第15回産業日本語研究会講演・2024年・2月20日
オンラインwebinar会場

大規模言語モデルの持つ言語知識とコミュニケーション

青山学院大学名誉教授・近藤泰弘

内容

1. **生成AIの仕組み**
2. 埋め込みベクトル
3. 単語ベクトル
4. 文ベクトル
5. 文ベクトルの応用

生成AIの仕組み

decoder のみの transformer

事前学習時

CLM（次の単語を予測する）方法で学習する

「吾輩は猫で」「ある」

前の文脈だけを見て、次と合致するかで正解を決めて学習していく。

モデルの訓練

テキストを「トークン」（単語）に分割し、これを位置などを配慮したベクトル（数値）に変換。

文脈を考慮（アテンション）しつつ、モデルに学習させる。結果はベクトル値としてネットワークの各層に保存される。

生成時

プロンプトが初期入力となり、次の単語を繰り返し予測生成する。

「吾輩は猫で」

「あるは夏目漱石の作品として有名である。漱石は明治時代の文豪として、森鴎外などと並び称される。」

議題のページに戻る

生成AIの仕組み

あなた

吾輩は猫で

ChatGPT Classic

ある。名前はまだない。

どこで生まれたかとんと見当がつかぬ。何でも薄暗いじめじめした所でニャーニャー泣いていた事だけは記憶している。吾輩はここで始めて人間というものを見た。そして、ここで人間というものがどんなものかを少しは知ることができた。

夏目漱石の『吾輩は猫である』の冒頭部分ですね。続きや、この作品について何か知りたいことがありますか？

議題のページに戻る

生成AIの仕組み

あなた
吾輩は猫であるは

ChatGPT Classic
『吾輩は猫である』は、夏目漱石によって1905年から1906年にかけて「ホトトギス」誌上に連載され、その後、1907年に単行本として出版された長編小説です。日本の近代文学を代表する作品の一つであり、漱石のデビュー作としても知られています。

この小説は、名前のない猫が語り手となり、東京のある家庭に迷い込んでからの出来事を、独特の視点とユーモラスな語り口で描き出します。人間社会の風刺や、当時の日本の知識人たちの生活や思想が、この猫の目を通して描かれているのが特徴です。社会の矛盾や人間の愚かさ、そして人間と動物との関係など、多様なテーマが織り交ぜられています。

『吾輩は猫である』は、そのユニークな視点と文体、そして深い洞察力で、多くの読者に愛され続けている作品です。夏目漱石の作品の中でも、特に人気が高く、日本文学の中で重要な地位を占めています。

議題のページに戻る

生成AIの仕組み

あなた

吾輩は猫である

ChatGPT Classic

「吾輩は猫である」は、夏目漱石の代表作の一つで、日本の近代文学の古典とされています。この作品は、猫が語り手となり、人間社会を独自の視点から観察し、その愚かさや矛盾を風刺的に描き出しています。漱石の洗練された文体とユーモアが光る作品で、多くの読者に愛され続けています。

議題のページに戻る

生成AIの仕組み

あなた
舞姫、吾輩は猫である

ChatGPT Classic
『舞姫』と『吾輩は猫である』は、明治時代を代表する文学作品であり、それぞれが異なるテーマを扱っていますが、日本の近代化と西洋文化の影響を背景に持っています。

舞姫

森鴎外の『舞姫』は、主に留学生の視点からドイツでの生活、恋愛、そして帰国後の文化的な断絶を描いています。この物語は、日本人留学生とドイツ人女性との恋愛を中心に展開し、恋愛関係の終焉と主人公の内面的葛藤を通じて、異文化間の交流の複雑さと、当時の日本社会の価値観との間に存在する矛盾を浮かび上がらせます。『舞姫』は、個人の感情と社会的期待との間の葛藤を描くことで、明治時代の日本人のアイデンティティの探求を描き出しています。

吾輩は猫である

夏目漱石の『吾輩は猫である』では、名前のない猫が語り手となり、その視点から見た人間社会

議題のページに戻る

生成AIの仕組み

あなた

次の語句を現代の非常にくだけた若者言葉にしてください。「吾輩は猫である。名前はまだない。」

ChatGPT Classic

「俺、猫っす。名前？まだなんも決まってねーんだわ。」

In Context Learning　（ICL）　ゼロ・ショット

議題のページに戻る

生成AIの仕組み

あなた

次の語句を絵文字にしてください。「吾輩は猫である。名前はまだない。」

ChatGPT Classic

「👉😸。名前は❓🚫」

議題のページに戻る

内容

1. 生成AIの仕組み
2. **埋め込みベクトル**
3. 単語ベクトル
4. 文ベクトル
5. 文ベクトルの応用

埋め込みベクトルの考え方1

	高い	長い	登る	流れる	頂上	橋	
山 [	1	0	1	0	1	0	]
川 [	0	1	0	1	0	1	]

山　[101010]
川　[010101]
（6次元の共起行列による
ベクトル表現）

丘　[101000]
「川」より「山」に近い。
実質的に意味を示している。

議題のページに戻る

埋め込みベクトルの考え方2

ニューラルネットワークの中に埋め込みベクトルを作成

Word2vec, FastText

100次元程度のベクトルで、前後に出現する他の単語との共起情報を用いて、それぞれの単語の意味を表現する。意味の加算、減算等が可能。

Transformer による埋め込みベクトル

BERT, GPT

数百から数千次元程度のベクトル。トランスフォーマーアーキテクチャから、埋め込みベクトルを作成する。

それぞれの単語ごとに与えられるベクトル値は、意味・統語情報・語用論など非常に多岐に渡る情報を持っている。

議題のページに戻る

単語ベクトルと文ベクトル

単語ベクトル

各トークンそのものあるいは、その合成されたベクトル

文ベクトル

文を構成するトークンのベクトルの平均値など

ファインチューニングによる文ベクトル

似たような文について同じ文ベクトルを返すように、ファインチューニングをする

議題のページに戻る

OpenAIの提供する埋め込みベクトル

- 今回は、OpenAIの提供する**text-embedding-ada-002**および最新の**text-embedding-3-large**というモデルを用いる。OpenAIのクラウドからAPIとしてアクセスできる。それぞれ1536次元及び3072次元のベクトルが得られる。GPT3あるいは3.5相当の数10〜数100億パラメタモデルを利用しているものと思われる。実質的にChatGPTの意味辞書と等価であるとみなすことができる。30カ国語以上の言語で同時学習している。

議題のページに戻る

OpenAIの提供する埋め込みベクトル

OpenAIの埋め込みベクトル

embedding": [-0.0017322878120467067, -0.01955992728471756,
0.01636701636016369, 0.007083430420607328,
0.005423251539468765, 0.012832009233534336, -0.011832548305392265,
-0.01664874516427517, -0.018594004213809967, -0.013361925259232521,
0.014354678802192211, 0.005547345615923405, -0.009840333834290504,
0.009766547940671444, -0.0065032062120735645, -0.01993556320667267,
0.024657849222421646, -0.01984165422618389,
, 0.011309340596199036, -0.0325193852186203, 0.017292693257331848,
(以下略）

議題のページに戻る

OpenAIの提供する埋め込みベクトル

OpenAIのクラウド

text-embedding-ada-001
text-embedding-3-large

← 単語やセンテンス

(API)

→ 埋め込みベクトル

議題のページに戻る

OpenAIのAPI呼び出し

```
url = "https://api.openai.com/v1/embeddings"
headers = {
   "Content-Type": "application/json",
   "Authorization": f"Bearer {openai.api_key}"
}
model_name = "text-embedding-3-large"
data = {
      "input" : sample_text,
      "model" : model_name,
      "dimensions" : 3072
}
response = requests.post(url, headers=headers, json=data)
response_data = response.json()
```

議題のページに戻る

OpenAIの埋め込みベクトル

embedding": [-0.0017322878120467067, -0.0195599272847175б,
0.016367016360163б9, 0.0070834304206073 28,
0.0054232515394687б5, 0.012832009233534336, -0.011832548305392265,
-0.016648745164275 17, -0.018594004213809967, -0.013361925259232521,
0.014354678802192211, 0.005547345615923405, -0.009840333834290504,
0.009766547940671444, -0.006503206212073 5645, -0.019935563206672б7,
0.024657849222421646, -0.019841654226183 89,
, 0.011309340596199036, -0.03251938521 86203, 0.017292693257331848,
(以下略)

