可能エネルギー" "北極"   "実施"
 [2,] "保全"       "プラスチックごみ" "保全"     "水産資源" "温室効果ガス" "災害"   "風力発電"           "北極圏" "原子力"
 [3,] "適応"       "マイクロプラスチック" "経済" "漁業資源" "CO2"         "津波"   "風車"               "開発"   "自然.1"
 [4,] "影響"       "海洋プラスチック" "利用"     "漁船"     "低炭素"     "地震"   "被害"               "日本"   "保全"
 [5,] "環境省"     "調査"           "世界"     "漁獲"     "二酸化炭素" "大震災" "支援"               "中国"   "福島"
 [6,] "支援"       "情報"           "日本"     "水産"     "実施"       "防災"   "東日本"             "実施"   "規制"
 [7,] "経済"       "観測"           "社会"     "漁業者"   "廃棄物"     "被災"   "施設"               "海域"   "委員会"
 [8,] "連携"       "技術"           "目標"     "管理"     "対策"       "管理"   "発生"               "計画"   "発電"
 [9,] "持続可能"   "日本"           "生態系"   "養殖"     "処理"       "基本"   "調査"               "開催"   "法律"
[10,] "実施"       "必要"           "自然.1"   "減少"     "技術"       "計画"   "被災地"             "会議"   "活用"
```

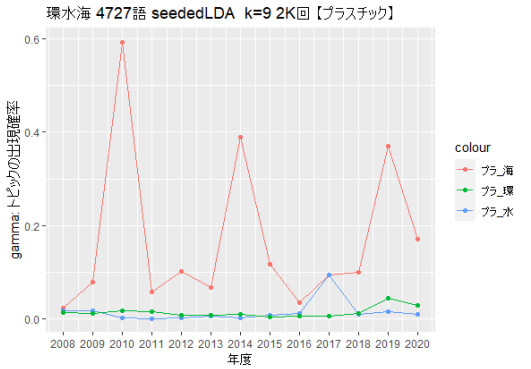

環水海 4727語 seededLDA k=9 2K回【気候変動】
gamma: トピックの出現確率
年度
colour
気候_海
気候_環
気候_水

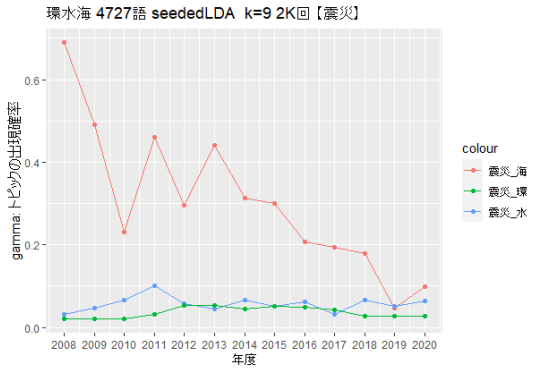

