確率的潜在意味スケーリング

持橋大地
統計数理研究所 数理・推論研究系
daichi@ism.ac.jp

情報処理学会 第249回自然言語処理研究会
2021-7-28 (水)

はじめに

- テキストを連続的な尺度で測りたい場合は、多くある
  - 住民へのアンケートがどれくらい肯定的か、否定的か
  - ホテルの評価で特に否定的(肯定的)なレビューはどれか
  - ある政治家の発言や法案がどれくらい保守的か/革新的か
  - 書簡から伺える作家の分裂症気質がどれくらいか

テキスト分類とは異なる!

- 1/0の単純なテキスト分類では対処できない
  (自民党ならば右派なのは自明なので、その中の違いが重要)
- 人間が事前に付けたラベルではなく、自由な軸でテキストの連続的なスケールを測りたい
  (英語は南北戦争前後でどのように変わったのか?)
- 教師データは一般に存在しない→教師なし学習

政治学方法論 (Political Methodology)

- 政治学に関する計量分析とその方法論の分野
  - 政治に関して、科学的・客観的な言明を行いたい
  - 文系分野としてはかなり高度な統計学・機械学習
- 特に、テキストの分析は中でも大きな分野
  - 背景：政治的なテキストのデータ化が進んでいる (国会会議録、地方議会議事録、法案のテキスト、外交文書、共産党機関紙、...)
  - 2019年秋に、POLTEXT 2019に招待されて参加

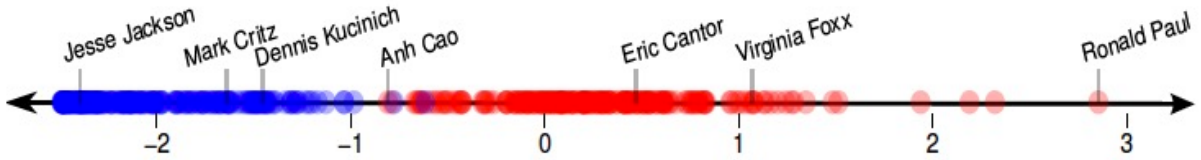
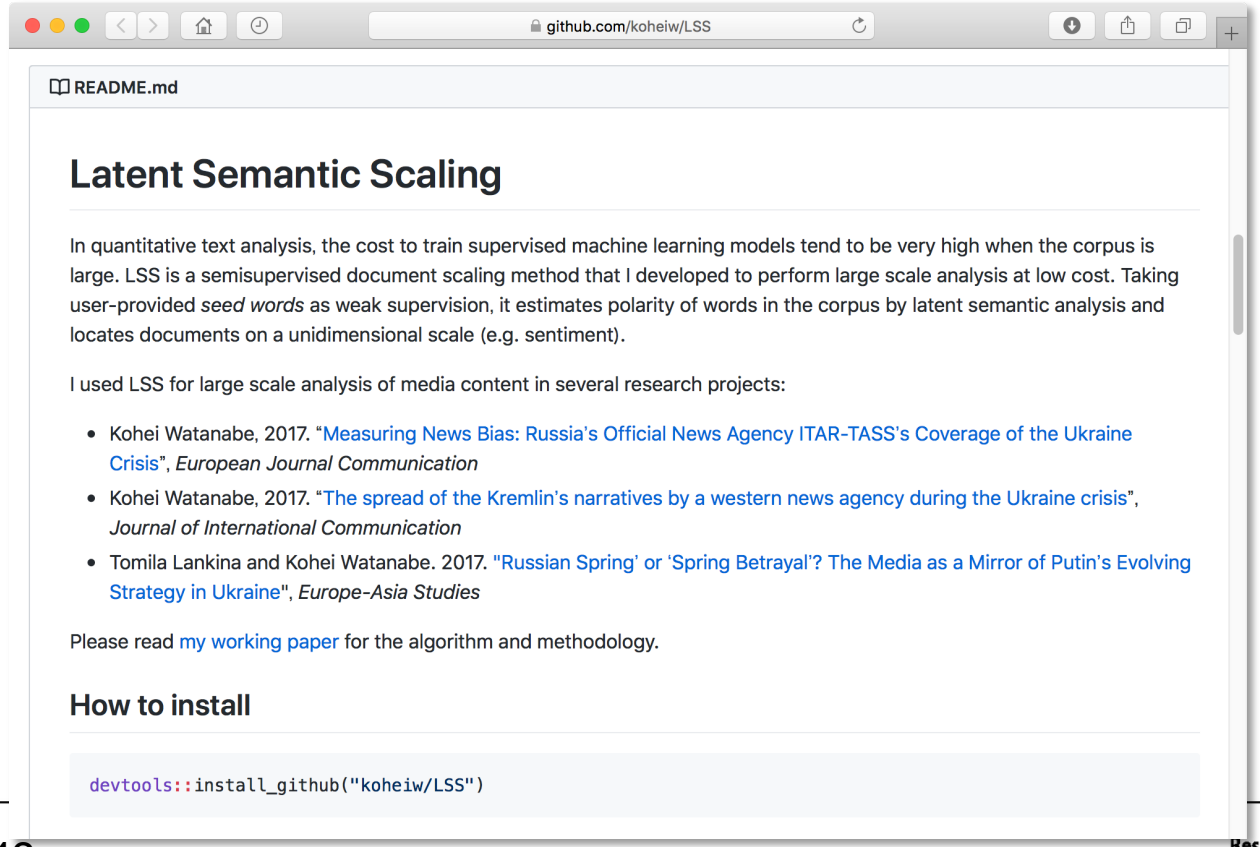


Figure 1: Traditional ideal points separate Republicans (red) from Democrats (blue).