議題のページに戻る

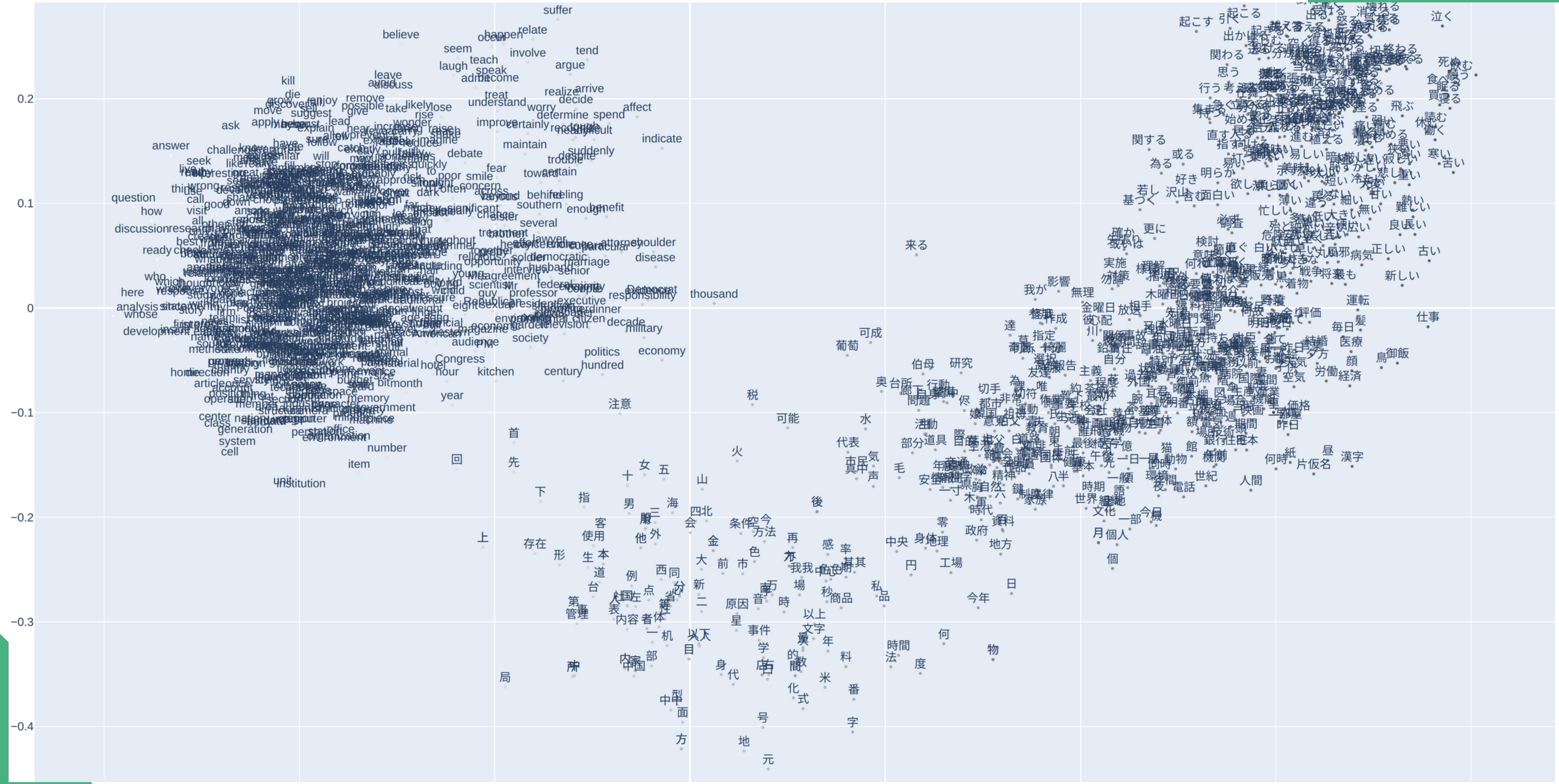
内容

1. 生成AIの仕組み
2. 埋め込みベクトル
3. **単語ベクトル**
4. 文ベクトル
5. 文ベクトルの応用

単語ベクトルの状態・多言語対応

- **PCA（主成分分析）**によって、高次元のベクトルを次元削減して、２次元にすることで、単語ベクトルのうち、もっとも重要な特徴が抽出可能。
- この単語埋め込みベクトルでは、まずX軸（第１次元・横軸）に**「言語の差」**が出てくる。当然、AIによる生成で、もっとも重要な点は、言語が統一されていることなので、当然である。
-
- Y軸（第２次元・縦軸）には、**「動詞か、それ以外か」**が出てくる。これは、確かに、言語にとって非常に重要な差異であり、生成AIにとっても重要なポイントとなる。