環水海 4727語 seededLDA k=9 2K回【生物多様性】
gamma: トピックの出現確率
年度
colour
生物_海
生物_環
生物_水

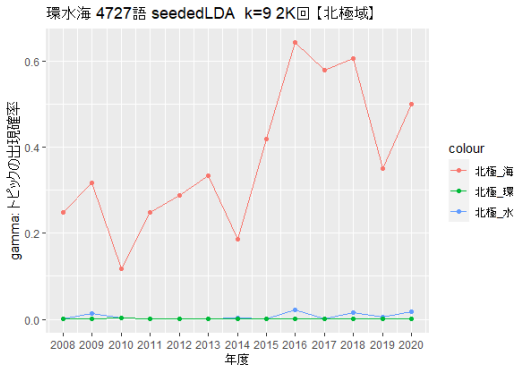

環水海 4727語 seededLDA k=9 2K回【温暖化】
gamma: トピックの出現確率
年度
colour
温暖化_海
温暖化_環
温暖化_水

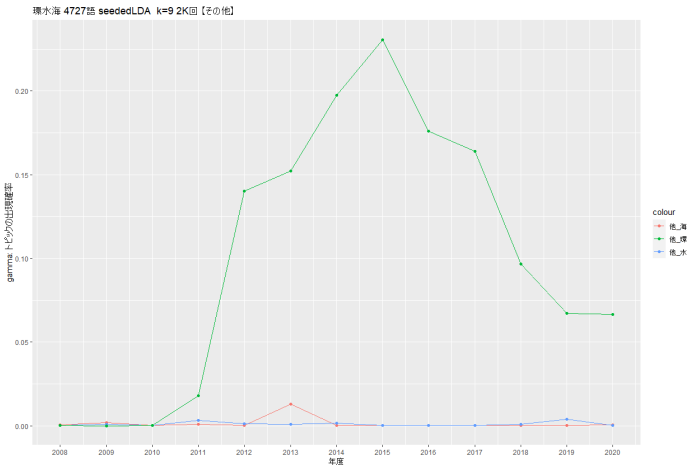

環水海 4727語 seededLDA k=9 2K回【水産資源】
gamma: トピックの出現確率
年度
colour
水産物_海
水産物_環
水産物_水


環水海 4727語 seededLDA k=9 2K回【エネルギー】
gamma: トピックの出現確率
年度
colour
エネ_海
エネ_環
エネ_水


環水海 4727語 seededLDA k=9 2K回【プラスチック】
gamma: トピックの出現確率
年度
colour
プラ_海
プラ_環
プラ_水

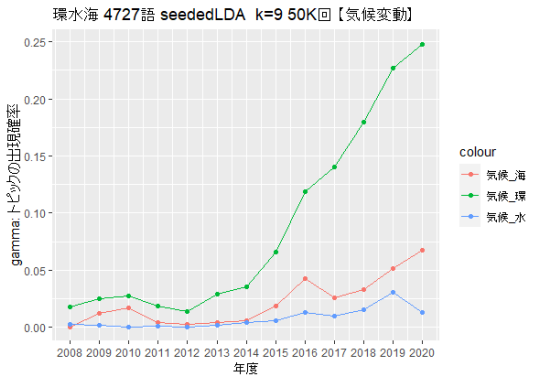

環水海 4727語 seededLDA k=9 2K回【震災】
gamma: トピックの出現確率
年度
colour
震災_海
震災_環
震災_水

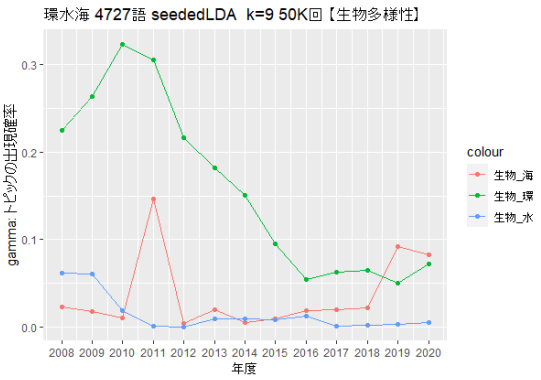

環水海 4727語 seededLDA k=9 2K回【北極域】
gamma: トピックの出現確率
年度
colour
北極_海
北極_環
北極_水

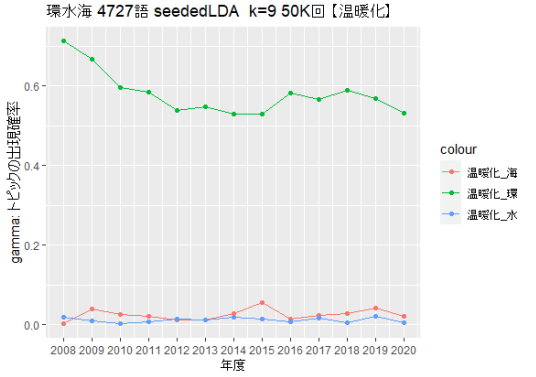


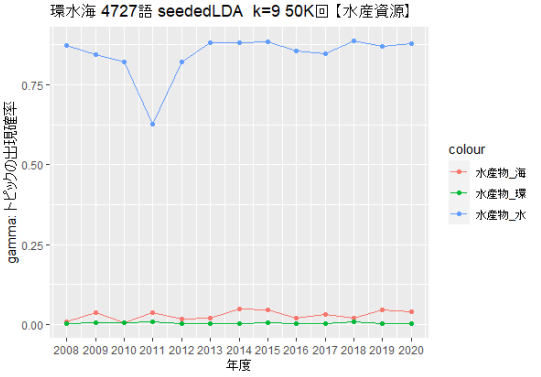
■ 4727 語 8 類辞書 50K 回

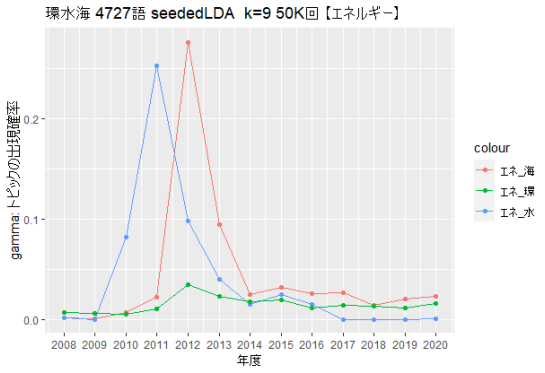
```
> seededlda::terms(環水海4727_seededLDA_50k,10)↓
      気候変動     プラスチック     生物多様性   水産資源   地球温暖化   震災       再可エネルギー       北極域   other