Sean Gerrish and David M. Blei. “How They Vote: Issue-Adjusted Models of Legislative Behavior”, NIPS 2012

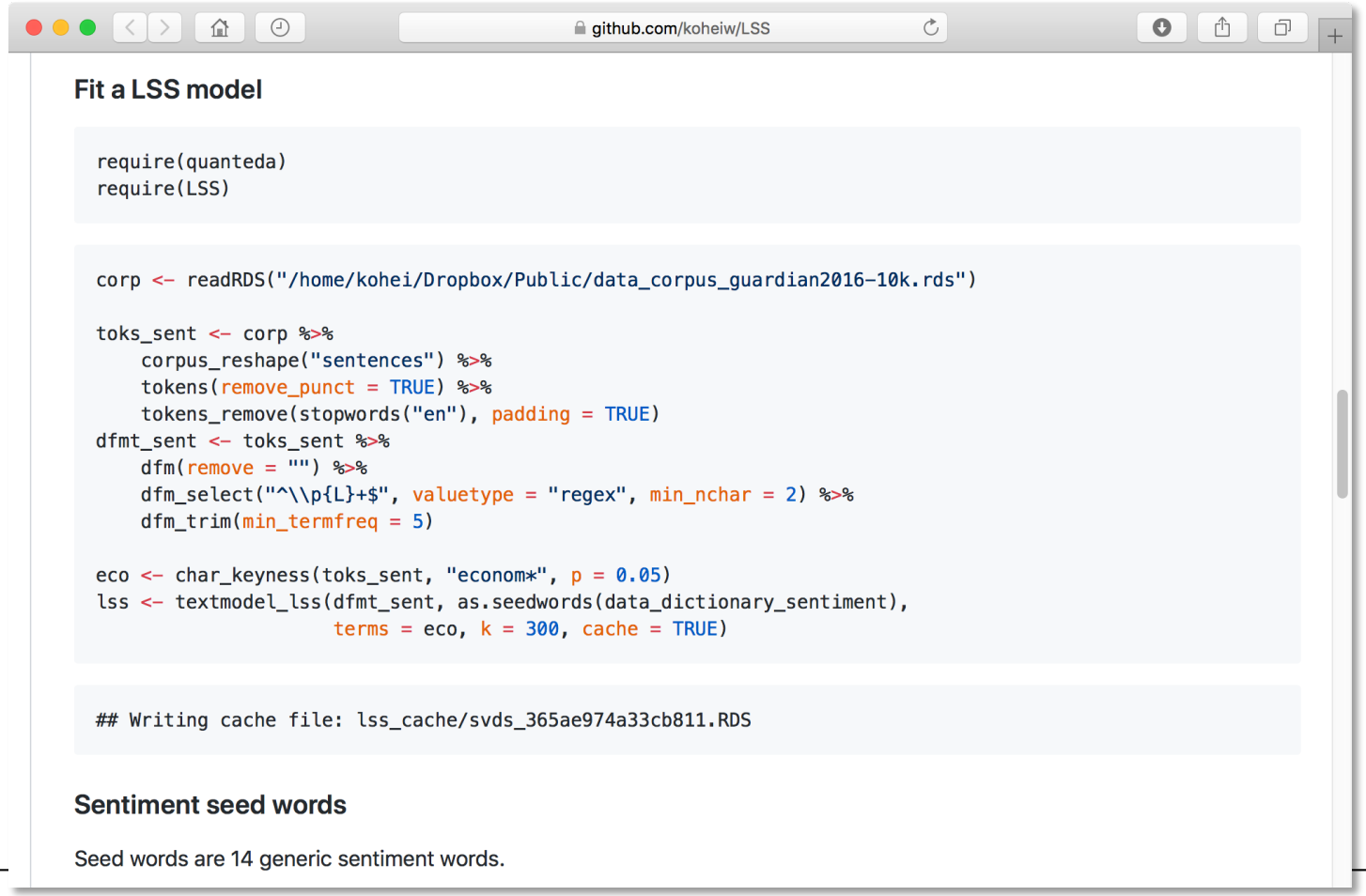
LSS (Latent Semantic Scaling)

