確率-ib によるオンラインニュースのトピックのランク付けと BERTopic の比較

Stochastic method of topic ranking and comparison with BERTopic

山内康英／Yasuhide YAMANOUCHI、永井拓史／Takushi NAGAI、酒井紀之／Noriyuki SAKAI
多摩大学情報社会学研究所教授、研究開発工房 tn-works 代表、ソフトウェア開発代表取締役社長

[Abstruct] We developed the system that fits the PDF of Stable Distribution to the linguistic data obtained from Yahoo! News and calculated the word importance and topic ranking using MF-ib and 2TF-ib methods. The word frequency in documents follows Zipf's law, which is a rank-frequency plot. Our group has indicated in previous papers that the Zipf's law corresponds to Stable Distribution of word frequency. MF-ib and 2TF-ib are methods for calculating word importance using the probability, instead of word frequency as the tf-idf (term frequency-inverse document frequency), which is widely used for the measure of word importance in NLP. In addition, we compared the topic ranking by Grootendorst's method applying to the same data and found that the rankings from both methods matched fairly well.

［キーワード］ フォーカル・トピックのランク付け、tf-idf、確率-ib、MF-ib および 2TF-ib、BERTopic

1. はじめに[1]

Karen Sparck Jones が 1972 年に開発した tf-idf 法（term frequency-inverse document frequency）[2]は、一定量のコーパスを前提として、そのなかの特定の文書の単語の重要度を測る尺度である。本研究では、tf-idf の単語の出現回数に換えて、出現頻度の確率密度関数を用いたトピックのランク付けを検討した。本稿では、この確率-ib 法を Yahoo! ニュースの 8 週間分のデータに適用し、その結果を BERTopic 法[3]によるトピックのランク付けと比較した。

本研究グループでは、SNS の単語の出現頻度について確率論的な研究を行っている。文書中の単語の出現頻度は、ランク図として Zipf 則に従う。本研究グループは、単語の出現頻度の Zipf 則が、確率分布としては安定分布[4]（Stable Distribution）であることを示し、Twitter の API から取得した言語データを、安定分布の確率密度関数で連続的に fit するシステムを運用してきた。[5] 本研究では、このシステムを利用し、特定した確率密度関数から「出現確率×出現回数」を計算して、Yahoo! ニュースに出現する単語の重み付けを行った。この確率論的な手法では、オンラインニュースのトピックを、コーパス全体の単語の出現頻度と比べた特定文書の確率的な偏りの時間変化として考えることになる。本稿では、確率-ib 法の具体的なアルゴリズムおよび試験運用の結果を中心に記述した。

2. 方法

2.1 今回利用したデータについて

本研究グループでは、2023 年度後半に言語リソースを Twitter から Yahoo!ニュースに変更し、単語の出現頻度に確率密度関数を fit するシステムの運用を再開した。この変更は、X 社の経営方針の変更にともない、Twitter が 2023 年 8 月からアカデミック API の運用を停止したからである。今回の研究では、Yahoo!ニュースの 1 月 21 日から 3 月 23 日まで 8 週間のデータを利用した。利用したデータは合計 4,771,837 文字（改行を含む）、ニュースは 4,616 件、データ量は約 14 メガバイトとなった。

文書のトピックを、単語の大きさや相互の位置で視覚化する一般的な方法として、いわゆるワードクラウドがあるが、これはトピックの定量的なランキングにはなっていない。確率-ib と BERTopic を併用したリアルタイムのトピック分析の利用法として、世論や社会的関心の動的な追跡、UGC のコメントのトレンドの決定、新聞社による「年間 10 大ニュース」の決定、など[6]が考えられる。（「図 1」）

【図 1: Yahoo!ニュースのコメントのトレンドとワードクラウド】

2.2 関連研究の動向と本研究の関係

tf-idf 法を使った単語の重要度計算は、自然言語解析の要素技術として、検索のランキングの尺度やトピック分析に広く用いられている。[7] これとは別に、Transformer 技術の普及にともなって、トピック分析にもニューラル・ネットワークの応用が始まった。Grootendorst は、2022 年の論文 “BERTopic: Neural topic modeling with a class-based TF-IDF procedure”で、Transformer 技術を使ったトピック分析を提案している。BERTopic では、言語データをトピックのクラスに分解した後、重要度計算の方法として tf-idf を

利用している。中里は、2024 年の論文で、BERTopic を COVID-19 のワクチンに関する SNS のトピック分析に利用し、これを道徳基盤理論の定量分析に結び付けた。[8] 本研究では、Grootendorst（2022）にしたがって BERTopic のシステムを構築した。これを使って Yahoo!ニュースの同一データを解析し、確率-ib と BERTopic のランク付けの結果を比較した。

2.3 tf-idf 法の構成

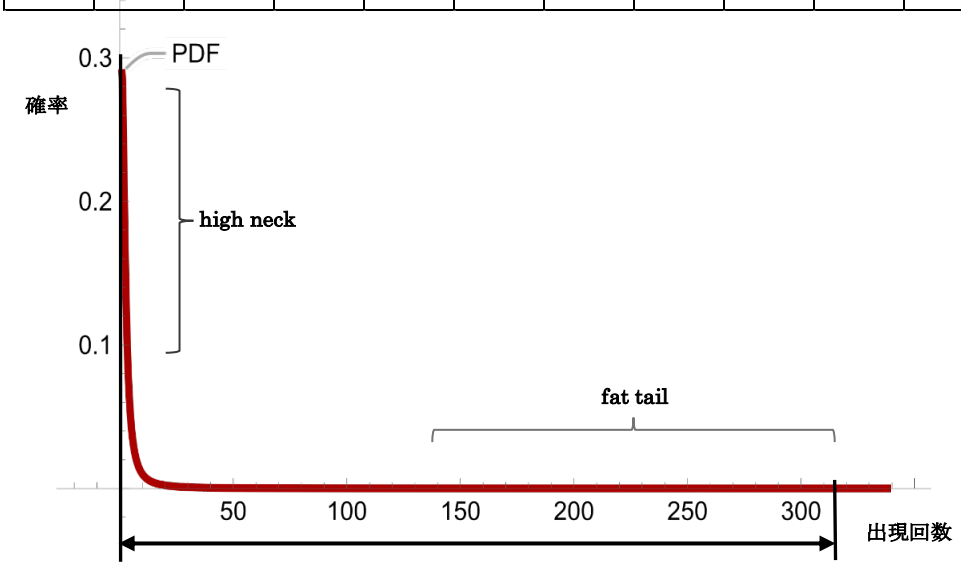
tf-idf 法では、特定文書内の単語の出現回数（term frequency: tf）を、その特定文書を含むコーパス内の、その単語（＝word）を含む文章（＝document）の数の逆数（idf : inverse document frequency）を用いてオフセットする。この操作は、どのような意味をもつのであろうか。高い出現頻度を有する単語の上位は、多くの場合、その文書のトピックとは関係のない一般的な頻出語である。たとえば、特定文書内で｛今日｝｛岸田首相｝｛大谷｝｛日本｝という 4 つの単語の出現頻度が高いとする。ここから一般的な頻出語である｛今日｝｛日本｝を省いて、｛岸田首相｝｛大谷｝に、この文書での適切な重要度を付与するために、tf-idf 法では次のようにする。まず、①｛今日｝｛岸田首相｝｛大谷｝｛日本｝の 4 つの単語の特定文書内の出現頻度を計量する。つぎに、②コーパスを構成する文書全体のなかで 4 つの単語を含む文書の数を計算する。｛今日｝もしくは｛日本｝を含む文書の数は多いので、その逆数[9]を各単語の出現頻度に掛ければ、特定文書の｛岸田首相｝および｛大谷｝が顕在化する。tf-idf 法は、図書館情報学や情報検索の分野で、文書（＝図書）のインデックスの抽出手法として発展した。以上の tf-idf 法の検討から、①特定文書内での単語の重要度、②単語のコーパス全体での一般性、という 2 つの指標化およびその組み合わせには、さまざまな方法のあることがわかる。