 [1,] "気候変動"   "プラスチック"   "生物多様性" "水産物"   "温暖化"     "復興"     "再生可能エネルギー" "北極"   "実施"
 [2,] "適応"       "プラスチックごみ" "保全"     "水産資源" "温室効果ガス" "災害"     "風力発電"           "北極圏" "原子力"
 [3,] "影響"       "マイクロプラスチック" "経済"   "漁業資源" "CO2"        "津波"     "風車"               "開発"   "自然.1"
 [4,] "保全"       "海洋プラスチック" "世界"       "漁船"     "低炭素"     "地震"     "被害"               "日本"   "保全"
 [5,] "目標"       "調査"           "社会"       "漁獲"     "二酸化炭素" "大震災"   "施設"               "海域"   "福島"
 [6,] "支援"       "情報"           "利用"       "水産"     "実施"       "防災"     "東日本"             "実施"   "規制"
 [7,] "持続可能"   "観測"           "目標"       "漁業者"   "対策"       "被災"     "支援"               "中国"   "発電"
 [8,] "環境省"     "技術"           "日本"       "養殖"     "廃棄物"     "管理"     "被災地"             "計画"   "調査"
 [9,] "実現"       "日本"           "自然.1"     "管理"     "処理"       "基本"     "福島"               "国際"   "委員会"
[10,] "連携"       "データ"         "活動"       "減少"     "利用"       "計画"     "発生"               "開催"   "経済"
```

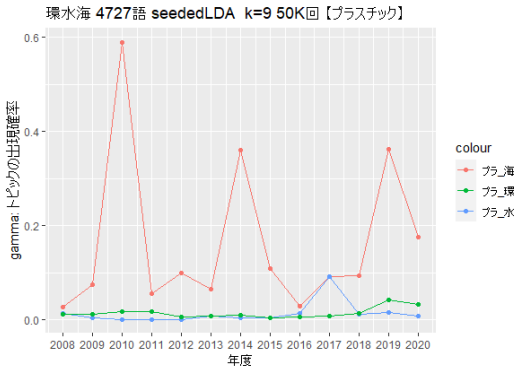


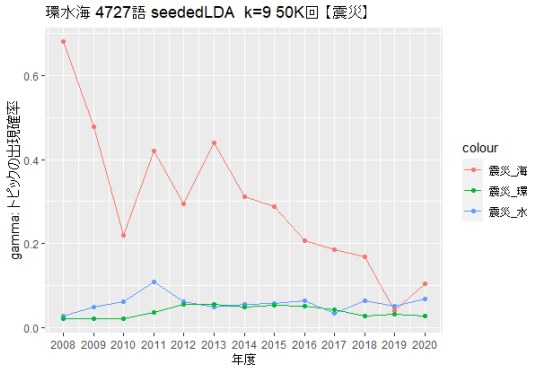


