言葉をベクトル化する

持橋大地
統計数理研究所
daichi@ism.ac.jp

都立小石川中等教育学校SSH
2025-9-5 (金)

自己紹介

- 持橋大地
  国立統計数理研究所 教授／
  国立国語研究所 次世代言語科学研究センター教授 (兼務)
  専門：自然言語処理、機械学習、人工知能

統計学の
国立研究所
(立川市)

[大学院から入学できます]

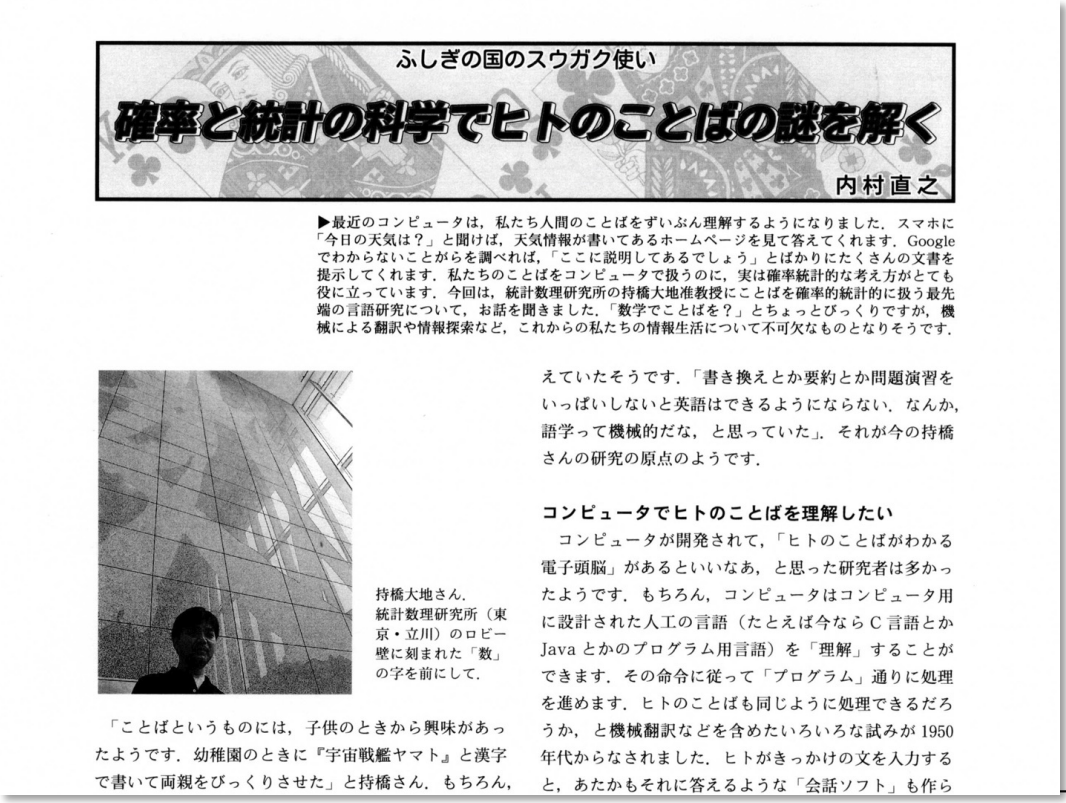
経歴

NAISTの様子

- 小石川高校卒 44期
  - 音楽研究会でした (指揮者)
- これまで：
  - 東大文科三類入学
  - 教養学部 基礎科学科第二に進学 (理転、文科から2名)
  - 東大の院を辞退して、奈良先端科学技術大学院大学 (NAIST) に進学
  - ATR 音声言語研究所
  - NTT コミュニケーション科学基礎研究所
  - 国立・統計数理研究所 (2011年〜)

小石川高校時代 (1)

- 予備校は行かず、Z会と月刊『大学への数学』などで勉強
- 『大学への数学』2014年9月号
  「ふしぎの国のスウガク使い」で紹介していただきました



ふしぎの国のスウガク使い

確率と統計の科学でヒトのことばの謎を解く

内村直之

▶最近のコンピュータは，私たち人間のことばをずいぶん理解するようになりました．スマホに「今日の天気は？」と聞けば，天気情報が書いてあるホームページを見て答えてくれます．Google でわからないことがらを調べれば，「ここに説明してあるでしょう」とばかりにたくさんの文書を提示してくれます．私たちのことばをコンピュータで扱うのに，実は確率統計的な考え方がとても役に立っています．今回は，統計数理研究所の持橋大地准教授にことばを確率的統計的に扱う最先端の言語研究について，お話を聞きました．「数学でことばを？」とちょっとびっくりですが，機械による翻訳や情報探索など，これからの私たちの情報生活について不可欠なものとなりそうです．

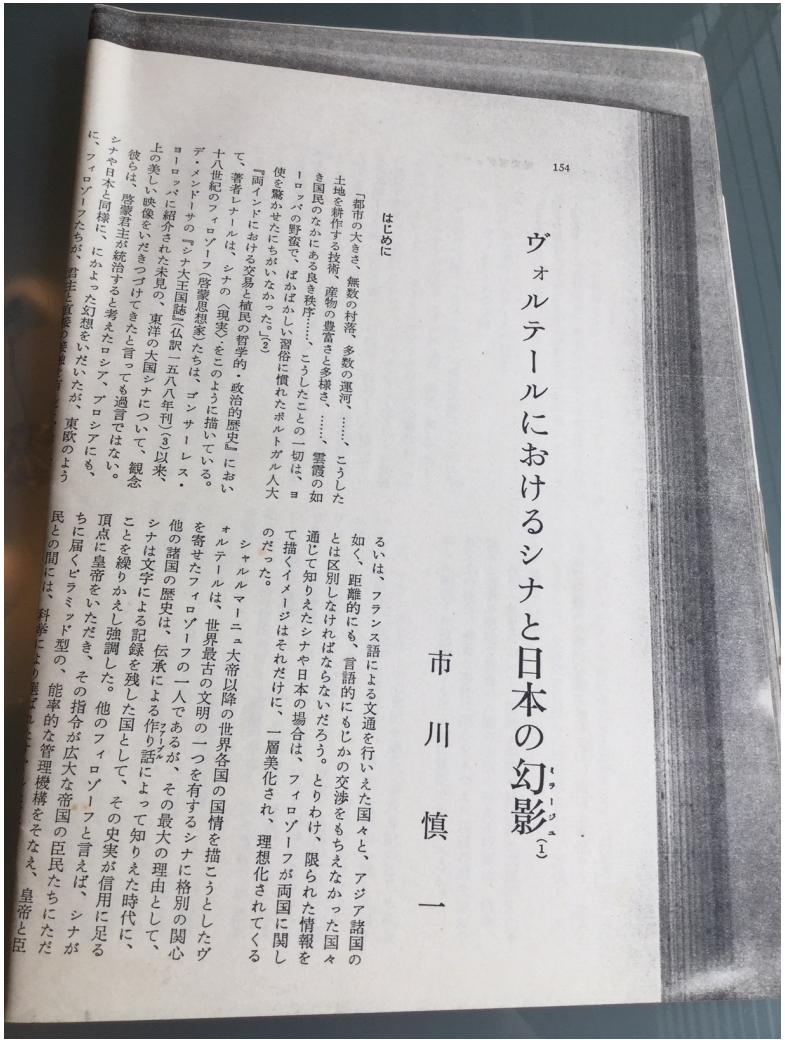
「ことばというものには，子供のときから興味があったようです．幼稚園のときに『宇宙戦艦ヤマト』と漢字で書いて両親をびっくりさせた」と持橋さん．もちろん，

えていたそうです．「書き換えとか要約とか問題演習をいっぱいしないと英語はできるようにならない．なんか，語学って機械的だな，と思っていた」．それが今の持橋さんの研究の原点のようです．

コンピュータでヒトのことばを理解したい

コンピュータが開発されて，「ヒトのことばがわかる電子頭脳」があるといいなあ，と思った研究者は多かったようです．もちろん，コンピュータはコンピュータ用に設計された人工の言語（たとえば今ならＣ言語とか Java とかのプログラム用言語）を「理解」することができます．その命令に従って「プログラム」通りに処理を進めます．ヒトのことばも同じように処理できるだろうか，と機械翻訳などを含めたいろいろな試みが 1950 年代からなされました．ヒトがきっかけの文を入力すると，あたかもそれに答えるような「会話ソフト」も作ら

小石川高校時代 (2)



154

ヴォルテールにおけるシナと日本の幻影(ファンタスム)(1)

市川慎一

はじめに

「都市の大きさ、無数の村落、多数の運河、……、こうした土地を耕作する技術、産物の豊富さと多様さ、……、雲霞の如き国民のなかにある良き秩序……、こうしたことの一切は、ヨーロッパの野蛮で、ばかばかしい習俗に慣れたポルトガル人大使を驚かせたにちがいなかった。(2)

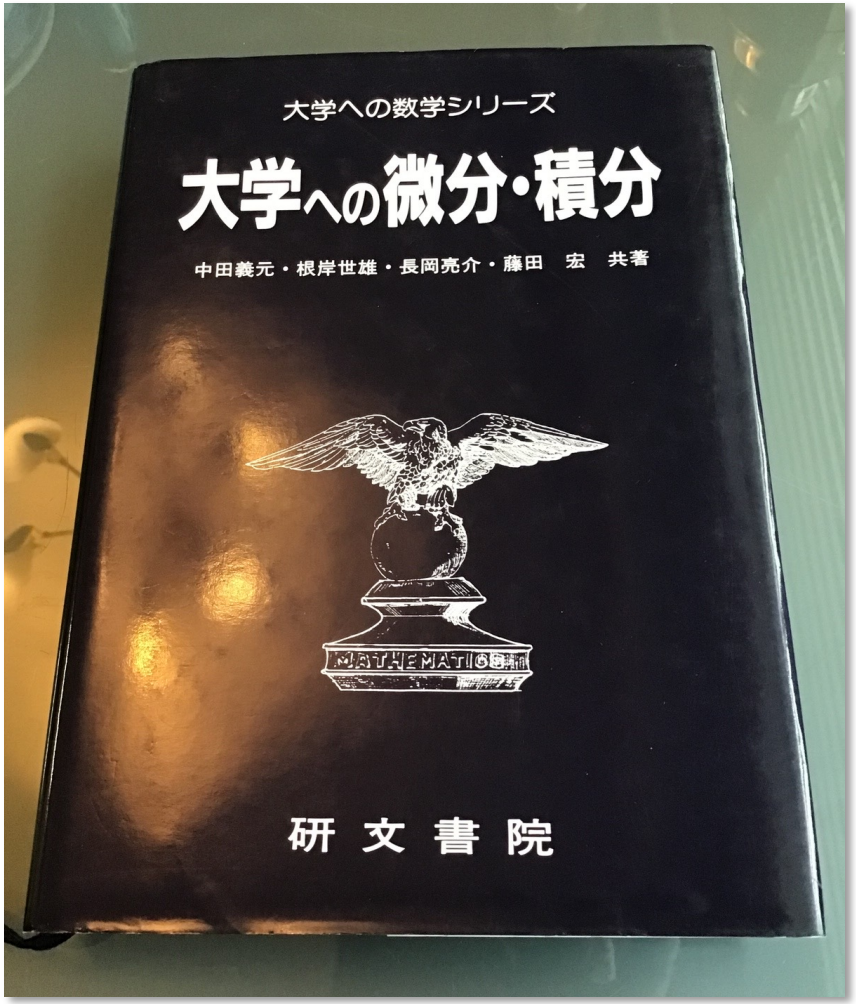
「両インドにおける交易と植民の哲学的・政治的歴史」において、著者レナールは、シナの〈現実〉をこのように描いている。十八世紀のフィロゾーフ(啓蒙思想家)たちは、ゴンサーレス・デ・メンドーサの『シナ大王国誌』(公訳一五八八年刊)(3)以来、ヨーロッパに紹介された未見の、東洋の大国シナについて、観念上の美しい映像をいだきつづけてきたと言っても過言ではない。彼らは、啓蒙君主が統治するシナや日本と同様に、にかよった幻想をいだいたが、東欧のようにフィロゾーフたちが、君主と直接の関係を持つ