2.4 単語の出現頻度の確率密度関数

単語の出現頻度は、Zipf 則あるいはこれと同等の確率分布になっている。3 日 17 日～3 月 23 日の 1 週間を例に取れば、Yahoo! ニュースを形態素解析した名詞の出現単語数は 15,082 語と多い。「図 2」は、この週間データを安定分布の確率密度関数で fit したものである。最大の出現頻度の単語は、338 回の｛日本｝であって fat tail に位置する。これに対して大多数の単語は、確率密度関数の high neck に位置している。Zipf 則という単語の出現頻度のこの確率的なパタンは、（すべての言語の）すべての文書と、その Bag of Words[10]に適合すると考えられている。

【表 1： Yahoo!ニュースの単語の出現頻度】

日本	大谷	発表	選手	環境	必要	確認	東京	以上	ロシア	
338	212	199	178	171	162	153	147	147	140	……
写真集	支払い	銀メダル	退院	スイーツ	犠飛	選抜高校野	オブ	お腹	福島県	
6	6	6	6	6	6	6	6	6	6	……
日本製鉄	白岡市	喜田	M＆A	マナー	金髪	ラッキー	伊東純也	再始動	児童生徒	
2	2	2	2	2	2	2	2	2	2	……
東成区	市場価格	芸術	平塚市	清掃	ゆとり教育	海洋学部	ロシア語	治安部隊	パパ	
1	1	1	1	1	1	1	1	1	1	……
宗基	年初来安値	屋根裏	ゴールネッ	間柄	利害	ホワイトソ	ブロンズ	改訂版	最古	
1	1	1	1	1	1	1	1	1	1	……



【図 2：語の度数間の移動】

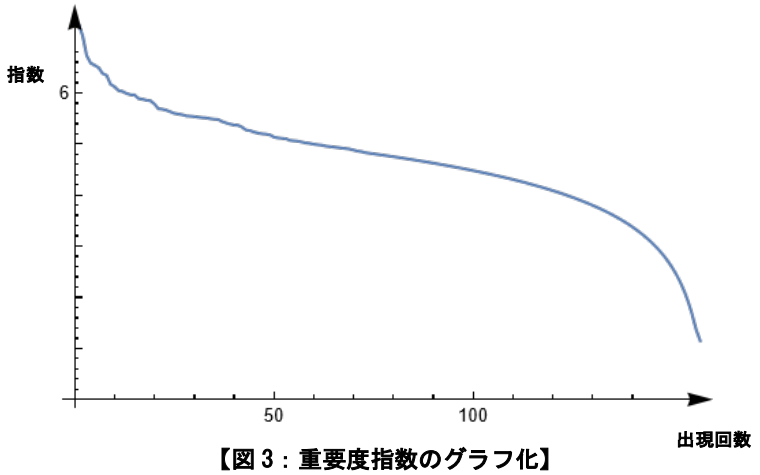
2.5 トピックの出現と確率の移動

この週間データの確率密度関数（PDF）が示すように、多くの単語の出現回数は、1 回から数回で、PDF では左側の high neck にある。しかし「MBL のソウル開幕戦」といったフォーカル・トピックが社会に登場すれば、{大谷} が fat tail にあらわれる。社会的な焦点化の終了後、この単語は high neck の通常の位置に復帰する。[11]

このように社会的なトピックとは、度数から見た high neck と fat tail の間の単語の往復を意味している。トピックの出現にともなう語の度数間の移動を、「図 2」の矢印で示した。確率-ib 法は、フォーカル・ポイントの出現にともなって生ずる確率密度曲線上の単語の度数間の移動を前提として、単語のコーパス全体の出現確率と比べた特定文書の確率的な偏りの変化としてトピックを捉えることになる。

2.6 MF-ib 法について

単語の「出現確率×出現回数」の和の正規化は、期待値すなわち確率関数の 1 次モーメントになる。したがって個々の単語の重み付けとして、これを用いるのは自然である。MF-ib 法は、単語の出現頻度の確率密度関数を前提として、「出現確率×出現回数」を文書内の重要度の指標として用いる。単語の出現頻度は、極端に裾の厚い（＝長い）正値の確率密度関数を有している。出現頻度の高い重要語は、確率変数の右端（＝長い裾）に現れるので、この指標作成方法では、重要度が大きいほど数値が小さくなる。このため MF-ib 法の重要度指数として、「出現確率×出現回数」の逆数の対数を用いる。この指標化は、オンラインの言語データに対して「図 3」のような、すべての度数に対して通常 1～10 前後の連続的な重要度指標を与える。



【図 3：重要度指数のグラフ化】

2.7 2TF-ib 法について

2 語共起による指標化（2TF）として、同一センテンスに出現するすべての 2 語の組み合わせを計算して、その出現頻度を計算する。共起の出現頻度をとは、2 つの名詞が同一の記事内に共起（co-occurrence）、つまり該当する 2 語が同一の記事内に出現している状態を指す。（順番は問わない。）確率的には独立の事象の条件付き確率となる。

【表 2：2TF の計算例】

02024/03/21	75	該当数	出現1	出現2	出現1-IB	出現2-IB	2TF-IB
大谷・ドジャース	14.66667	11	55	29	1	1	14.66667
大谷・通訳	13.33333	10	55	37	1	1	13.33333
大谷・翔平	13.33333	10	55	17	1	1	13.33333
通訳・ドジャース	13.33333	10	37	29	1	1	13.33333
ドジャース・翔平	13.33333	10	29	17	1	1	13.33333
通訳・翔平	12.00000	9	37	17	1	1	12
水原・翔平	12.00000	9	42	17	1	1	12
大谷・水原	12.00000	9	55	42	1	1	12
水原・通訳	12.00000	9	42	37	1	1	12
水原・ドジャース	12.00000	9	42	29	1	1	12
大谷・一平	10.66667	8	55	18	1	1	10.66667
大谷・パドレス	10.66667	8	55	18	1	1	10.66667
大谷・投手	10.66667	8	55	16	1	1	10.66667
水原・一平	10.66667	8	42	18	1	1	10.66667