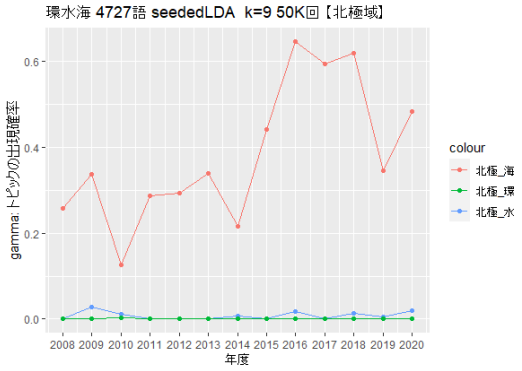


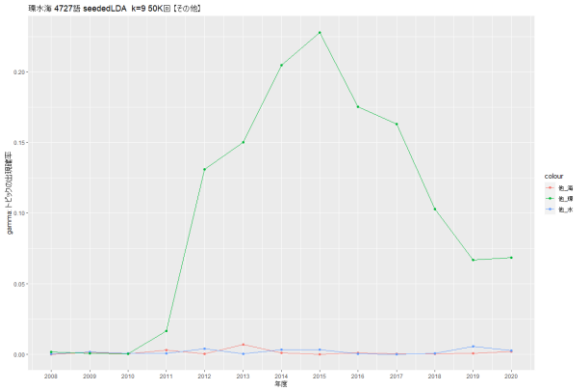



環水海 4727語 seededLDA k=9 50K回【エネルギー】
gamma: トピックの出現確率
年度
colour
エネ_海
エネ_環
エネ_水


環水海 4727語 seededLDA k=9 50K回【プラスチック】
gamma: トピックの出現確率
年度
colour
プラ_海
プラ_環
プラ_水

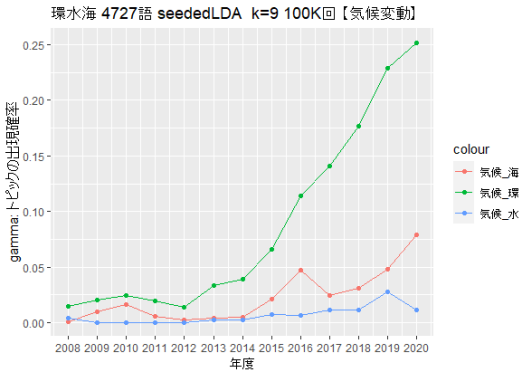

環水海 4727語 seededLDA k=9 50K回【震災】
gamma: トピックの出現確率
年度
colour
震災_海
震災_環
震災_水

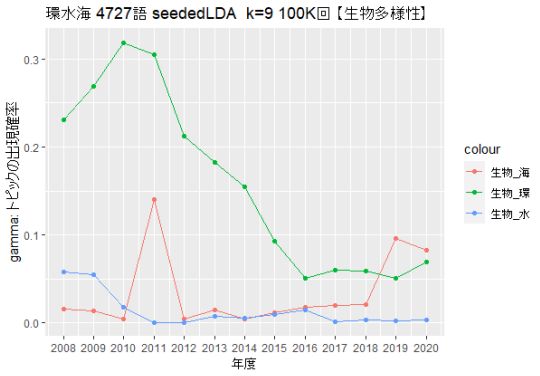

環水海 4727語 seededLDA k=9 50K回【北極域】
gamma: トピックの出現確率
年度
colour
北極_海
北極_環
北極_水

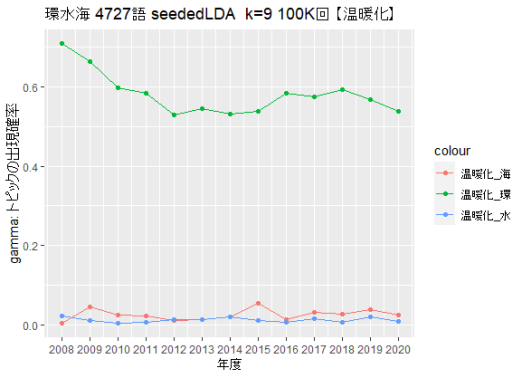


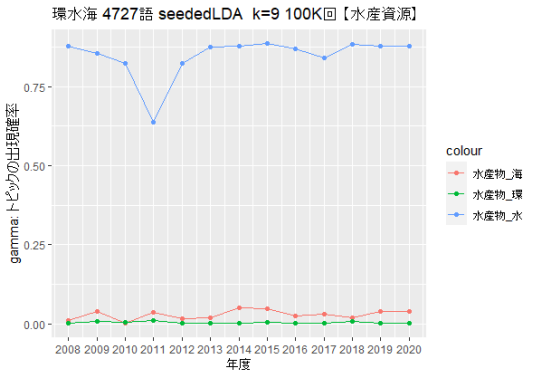
■ 4727 語 8 類辞書 100K 回

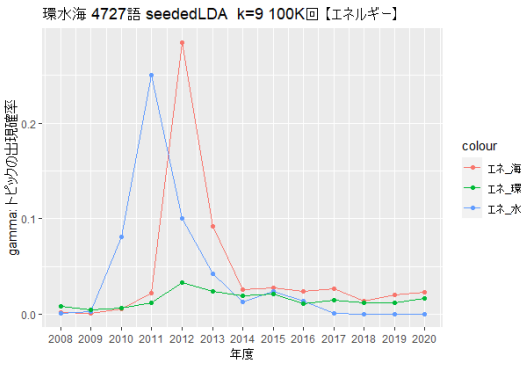
```
> seededlda::terms(環水海4727_seededLDA_100k,10)↓
      気候変動     プラスチック       生物多様性   水産資源   地球温暖化   震災     再可エネルギー       北極域   other
 [1,] "気候変動"   "プラスチック"     "生物多様性" "水産物"   "温暖化"     "復興"   "再生可能エネルギー" "北極"   "原子力"
 [2,] "保全"       "プラスチックごみ" "保全"       "水産資源" "温室効果ガス" "災害"   "風力発電"           "北極圏" "保全"
 [3,] "適応"       "マイクロプラスチック" "世界"     "漁業資源" "CO2"        "津波"   "風車"               "開発"   "自然.1"