単語ベクトルの状態・日英語



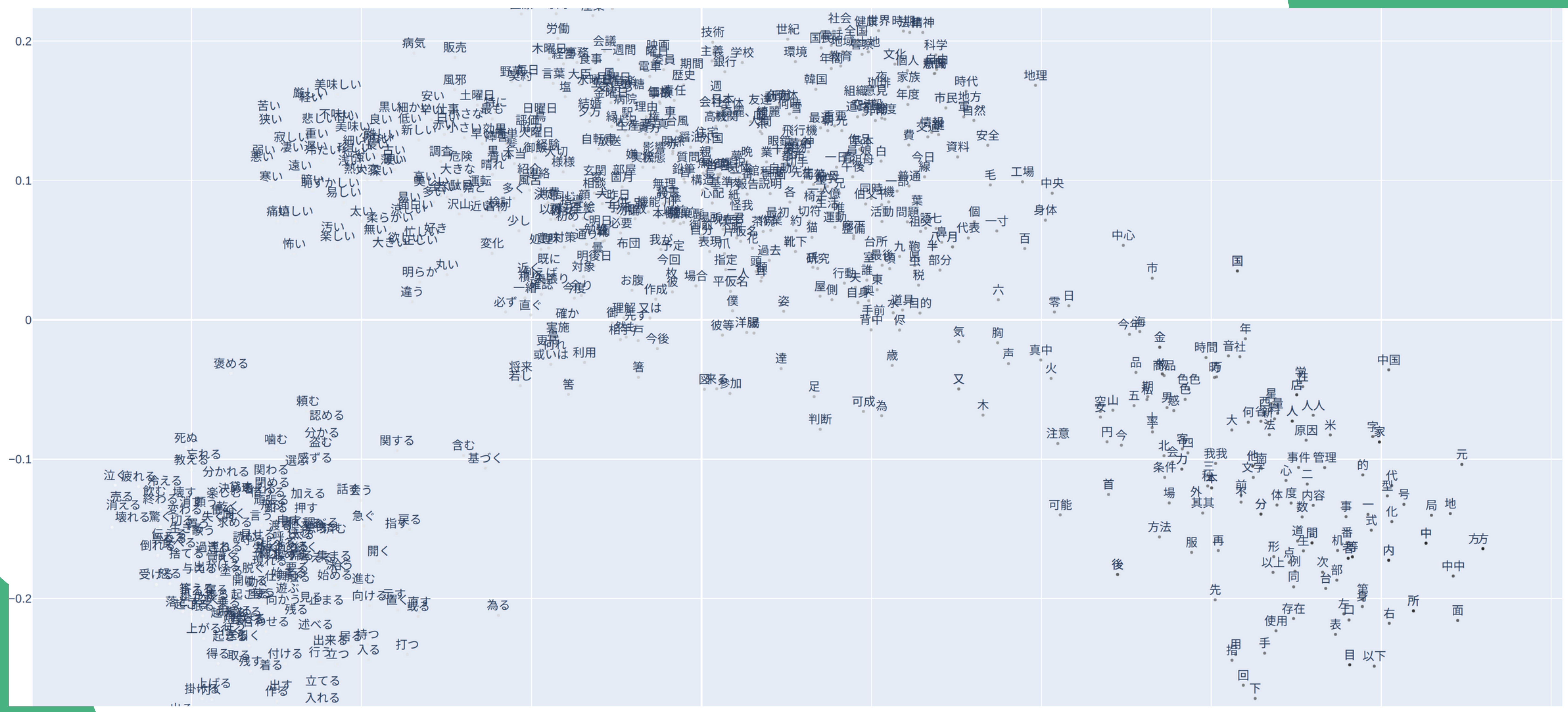
単語ベクトルの状態・日本語品詞

- 日本語だけに限定して主成分分析を行うと、品詞がかなりきれいに分離する。
- X軸は、「動詞か、そうでないか」であるが、具体的には

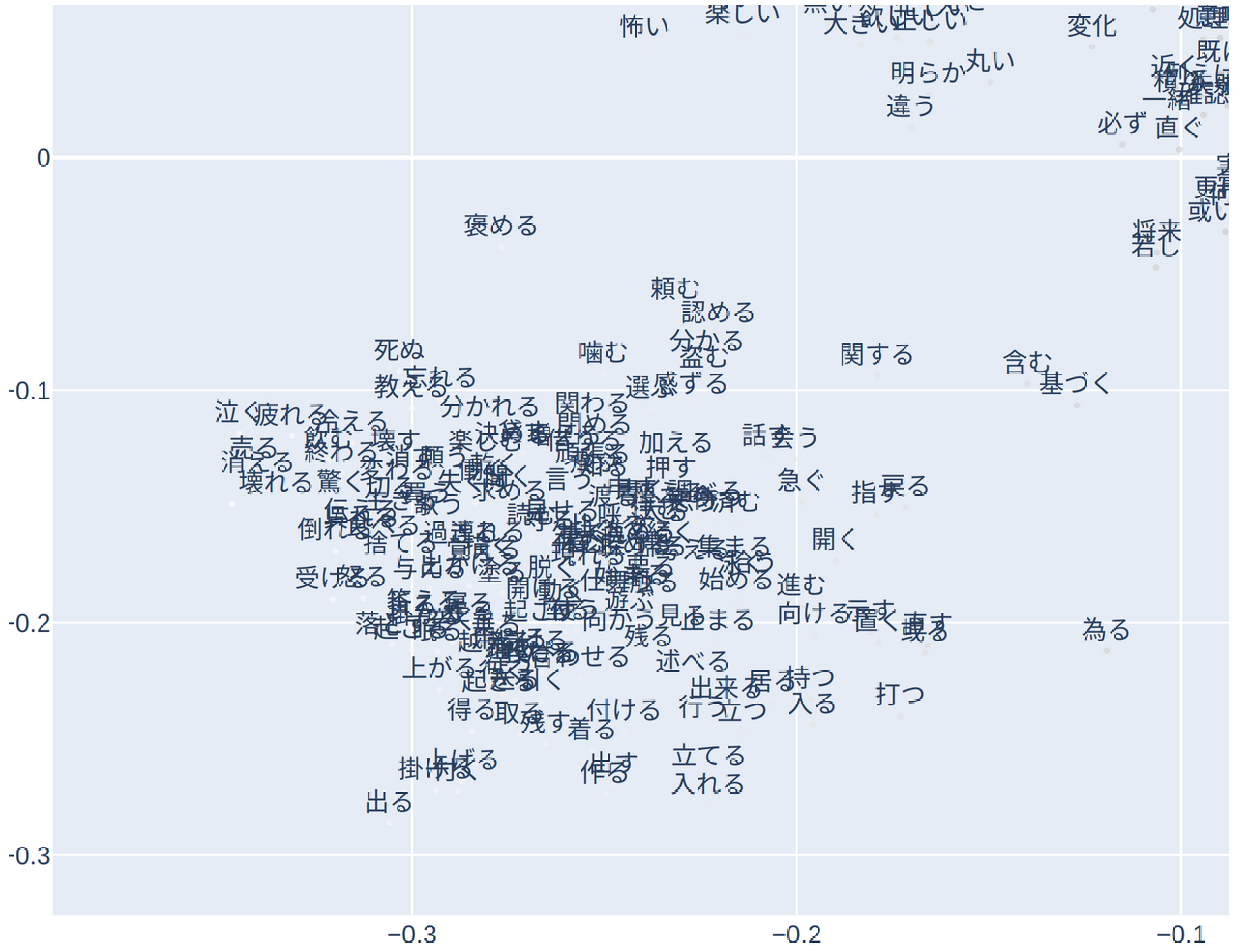
動詞ー形容詞ー名詞ー接辞

の形で品詞が分離する。

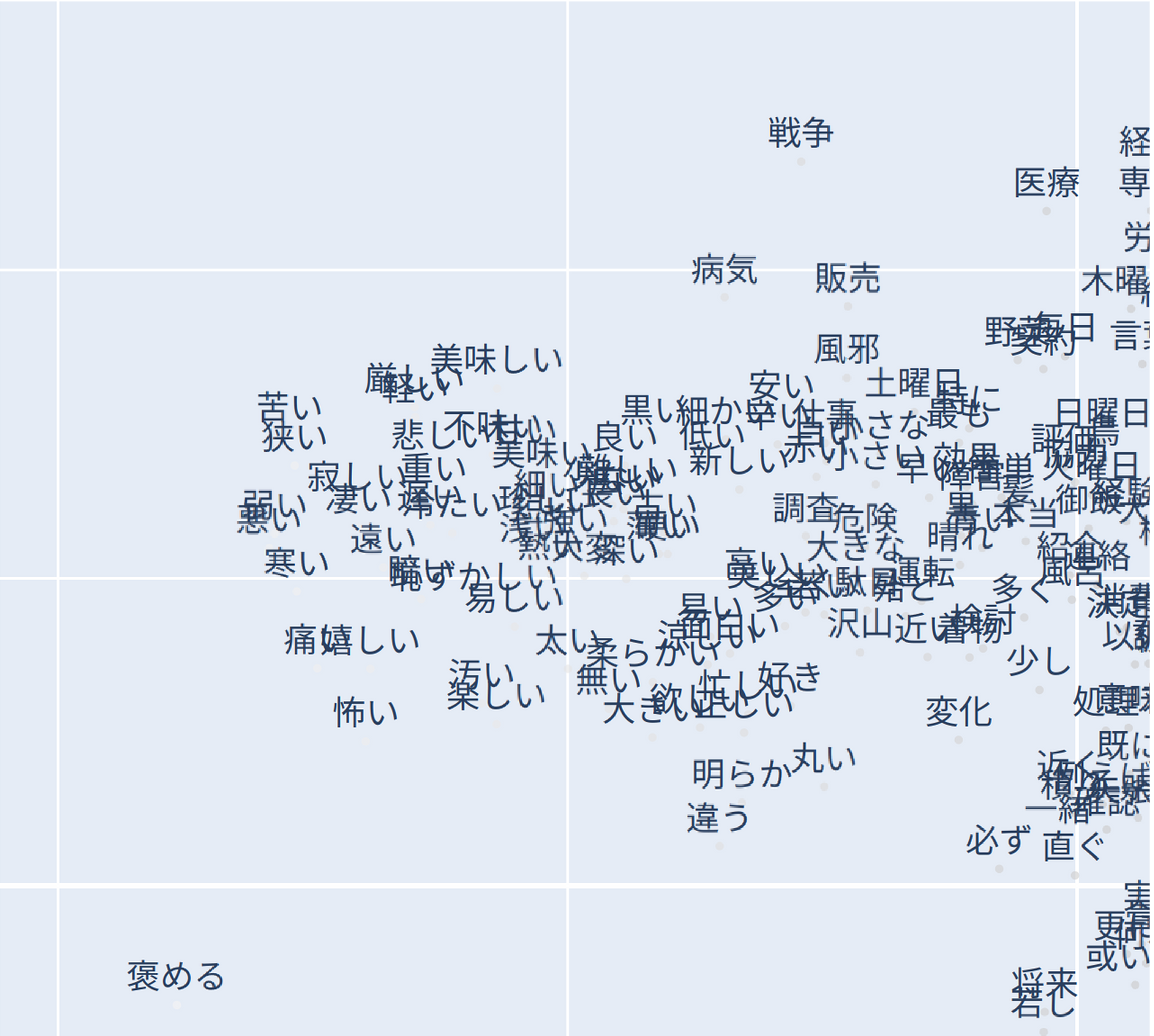
単語ベクトルの状態・日本語品詞



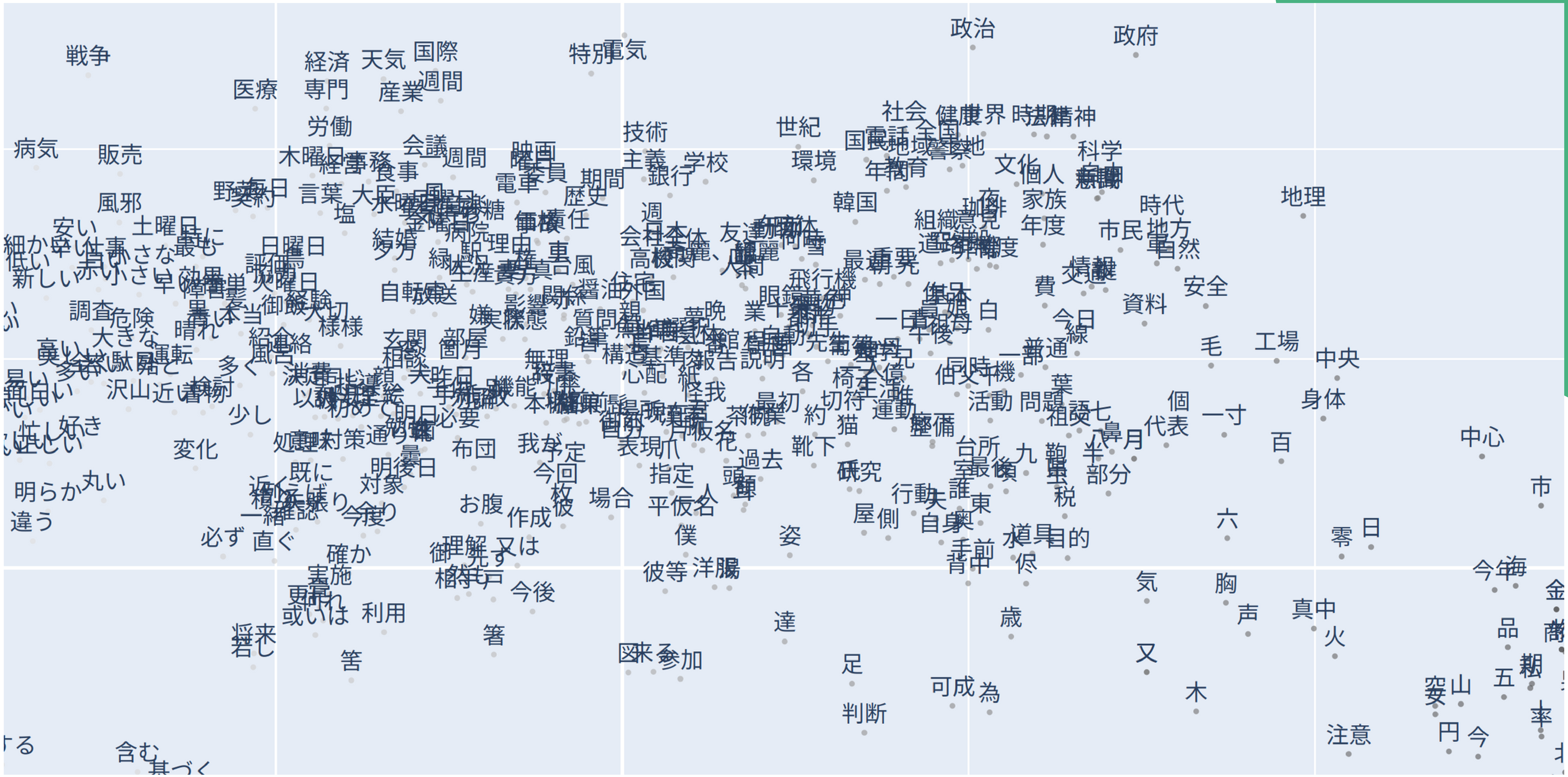
単語ベクトルの状態・動詞



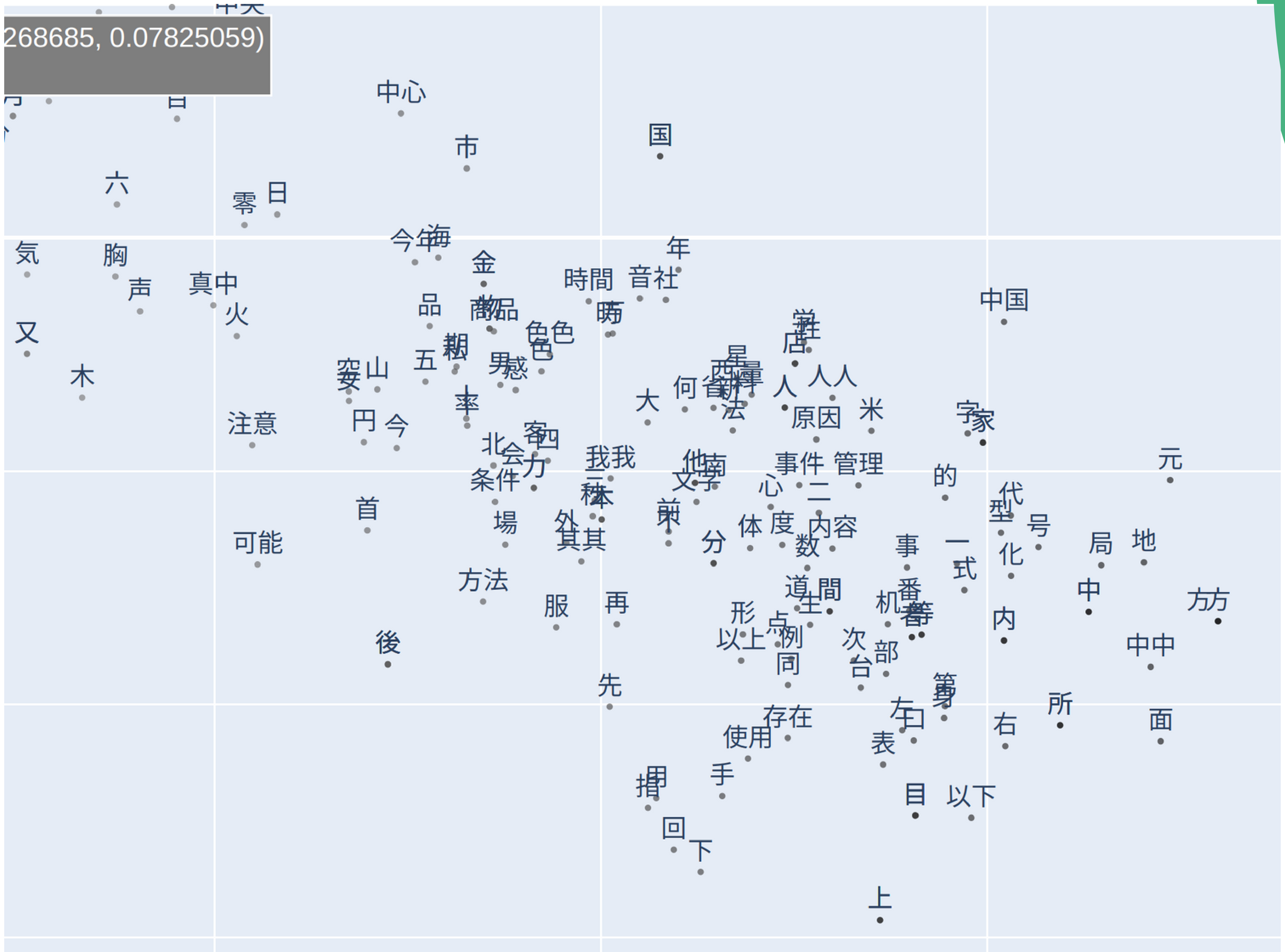
単語ベクトルの状態・形容詞