- LSE(当時)の渡辺さんが開発 (Watanabe 2015/2020)
- テキストをある視点で時系列分析するためのRパッケージ



LSS/quanteda

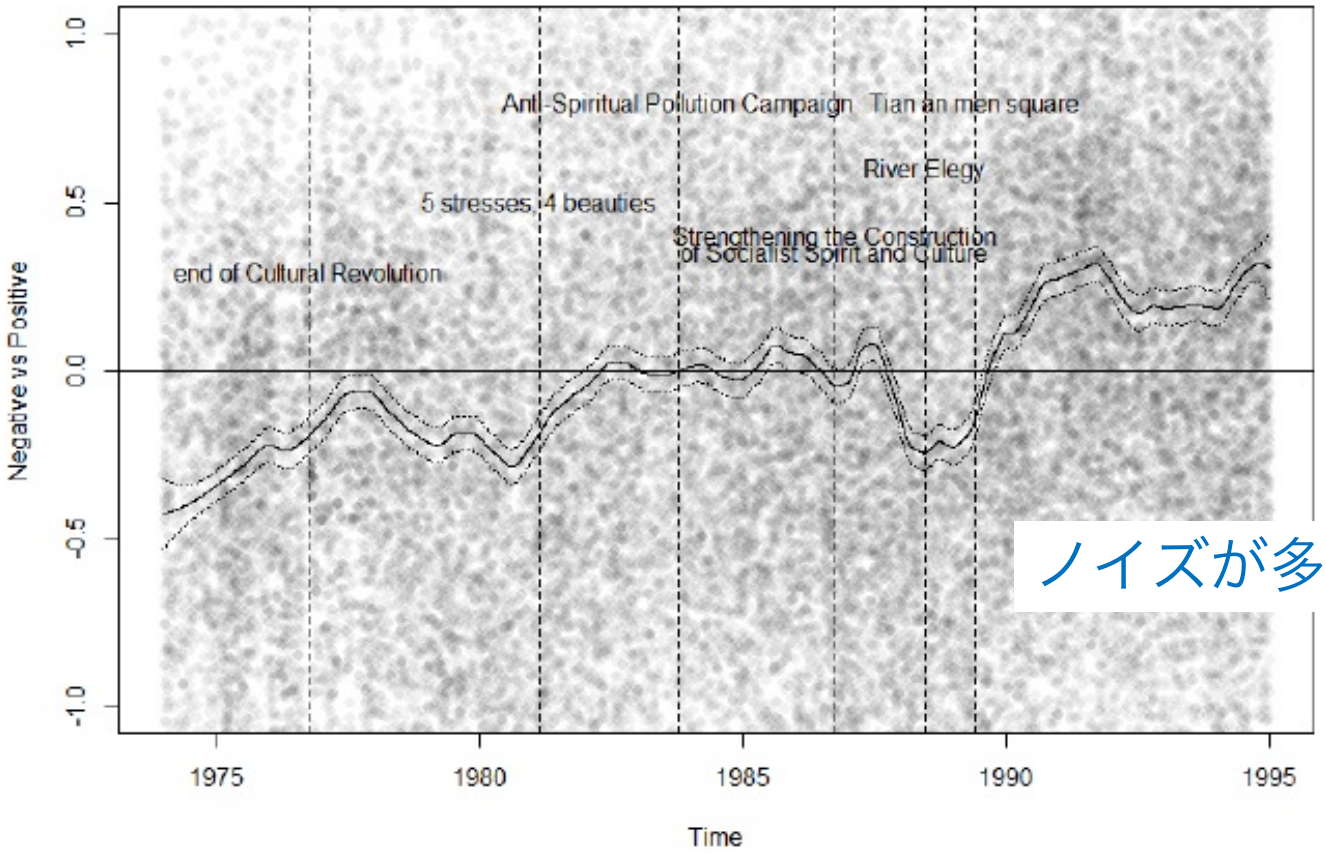
- Rを使って手軽に分析できるため、世界的に使われている



LSSによる分析例

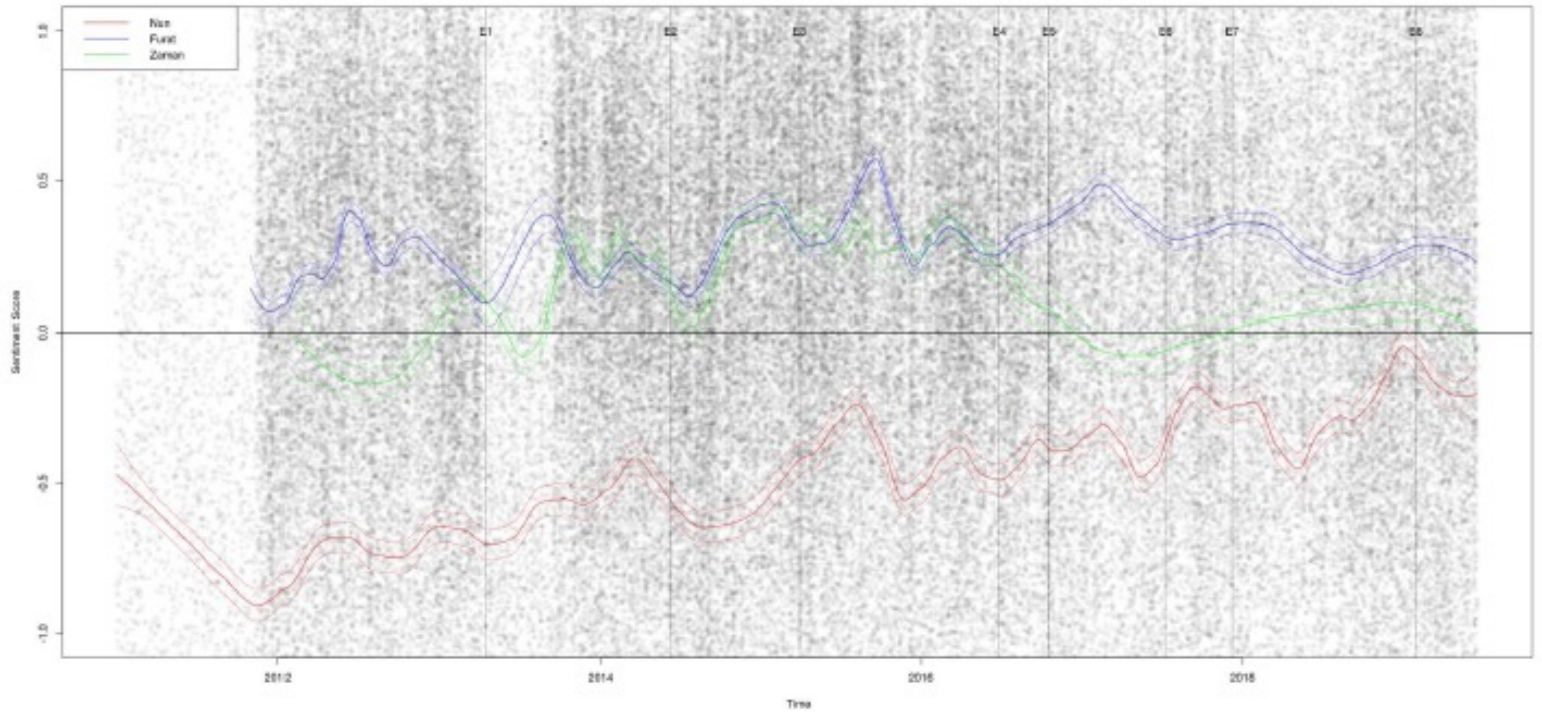
- 御器谷さん(慶応大学博士課程)による、中国の人民日報の中での「伝統」の使われ方の分析 @POLTEXT 2019