 [4,] "影響"       "海洋プラスチック" "経済"       "漁船"     "低炭素"     "地震"   "被害"               "日本"   "実施"
 [5,] "支援"       "情報"             "利用"       "漁獲"     "二酸化炭素" "大震災" "施設"               "中国"   "規制"
 [6,] "環境省"     "調査"             "自然.1"     "水産"     "実施"       "防災"   "東日本"             "実施"   "福島"
 [7,] "企業"       "観測"             "日本"       "漁業者"   "廃棄物"     "被災"   "支援"               "海域"   "経済"
 [8,] "連携"       "技術"             "目標"       "管理"     "対策"       "管理"   "発生"               "開催"   "調査"
 [9,] "持続可能"   "必要"             "企業"       "養殖"     "処理"       "基本"   "被災地"             "計画"   "委員会"
[10,] "SDGs"       "計画"             "社会"       "資源"     "利用"       "計画"   "調査"               "国際"   "法律"
```

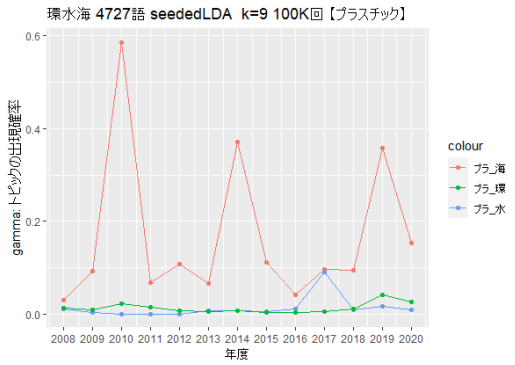


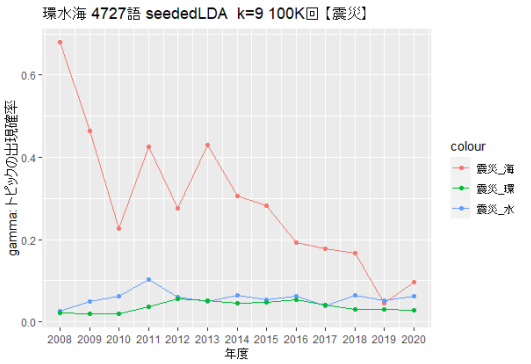


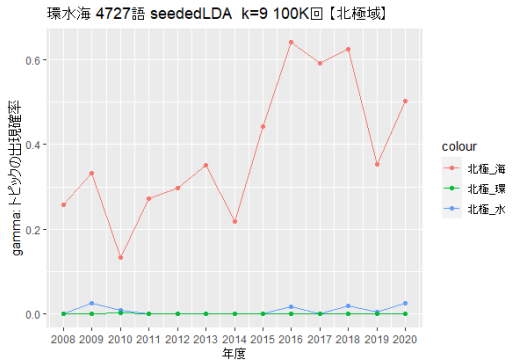


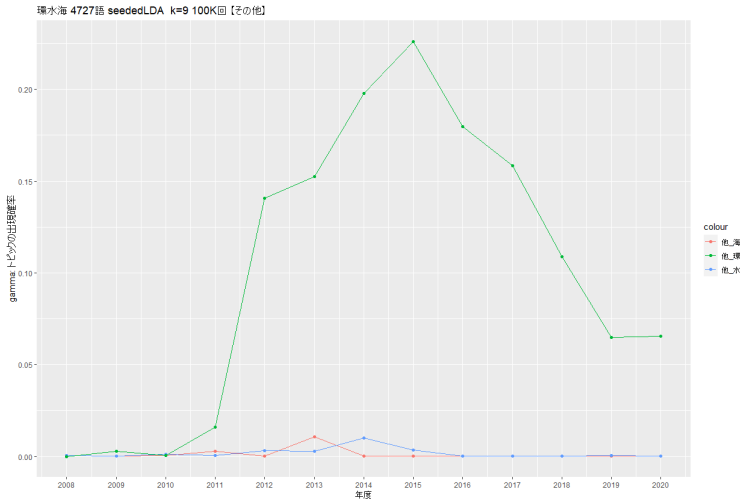



環水海 4727語 seededLDA k=9 100K回【エネルギー】
gamma: トピックの出現確率
年度
colour
エネ_海
エネ_環
エネ_水


環水海 4727語 seededLDA k=9 100K回【プラスチック】
gamma: トピックの出現確率