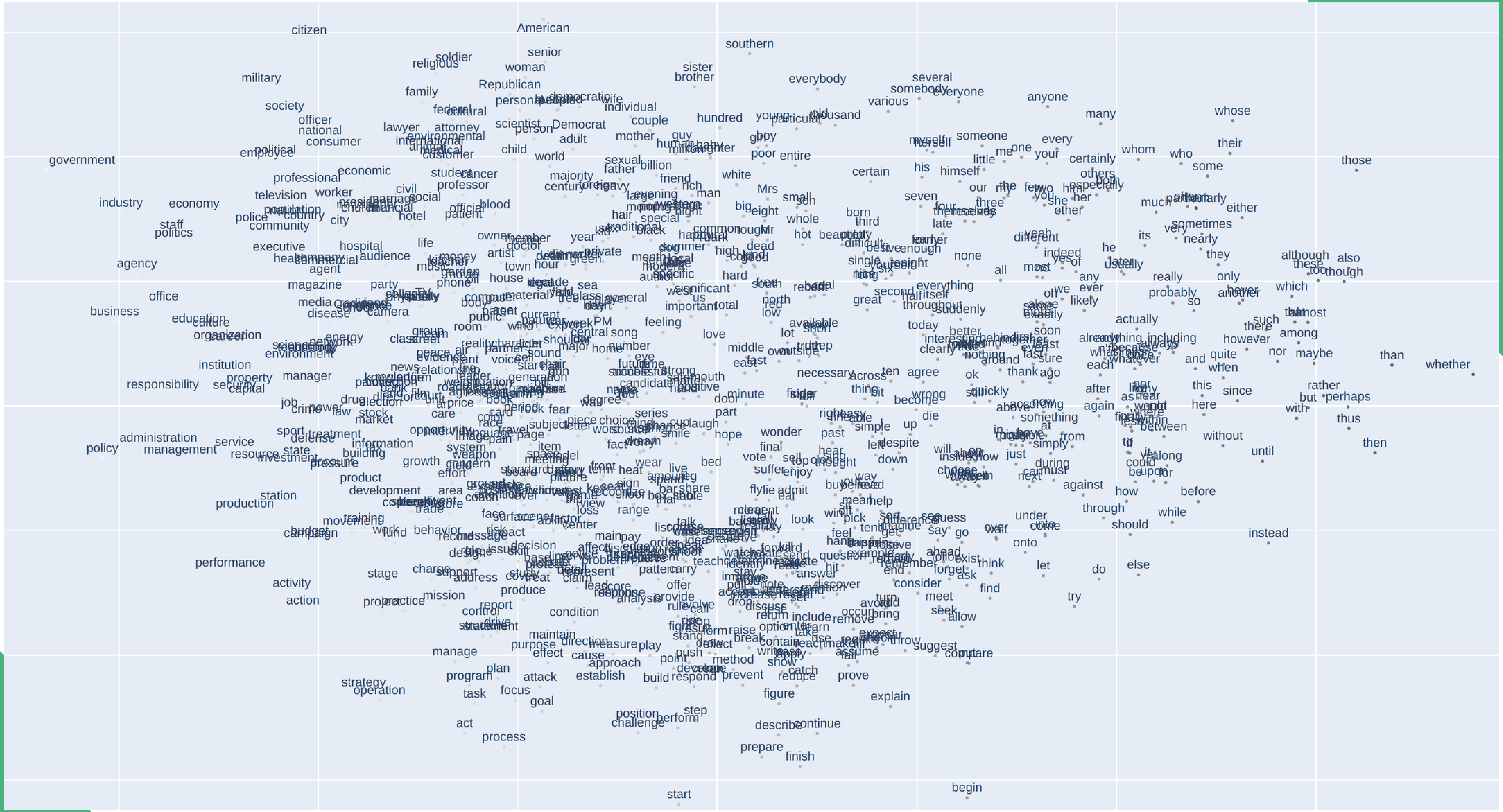
単語ベクトルの状態・名詞



単語ベクトルの状態・接辞



英語の単語ベクトルの状態



内容

1. 生成AIの仕組み
2. 埋め込みベクトル
3. 単語ベクトル
4. **文ベクトル**
5. 文ベクトルの応用

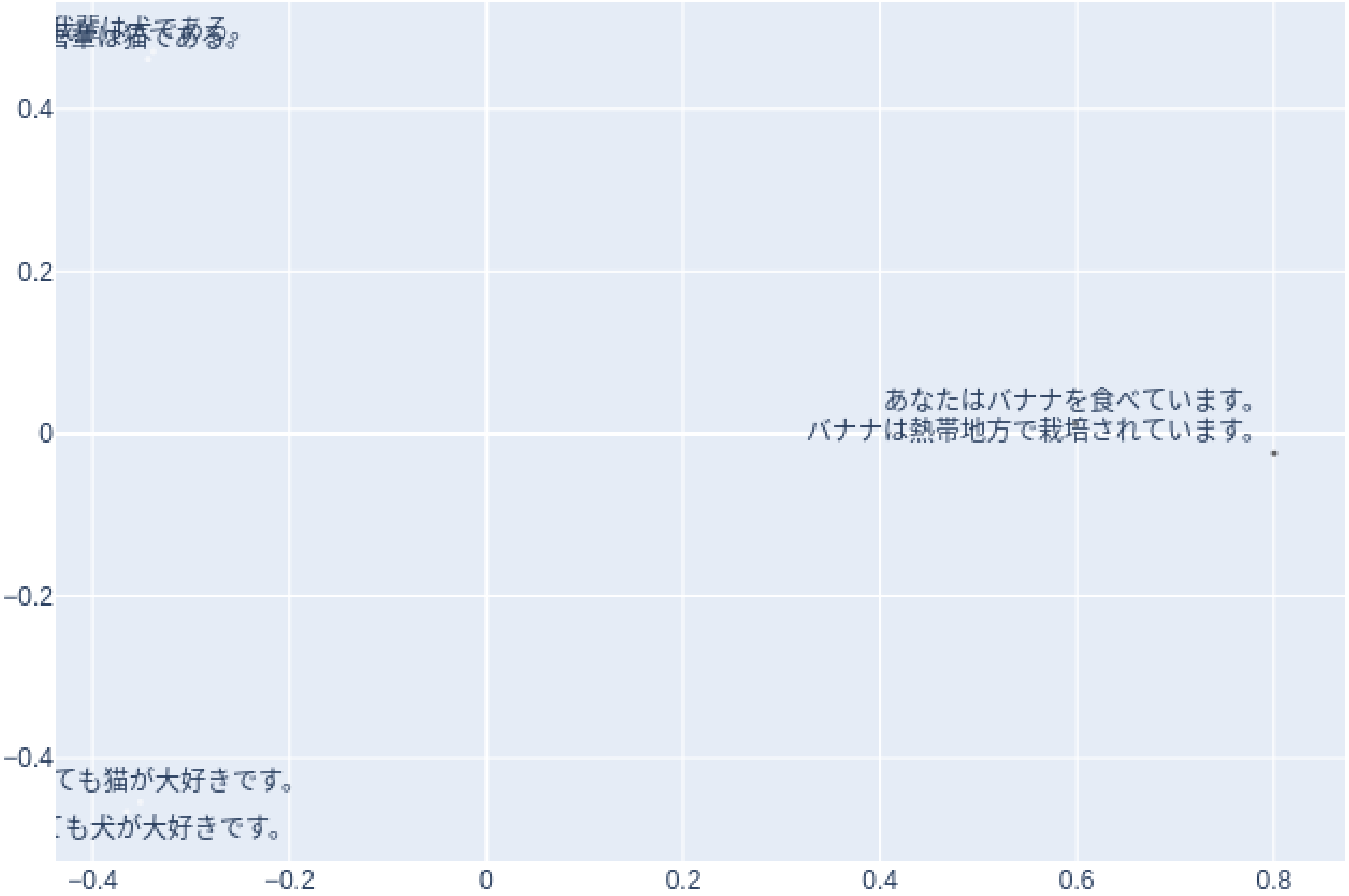
文ベクトルの利用

以上のように、埋め込みベクトルを次元圧縮をして**単語ベクトルの遠近関係**を見てきたが、文ベクトルにも同じようなことが可能である。**文と文との近さ**を測り、マッピングすることもできる。

文ベクトルの利用

1,私はとても猫が大好きです。
2,私はとても犬が大好きです。
3,吾輩は猫である。
4,我輩は犬である。
5,あなたはバナナを食べています。
6,バナナは熱帯地方で栽培されています。

文ベクトルの利用