るいは、フランス語による文通を行いえた国々と、アジア諸国の如く、距離的にも、言語的にもじかの交渉をもちえなかった国々とは区別しなければならないだろう。とりわけ、限られた情報を通じて知りえたシナや日本の場合は、フィロゾーフが両国に関して描くイメージはそれだけに、一層美化され、理想化されてくるのだった。

シャルルマーニュ大帝以降の世界各国の国情を描こうとしたヴォルテールは、世界最古の文明の一つを有するシナに格別の関心を寄せたフィロゾーフの一人であるが、その最大の理由として、他の諸国の歴史は、伝承による作り話(ファーブル)によって知りえた時代に、シナは文字による記録を残した国として、その史実が信用に足ることを繰りかえし強調した。他のフィロゾーフと言えば、シナが頂点に皇帝をいただき、その指令が広大な帝国の臣民たちにただちに届く、ピラミッド型の、能率的な管理機構をそなえ、皇帝と臣民との間には、科挙により選ばれた

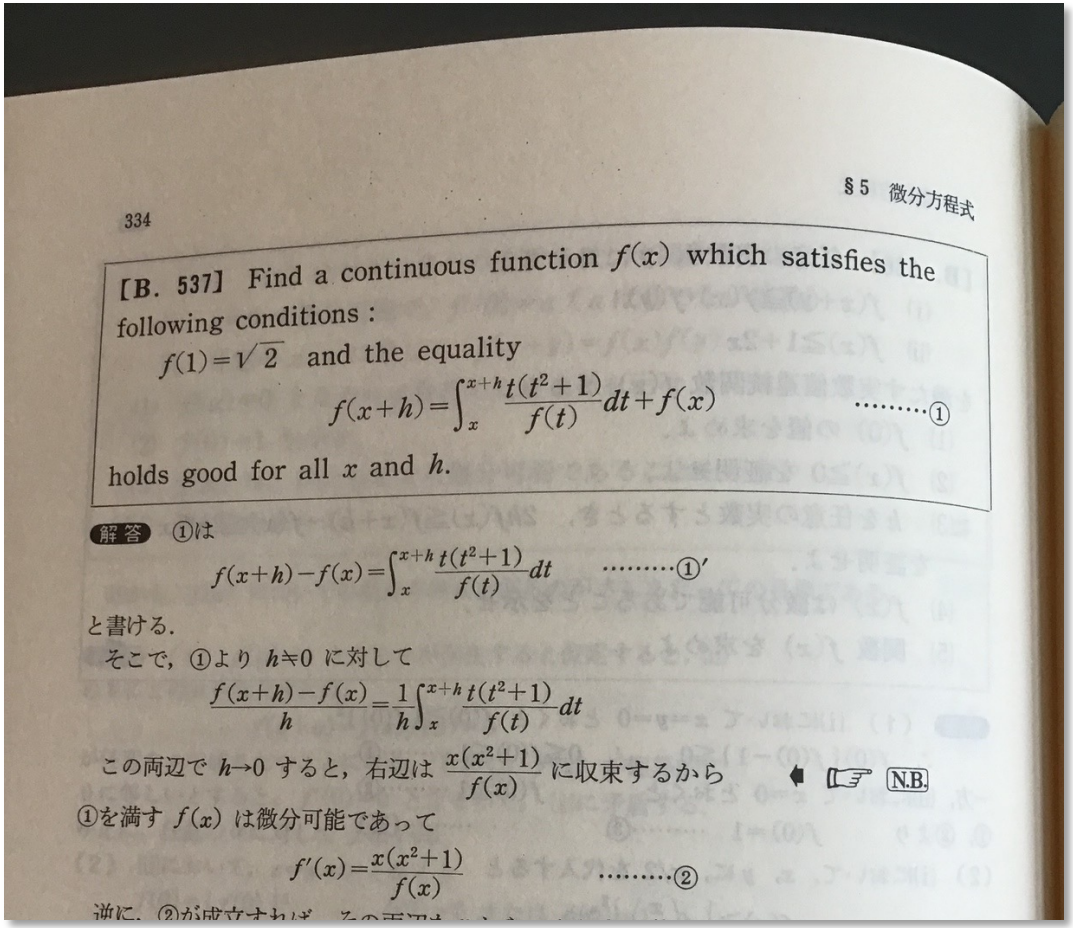
- 高3で世界史を勉強している際、フランス啓蒙思想のヴォルテールが中国に興味を持った、という話を教科書で読み、世界史科を訪ねると、この「ヴォルテールにおけるシナと日本の幻影」をいただいた
- 高3の5月くらいに、小石川の藤棚の下で読みました

小石川高校時代 (3)



- 高校数学史上最高の参考書といわれる、通称『黒大数』
- 著者の中心となっている藤田先生は、東大数学科の名誉教授
- 現在は絶版 (図書館・中古で入手できます)

小石川高校時代 (3)



334 §5 微分方程式

[B. 537] Find a continuous function $f(x)$ which satisfies the following conditions :

$f(1)=\sqrt{2}$ and the equality

$$f(x+h)=\int_x^{x+h}\frac{t(t^2+1)}{f(t)}dt+f(x) \qquad \cdots\cdots\cdots①$$

holds good for all x and h.

解答 ①は

$$f(x+h)-f(x)=\int_x^{x+h}\frac{t(t^2+1)}{f(t)}dt \qquad \cdots\cdots\cdots①'$$

と書ける.

そこで, ①より $h\neq 0$ に対して

$$\frac{f(x+h)-f(x)}{h}=\frac{1}{h}\int_x^{x+h}\frac{t(t^2+1)}{f(t)}dt$$

この両辺で $h\to 0$ すると, 右辺は $\dfrac{x(x^2+1)}{f(x)}$ に収束するから ☛ ☞ N.B.

①を満す $f(x)$ は微分可能であって

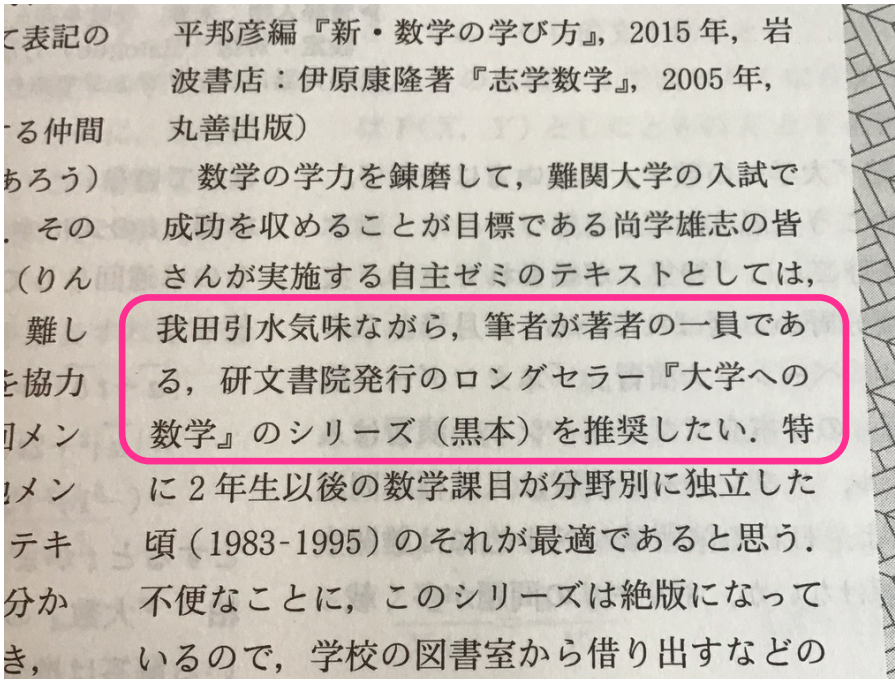
$$f'(x)=\frac{x(x^2+1)}{f(x)} \qquad \cdots\cdots\cdots②$$

逆に, ②が成立すれば その両辺は

- 当たり前のように、英語で書かれた問題が入っている!
- 解説は普通に日本語
- このことの意味は..

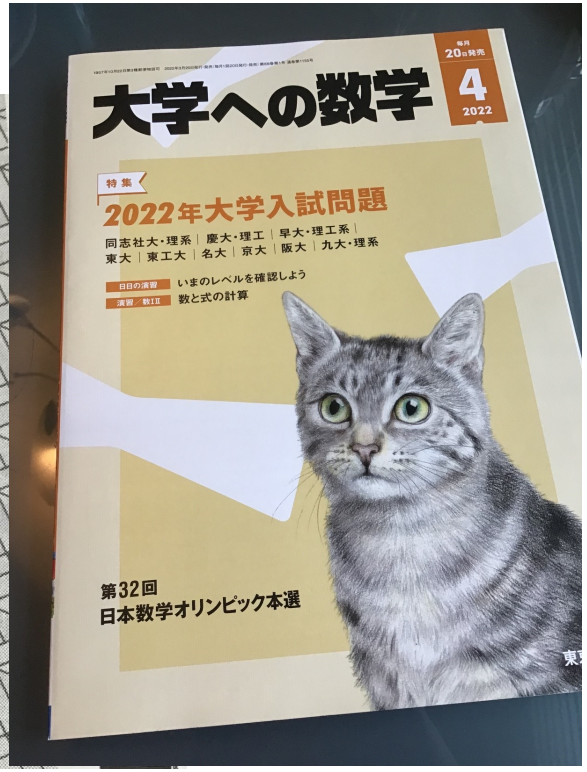
小石川高校時代 (3)

- 『大学への数学』2022年4月号・巻頭言でも、この本が推薦されているようです



平邦彦編『新・数学の学び方』，2015 年，岩波書店：伊原康隆著『志学数学』，2005 年，丸善出版）

数学の学力を錬磨して，難関大学の入試で成功を収めることが目標である尚学雄志の皆さんが実施する自主ゼミのテキストとしては，我田引水気味ながら，筆者が著者の一員である，研文書院発行のロングセラー『大学への数学』のシリーズ（黒本）を推奨したい．特に 2 年生以後の数学課目が分野別に独立した頃（1983-1995）のそれが最適であると思う．不便なことに，このシリーズは絶版になっているので，学校の図書室から借り出すなどの



今日の話

言葉をベクトル化する

- “ベクトル”とは？ (高2で勉強します)
- 言葉を数値で表して、何が嬉しいの？
- 言葉のベクトル化は、現在の人工知能の基礎です

言葉のベクトル化

- 言葉を “数値の組” で表す

東京　　→ [1.2, -0.7, 3.3, 0.4, -1.6]
エジプト → [0.8, -2.5, 2.1, -4.4, 3.3]
お茶漬け → [-2.2, 0.1, 0.2, 2.7, -0.8]
ささやく → [-0.6, -1.5, 1.1, 2.6, 0.9]
　　:

- ChatGPTなどの人工知能(AI)では、言葉はすべて
こうしてベクトル化されています

実際の単語ベクトル (一部)