2.8 頻出語のオフセットと「ib」の計算

MF および 2TF を用いた①特定文書内での単語の重要度に続いて、②単語のコーパス全体の一般性をもちいたオフセットを行う。頻出語の影響をオフセットする手法として idf がある。[12] 本研究が取り扱うオンラインニュースや UGC の言語リソースでは、図書館情報学のように文書（＝図書）を単位とした数を用いることができない。そこでオンラインニュースの言語リソースを Bag of Words として、コーパス全体での直接的な一般性を考え、ib（inverse bottom）の概念を導入する。ib は、底（＝基準となる頻

出語）となる語の出現回数を用いて、単語の出現頻度の逆を成す指標化を行う。今回は 3 ヶ月間の Yahoo!ニュースの出現頻度上位 10 万語からなるコーパス（＝C）を用いて以下の 2 つの ib を作った。具体例を「表 3」に示した。

【表 3： ib の計算例】

（1）頻出コーパスCと逆重付指標（ib）

語句	出現数	ib
日本	3033	0.17
発表	1799	0.29
女性	1582	0.33
以上	1480	0.36
必要	1422	0.37
男性	1407	0.37
撮影	1326	0.40
東京	1252	0.42
昨年	1212	0.43
……		
関係	553	0.95
全国	548	0.96
人気	541	0.97
事故	538	0.98
予想	538	0.98
チーム	529	0.99
岸田	529	0.99
経験	526	1.00

2.8.1 MF で利用した ib

この場合は｛経験｝を区切り語（bottom of list）とする。区切り語の出現回数を k（＝526）、対象語句の出現回数を t として、

① t∋C のとき ib=k／t

② t∋C でないとき ib=1

このとき、ib は頻出語であるほど 1 より小さな値、t＝k のとき 1.0、出現語がコーパス C に含まれないときに 1.0 の値を取ることになりコーパス全体の一般性によるオフセットという条件を満たしている。

③単語の重要度と頻出語によるオフセットを次のようにして語の重要度を計算する。MF に ib を組み合わせる計算式：語の重要度＝MF×ib

2.8.2 2TF で利用した ib

区切り語（bottom of list）の出現回数を k（＝526）、対象語句の出現回数を t として、

① t∋C のとき ib ＝k／(t+k)

② t∋C でないとき ib ＝1

このとき、ib は頻出語であるほど 1 より小さな値、t＝k のとき 1.0、出現語がコーパス C に含まれないときに 1.0 の値を取ることになりコーパス全体の一般性によるオフセットという条件を満たしている。

③単語の重要度と頻出語によるオフセットを次のように適用して語の重要度を計算する。2TF に ib を組み合わせる計算式：語の重要度＝2TF×ib[13]

3. 結果

3.1 単語の重要度計算の概要

「表 4」に、MF-ib による単語の重要度計算の概要を示した。まず、(2) MF 重要度による 1 週間の言語データの指標化を行う。これに対して (1) コーパス全体の頻出語から底となる単語を決めて逆重み付け指数を計算する。単語毎に逆重み付け指標を乗じて、(3) 並べ替えをすれば、オフセットされた単語と MF-ib 値を決定することができる。この週間データでは｛大雪｝が顕在化している。

【表 4：週間の MF-ib 指標の計算例】

Yahoo!ニュース　週間重要度指数の計算　2月4日（日）～2月10日（土）

1.　週間のMF-ib指標

①1次M重要度の下限を4.5とすると（2）で「首相」となる.②頻出語コーパスCで底（bottom）に「世界」を取り、逆重付指標 ib＝t/k を算出する。K＝675、t=出現数.③出現語句ごとに1次M重要度と逆重付指標（ib）を並べる.④「1次M重要度×ib」を計算して降順に並べ替える。【大雪／5.22】

（1）頻出コーパスCと逆重付指標（ib）

	語句	出現数	ib
1	日本	3033.00	0.22
2	発表	1799.00	0.38
3	女性	1582.00	0.43
4	以上	1480.00	0.46
5	必要	1422.00	0.47
6	男性	1407.00	0.48
7	撮影	1326.00	0.51
8	東京	1252.00	0.54
9	昨年	1212.00	0.56
10	確認	1168.00	0.58
11	避難	1142.00	0.59
12	問題	1069.00	0.63
13	午後	1056.00	0.64
14	可能性	1041.00	0.65
15	子ども	1038.00	0.65
16	今回	1037.00	0.65
17	当時	1010.00	0.67
18	選手	1006.00	0.67
19	影響	993.00	0.68

（2）1次M重要度に対する逆重付乗指標（ib）

単語	1次M重要度	ib
日本	6.01	0.22
発表	5.71	0.38
以上	5.40	0.45
昨年	5.33	0.56
東京	5.32	0.54
必要	5.31	0.47
男性	5.24	0.48
大雪	5.22	1.00
撮影	5.19	0.51
今回	5.18	0.65
センチ	5.16	1.00
避難	5.15	0.59
可能性	5.15	0.65
選手	5.13	0.67
代表	5.10	0.82
確認	5.09	0.58
状況	5.01	0.68
現在	5.00	0.75
積雪	4.95	1.00

（3）逆重付指標による並替

単語	重要度	ib値降順並替
大雪	5.22	5.22
センチ	5.16	5.16
積雪	4.95	4.95
状態	4.77	4.77
監督	4.81	4.76
言葉	4.75	4.75
場所	4.79	4.74
地域	4.69	4.69
会社	4.67	4.67
予想	4.65	4.65
イラン	4.62	4.62
経験	4.58	4.58
ファン	4.57	4.57
世界	4.61	4.56
五輪	4.55	4.55
ドラマ	4.53	4.53
中心	4.52	4.52
中国	4.80	4.51
生活	4.51	4.51

3.2　Yahoo!ニュースの各週の社会的なトピック

2024 年 1 月 21 日から 3 月 23 日にかけて、Yahoo!ニュースの各週の社会的なトピックスの推移は、以下のようになっていた。

（1）逆重付指標による並替

単語	重要度	ib値降順並替
能登半島地震	5.02	5.02
現在／韓国	4.91	4.91
1日	4.85	4.85
店舗	4.73	4.73
ポール／桐島	4.58	4.58
ドラマ	4.55	4.55

02024/01/31	82	該当数	出現2	出現1-IB	出現2-IB	2TF-IB
地震・能登半島地震	15.85	13	18	1	1	15.85
地震・道路	9.76	8	18	1	1	9.76
地震・発生	9.76	8	20	1	1	9.76
発生・能登半島地震	9.76	8	18	1	1	9.76
道路・能登半島地震	9.76	8	18	1	1	9.76