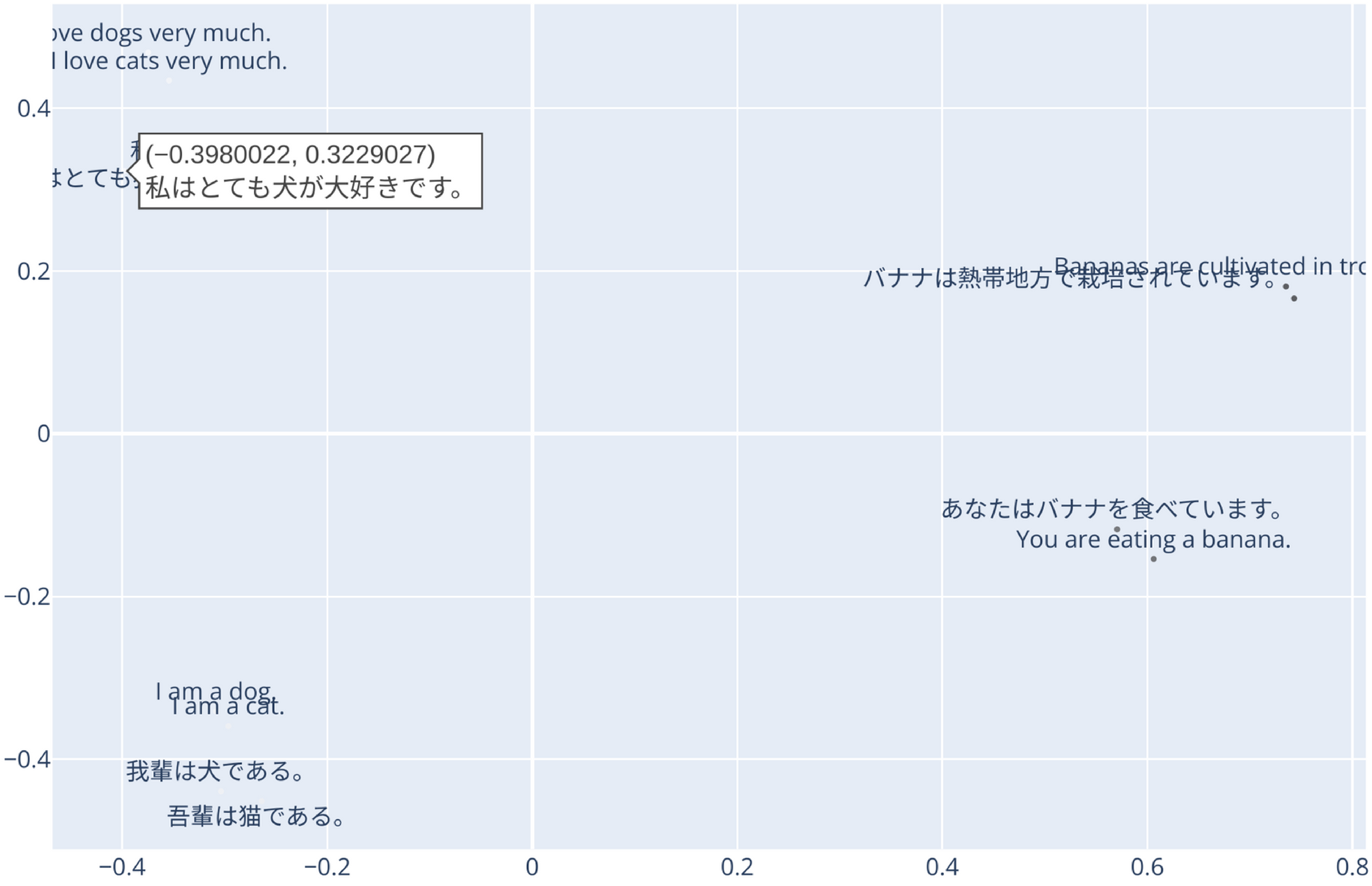


あなたはバナナを食べています。
バナナは熱帯地方で栽培されています。
ても猫が大好きです。
も犬が大好きです。
0.4
0.2
0
−0.2
−0.4
−0.4
−0.2
0
0.2
0.4
0.6
0.8

文ベクトルの利用

1,私はとても猫が大好きです。
2,私はとても犬が大好きです。
3,吾輩は猫である。
4,我輩は犬である。
5,あなたはバナナを食べています。
6,バナナは熱帯地方で栽培されています。
7,I love cats very much.
8,I love dogs very much.
9,I am a cat.
10,I am a dog.
11,You are eating a banana.
12,Bananas are cultivated in tropical regions.