正義　[0.002604 0.00062024 -0.025644 -0.082722 0.077728 0.0087165
0.049547 -0.025976 -0.095092 -0.067749 0.013562 0.060091 0.053064
.. 0.0228 -0.041253 0.1974 -0.10254 -0.029888 0.033918 0.21347]
ディレクター　[0.0015758 -0.15125 0.21817 -0.12623 0.13215 0.10187
0.080754 -0.024508 -0.048359 -0.13401 -0.14256 0.083109 0.30153
.. -0.077481 -0.15859 0.1278 0.028061 -0.15356 -0.11111 0.29463]
幕下　[-0.18127 -0.16785 0.0039202 0.17829 -0.21384 0.094747
0.29718 0.5263 -0.30177 0.11019 0.16902 -0.30849 -0.19053 -0.11944
.. 0.093231 0.87569 0.025137 0.0011367 -0.025216 -0.12178 0.1811]
静岡　[0.039938 0.10639 -0.021873 -0.014316 0.015339 0.044891
0.0024299 -0.0086346 -0.054358 -0.022956 -0.058226 -0.011501
.. -0.005193 -0.081363 -0.033958 -0.058588 -0.080491 0.09922]
高まる　[0.036186 0.033845 -0.018044 -0.034166 -0.021112 0.0028494
0.12283 -0.090323 -0.029995 0.037449 -0.030411 0.014132 0.040336
.. 0.00908 0.0084737 -0.007582 -0.14347 0.029086 0.14277 0.11097]

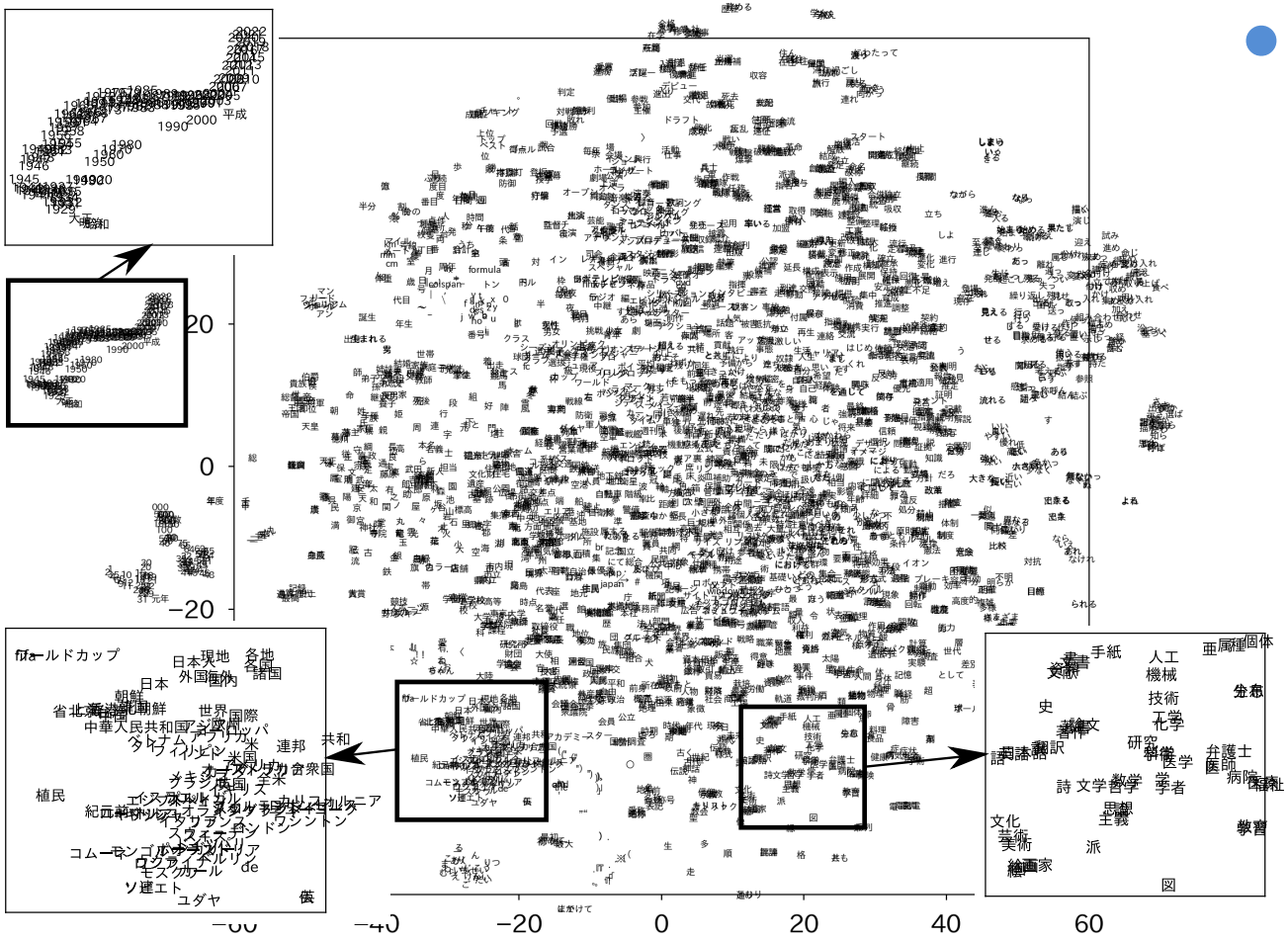
言葉のベクトル化 (2)

- 次元が高すぎるので、可視化のために2次元に圧縮してみる (*t*-SNEという方法)

ディレクター → [30.8063 -21.2218]
正義 → [9.0404 -8.6712]
幕下 → [-4.8868 10.0548]
高まる → [-22.8767 -12.3432]
静岡 → [35.9723 31.3486]
:

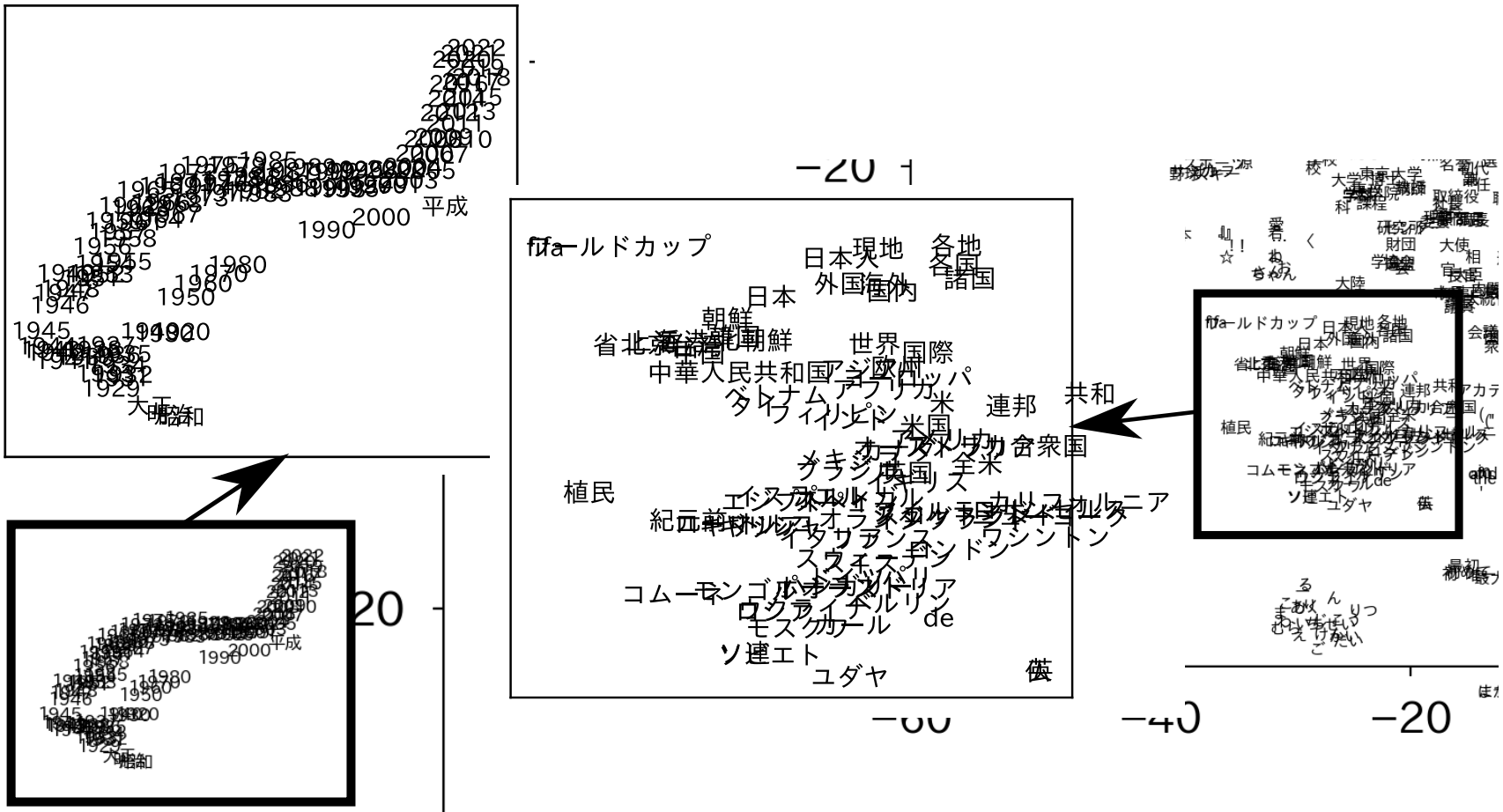
- xy 座標にプロットできる！

日本語単語ベクトルの分布



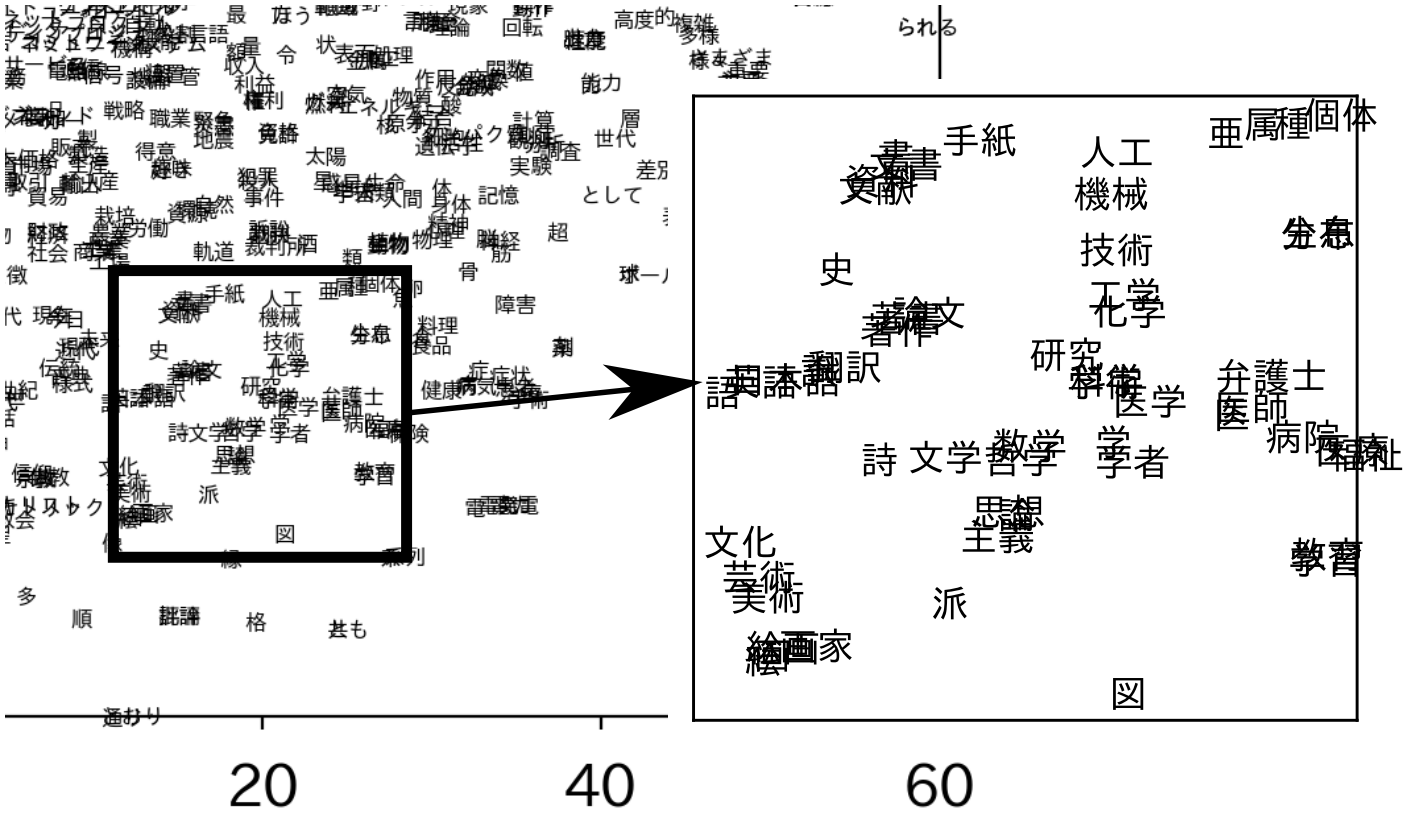
- 日本語のWiki-pediaから計算した400次元のベクトルを、*t*-SNEで2次元に可視化したもの

日本語単語ベクトルの分布 (2)



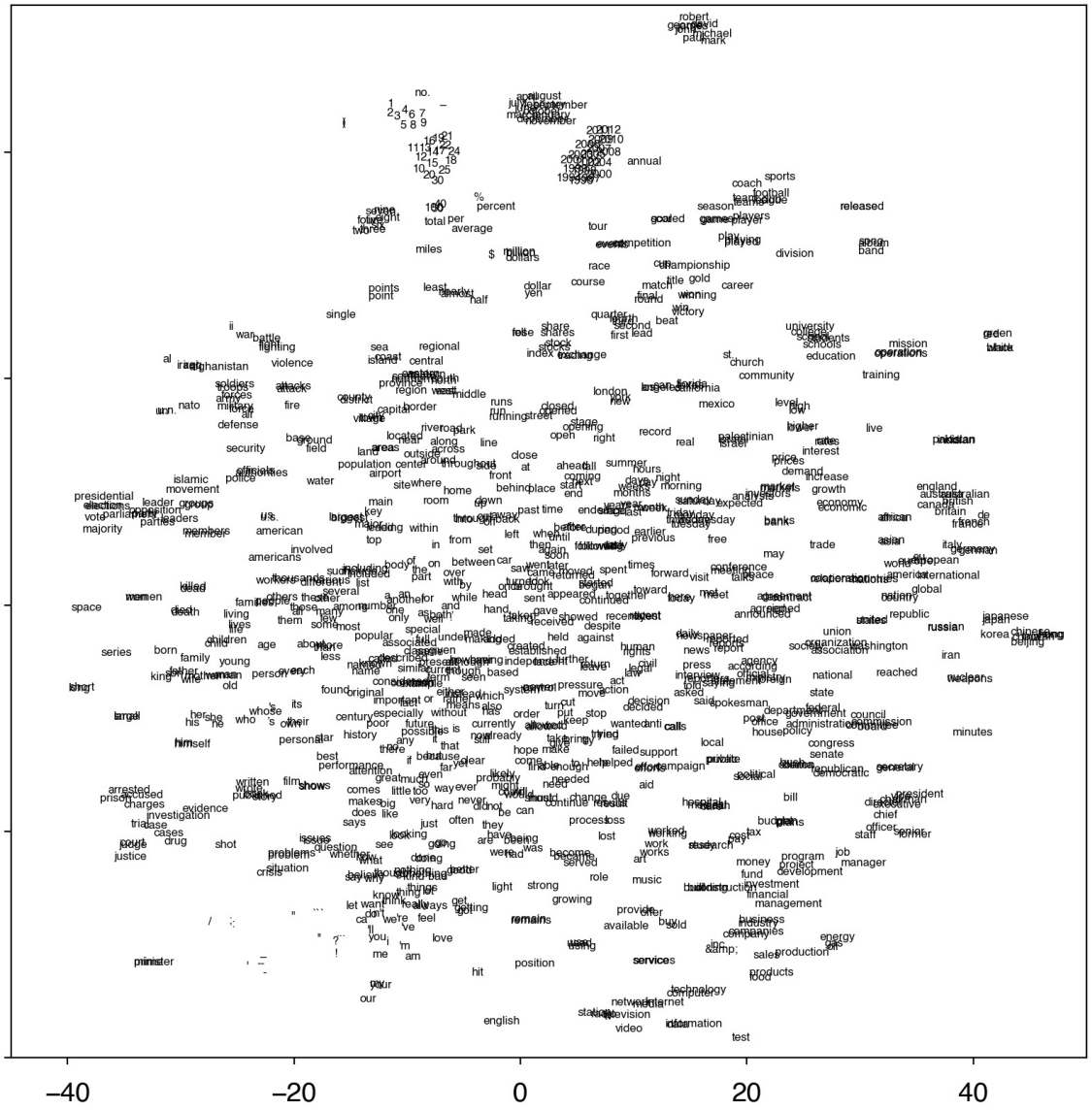
- 西暦や年号、国名などがまとまって分布している

日本語単語ベクトルの分布 (3)



- 意味的に関連する単語が、自動的に近くに配置されている

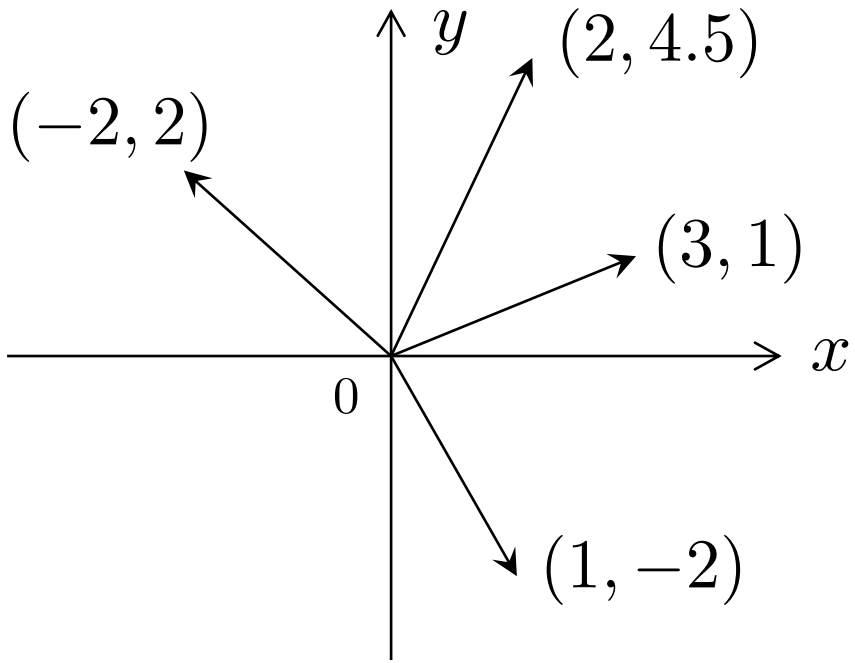
単語ベクトルの分布 (英語)



- 300次元の単語ベクトルの上位1000語を、*t*-SNEで2次元に描画
- 同様に意味的なかたまりがみられる

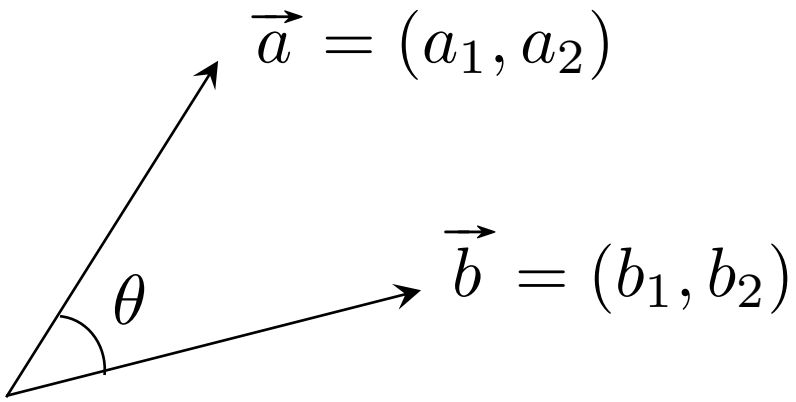
「ベクトル」とは？

- 簡単にいうと、“数値の組”のこと (数学Bですぐ習います)



  - Pythonならば、リスト [1,-2] など
  - 原点からのびた矢印だと思ってもよい

ベクトルの性質

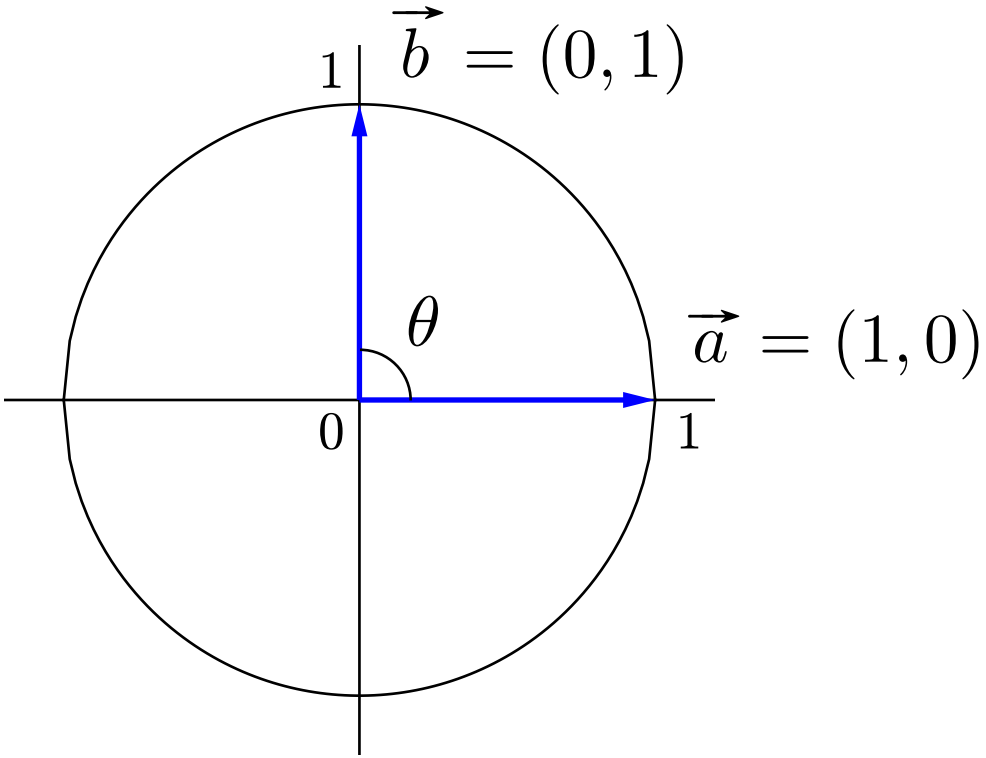


- 2つのベクトル $\vec{a}, \vec{b}$ の長さがどちらも1のとき、$\vec{a}$ と $\vec{b}$ のなす角 θ の余弦は

$$\cos\theta = \vec{a} \cdot \vec{b} = a_1 b_1 + a_2 b_2$$

- この $\vec{a} \cdot \vec{b}$ を $\vec{a}$ と $\vec{b}$ の内積という
  - 高次元の場合は、$\vec{a} \cdot \vec{b} = a_1 b_1 + a_2 b_2 + \cdots + a_N b_N$