02024/01/29	89.00	該当数	出現2	出現1-IB	出現2-IB	2TF-IB
容疑者・桐島	8.99	8	43	1	1	8.99
容疑者・事件	8.99	8	31	1	1	8.99
容疑者・桐島聡	8.99	8	28	1	1	8.99

(1) 1月28日（日）～2月3日（土）

①能登半島地震１ヶ月　【5.02／15.85】

『朝日新聞デジタル、地域・主要、能登半島地震から 1 カ月、復興の道筋みえず、被害把握もいまだ不十分　2024 年 02 月 01 日 00:00:10』

②東アジア半日武装戦線の桐島聡容疑者の出頭【4.58／8.99】

『読売新聞オンライン、国内、連続企業爆破事件の桐島容疑者名乗る男、神奈川の土木会社に長期間勤務、2024 年 01 月 28 日 21:25:49』

(2) 逆重付指標による並替

単語	重要度	ib値降順並替
大雪	5.22	5.22
センチ	5.16	5.16
対応／積雪	4.95	4.95

02024/02/05	83	該当数	出現1	出現2	出現1-IB	出現2-IB	2TF-IB
大雪・注意	15.66	13.00	93.00	39.00	1.00	1.00	15.66
大雪・積雪	14.46	12.00	93.00	61.00	1.00	1.00	14.46
大雪・警報	14.46	12.00	93.00	46.00	1.00	1.00	14.46

(2) 2月4日（日）～2月10日（土）

①大雪・注意【5.22／15.66】

『毎日新聞、国内・主要、関東甲信、5 日昼過ぎ～6 日に大雪の恐れ、東京 23 区も積雪見込み、2024 年 02 月 05 日 00:20:25』

(3) 逆重付指標による並替

単語	重要度	ib値降順並替
子ども	5.27	5.11
事件	5.10	5.10
容疑者	5.05	5.05

02024/02/14	91	該当数		出現1	出現2	出現1-IB		出現2-IB	2TF-IB
事件・逮捕	13.19	12		42	27	1		1	13.19
事件・午前	10.99	10		42	25	1		1	10.99
事件・捜査	10.99	10		42	19	1		1	10.99

(3) 2月11日（日）～2月17日（土）

①4 歳次女を中毒死 両親を逮捕【5.05／13.19】

『読売新聞オンライン、地域主要、４歳次女殺害容疑の両親、複数の有毒物質をネット検索…スマホ・ＰＣに履歴、2024 年 02 月 16 日 12:00:19』

(4) 逆重付指標による並替

単語	重要度	降順並替
ウクライナ	5.81	5.81
ロシア	5.66	5.66

02024/02/24	76	該当数		出現1	出現2	出現1-IB		出現2-IB	2TF-IB
ウクライナ・ロシア	15.79	12		64	43	1		1	15.79
ウクライナ・侵攻	13.16	10		64	21	1		1	13.16
ロシア・侵攻	13.16	10		43	21	1		1	13.16

(4) 2月18日（日）～2月24日（土）

①ウクライナ侵攻が３年目突入甚大な犠牲と細る国際的支援で疲弊【5.81／15.79】

『Yahoo!ニュース オリジナル 特集、主要,国際、侵攻から２年、終わりの見えない戦闘、ウクライナ人の領土断念という「不確かな選択肢」の意味 、2024 年 02 月 22 日 18:20:46』

(5) 逆重付指標による並替

単語	重要度	降順並替
大谷	5.88	5.88
首相	5.88	5.88
問題	5.93	5.75

02024/02/28	80	該当数	出現1	出現2	出現1-B	出現2-B	2TF-B
出席・開催	11.25000	9	50	23	1	1	11.25
公開・審査会	10.00000	8	37	27	1	1	10
公開・政倫	10.00000	8	37	25	1	1	10
出席・公開	10.00000	8	50	37	1	1	10

大谷・自身	7.86517	7	51	23	1	1	7.865169
結婚・報告	7.86517	7	37	22	1	1	7.865169
翔平・自身	7.86517	7	24	23	1	1	7.865169

(5) 2月25日（日）～3月2日（土）

①岸田首相の政治倫理審査会出席【5.88／11.25】

『FNNプライムオンライン（フジテレビ系）、国内主要、【速報ライブ】岸田首相政倫審出席「今の政治を未来に引き継げるか申し訳ない」、2024年02月29日 15:20:15』【】

②大谷選手の結婚公表【5.88／7.87】

(6) 逆重付指標による並替

単語	重要度	降順並替
トランプ氏	4.24	4.24

02024/03/05	87	該当数		出現1	出現2	出現1-B		出現2-B	2TF-B
トランプ氏・共和党	5.75	5		31	25	1		1	5.75
ヘイリー・トランプ氏	4.60	4		33	31	1		1	4.60
ヘイリー・共和党	4.60	4		33	25	1		1	4.60

(6) 3月3日（日）～3月9日（土）

①トランプ氏共和党大統領指名獲得 【4.24／4.60】

『ロイター主要,国際 ヘイリー氏指名争い撤退表明、米大統領選バイデン対トランプ確実に 2024 年 03 月 07 日 03:20:08』

(7) 逆重付指標による並替

単語	重要度	降順並替
映像	5.24	5.24
被告	5.18	5.18
津波	4.95	4.95

02024/03/11	77	該当数		出現1	出現2	出現1-B		出現2-B	2TF-B
津波・東日本大震災	20.78	16		66	39	1		1	20.78
震災・東日本大震災	20.78	16		49	39	1		1	20.78
地震・東日本大震災	20.78	16		46	39	1		1	20.78
東日本大震災・発生	19.48	15		39	37	1		1	19.48

02024/03/14	84	該当数		出現1	出現2	出現1-B		出現2-B	2TF-B
判決・違反	9.52	8		30	21	1		1	9.52
判断・判決	8.33	7		42	30	1		1	8.33
判断・同性婚	5.95	5		42	20	1		1	5.95

(7) 3月10日（日）～3月16日（土）

①東日本大震災 13 年経過【4.95／20.78】『tenki.jp、科学主要、東日本大震災から 13 年、大地震や津波発生で慌てないために普段からの心構えを、2024年03月11日 15:55:10』

②裁判関係の記事重複【5.88／7.87】『同性婚認めぬ規定「違憲状態」賠償請求は棄却 全国で判断割れる・東京地裁』

『エラーコード：1001101 ご利用の環境では映像を視聴できません。 映像視聴における推奨環境はこちらをご確認ください』とのエラーコードの取得が 3 月 1 日から始まっている。このため単語「映像」が頻出コーパス C の順位に反映されていないため逆重付指標による並替で高い順位になっている。

(8) 逆重付指標による並替

単語	重要度	降順並替
環境	5.45	5.45
映像	5.23	5.23
ドジャース／試合／取材	5.04	5.04
可能性／韓国／水原	5.02	5.02
視聴	4.97	4.97
金利	4.92	4.92