文ベクトルの利用


ove dogs very much.
I love cats very much.
(−0.3980022, 0.3229027)
私はとても犬が大好きです。
Bananas are cultivated in tro
バナナは熱帯地方で栽培されています。
あなたはバナナを食べています。
You are eating a banana.
I am a dog.
I am a cat.
我輩は犬である。
吾輩は猫である。
0.4
0.2
0
−0.2
−0.4
−0.4
−0.2
0
0.2
0.4
0.6
0.8

文ベクトルの利用・英語で古語

- このように、文ベクトルは、文全体の特徴を反映するので、言語が異なっても、そのベクトル（数値）の比較で、類似を判定可能。
- 次の例は、英語と、日本古典語（古今集）をそれぞれベクトル化して、比較して、ベクトルのコサイン類似度のみで、もっとも適切な和歌を検出した。

文ベクトルの利用・英語と古語

質問文: I lament how quickly the seasons change

順位: 本文(コサイン類似度)

1: 秋風にあふ田の実こそ悲しけれわが身むなしくなりぬと思へば (0.8152505644733212)

2: 物ごとに秋ぞ悲しきもみぢつつ移ろひゆくを限りと思へば (0.8101851226217894)

3: 秋来ぬと目にはさやかに見えねども風の音にぞおどろかれぬる (0.8078393920033705)

4: おほかたの秋くるからにわが身こそ悲しきものと思ひ知りぬれ (0.8075361576183776)

5: 夕さればいとど干がたきわが袖に秋の露さへ置き添はりつつ (0.8066052282284499)

文ベクトルの利用・要約

質問文: 天皇の仕事はなんですか？
順位:番号・本文（コサイン類似度）
1: 11 第一章　天皇第一条　天皇は、日本国の象徴であり日本国民統合の象徴であつて、この地位は、主権の存する日本国民の総意に基く (0.5512648920323997)
2: 20 第七条　天皇は、内閣の助言と承認により、国民のために、左の国事に関する行為を行ふ (0.5440142478820604)
3: 14 第四条　天皇は、この憲法の定める国事に関する行為のみを行ひ、国政に関する権能を有しない (0.5401211034076915)
4: 15 (2)天皇は、法律の定めるところにより、その国事に関する行為を委任することができる (0.5299827335152074)
5: 13 第三条　天皇の国事に関するすべての行為には、内閣の助言と承認を必要とし、内閣が、その責任を負ふ (0.5141027807756001)

文ベクトルの利用・外部知識

- 先のように、本文を文ベクトルにしておくことで、類似・関連「検索」がベクトルの一致度を計測するだけの作業に単純化できる。
- ChatGPTで応答システムを作る場合でも、このように手元のデータをベクトルDBとして整備しておけば、この外部知識をChatGPTに与えることで、より有益な回答を得ることができる。
- このようなシステムを**RAG（Retrieval Augmented Generative)** と呼ぶ。もちろん、外部知識を文字列一致や従来のSQL文だけでキー検索することも可能だが、このようなベクトル検索を組み込むことでより柔軟なシステムが可能になる。前記の**ICL(In Context Learning)**と並んで有効な方法

文ベクトルの状態・絵文字

あなた

「国民の生活に関する条文はなんですか」という文を、20字以内の絵文字に翻訳してください。

ChatGPT Classic

👨‍👩‍👧‍👦📜➡️🏠💼🍲💊

文ベクトルの状態・絵文字検索

質問文: 👨‍👩‍👧‍👦📜➡🏠💼🍲💊
順位:番号・本文（コサイン類似度）
1: 66 第二四条　婚姻は、両性の合意のみに基いて成立し、夫婦が同等の権利を有することを基本として、相互の協力により、維持されなければならない（0.2451230865412084）
2: 67 (2)配偶者の選択、財産権、相続、住居の選定、離婚並びに婚姻及び家族に関するその他の事項に関しては、法律は、個人の尊厳と両性の本質的平等に立脚して、制定されなければならない（0.23684900975383597）
3: 71 (2)すべて国民は、法律の定めるところにより、その保護する子女に普通教育を受けさせる義務を負ふ（0.19564681423938468）
4: 69 (2)国は、すべての生活部面について、社会福祉、社会保障及び公衆衛生の向上及び増進に努めなければならない（0.17349794948835515）
5: 30　十　儀式を行ふこと（0.16154755580873012）
6: 63 第二二条　何人も、公共の福祉に反しない限り、居住、移転及び職業選択の自由を有する（0.15772092911770874）
7: 21　一　憲法改正、法律、政令及び条約を公布すること（0.15120280654956997）
8: 75 (3)児童は、これを酷使してはならない（0.14950980831379973）
9: 159 二　外交関係を処理すること（0.14348970659460783）
10: 86 第三五条　何人も、その住居、書類及び所持品について、侵入、捜索及び押収を受けることのない権利は、第三十三条の場合を除いては、正当な理由に基いて発せられ、且つ捜索する場所及び押収する物を明示する令状がなければ、侵されない（0.1433156619666082）