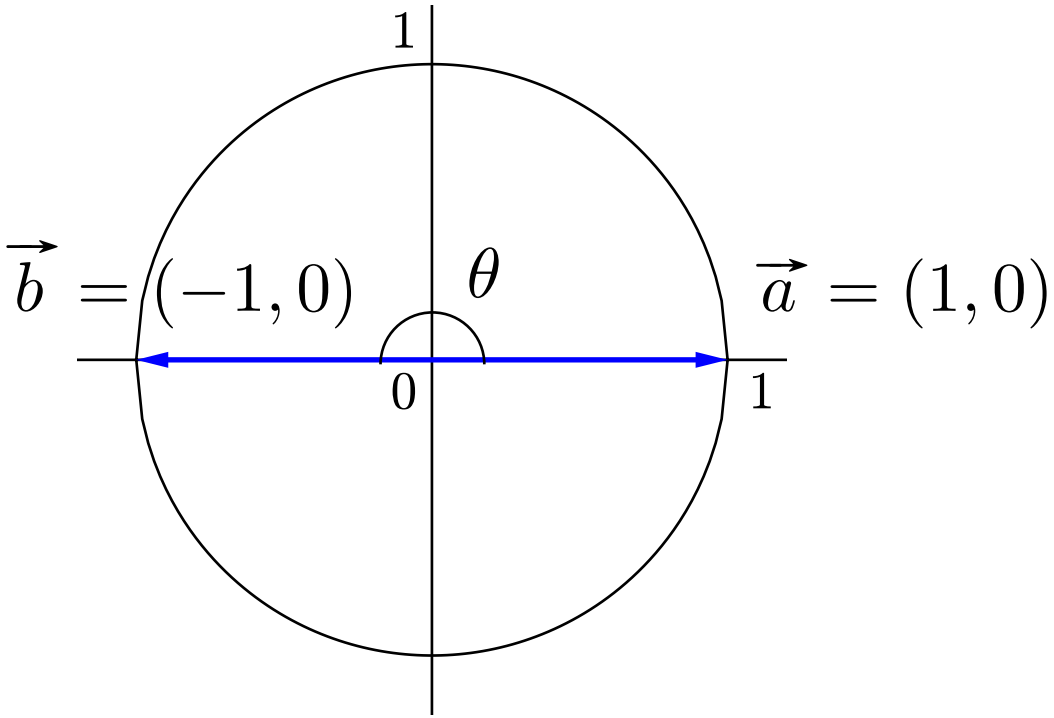
ベクトルのなす角の計算 (1)



- $\cos\theta = \vec{a} \cdot \vec{b} = (1, 0) \cdot (0, 1) = 0$

  $\therefore\ \theta = 90^\circ \left(\frac{\pi}{2}\right)$

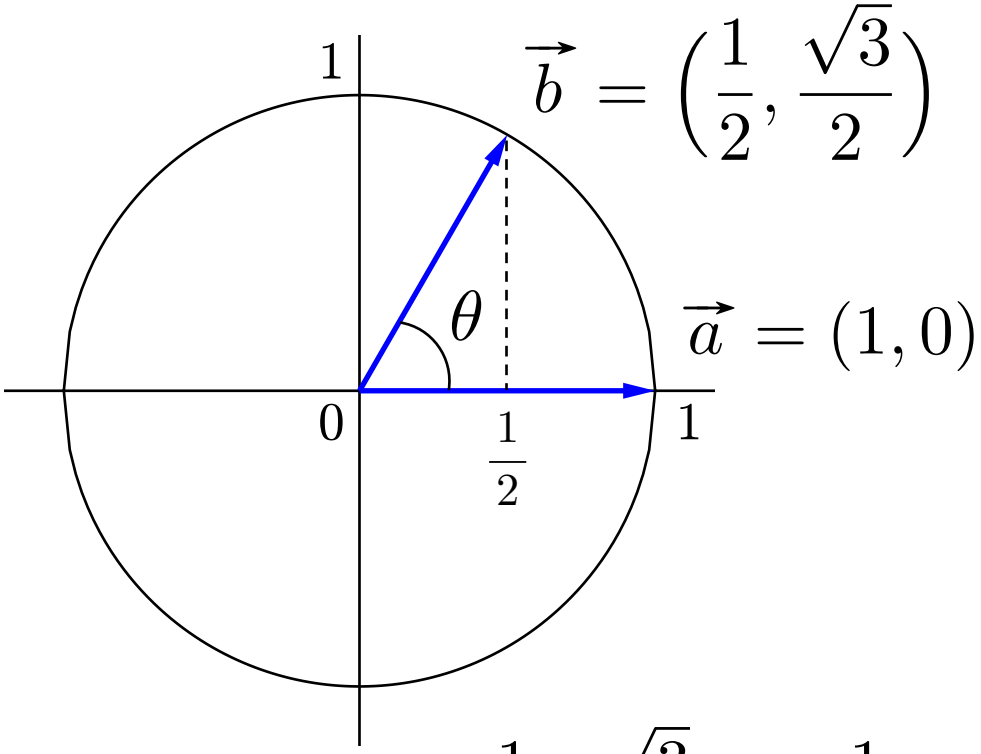
ベクトルのなす角の計算 (2)



- $\cos\theta = \vec{a} \cdot \vec{b} = (1, 0) \cdot (-1, 0) = -1$

  $\therefore\ \theta = 180^\circ\ (\pi)$

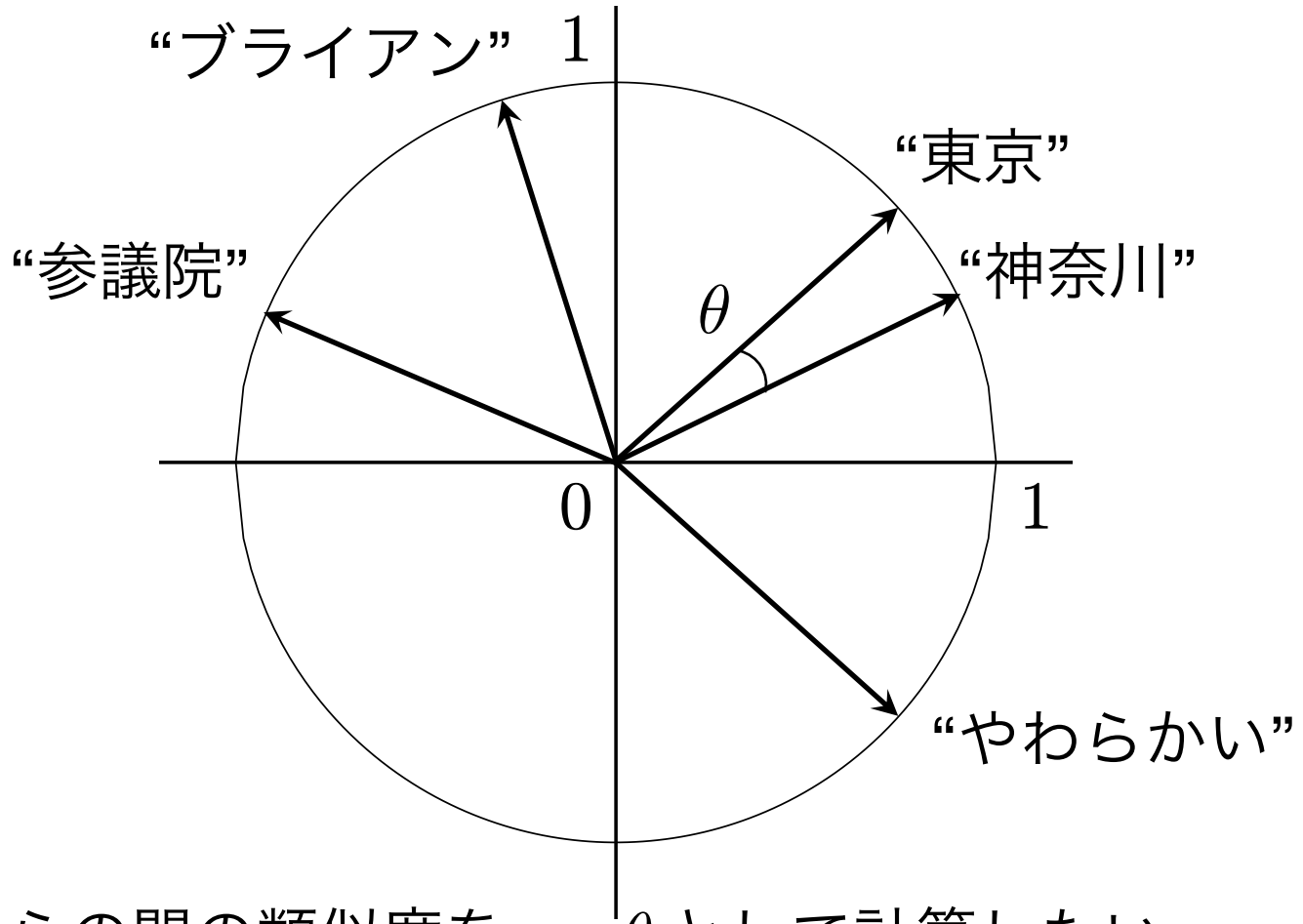
ベクトルのなす角の計算 (3)



- $\cos\theta = \vec{a} \cdot \vec{b} = (1, 0) \cdot \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right) = \frac{1}{2}$