年度
colour
プラ_海
プラ_環
プラ_水

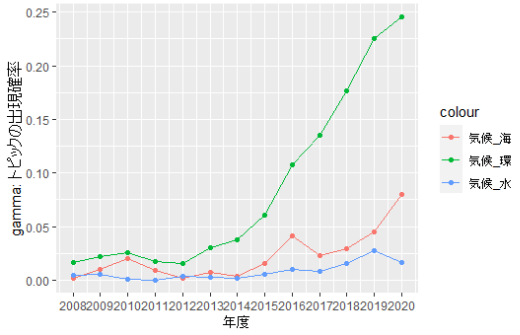

環水海 4727語 seededLDA k=9 100K回【震災】
gamma: トピックの出現確率
年度
colour
震災_海
震災_環
震災_水

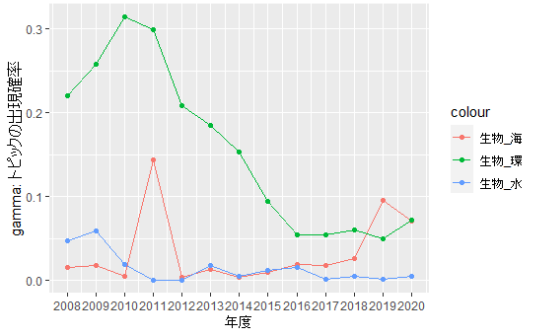

環水海 4727語 seededLDA k=9 100K回【北極域】
gamma: トピックの出現確率
年度
colour
北極_海
北極_環
北極_水

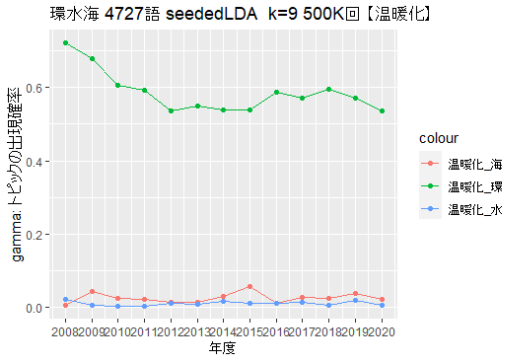

環水海 4727語 seededLDA k=9 100K回【その他】
gamma: トピックの出現確率
年度
colour
他_海
他_環
他_水

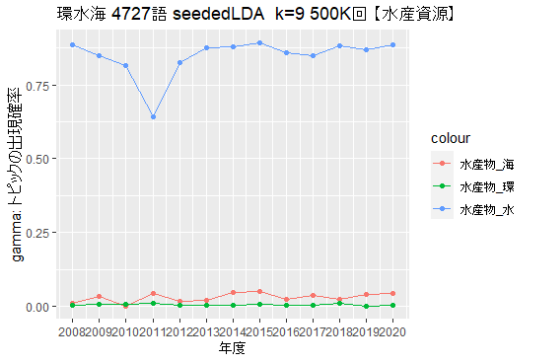
■ 4727 語 8 類辞書 500K 回

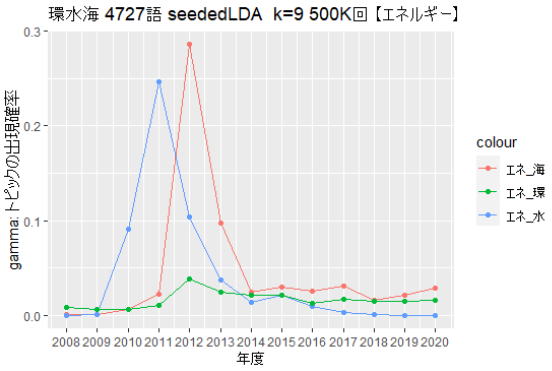
```
> seededlda::terms(環水海4727_seededLDA_500kx,10)↵
      気候変動     プラスチック       生物多様性     水産資源     地球温暖化   震災       再可エネルギー         北極域   other
 [1,] "気候変動"   "プラスチック"     "生物多様性"   "水産物"     "温暖化"     "復興"     "再生可能エネルギー"   "北極"   "実施"
 [2,] "保全"       "プラスチックごみ" "保全"         "水産資源"   "温室効果ガス" "災害"     "風力発電"             "北極圏" "原子力"
 [3,] "適応"       "マイクロプラスチック" "経済"     "漁業資源"   "CO2"        "津波"     "風車"                 "開発"   "保全"
 [4,] "影響"       "海洋プラスチック" "利用"         "漁船"       "低炭素"     "地震"     "被害"                 "日本"   "自然.1"