02024/03/21	75.00	該当数		出現1	出現2	出現1-IB		出現2-IB	2TF-IB
大谷・ドジャース	14.67	11		55	29	1		1	14.67
大谷・通訳	13.33	10		55	37	1		1	13.33
大谷・翔平	13.33	10		55	17	1		1	13.33
水原・翔平	12.00	9		42	17	1		1	12.00
大谷・水原	12.00	9		55	42	1		1	12.00

02024/03/19	82	該当数		出現1	出現2	出現1-IB		出現2-IB	2TF-IB
金利・政策	13.41	11		85	38	1		1	13.41
金利・解除	13.41	11		85	32	1		1	13.41
金利・マイナス金	13.41	11		85	23	1		1	13.41
環境・政策	13.41	11		49	38	1		1	13.41
環境・目標	13.41	11		49	22	1		1	13.41

(8) 3 月 17 日（日）～3 月 23 日（土）

①MLB ソウル開幕戦【ドジャース・試合・取材／5.04、大谷・ドジャース／14. 67】 【可能性・韓国・水原／5.02】【大谷／ 4.54】

『3 月 20 日(水)ソウルで MLB きょう開幕!初対決へ、大谷翔平 「思い入れがあるので楽しみ 」ダルビッシュ「私情を入れずに」、2024 年 03 月 20 日 09:20:43』『スポーツ報知、スポーツ、大谷翔平の水原一平通訳がドジャース解雇…米報道、大谷資金で巨額の賭博疑惑 開幕戦翌日に衝撃、2024 年 03 月 21 日 09:20:11』

②日銀が 17 年ぶり利上げ決定【金利＝4.92／金利・政策＝13.41】

『Bloomberg 経済,主要 日銀が 17 年ぶり利上げ決定、世界最後のマイナス金利に幕 2024 年 03 月 19 日 15:20:59』

3.3 語の重要度と社会的トピックスの決定

8 週間のトピックを時系列にリスト化して、係数付き重付指数和＝MF-ib×2＋2TF-ib を計算し、その結果を「表 5」にまとめた。

【表 5： トピックスの時系列リスト】

	期間	トピックス	MF-ib	2TF-ib	重付指数	記事の内容
(1)	1 月 28 日 ～2 月 3 日	能登半島地震 1 ヶ月	5.02	15.85	25.89	『朝日新聞デジタル、地域・主要、能登半島地震から 1 カ月、復興の道筋みえず、被害把握もいまだ不十分 2024 年 02 月 01 日 00:00:10』
		東アジア半日武装戦線の桐島聡容疑者の出頭	4.58	8.99	18.15	『読売新聞オンライン、国内、連続企業爆破事件の桐島容疑者名乗る男、神奈川の土木会社に長期間勤務、2024 年 01 月 28 日 21:25:49』
(2)	2 月 4 日 ～2 月 10 日	大雪・注意	5.22	15.66	26.1	『毎日新聞、国内・主要、関東甲信、5 日昼過ぎ～6 日に大雪の恐れ、東京 23 区も積雪見込み、2024 年 02 月 05 日 00:20:25』

(3)	2月11日 ～2月17日	4歳次女を中毒死 両親を逮捕	5.05	13.19	23.29	『読売新聞オンライン、地域主要、4歳次女殺害容疑の両親、複数の有毒物質をネット検索…スマホ・ＰＣに履歴、2024年02月16日12:00:19』
(4)	2月18日 ～2月24日	ウクライナ侵攻が3年目突入甚大な犠牲と細る国際的支援で疲弊	5.81	15.79	27.41	『Yahoo!ニュース オリジナル 特集、主要,国際、侵攻から2年、終わりの見えない戦闘、ウクライナ人の領土断念という「不確かな選択肢」の意味 、2024年02月22日18:20:46』
(5)	2月25日 ～3月2日	岸田首相の政治倫理審査会出席	5.88	11.25	23.01	『FNN プライムオンライン（フジテレビ系）、国内主要、【速報ライブ】岸田首相政倫審出席「今の政治を未来に引き継げるか申し訳ない」、2024年02月29日15:20:15』
		大谷選手の結婚公表	5.88	7.87	19.63	『大谷翔平 大リーグ・ドジャースの大谷翔平投手（２９）が２９日に自身のインスタグラムで結婚を発表した。【写真】大谷の結婚発表投稿　ともに歩むパートナーの姿がチラリ「いつも温かい応援をいただきありがとうございます」と書き出した文章をアップ』
(6)	3月3日 ～3月9日	トランプ氏共和党大統領指名獲得	4.24	4.6	13.08	『ロイター主要,国際 ヘイリー氏指名争い撤退表明、米大統領選バイデン対トランプ確実に2024年03月07日03:20:08』
(7)	3月10日 ～3月16日	東日本大震災13年経過	4.95	20.78	30.68	『tenki.jp、科学主要、東日本大震災から13年、大地震や津波発生で慌てないために普段からの心構えを、2024年03月11日15:55:10』
		裁判関係の記事重複	5.88	7.87	19.63	『同性婚認めぬ規定「違憲状態」 3例目、賠償請求は棄却　全国で判断割れる・東京地裁』
(8)	3月17日～3月23日	MLBソウル開幕戦	5.04	14.67	24.75	3月20日(水)ソウルでMLBきょう開幕!初対決へ、大谷翔平 「思い入れがあるので楽しみ 」ダルビッシュ「私情を入れずに」、2024年03月20日 09:20:43 『スポーツ報知 スポーツ, 大谷翔平の水原一平通訳がドジャース解雇…米報道　大谷資金で巨額の賭博疑惑　開幕戦翌日に衝撃 2024年03月21日09:20:11』
		日銀が17年ぶり利上げ決定	4.92	13.41	23.25	（2）『Bloomberg 経済,主要 日銀が17年ぶり利上げ決定、世界最後のマイナス金利に幕 2024年03月19日15:20:59』

3.4　語の重要度と社会的トピックスの順位

係数付き重付指数和を用いてトピックスの順位を並べ替えて「表6」を作成した。

【表6： Yahoo!ニュースの週間トピックスの順位】

重付指数	トピックス	MF-ib	2TF-ib	期間
30.68	東日本大震災13年経過	4.95	20.78	3月10日～3月16日
27.41	ウクライナ侵攻が3年目突入、甚大な犠牲と細る国際的支援で疲弊	5.81	15.79	2月18日～2月24日
26.1	大雪・注意	5.22	15.66	2月4日～2月10日
25.89	能登半島地震１ヶ月	5.02	15.85	1月28日～2月3日
24.75	MLBソウル開幕戦	5.04	14.67	3月17日～3月23日
23.29	4歳次女を中毒死 両親を逮捕	5.05	13.19	2月11日～2月17日