Figure 2 Valuation of tradition in the People's Daily



ノイズが多い！

LSSによる分析例 (2)



- 山尾さん (九大) / イスラム国の異なる宗派主義の時間変化, イラクの3種類の新聞記事(赤青緑) @POLTEXT
- 各文書での分散が非常に大きい

LSSの分析方法

- テキストの集合から、与えられた尺度の変化を抽出する方法 (古典的なベクトル空間モデル)
- 尺度：極性語とのベクトル空間での近さ
  (ex: positive-negative, left-right)

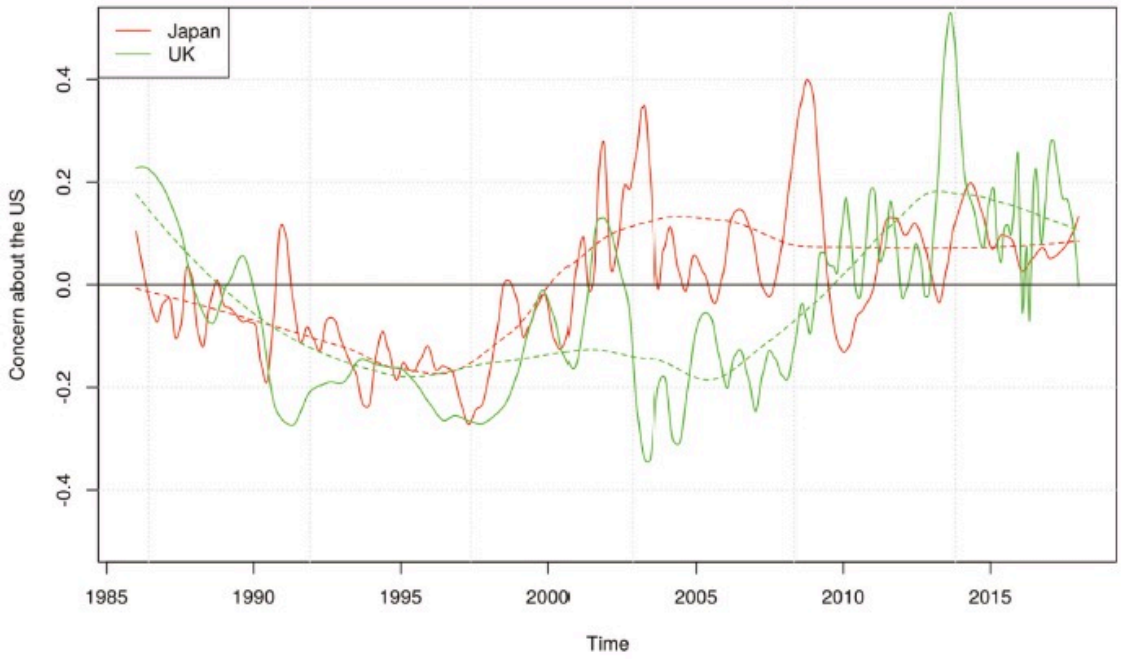


図2　日本とイギリスの新聞で見るアメリカに対する懸念

LSSによる分析方法 (1)

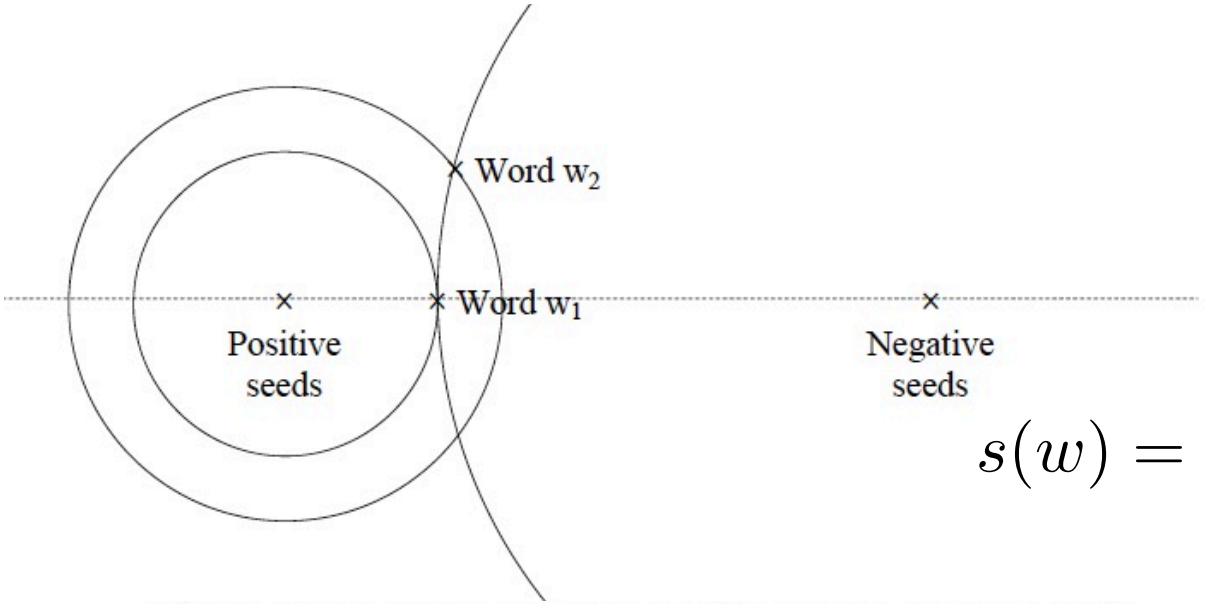
- 目標語 *w*との共起に基づいて、解析に用いる特徴語 *c* を選ぶ
- 例) 目標語 “econom*” との共起から選ばれた特徴語