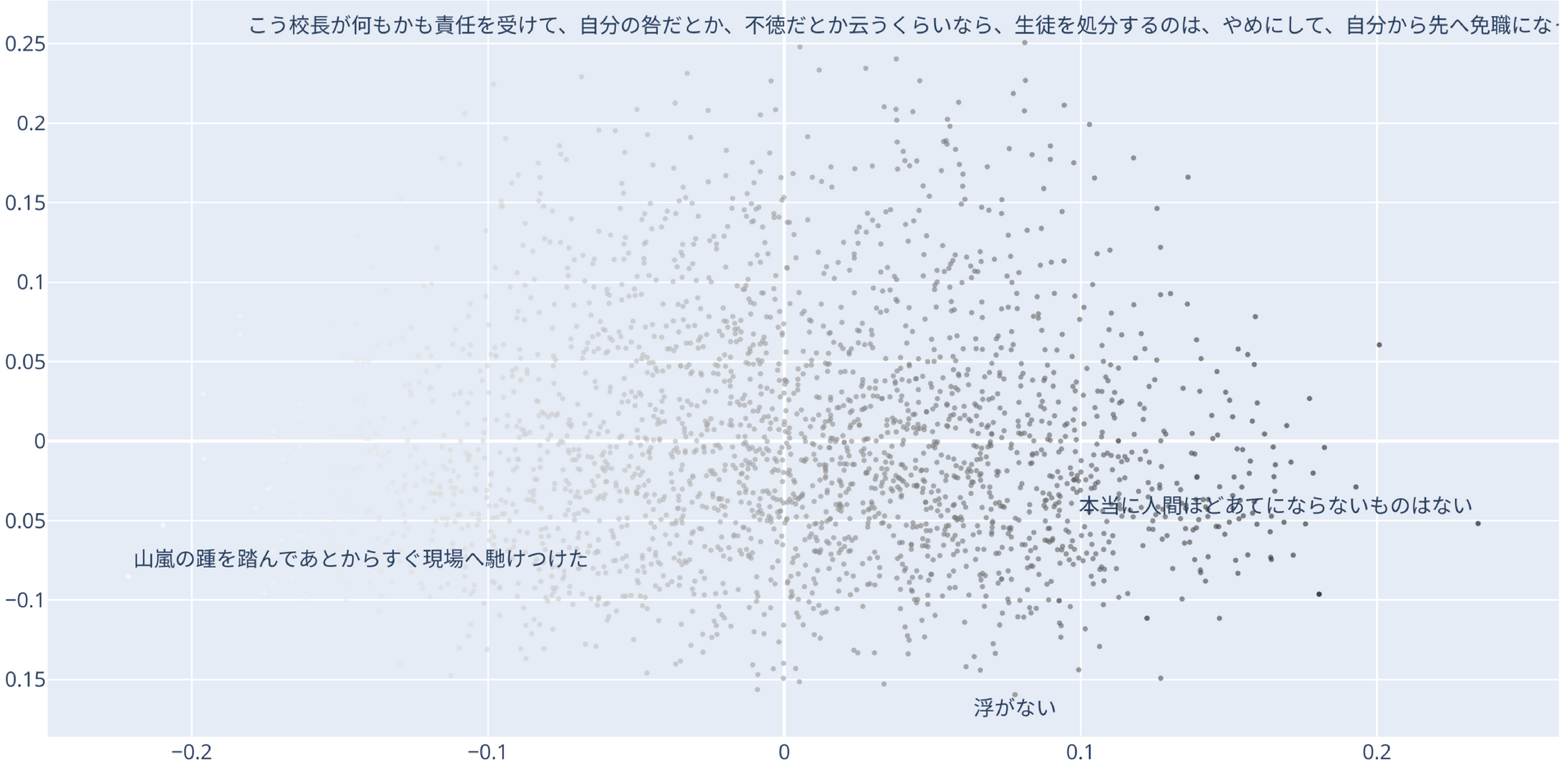
内容

1. 生成AIの仕組み
2. 埋め込みベクトル
3. 単語ベクトル
4. 文ベクトル
5. **文ベクトルの応用**

文ベクトルによる文章の解析

- 単語ベクトルと同じように、文ベクトルもその意味を反映するので、ひとつの文章内の**文ベクトルを主成分分析する**ことで、その文章の中で、どのような文が、中核的な構造を持っているかを調べることができる。
- この、夏目漱石の「坊っちゃん」の例だと、X軸は、「１人称かかそうでないか」で区別されている。

文ベクトルによる文章の解析



文ベクトルによる文章の解析

X-axis Max (Rank 1): 本当に人間ほどあてにならないものはない
X-axis Max (Rank 2): 議論のいい人が善人とはきまらない
X-axis Max (Rank 3): いくら人間が卑怯だって、こんなに卑怯に出来るものじゃない
X-axis Max (Rank 4): あんまり喜んでもいないでしょういえ、お世辞じゃない
X-axis Max (Rank 5): 面白くもない

————————————————————————————————————

X-axis Min (Rank 1): 山嵐の踵を踏んであとからすぐ現場へ馳けつけた
X-axis Min (Rank 2): その晩母が山城屋に詫びに行ったついでに袷の片袖も取り返して来た
X-axis Min (Rank 3): 控所へ帰って来たら、山嵐がどうだいと聞いた
X-axis Min (Rank 4): お婆さん、東京へ行って奥さんを連れてくるんだと答えて勘定を済まして、すぐ汽車へ乗って浜へ来て港屋へ着くと、山嵐は二階で寝ていた
X-axis Min (Rank 5): 山嵐はおい君どこに宿ってるか、山城屋か、うん、今に行って相談すると云い残して白墨を持って教場へ出て行った

文ベクトルによる文章の解析

X-axis Max (Rank 1): メロスは腕に唸りをつけてセリヌンティウスの頬を殴った
X-axis Max (Rank 2): メロスは激怒した
X-axis Max (Rank 3): メロスは、王の前に引き出された
X-axis Max (Rank 4): メロスは、ざんぶと流れに飛び込み、百匹の大蛇のようにのた打ち荒れ狂う浪を相手に、必死の闘争を開始した
X-axis Max (Rank 5): メロスは口惜しく、地団駄踏んだ

———————————————————————————————

X-axis Min (Rank 1): 私は、信じられている
X-axis Min (Rank 2): 君だけは私を信じてくれるにちがい無い
X-axis Min (Rank 3): どうでも、いいのだ
X-axis Min (Rank 4): ああ、待っているだろう
X-axis Min (Rank 5): おまえに言いたいのは、それだけだ

文ベクトルによる文章の解析

- 『走れメロス』（太宰治）の場合も、同様な分析をすることで、X軸は、「**１人称者か、三人称者か**」という区別を反映していることがわかる。
- この区別は、日本の現代小説の基本構造となっている。（文学作品がすべてそうではない。）

文ベクトルによる文章の解析

X-axis Max (Rank 1): 但し、これがため、訴追の権利は、害されない
X-axis Max (Rank 2): 第一九条　思想及び良心の自由は、これを侵してはならない
X-axis Max (Rank 3): (3)何人も、自己に不利益な唯一の証拠が本人の自白である場合には、有罪とされ、又は刑罰を科せられない
X-axis Max (Rank 4): 又、犯罪に因る処罰の場合を除いては、その意に反する苦役に服させられない
X-axis Max (Rank 5): 但し、人種、信条、性別、社会的身分、門地、教育、財産又は収入によつて差別してはならない

ーー

X-axis Min (Rank 1): 第五三条　内閣は、国会の臨時会の召集を決定することができる
X-axis Min (Rank 2): (2)衆議院と参議院とが異なつた指名の議決をした場合に、法律の定めるところにより、両議院の協議会を開いても意見が一致しないとき、又は衆議院が指名の議決をした後、国会休会中の期間を除いて十日以内に、参議院が、指名の議決をしないときは、衆議院の議決を国会の議決とする
X-axis Min (Rank 3): 第六七条　内閣総理大臣は、国会議員の中から国会の議決で、これを指名する
X-axis Min (Rank 4): (2)予算について、参議院で衆議院と異なつた議決をした場合に、法律の定めるところにより、両議院の協議会を開いても意見が一致しないとき、又は参議院が、衆議院の可決した予算を受け取つた後、国会休会中の期間を除いて三十日以内に、議決しないときは、衆議院の議決を国会の議決とする
X-axis Min (Rank 5): 第五四条　衆議院が解散されたときは、解散の日から四十日以内に、衆議院議員の総選挙を行ひ、その選挙の日から三十日以内に、国会を召集しなければならない

文ベクトルによる文章の解析

あなた

日本国憲法の条文を大きく二種類に分類するとすると、どういうことと、どういうことに分けられるでしょうか？

ChatGPT Classic

日本国憲法の条文を大きく二種類に分類すると、一般的には「基本的人権の保障」に関するものと、「国の組織及び運営」に関するものに分けられます。

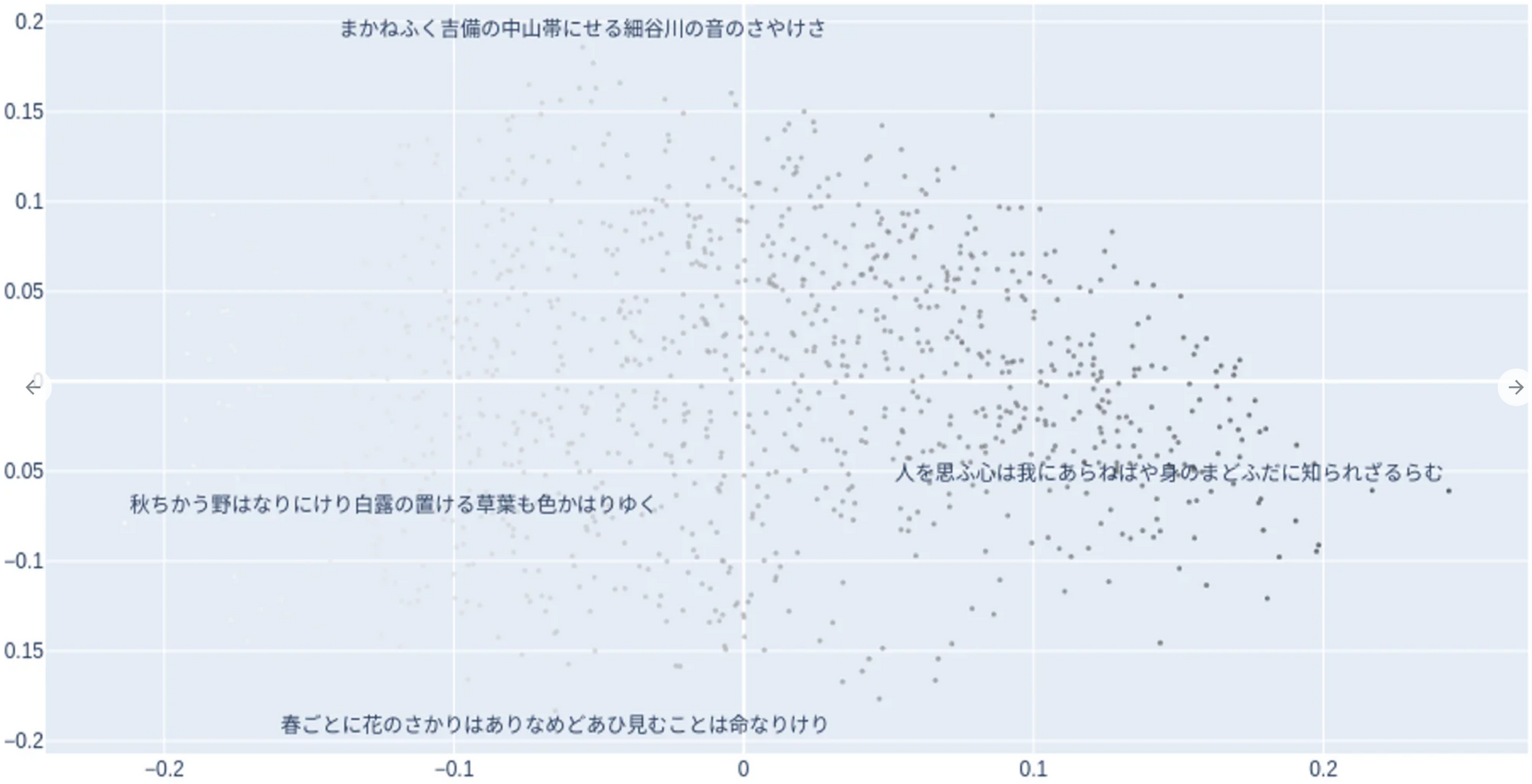
1. **基本的人権の保障**: この部分では、個人の尊厳と自由、平等原則、思想・信教の自由、表現の自由など、国民の基本的人権について規定しています。第11条から第40条までがこれにあたります。
2. **国の組織及び運営**: これには天皇の地位や国会、内閣、裁判所など国の組織に関する規定や、これらの機関の権能と運営の仕組みについて定めたものが含まれます。第41条から第103条まで、そして第104条から第118条までがこれに該当します。