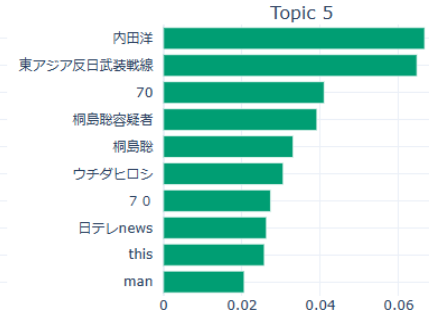
23.25	日銀が 17 年ぶり利上げ決定	4.92	13.41	3 月 17 日～3 月 23 日
23.01	岸田首相の政治倫理審査会出席	5.88	11.25	2 月 25 日～3 月 2 日
19.63	大谷選手の結婚公表	5.88	7.87	2 月 25 日～3 月 2 日
19.63	裁判関係の記事重複	5.88	7.87	3 月 10 日～3 月 16 日)
18.15	東アジア半日武装戦線の桐島聡容疑者の出頭	4.58	8.99	1 月 28 日～2 月 3 日
13.08	トランプ氏共和党大統領指名獲得	4.24	4.6	3 月 3 日～3 月 9 日

3.5 BERTopic 法について

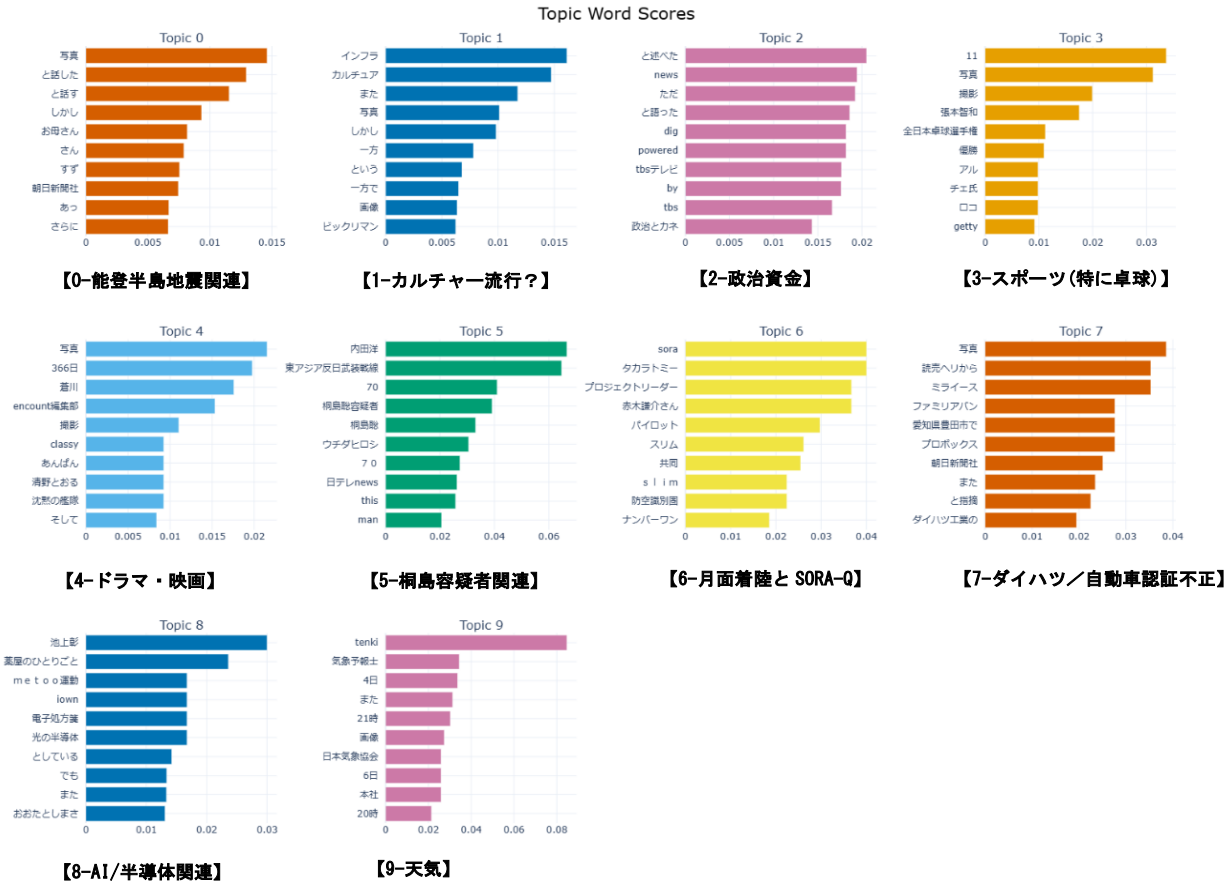
3.5.1 トピック抽出の手順

確率-ib と比較するために、BERTopic を利用して同一データからトピックを抽出した。BERTopic は、Google の開発した深層学習モデル Transformer を文書の分類に利用する手法である。Grootendorst の 2022 年の論文に github 上のリソースが記載されている。本研究グループは、このシステムを Google Colab 上に構築した。sentence-transformer モデルとして、多言語モデルである paraphrase-multilingual-MiniLM-L12-v2 を用いた。BERTopic のトピック抽出の手順は次のようになる。

【表 7： 1 月 28 日～2 月 3 日の Topic 5 の単語の重要



トピック表現	TF-IDF
内田洋	0.066618592
東アジア反日武装戦線	0.064653205
70	0.040980171
桐島聡容疑者	0.039122973
桐島聡	0.03305192
ウチダヒロシ	0.030514039
７０	0.027294122
日テレnews	0.026244755
this	0.025707049
man	0.020566045



【0-能登半島地震関連】 【1-カルチャー流行？】 【2-政治資金】 【3-スポーツ(特に卓球)】

【4-ドラマ・映画】 【5-桐島容疑者関連】 【6-月面着陸と SORA-Q】 【7-ダイハツ／自動車認証不正】

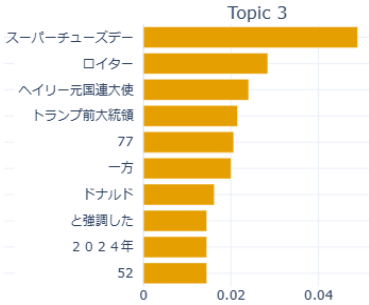
【8-AI/半導体関連】 【9-天気】

【図 4： 1 月 28 日～2 月 3 日の各クラスのトピック表現／重要度順／結果一覧】

① BERT を使って文章をベクトル化する
② 文書ベクトルを次元圧縮(UMAP)する
③ HDBSCAN によって、文書ベクトルをクラスタリングしてトピックのクラスを作成する
④クラス毎に含まれる単語をすべて結合し、1 つの文章として扱ってクラス単位の TF-IDF を計算して重要度を決定する