Table 1: Entry words selected by econom*

Rank	Collocates	LL
1	growth	9969.0
2	social	5219.6
3	recovery	2264.4
4	political	2237.9
5	chief	1919.3
6	data	1518.5
7	inflation	1494.2
8	cycle	1462.6
9	rates	1346.8

LSSによる分析方法 (2)

- 特徴語を文書-単語行列の主成分分析で求めた、極性語とのコサイン距離でスコアリング



極性語 v_i は複数あるので、cos類似度を平均

$$s(w) = \frac{1}{M} \sum_{i=1}^{M} \cos(\vec{w}, \vec{v}_i) \cdot \mathrm{score}(v_i)$$

Table 2: positive and negative economic words

Rank	Positive words		Negative words	
1	good	100.00	poor	-33.75
2	achieving	71.68	damage	-31.46
3	prospects	71.01	politics	-27.49
4	role	68.77	news	-26.39
5	successful	67.84	yen	-24.78
6	strategic	64.17	devalue	-24.08

Wordfish (Slapin and Proksch 2008)

- ポアソン確率モデルに基づく教師なしIRT
  - y_{ivt} を時間 t で政党 i が単語 v を使った頻度として、

$$p(Y) = \prod_t \prod_i \prod_v \mathrm{Po}(\exp(\alpha_{it} + \psi_v + \beta_v * \theta_{it}))$$

    を最大化する政党の位置θと単語パラメータβを求める
- 完全な教師なし学習なので、得られる軸が分析したい軸と一致する保証がない
- 分析するテキスト集合に依存→選択の方法はない
- ポアソンモデルは、テキストの長さに依存

現在のテキスト分析の課題

- LSS: 特徴語・種語の選択や前処理に人手が介在
  - ヒューリスティックにより、結果が変わってしまう
  - 統計的に何を計算しているのか、の基準が曖昧
- Wordfish: 抽出される軸が分析したい軸と一致している保証がない

Wordfishと項目反応理論(IRT)をベースに、LSSを確率化したい…Probabilistic LSS (PLSS)

PLSSと項目反応理論 (IRT)

- 項目反応理論 (IRT) : 心理統計学における、
  被験者の潜在的特性を推定するための統計モデル
- SlapinのWordfish (ポアソンGLM)

$$p(y_{ivt} = k) = \mathrm{Po}(\exp(\alpha_{it} + \psi_v + \beta_v * \theta_{it}))$$

  の代わりに、次の多項モデルを考える

- 多項分布 IRT:

$$p(v|\theta, \phi) \propto p(v) \exp(\theta \cdot \phi_v)$$

$$= \frac{\exp(\log p(v) + \theta \cdot \phi_v)}{\sum_{v=1}^{V} \exp(\log p(v) + \theta \cdot \phi_v)}$$

- どんなモデル?

PLSS: Simple multinomial IRT

$$p(v|\theta,\phi) \propto p(v)\exp\left(\theta\cdot\phi_v\right)$$
$$= \frac{\exp\left(\log p(v)+\theta\cdot\phi_v\right)}{\sum_{v=1}^{V}\exp\left(\log p(v)+\theta\cdot\phi_v\right)}$$

未知

未知

- テキストに、潜在的な極性 $\theta \sim \mathcal{N}(0,1)$ が存在すると仮定
- 単語 v の確率は、ユニグラム確率 $p(v)$ に因子 $\exp(\theta\cdot\phi_v)$ が掛け合わされて決まる
  - θ と ϕ_v の正負が一致すると高い確率→ ϕ_v は単語 v の極性
- θ も ϕ_v も未知の場合の教師なしロジスティック回帰と考えてもよい

PLSS: 単語ベクトルの利用

- 単語に関する言語的知識を使うために、単語ベクトルの利用を考える
- GloVeやWord2vec等で事前に計算された単語ベクトルを利用 (手元のテキストで計算してもよい)
- ニューラル単語ベクトルの空間に、以下で示すような部分空間が存在する: Ultradense Embedding (Rothe+ 2016)
  - DensRay (Dufter and Schutze 2019)は、以下の簡単な手法より明らかに性能が悪かった (詳しい議論は論文を参照)

PLSS: 単語ベクトルの利用 (2)