この分類は大まかなものであり、憲法全体を通じてこれらの要素が相互に関連し合っています。

文ベクトルによる文章の解析

- 「日本国憲法」全文を各条ごとにベクトル化して、それを主成分分析で、重要な文の意味の構造を調べてみると、**「個人の権利」と「国の組織」**というのが、もっとも重要な軸（X軸・第１次元）になっているのがわかる。これは、一般的常識とも合致する。
- このように、文ベクトルは、単に文の類似や分類だけに役立つのではなく、**ある文章がどのようなことを主題としているかを見分けるための非常に有力なツールとなることがわかる。コミュニケーション上、何が重要になるか**なども見ることができる。

ベクトルによる古今和歌集の解析



文ベクトルによる古文の解析

第1次元・上位・人間の感情

(Rank 1): 人を思ふ心は我にあらねばや身のまどふだに知られざるらむ

(Rank 2): 思ひけむ人をぞともに思はましまさしやむくいなかりけりやは

(Rank 3): 身を捨ててゆきやしにけむ思ふよりほかなるものは心なりけり

第1次元・下位・自然の風景

(Rank 1): 秋ちかう野はなりにけり白露の置ける草葉も色かはりゆく

(Rank 2): 秋の月山辺さやかに照らせるは落つる紅葉のかずを見よとか

(Rank 3): 秋風の吹きと吹きぬる武蔵野はなべて草葉の色かはりけり

文ベクトルによる古文の解析

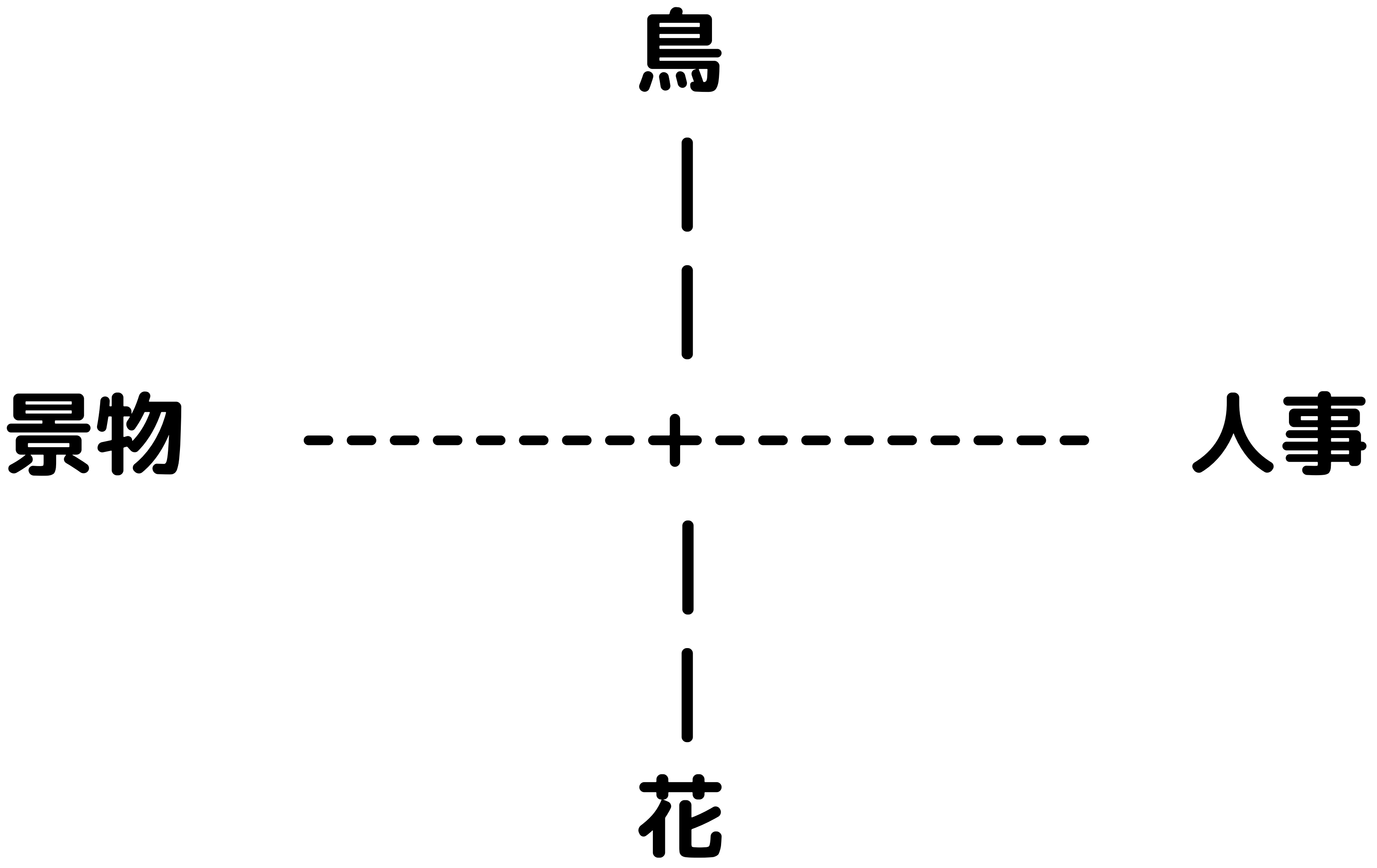
第2次元・上位・聴覚の美＝鳥

(Rank 1): まかねふく吉備の中山帯にせる細谷川の音のさやけさ
(Rank 2): 郭公声もきこえず山彦は外に鳴く音をこたへやはせぬ
(Rank 3): しほの山さしでの磯にすむ千鳥君が御代をば八千代とぞ鳴く

第2次元・下位・視覚の美＝花

(Rank 1): 春ごとに花のさかりはありなめどあひ見むことは命なりけり
(Rank 2): 色見えで移ろふものは世の中の人の心の花にぞありける
(Rank 3): 花見れば心さへにぞ移りける色にはいでじ人もこそ知れ

文ベクトルによる古文の解析



「人事」と「景物」、そして「鳥」と「花」が『古今集』の主要な意味構造の軸となっていることが、LLMの埋め込みベクトルで実証できる。これ自体は古典研究の常識だが、AIがそれを「読めている」ことは驚嘆に値する。

「花鳥風月」というが、「花・月」が視覚、「鳥・風」が聴覚であることもここからわかる。また、これは実は漢詩の影響であることも実証できる。

結論

埋め込みベクトルの有用性

高性能多言語埋め込みベクトルを提供するOpenAIの埋め込みベクトルを利用するのがユーザーとしてはメリットがある。AIの知識を知る第一の手段。

単語ベクトルの利用

各言語の意味の体系を探求することが可能。また、言語による単語の性格の差などを新しい角度から研究できる。

文ベクトルの利用

文ベクトルはRAGなどのビジネス用途にまず第一に便利。言語学的には、多言語に渡る文の意味の解析や、文章の構造分析やコミュニケーション研究に応用可能。

議題のページに戻る

その先へ

現状、AIの能力の全部を引き出せていない

生成AIは、これらの内部表現を用いて文を生成しているが、自覚的に利用できるわけではない。

高次元ベクトルの不思議

言語ごとの意味・文法・語用論など、多くの情報が含まれているが、大半は不明のまま。

AIとの協業による解明

AIによってAIの内部の研究を行うことで、さらに高度な応用が可能になると思われる。

議題のページに戻る