  $\therefore \quad \theta = 60^\circ \ \left(\frac{\pi}{3}\right)$

言葉のベクトルの間の角度



- これらの間の類似度を $\cos\theta$ として計算したい

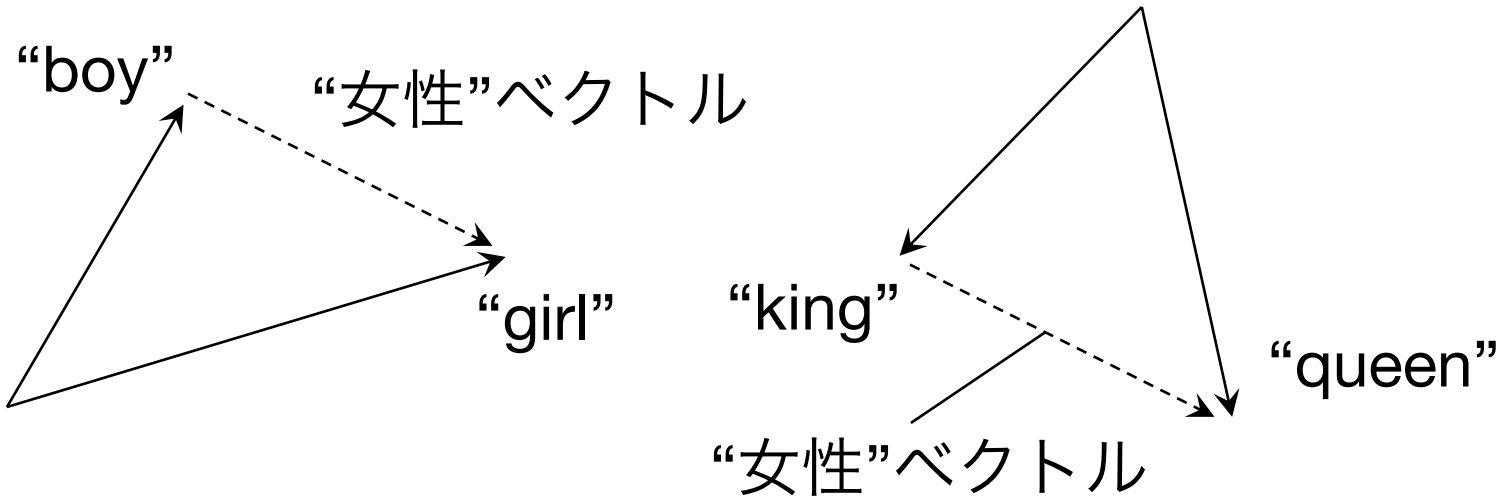
言葉の類似度の計算

- 単語ベクトルが似ていれば、意味が似ている
  → ベクトルの内積を計算すれば、意味の似ている単語がわかる

```
% similar.py model 静岡
   静岡 ->  1.0000
   滋賀 ->  0.8056
   島根 ->  0.7873
   岐阜 ->  0.7834
 神奈川 ->  0.7770
   大分 ->  0.7758
   山梨 ->  0.7742
   岡山 ->  0.7680
   鳥取 ->  0.7622 ...
```

```
% similar.py model 正義
   正義 ->  1.0000
   幸福 ->  0.7800
   唱え ->  0.7677
   なき ->  0.7549
   理想 ->  0.7538
われわれ ->  0.7431
 信じる ->  0.7320
   連帯 ->  0.7303
      観 ->  0.7095 ...
```

言葉のベクトルの「計算」



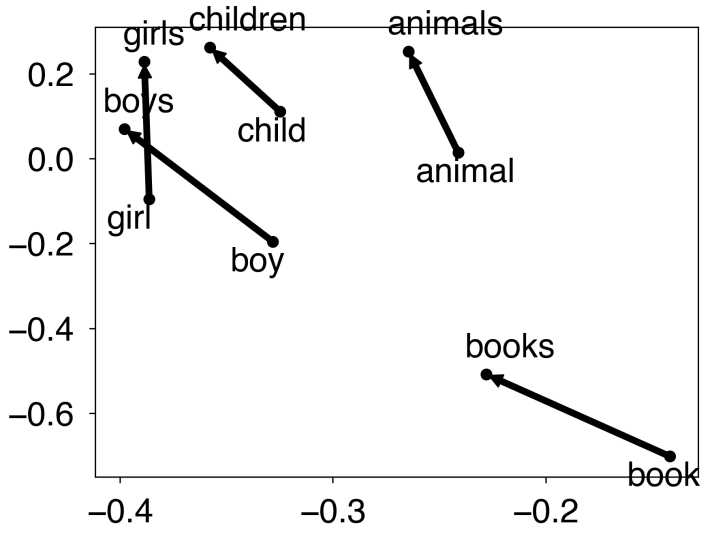
- ベクトルの「引き算」をして別の言葉のベクトルに足すことで、言葉の意味の「計算」ができる！
- 言葉のベクトルの間で、

$$\overrightarrow{\text{boy}} : \overrightarrow{\text{girl}} = \overrightarrow{\text{king}} : \overrightarrow{\text{queen}}$$

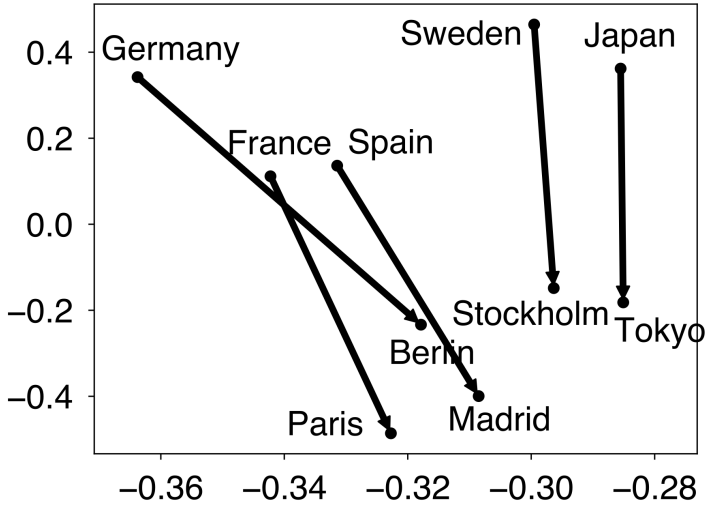
が成り立つ

言葉のベクトルの「計算」

- 実際のテキストから言葉のベクトルを作って
  計算してみたもの (持橋(2025)から抜粋)



“複数形”の関係



“首都”の関係

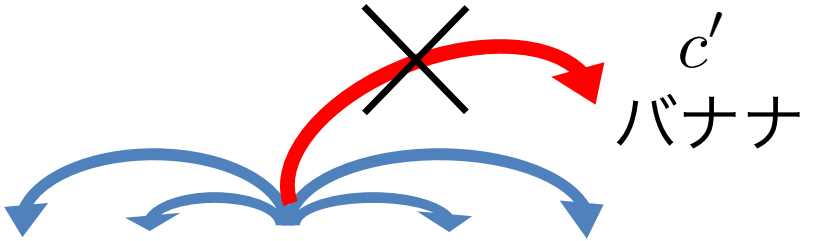
単語の「比例関係」の計算

- 単語ベクトルの間には、$\vec{a} : \vec{b} = \vec{c} : \vec{d}$ の比例関係が成り立つことが知られている

```
contrast (vectors9, "日本", "東京", "フランス")
⇒ パリ -> 0.5812
  リヨン -> 0.5394
  マルセイユ -> 0.5321
  トゥールーズ -> 0.5123
  ニース -> 0.5098
  ストラスブール -> 0.4985
  ナント -> 0.4955
  ディジョン -> 0.4722
  ブリュッセル -> 0.4655
  ボルドー -> 0.4468
  マドリード -> 0.4367
```

```
contrast (vectors9, "日本", "聖子", "アメリカ")
⇒ リンダ -> 0.4708
  ジャネット -> 0.4679
  マライア -> 0.4667
  マリリン -> 0.4603
  ジョニー -> 0.4508
  ビリー -> 0.4478
  テイラー -> 0.4402
  ロジャー -> 0.4344
  パウエル -> 0.4299
  タウンゼント -> 0.4274
  ディーン -> 0.4265
```

どうやって言葉のベクトルを学習？



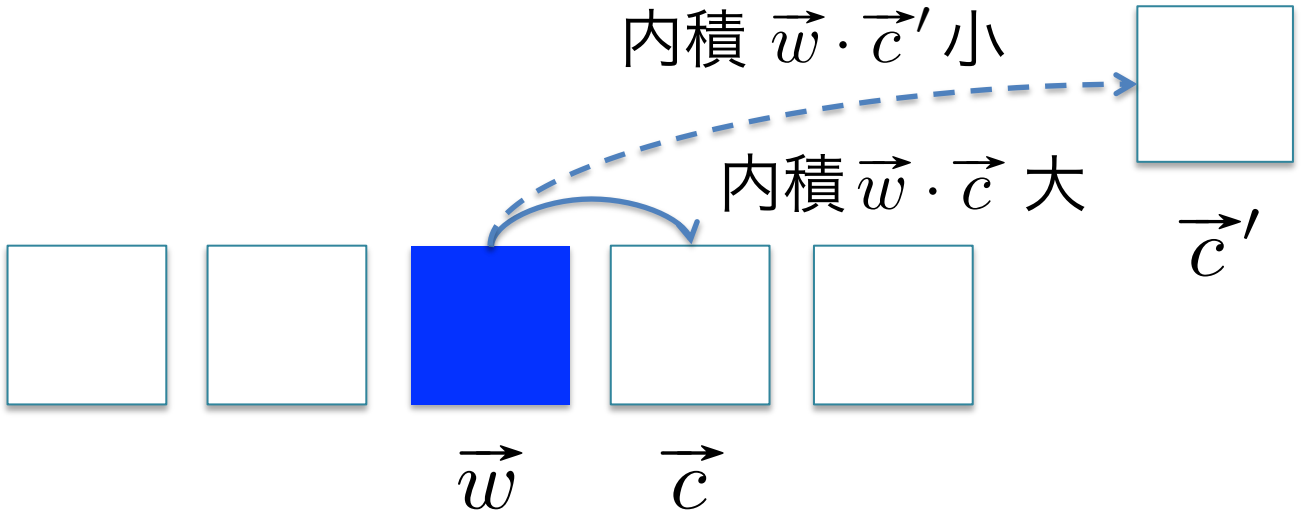
… 一方 で 東京都 の 小石川 中等 教育 学校 の …

c w c

- それぞれの単語 w の周囲の語 c は、意味的に関係
- ベクトル $\vec{w}$ から $\vec{c}$ を予測する確率が高くなるように、$\vec{w}$ を最適化する
  - 周囲に出てこない単語のベクトル $\vec{c}'$ の予測確率は、低くなるようにする