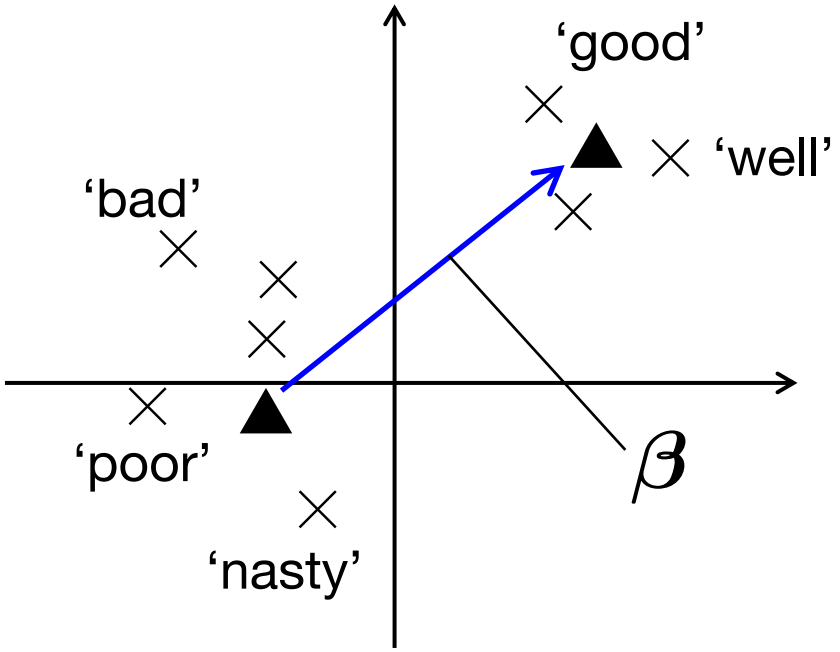
- 単語 v のベクトルを $\vec{v}$ と書くことにすると、単語毎のバイアス ϕ_v に代えて、次のモデルを考える
- 単語ベクトルに対する回帰係数 $\boldsymbol{\beta}$ を使って、

$$p(v|\theta, \boldsymbol{\beta}) = \frac{\exp\left(\ell_v + \theta \cdot \boldsymbol{\beta}^T \vec{v}\right)}{\sum_v \exp\left(\ell_v + \theta \cdot \boldsymbol{\beta}^T \vec{v}\right)}$$

  - $\boldsymbol{\beta}$ と $\vec{v}$ の内積 $\boldsymbol{\beta}^T \vec{v}$ が、ϕ_v と同じ働き
  - LSSで使われているコサイン類似度の確率モデル化
- 単語ベクトルと同じ次元のパラメータ $\boldsymbol{\beta}$ を、1つ推定すればよい

βの推定

- 最も簡単には、＋ー を与えた少数の極性語辞書の単語ベクトルの平均の差を計算



  - βのノルムは1にする→極性語として極端な単語を与える必要がない (LSSとの違い)

βの推定

- 各単語ベクトル $\vec{v}$ について、$\phi_v = \beta^T \vec{v}$ が軸 βに沿った「単語の極性」を表す
- 例: 英語の一般的なpositive-negative辞書を使った場合

```
# general positive-negative words, same as LSS
positive        good nice excellent positive fortunate correct superior
negative        bad nasty poor negative unfortunate wrong inferior
```

計算された単語の極性 $\phi_v = \boldsymbol{\beta}^T \vec{v}$ (上位)

honored	0.4658	guest	0.3765	unique	0.3533
selected	0.4635	custom	0.3731	dedicated	0.3530
coveted	0.4254	suites	0.3719	lounge	0.3468
suite	0.4189	designing	0.3716	certified	0.3459
invited	0.4185	terrific	0.3708	wines	0.3451
excellent	0.4183	booked	0.3686	via	0.3447
tasting	0.4146	flexible	0.3654	elegant	0.3414
craft	0.4117	spa	0.3605	installed	0.3403
chosen	0.4116	wonderful	0.3579	happy	0.3402
nice	0.4053	slated	0.3574	venue	0.3400
designed	0.3884	plus	0.3558	prepared	0.3393
available	0.3858	buick	0.3554	design	0.3383
pleasant	0.3849	indoor	0.3552	preparing	0.3370
fortunate	0.3780	accessible	0.3544	fine	0.3369

計算された単語の極性 $\phi_v = \boldsymbol{\beta}^T \vec{v}$ (下位)

greed	-0.5244	anger	-0.4404	rumors	-0.4045
rampant	-0.5135	scandals	-0.4394	backlash	-0.4031
panic	-0.4923	violence	-0.4367	reckless	-0.4018
fear	-0.4839	catastrophe	-0.4350	misery	-0.4018
racism	-0.4753	rage	-0.4317	enron	-0.4013
ignoring	-0.4693	violent	-0.4298	tide	-0.4010
unrest	-0.4618	abuses	-0.4234	depression	-0.3978
chaos	-0.4593	outrage	-0.4233	badly	-0.3972
intimidation	-0.4500	distress	-0.4202	confronting	-0.396
riots	-0.4532	abruptly	-0.4175	epidemic	-0.3940
subprime	-0.4514	blame	-0.4164	persistent	-0.3935
fears	-0.4489	bust	-0.4096	meltdown	-0.3921
collapse	-0.4429	resentment	-0.4065	deepening	-0.3886
fearful	-0.4418	anxiety	-0.4053	rhetoric	-0.3882

