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

企画営業部

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 21 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
強制抽出する語の指定	環境・水産・海洋-複合語 min.txt
使用しない語の指定	除外語_015_OPRI.txt

KH Coder の共起ネットワーク、対応分析、トピック分析機能を利用して、環境・水産・海洋白書、海洋基本計画の処理結果は、別途資料で提出済であり、その文書番号一覧を表 2-3 に示す。これら分析処理での分析対象語は、KH Coder による自動設定値と可視化された分析結果の判読での有効上限値 150 に近い値を設定して実施した。また、トピック分析（LDA）でのトピック数設定に当たっては、R 言語 ldatuning パッケージによる評価結果を参考にしている。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 とは異なるが、ldatuning パッケージでの評価パラメータは、Giffiths2004、Arun2010、CaoJuan2009、Deveaud2014 の 4 種が準備されている。Giffiths2004 では perplexity と同様に、トピック数の増加に対して滑らかな曲線となることが多く、トピック数の推定には窮する。一方、CaoJuan2009 の最大値と Deveaud2014 最小値によるトピック数の判定は、容易なことが多い。

表 2-3 KH Coder による処理結果一覧

分析対象文書	期間	報告書の文書番号
水産白書（施策編）	2007 年～2020 年	OPRI-21-023
水産白書	2007 年～2020 年	OPRI-21-019、OPRI-21-024
海洋基本計画	1 次、2 次、3 次	OPRI-21-021b
海洋白書	2004 年～2020 年	OPRI-21-017、OPRI-21-025
環境白書	2008 年～2020 年	OPRI-21-018、OPRI-21-026
環境・水産・海洋白書	2008 年～2020 年	OPRI-21-016d、OPRI-21-027

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）を固定化している。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 に固定しており、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 程度であれば問題ない。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、LDA 分析での Gibbs サンプリング数の影響を調査するため、R-Studio での topicmodels パッケージの処理結果と KH Coder 処理結果の整合性を確認した。それらの結果は、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（OPRI-21-028）
- 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（OPRI-21-029）

OPRI 殿より提示された「関心の高いキーワード」を軸にトピックを抽出するためには、R 言語の seedelda パッケージを利用した。上記作業により、KH Coder 処理結果と R-Studio 環境下で R 言

語 topicmodels パッケージの処理結果の整合性は確認済であるから、seedelda パッケージに用意された LDA 分析（textmodel_lda）結果との整合性についても調査した。その結果は、下記文書で報告した。

- 「こころ」 で KHCoder と R-Studio の LDA 処理結果を比較（OPRI-21-035）

seedelda パッケージでは、辞書として指定した「関心の高いキーワード」を軸に、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することで、省庁横断的な経年変化を把握した。
「関心の高いキーワード」として、下記 2 種について seededlda で処理を実施した。

表 2-4 「関心の高いキーワード」

	キーワードの規定コード
5 類辞書	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")))
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")))

これら辞書ファイルを指定した教師付き LDA の処理結果は、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（OPRI-21-033）
- 環境・水産・海洋白書（2008～2020） 教師付き LDA 分析処理結果（その 2 ）（OPRI-21-034）

４．今後の展望

作業初期段階の手法検討段階で、日本語テキストの形態素解析ツールの処理結果が、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手順を確定した。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いるので、処理コードの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。本作業では、KH Coder の形態素解析結果である DTM（Document Term Matrix）を csv 形式で取得し、そのデータを最新版 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。

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 か spacy/Ginza を使用する予定であった。KH Coder の Chasen と Ginza との比較はしていないので、Chasen の性能確認は未着手である。

また、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）との比較も重要である。