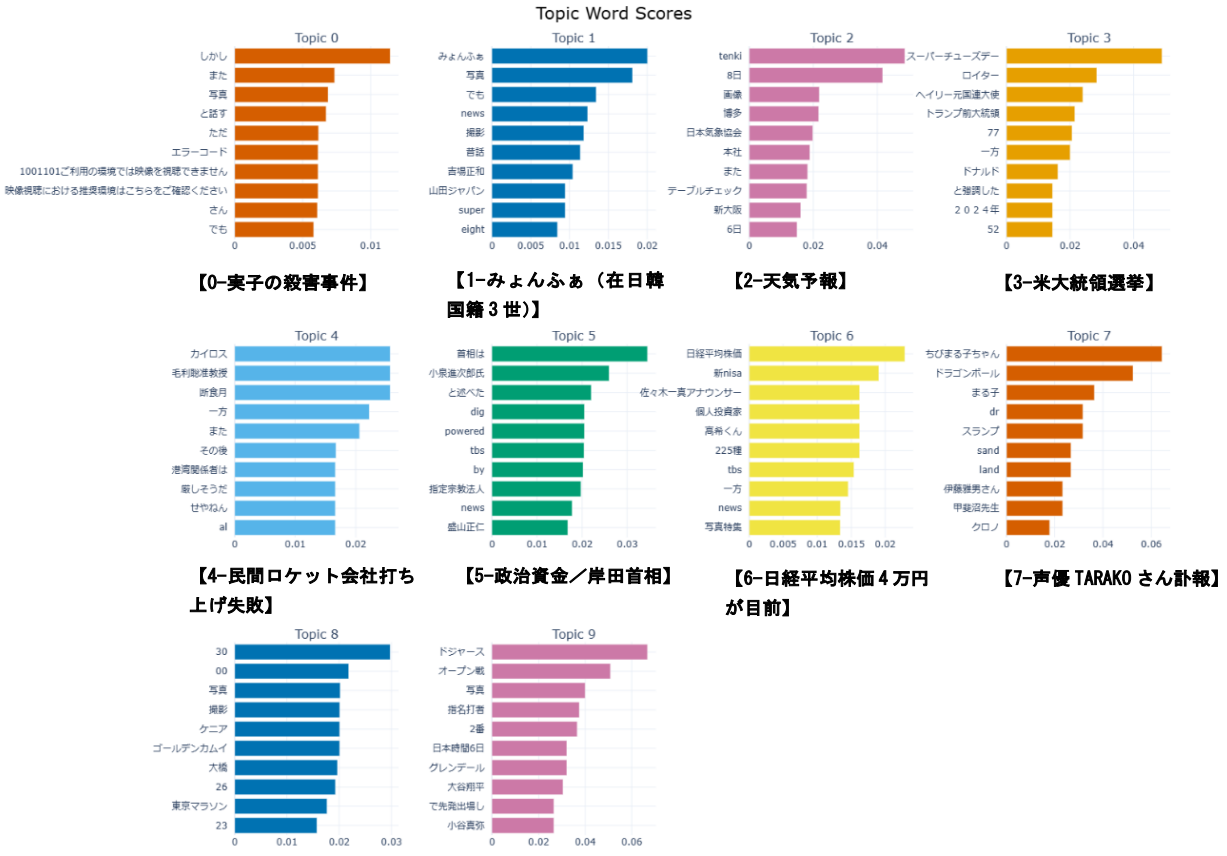
本稿では、1 月 28 日から 2 月 3 日および 3 月 3 日から 3 月 9 日の 2 つの分析結果を示した。1 月 28 日の週のデータから合計 10 個のクラスができた。(「図 4」)。その中で、Topic 3 について抜き出したのが「表 7」である。Topic 3 の内部では、「内田洋」「東アジア半日武装戦線」「70」「桐島聡容疑者」の重要度が高い。BARTopic は、クラス内の重要度の計算に tf-idf をもちいる。なお「内田洋」は桐島容疑者の変名である。Topic 3 は、確率-ib の「容疑者・桐島」に該当し、この週の「東アジア半日武装戦線の桐島聡容疑者の出頭」というトピックの重要性について両者は一致している。

【表 8：3 月 3 日～9 日の Topic 3 の単語の重要度】



トピック表現	TF-IDF
スーパーチューズデー	0.0489325
ロイター	0.028399139
ヘイリー元国連大使	0.024031227
トランプ前大統領	0.021491092
77	0.020618373
一方	0.020015211
ドナルド	0.016185879
と強調した	0.014466188
２０２４年	0.014466188
52	0.014466188

つづいて、3 月 3 日の週は合計 10 個のクラスができた。(「図 5」)そのなかで、Topic 5 について抜き出したのが「表 8」である。このクラスでは、「スーパーチューズデー」「トランプ前大統領」「ロイター」「ヘイリー元国連大使」といった単語の重要度が高い。Topic 3 は、確率-ib の「トランプ氏」に該当し、この週の「トランプ氏共和党大統領指名獲得」というトピックの重要性について両者は一致している。



【0-実子の殺害事件】 **【1-みょんふぁ（在日韓国籍 3 世）】** **【2-天気予報】** **【3-米大統領選挙】**

【4-民間ロケット会社打ち上げ失敗】 **【5-政治資金／岸田首相】** **【6-日経平均株価 4 万円が目前】** **【7-声優 TARAKO さん訃報】**

【8-東京マラソンとパリ五輪代表選考】 **【9-カブスの今永／ドジャースの大谷】**

【図 5：3 月 3 日～9 日の各クラスのトピック表現／重要度順／結果一覧】

3.5.2 BARTopic の特徴

BARTopic のトピック抽出は、確率-ib のような単語の出現頻度ではなく、確率的な距離を使ってベクトル化した単語の次元圧縮とクラスタリングを用いている。BARTopic では、単語のクラスタリングという手法を介するために、クラス間の横断的な重み付けができない。また、それぞれのクラスが、具体的にどのトピックと結び付いているのかは、この手法でも、元の記事から推測する必要がある。「図 4」および「図 5」では、これを【0-能登半島地震関連】のように記入した。本研究では、BARTopic によって抽出したトピックスが、確率-ib の重要度指数によるトピックスをおおむね包含していることを確認した。[14]

4. まとめにかえて

本研究では、オンラインニュースのトピック抽出を、確率-ib による重要度指数によって行い、その結果を BERTopic と比較した。また、確率-ib による重要度指数を Yahoo!ニュースの 8 週間分のデータに適用して、1 月 21 日から 3 月 23 日の間のトピックスの総体的な順位を決定した。この結果、「東日本大震災 13 年経過」および「ウクライナ侵攻 3 年目」といった総括的な記事が上位に入った。これは Yahoo!ニュースが、各紙の記事のクリッピングをしているために、情報の集積効果が働くためかもしれない。これに対して「日銀のゼロ金利解除」「MBL の大谷選手の活躍」「岸田首相の政治倫理審査会出席」といったトピックスの顕在化は、日々の感覚と一致していた。