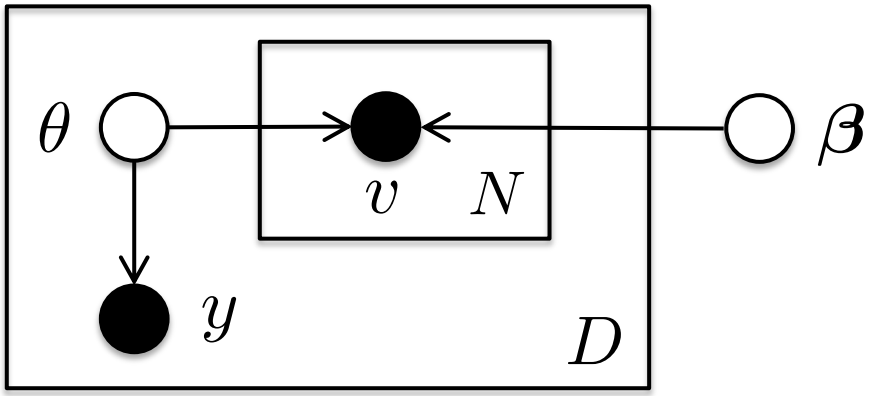
- 辞書でごく一部の単語を指定しただけで、多数の単語に適切な極性が与えられている！

βの半教師あり学習

- 辞書が分析前に自明に準備できるとは限らない
  → 辞書の代わりに、「正例のテキスト集合」「負例のテキスト集合」を少数だけ与える
- 辞書と同時に使うことも可能：辞書によるβを初期値にして学習
- データ:
  X_ℓ, Y_ℓ ：ラベルがある少数のテキストとラベル
  X_u ：ラベルのない多数のテキスト
  - 機械学習的には、半教師あり学習の枠組み

β の半教師あり学習 (2)

- $p(Y_\ell|X_\ell)$ (だけ)を最大化するようにβ を学習



$$p(y,d|\theta) = p(y|\theta)\prod_{v\in d} p(v|\theta,\beta)$$

$$p(v|\theta,\beta) = \frac{\exp(\ell_v + \theta\cdot\beta^T\vec{v})}{\sum_{v=1}^{V}\exp(\ell_v+\theta\cdot\beta^T\vec{v})}$$

未知

未知

- 問題：ラベル付きデータの θ はそもそも未知！

β の学習 (1)

- θ を積分消去して、β のみを学習する

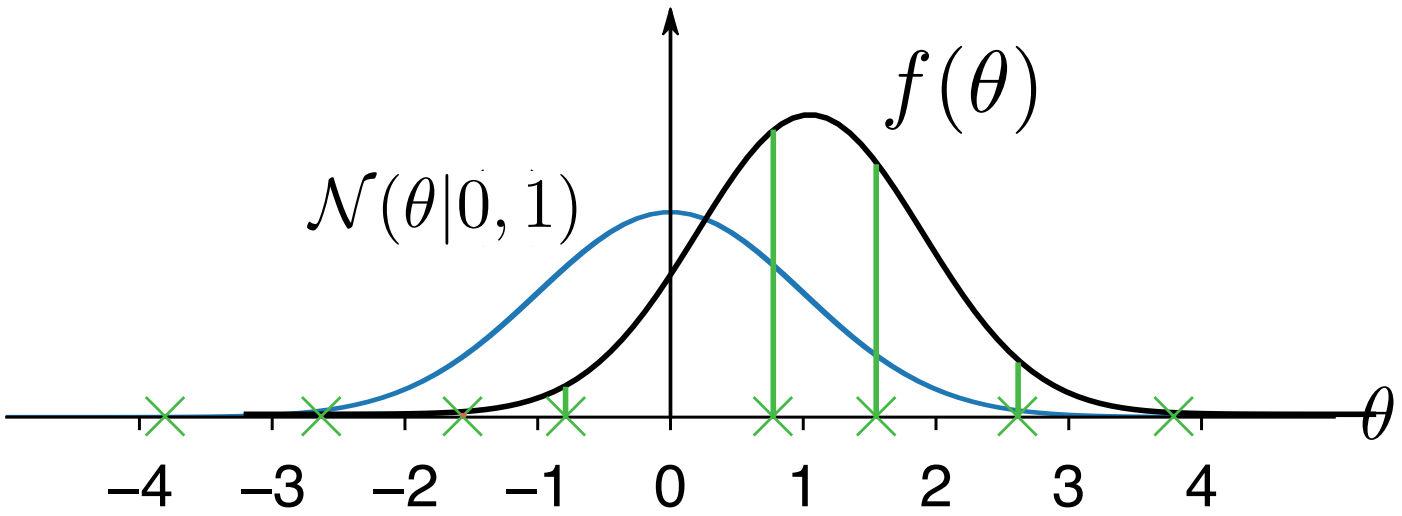
$$p(y, d|\boldsymbol{\beta}) = \int_{-\infty}^{\infty} p(y, d, \theta|\boldsymbol{\beta}) d\theta = \int_{-\infty}^{\infty} p(y|\theta) p(d|\theta, \boldsymbol{\beta}) p(\theta) d\theta$$

- $p(\theta)$ は標準正規分布 $\mathcal{N}(0, 1)$ なので、Gauss-Hermite 求積で上の式は高精度に数値積分できる

$$\int_{-\infty}^{\infty} f(\theta) \mathcal{N}(\theta|0, 1) d\theta \simeq \frac{1}{\sqrt{\pi}} \sum_{i=1}^{H} w_i f(\sqrt{2} x_i)$$

  - *H*個の分点 $x_1 \cdots x_H$ とその重み $w_1 \cdots w_H$ は直交エルミート多項式から事前に数学的に求まる
  - $f(\theta)$ を直交多項式で近似していることに相当