 [5,] "支援"       "調査"           "世界"         "漁獲"       "二酸化炭素" "大震災"   "東日本"               "中国"   "規制"
 [6,] "活用"       "情報"           "日本"         "水産"       "実施"       "防災"     "支援"                 "海域"   "評価"
 [7,] "環境省"     "観測"           "社会"         "漁業者"     "廃棄物"     "被災"     "施設"                 "実施"   "福島"
 [8,] "企業"       "技術"           "自然.1"       "管理"       "対策"       "管理"     "福島"                 "計画"   "処理"
 [9,] "実現"       "必要"           "目標"         "養殖"       "処理"       "基本"     "被災地"               "開催"   "発電"
[10,] "SDGs"       "船舶"           "企業"         "減少"       "技術"       "計画"     "発生"                 "国際"   "調査"
```

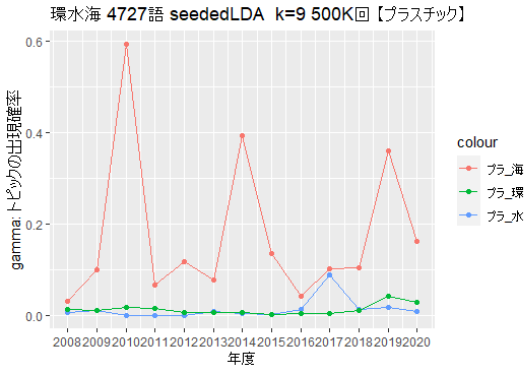


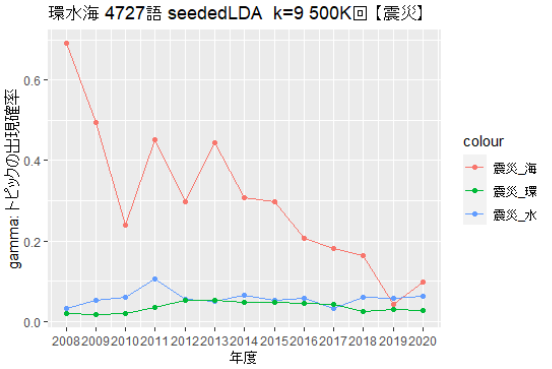


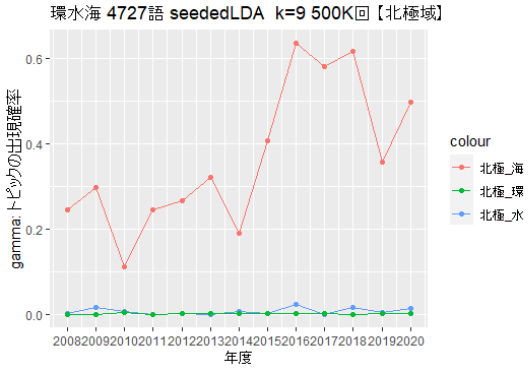


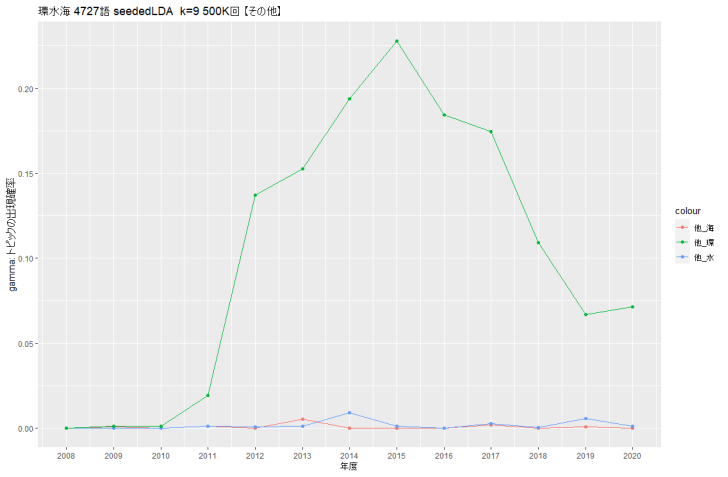



環水海 4727語 seededLDA k=9 500K回【エネルギー】
gamma:トピックの出現確率
年度
colour
エネ_海
エネ_環
エネ_水


環水海 4727語 seededLDA k=9 500K回【プラスチック】
gamma:トピックの出現確率
年度
colour
プラ_海
プラ_環
プラ_水


環水海 4727語 seededLDA k=9 500K回【震災】
gamma:トピックの出現確率