本研究で用いた「① 確率的な単語の重要度の指数化 ⇒ ② ib による頻出語のオフセット」という 2 段階の語の重要度の計算方法は、直感的に明らかなもので所期の結果を得た。この計算方法は、一般的な tf-idf 法の正当性を確率論的に補完するものである。[15] この計算方法では、オンラインニュースのトピックないしは単語の重要度を、コーパス全体の単語の出現頻度と比べた特定文書の確率的な偏りとして考えることになる。フォーカル・トピックの出現にともなって、ニュースの言語空間には、つねに単語の度数間の移動が生じている。これは「図 2」で示した確率密度曲線上の度数の移動として現れる。この確率の変動を、コーパス全体の出現頻度と比べた、特定文書の確率的な偏りと考えてトピックを検出する。言語データがつねに Zipf 則に従っているため、これは tf-idf 法についても同一だ、と考えることができる。

単語のコーパス全体の一般性として、日常の頻出語を見た場合、Twitter と Yahoo!ニュースの違いが明らかになる。Tweet は全角 140 文字の制限があり、また日常的な感情や感想の表明である。Twitter の頻出語と MF-ib による重要度指数は、以下のようになっている。

{今日, 8.09457}, {時間, 6.86797}, {自分, 6.80931}, {昨日, 6.63927}, {仕事, 6.54958}, {本当, 6.35079}, {楽しみ, 6.22546}, {ツイート, 6.16771}, {最近, 6.15571}, {本日, 6.15269}, {お疲れ様, 6.11559}, {フォロー, 6.10288}, {気持ち, 6.06366}, {明日, 6.0331}, {一日, 5.99791}, {無理, 5.96131}, {一緒, 5.91537},

これに対して、Yahoo!ニュースの頻出語と MF-ib 指数は以下の通りである。Yahoo!ニュースの記事は、Tweet よりも重み付け値の差が平坦である。つまり、「図 3」の重要度指数のグラフで左端の値が低いことになる。

{"日本", 6.00522}, {"発表", 5.70693}, {"以上", 5.40036}, {"影響／昨年", 5.32854}, {"東京", 5.31699}, {"必要", 5.3053}, {"男性", 5.2446}, {"撮影", 5.18645}, {"午後／今回", 5.17977}, {"可能性", 5.14561}, {"選手", 5.1316}, {"代表", 5.10295}, {"確認／画像", 5.08829}, {"状況", 5.01143}, {"現在", 5.00339}, {"問題", 4.90125}, {"活動", 4.89222},

MF-ib、2TF-ib および BERTopic の言語解析を組み合わせて、オンラインニュースや UGC のデータに適用し、フォーカル・トピックの時間的な推移やトピックの全体的な順位を決めることができる。確率-ib では、idf（＝ib）の適用によって日常的な頻出語が省かれるために、重要度指数の高い単語と、そのニュースの記事のトピックは同一だ、と考えている。これに対して、BARTopic では、データにクラスタリングを適用するために、より幅広いトピックを抽出し、同時に重要単語と記事のトピックの結び付きが（原理的には）明確になる。本研究の現段階では、「3.3 語の重要度と社会的トピックスの決定」で用いた「重付指数和＝MF-ib×2＋2TF-ib」 という 2 つの指数の加算の係数や、重付指数 ib の底の推定など、実験結果にもとづく経験則に拠った部分が多い。今後のリアルタイムの解析のためには、言語リソースに対応した、このような式や定数の決定を含むシステムの実装が課題になる。

[1] 本研究の実施に際して、株式会社構造計画研究所から貴重なご支援を戴いた。自然言語処理の確率論的研究の重要性に着目された服部正太代表執行役会長に御礼申し上げたい。なお本研究にあり得る瑕疵はすべて執筆者に属するものである。

[2] Karen Sparck Jones, “IDF term weighting and IR research lessons,” *Journal of Documentation,* 60(5), October 2004.

[3] Maarten Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv:2203.05794, 11 Mar, 2022. https://arxiv.org/abs/2203.05794

[4] 安定分布の紹介については Wikipedia の記事を参照。

[5] 本研究は、確率過程論国際学会会長を務められた飛田武幸教授と多摩大学情報社会学研究所の公文俊平所長のアイデアを発展させたものである。飛田武幸『確率論の基礎と発展』共立出版、2011 年、228 頁。山内康英、小松正、永井拓史「SNS の Zipf 則と安定分布」情報社会学会編『情報社会学会誌』2019 年、Vol.14 No.1。

[6] 『時事通信社が選ぶ 10 大ニュース（2023 年）特集』 https://www.jiji.com/jc/d4?p=jtn223-jpp046037384&d=d4_oldnews

[7] 黒橋禎夫『自然言語処理』放送大学、2023 年、181 頁。

[8] 中里朋楓、澁谷遊野、大西正輝、高木聡一郎「Twitter 上の COVID-19 ワクチンに関するトピック・道徳基盤と有害性の関係性」人工知能学会『2024 年度人工知能学会全国大会論文集』。https://www.jstage.jst.go.jp/article/pjsai/JSAI2024/0/JSAI2024_1L5OS4b01/_article/-char/ja

[9] 指数は分母が 0 にならないようにして対数をとる。またそれ以外にも各種の手法が考案されている。tf-idf については Wikipedia の記事を参照。

[10] Bag of Words とは、形態素解析をして文書をバラバラにし、単語の集まり（袋）として扱う手法の総称である。文書の文脈や語順を無視して単語の出現頻度だけに着目する。

[11] 社会のフォーカル・トピックが、政治的な争点となって繰り返しメディアに浮上するパタンを、社会学では「社会問題の研究（studies of social problems）」という。ジョエル・ベスト『社会問題とは何か―なぜ、どのように生じ、なくなるのか?』赤川学訳、筑摩書房、2020 年。

[12] 山本和英『テキスト処理の要素技術：実践・自然言語処理シリーズ』近代科学社 2021 年、第 5 章。

[13] この 2 つのオフセットの条件式は、MF の計算を永井が、2TF を酒井が担当した、という経緯から生じたもので本質的ではない。

[14] 各クラスが具体的にどのトピックに結び付くのかは、BARTopic でも元の記事と付き合わせて確認する必要がある。本研究では、クラス決定後の tf-idf の適用によって、上位に来る単語が、クラス内のトピックに結び付く可能性が高いと考え、オンラインニュースの該当期間の記事から【3-米大統領選挙】などを決定した。両手法とも記事とトピックの付き合わせは、今後の研究課題である。このため、ここでは「おおむね包含している」とした。

[15] 特徴語抽出法としての tf-idf 法のエントロピー概念を使った説明として以下を参照。「APPENDIX_7 特徴語の抽出」文部科学省 科学技術・学術政策研究所, サイエンスマップ 2016, NISTEP REPORT No.178, 2018 年 10 月。この論文によれば、tf-idf は平均相互情報量の特殊な形式になる。http://www.nistep.go.jp/wp/wp-content/uploads/NISTEP-NR178-Appendix7.pdf

付記：本稿の査読を担当された情報社会学会編集委員会のメンバーに御礼申し上げたい。執筆者のグループは、最初のバージョンの査読結果に対応するために、異なるトピック抽出手法の比較という新たな論点を原稿に導入した。また査読者のコメントに従って誤記や論文の体裁を修正した。

（2024 年 8 月 30 日受理）