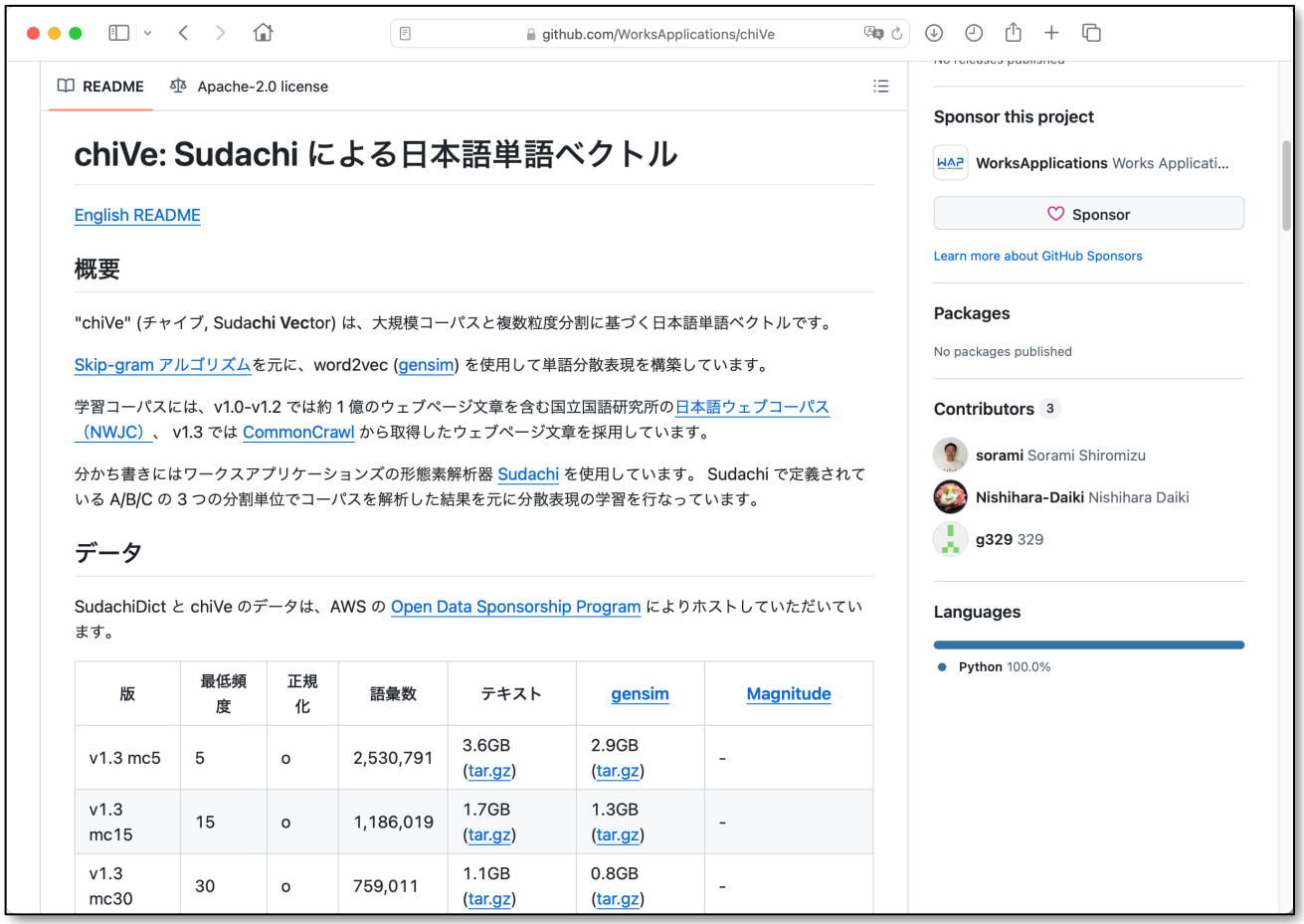
Word2Vecによる単語ベクトル

- Word2vec (Mikolov+ 2013)は、単語ベクトル $\vec{w}$ が
  (a) 前後の語 c と内積が大きく
  (b) ランダムな語 c' とは小さい
  ように $\vec{w}$ を学習する



学習済み単語ベクトルの例

- ChiVe：フリーで入手できる単語ベクトル



github.com/WorksApplications/chiVe

README　Apache-2.0 license

chiVe: Sudachi による日本語単語ベクトル

English README

概要

"chiVe" (チャイブ, Sudachi Vector) は、大規模コーパスと複数粒度分割に基づく日本語単語ベクトルです。

Skip-gram アルゴリズムを元に、word2vec (gensim) を使用して単語分散表現を構築しています。

学習コーパスには、v1.0-v1.2 では約 1 億のウェブページ文章を含む国立国語研究所の日本語ウェブコーパス（NWJC）、 v1.3 では CommonCrawl から取得したウェブページ文章を採用しています。

分かち書きにはワークスアプリケーションズの形態素解析器 Sudachi を使用しています。 Sudachi で定義されている A/B/C の 3 つの分割単位でコーパスを解析した結果を元に分散表現の学習を行なっています。

データ

SudachiDict と chiVe のデータは、AWS の Open Data Sponsorship Program によりホストしていただいています。

版	最低頻度	正規化	語彙数	テキスト	gensim	Magnitude
v1.3 mc5	5	o	2,530,791	3.6GB (tar.gz)	2.9GB (tar.gz)	-
v1.3 mc15	15	o	1,186,019	1.7GB (tar.gz)	1.3GB (tar.gz)	-
v1.3 mc30	30	o	759,011	1.1GB (tar.gz)	0.8GB (tar.gz)	-

Sponsor this project

WorksApplications Works Applicati...

Sponsor

Learn more about GitHub Sponsors

Packages

No packages published

Contributors 3

sorami Sorami Shiromizu

Nishihara-Daiki Nishihara Daiki

g329 329

Languages

Python 100.0%

学習済み単語ベクトルの例 (2)

- 朝日新聞社による単語ベクトル (要登録)

“文ベクトル”の計算

- 単語ベクトルと同じように、文についても“文ベクトル”を計算することができる
  - 計算には色々な方法があるが、たとえばOpenAI の提供しているPythonパッケージを使えば、

```
import openai
client = openai.OpenAI ()

def sent_embedding (text, model="text-embedding-3-small"):
    return client.embeddings.create (input=[text], model=model)
                .data[0].embedding
```

のようにして計算できる

文ベクトルの類似度

- 与えられた文のベクトルと、他の文のベクトルのなす角 θ の余弦 $\cos\theta$ を調べれば、意味の似た文が検索できる

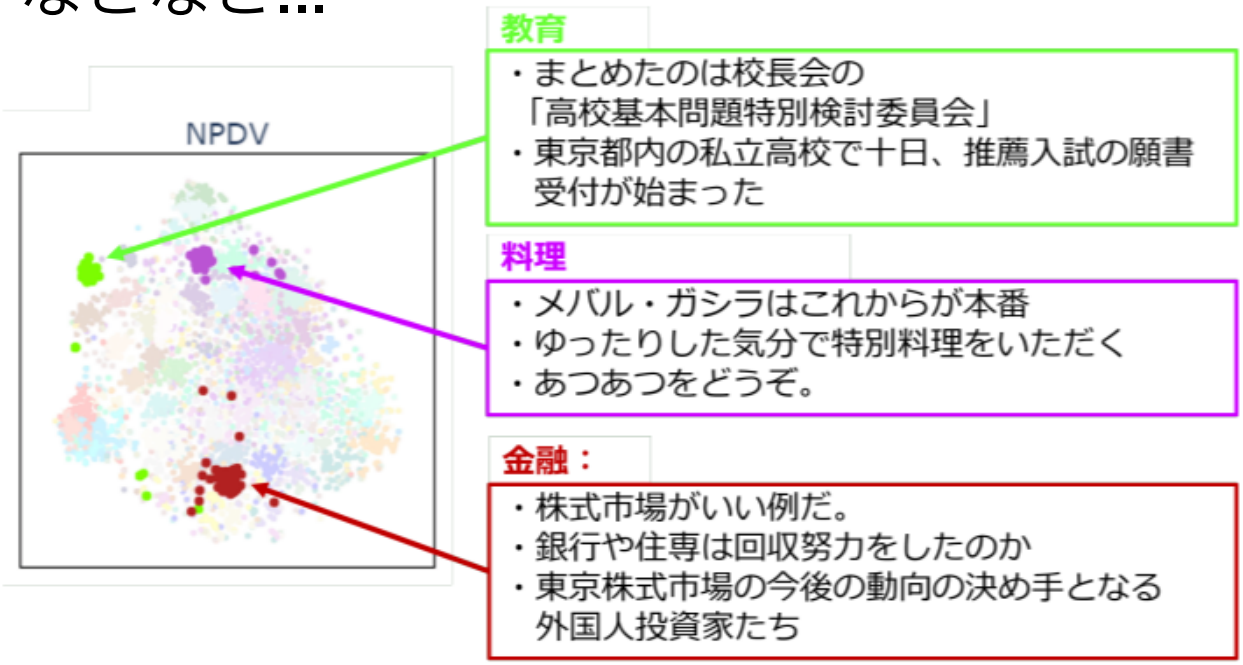
$\cos\theta$	文
(目標文)	だが、新しい後期高齢者医療制度では、介護..
0.8929	健康保険や介護保険、厚生年金、雇用保険、..
0.8705	国は患者が混合診療を受けた場合、「一体化..
0.8597	労働保険は、法人個人を問わず労働者を 1 人..
0.8313	また、短期入所や通所を受け入れる福祉施設..
0.8212	「住宅ローン控除」は、国内で一定の居住用..
0.8109	全体で５％アップと同水準だが、保険制度の..
0.8107	この免除の手続きをするだけで、保険料を払..
0.8098	「年金制度は世代間扶養の仕組みである」→..
0.8085	また、介護保険は対象外となっています。
0.8013	語学学校は特定商取引法の指定業務で、受講..
0.7946	機構や文部科学省によると、新制度は、悪質..
0.7937	連合はほかに「中低所得者層の所得税減税」..
0.7777	農水省案では、減反に加わる農家には生産量..
0.7740	厚生労働省は２６日、サービス事業者に支払..

$\cos\theta$	文
(目標文)	再生可能なファイル形式は、映像が MPEG..
0.9054	無線 LAN セキュリティは 64/128bit の W..
0.8790	その他の機能は地上デジタル/BS デジタル/..
0.8780	ネットワーク機能は 10/100/1000BASE-T..
0.8731	録音形式はリニア PCM で 16bit/44.1kHz..
0.8708	基本仕様は MP3/WMA/AAC 再生。
0.8627	その他の機能は、IEEE802.11b/g/n 対応..
0.8588	入出力端子には HDMI/コンポジットビデオ..
0.8486	同サービスは、i モード/EZweb/Yahoo!ケ..
0.8485	CG-BARPROG-X コレガは、WAN/LAN..
0.8417	対応 OS は、WindowsXP(SP2/SP3)/Vi..
0.8383	その他の機能は、10BASE-T/100BASE-T..
0.8377	「morawin[モーラウィン]forS!ミュージック..
0.8368	Blu-ray ディスク作成においては、1080i/7..
0.8339	対応ゲスト OS は OracleEnterpriseLinux..

($\cos\theta$ の大きい順に上位を表示)