年度
colour
震災_海
震災_環
震災_水


環水海 4727語 seededLDA k=9 500K回【北極域】
gamma:トピックの出現確率
年度
colour
北極_海
北極_環
北極_水


環水海 4727語 seededLDA k=9 500K回【その他】
gamma:トピックの出現確率
年度
colour
他_海
他_環
他_水

■ 4727 語 5 類辞書 500K 回

```
> seededlda::terms(環水海4727_seededLDA_500k6,10)
      気候変動     生物多様性   温暖化       水産物       エネルギー           other
 [1,] "気候変動"   "生物多様性" "温暖化"     "水産物"     "エネルギー"         "調査"
 [2,] "開発"       "管理"       "温室効果ガス" "漁船"     "再生可能エネルギー" "情報"
 [3,] "開催"       "基本"       "廃棄物"     "漁獲"       "実施"               "資源"
 [4,] "持続可能"   "計画"       "実施"       "水産"       "対策"               "技術"
 [5,] "日本"       "開発"       "利用"       "漁業者"     "支援"               "活動"
 [6,] "世界"       "海域"       "処理"       "管理"       "処理"               "産業"
 [7,] "実施"       "教育"       "社会"       "養殖"       "調査"               "日本"
 [8,] "目標"       "政策"       "対策"       "操業"       "情報"               "利用"
 [9,] "中国"       "関係"       "地球"       "減少"       "原子力"             "必要"
[10,] "国際"       "総合的"     "資源"       "資源管理"   "管理"               "影響"
```