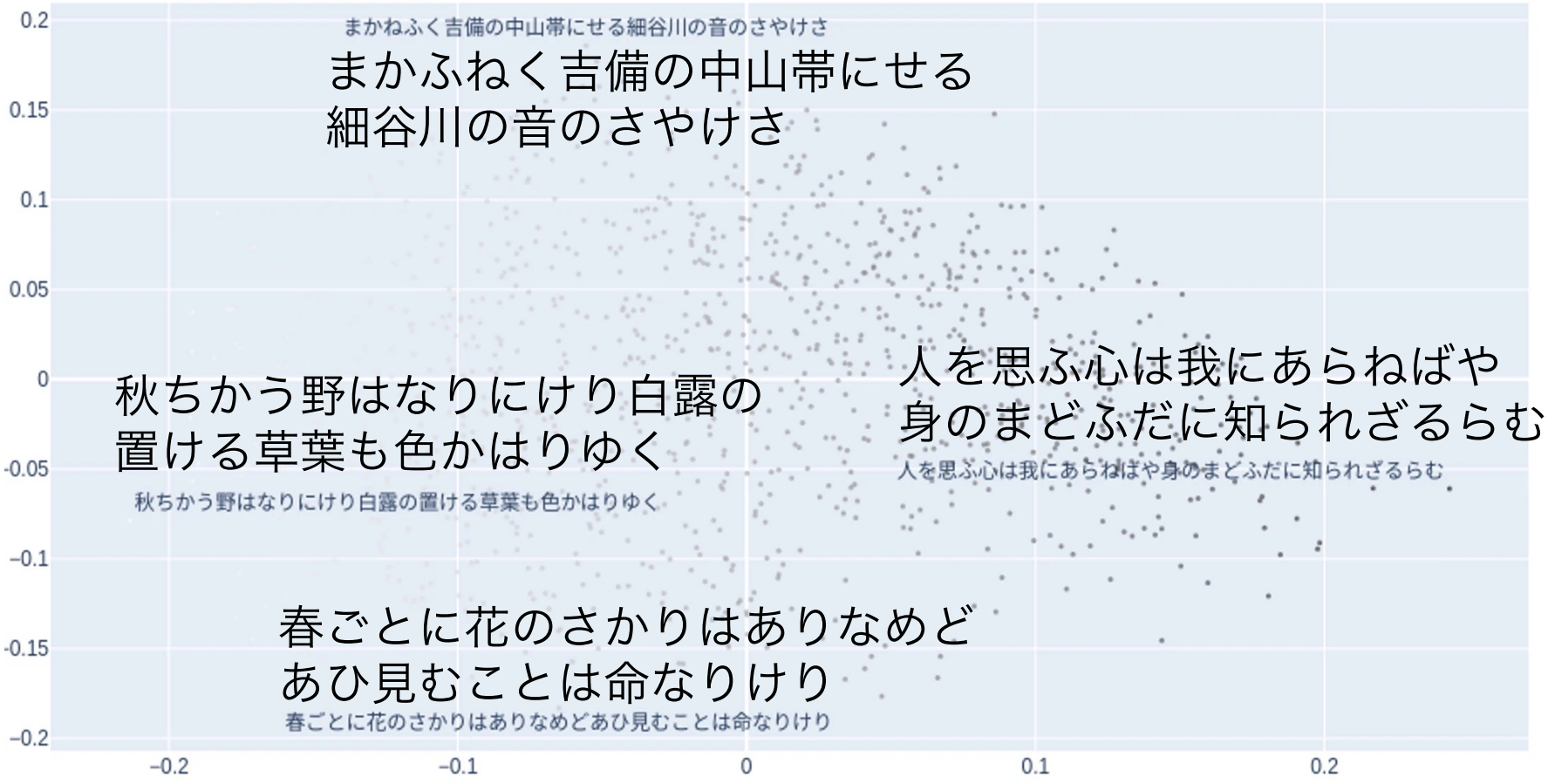
皆さんのデータ分析での応用

- 表面的な言葉が違っても、似た意味を考慮して分析できる (「考える」≈「考慮する」)
- アンケートなどで、文ベクトルを計算してどんな意見があるか、全体の様子を可視化してみる
- などなど...



毎日新聞の記事5,000個の各文のベクトルを2次元に可視化したもの
(石塚&持橋2022)

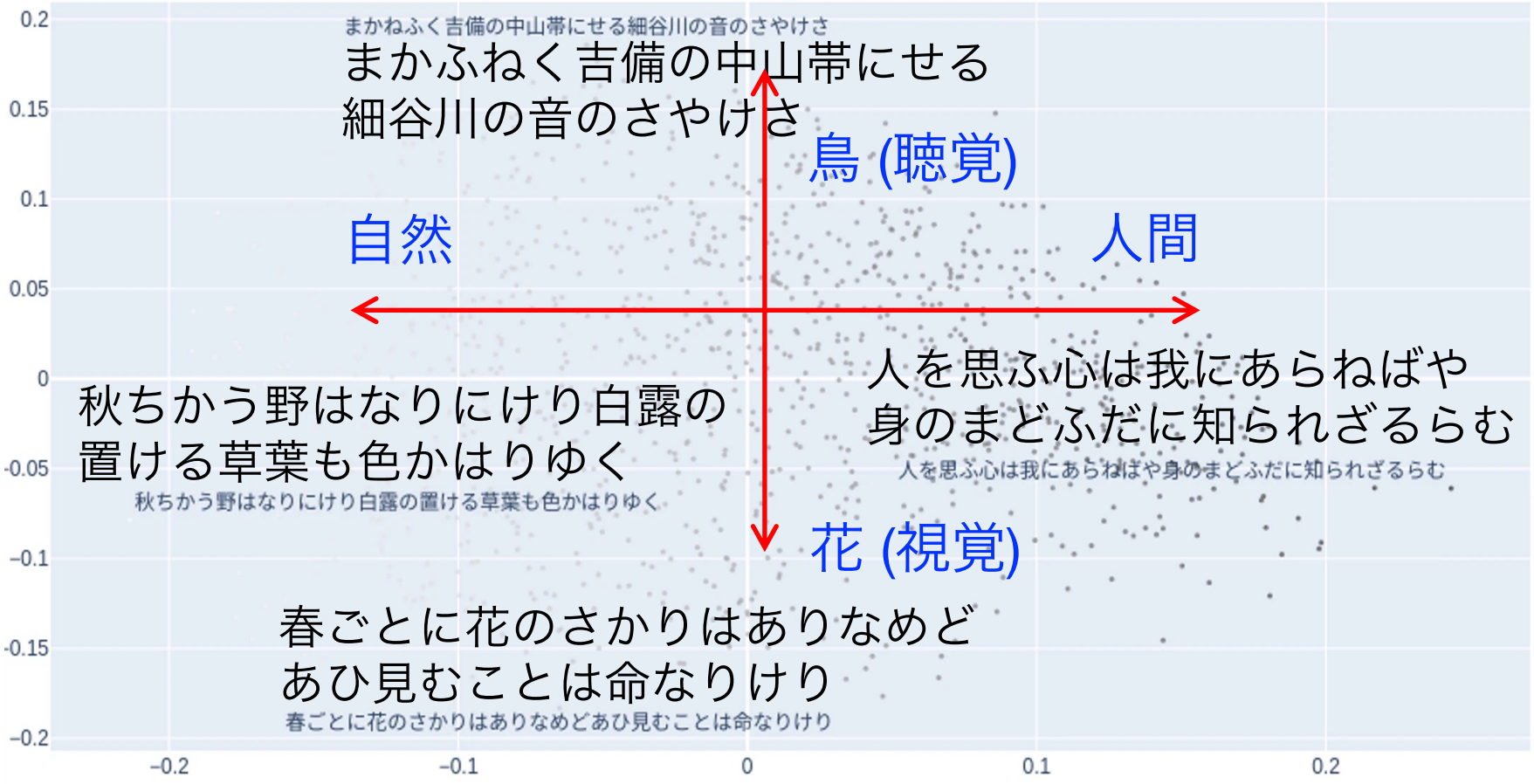
「古今和歌集」の和歌の文ベクトル



- 各点が1つの和歌 / 近藤(2023)による

近藤泰弘,「日本語の研究」第19巻3号 “和歌集の歌風の言語的差異の記述ー大規模言語モデルによる分析ー」

「古今和歌集」の和歌の文ベクトル

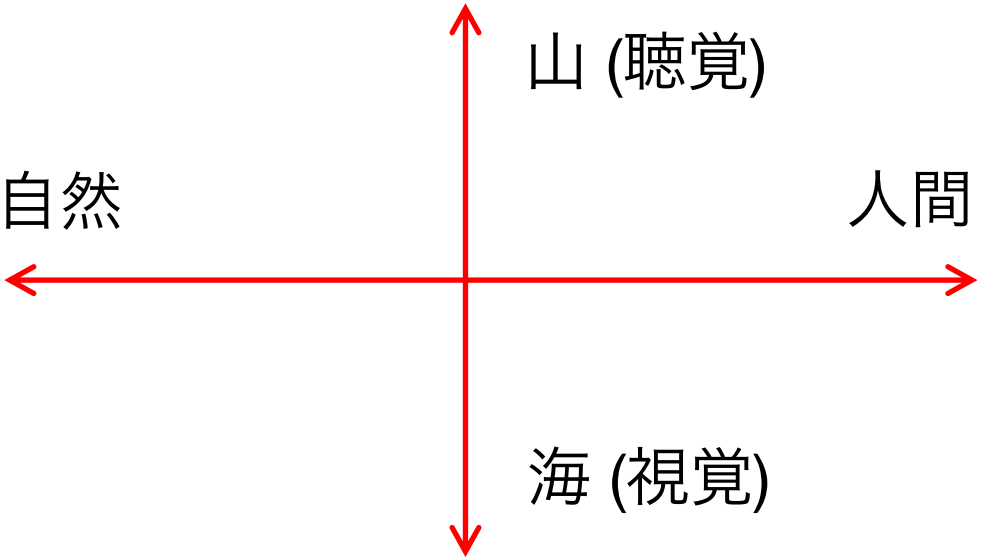


- 各点が1つの和歌 / 近藤(2023)による

近藤泰弘,「日本語の研究」第19巻3号 “和歌集の歌風の言語的差異の記述ー大規模言語モデルによる分析ー」

「万葉集」の和歌の文ベクトル

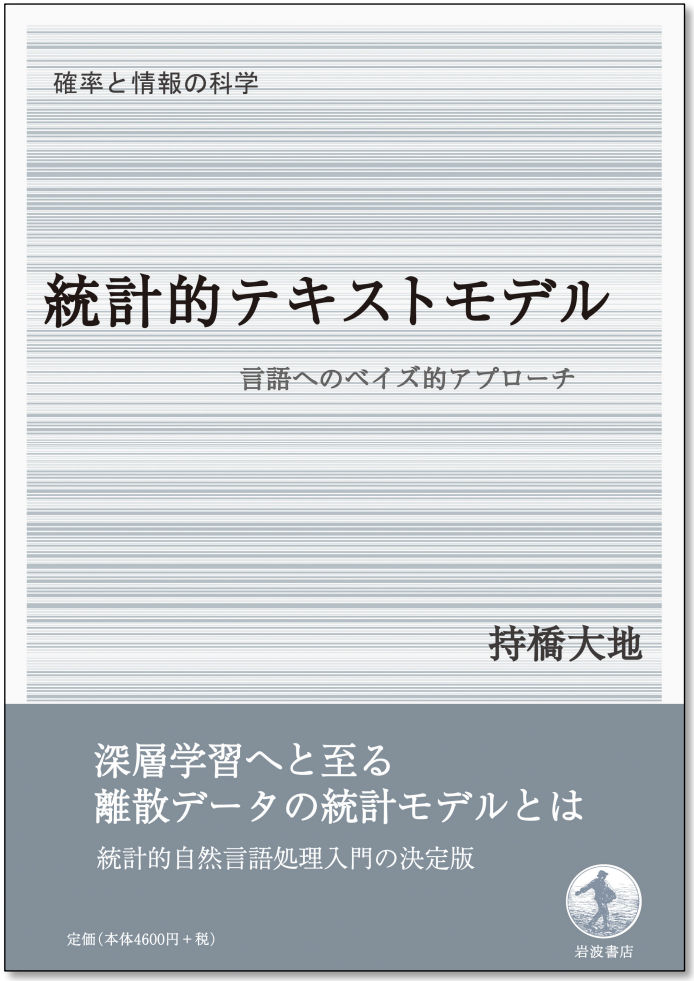
- 同様にして「万葉集」をプロットすると、



という、古今和歌集へとつながる同じ構造がある
ことがわかる

- 「後撰集」「和漢朗詠集」などでも同様の結果

教科書 (自然言語処理)



- 岩波書店から、今年7/1に発売
- 池袋ジュンク堂など、大きな書店には置いてあります
- 岩波書店の担当編集者は小石川の先輩、かつ同期のI君 (現役で先に文IIIに入学) のお姉様でした！

参考図書 (自然言語処理)

- 「岩波データサイエンス Vol.2 統計的自然言語処理」
- 私が特集担当と記事の執筆をしています
- 1500円で読みやすい読み物形式です
- 大きな書店には置いてあるはずです

まとめ

- 言葉はベクトル (数値の組) として表すことができ、様々な応用が可能
  - 言葉を扱う現在の人工知能の基礎
- 背後はすべて、数学でできている (数学は重要!)
- 皆さんのデータ分析の際にも、単語ベクトルや文ベクトルを活用することが有効
- 「自然言語処理」(＝計算言語学)には、ほかにも様々な面白い話題があります

私のホームページ

https://www.ism.ac.jp/~daichi/

- この講演のスライドも、Googleドライブ以外に上記のサイトにも置いておくようにします
- 他にも、様々なスライドや研究資料があります