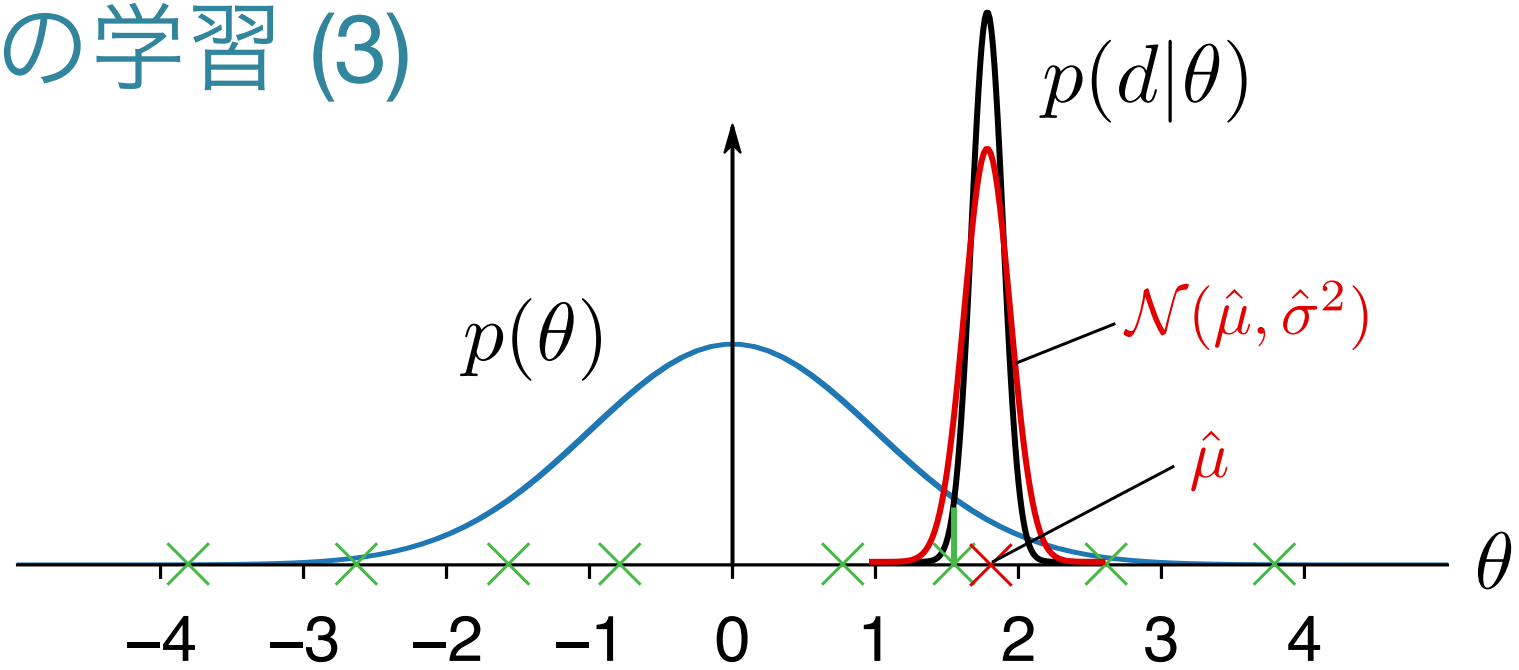
β の学習 (2)



- Gauss-Hermite求積: 緑の分点で計算するだけで、$\mathcal{N}(0,1)$ に関する積分は高精度に計算できる

$$\int_{-\infty}^{\infty} f(\theta)\mathcal{N}(\theta|0,1)d\theta \simeq \frac{1}{\sqrt{\pi}} \sum_{i=1}^{H} w_i f(\sqrt{2}x_i)$$

β の学習 (3)



- 問題: 実際のテキストでは、尤度 $p(d|\theta)$ は特定の θ の周りに集中→緑の分点がヒットしない
- 解決策: 積分を仮想的に $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ で行うように変数変換 (Liu+ 1994, Biometrika)
  - $\hat{\mu}$ は2分探索で、$\hat{\sigma}^2$ は数値二階差分で計算できる

β の学習 (4)

$$\log \int_{-\infty}^{\infty} e^{\ell(\theta)} \mathcal{N}(\theta|0,1) d\theta$$

$$\simeq \log \left[\frac{\sigma}{\sqrt{\pi}} \sum_{k=1}^{K} \exp \left\{ \log w_k + \ell(y_k) + \frac{1}{2} \left(\frac{1}{\sigma^2} (y_k - \mu)^2 - y_k^2 \right) \right\} \right]$$

β を含む

- 実際の計算は、上式のようにすべて対数で行う
- β はK次元のベクトルなので、さらにこの式を β で偏微分して勾配を計算→L-BFGSで最適化
  - Hamiltonian MCMCによるベイズ推定も試したが、最適化によるMAP推定の結果がよかった
- ベクトル化によるRの実装が大変。。

実験とデータ

- Young and Soroka (*Political Communication*, 2012) のテキストに対する人手のコーディングデータ

  <doc>
  The rate at which Americans bought new single family homes dropped 14 percent in September from the month before x the Government reported today . Resurging mortgage rates and a sharp jump in home prices apparently deterred potential ...
  <doc>
  EXCEPT for ‘’conservative ,‘’ which is the way Ronald Reagan describes himself x economic labels are difficult to pin on the President elect . One of the best ways to predict what he may do as President is to examine ...

- [“Neutral” “Negative” “Negative”], [“Neutral” “Neutral” “Neutral”] のような3人の評定を、5段階に変換
  - この生評定データを使った方がよい?

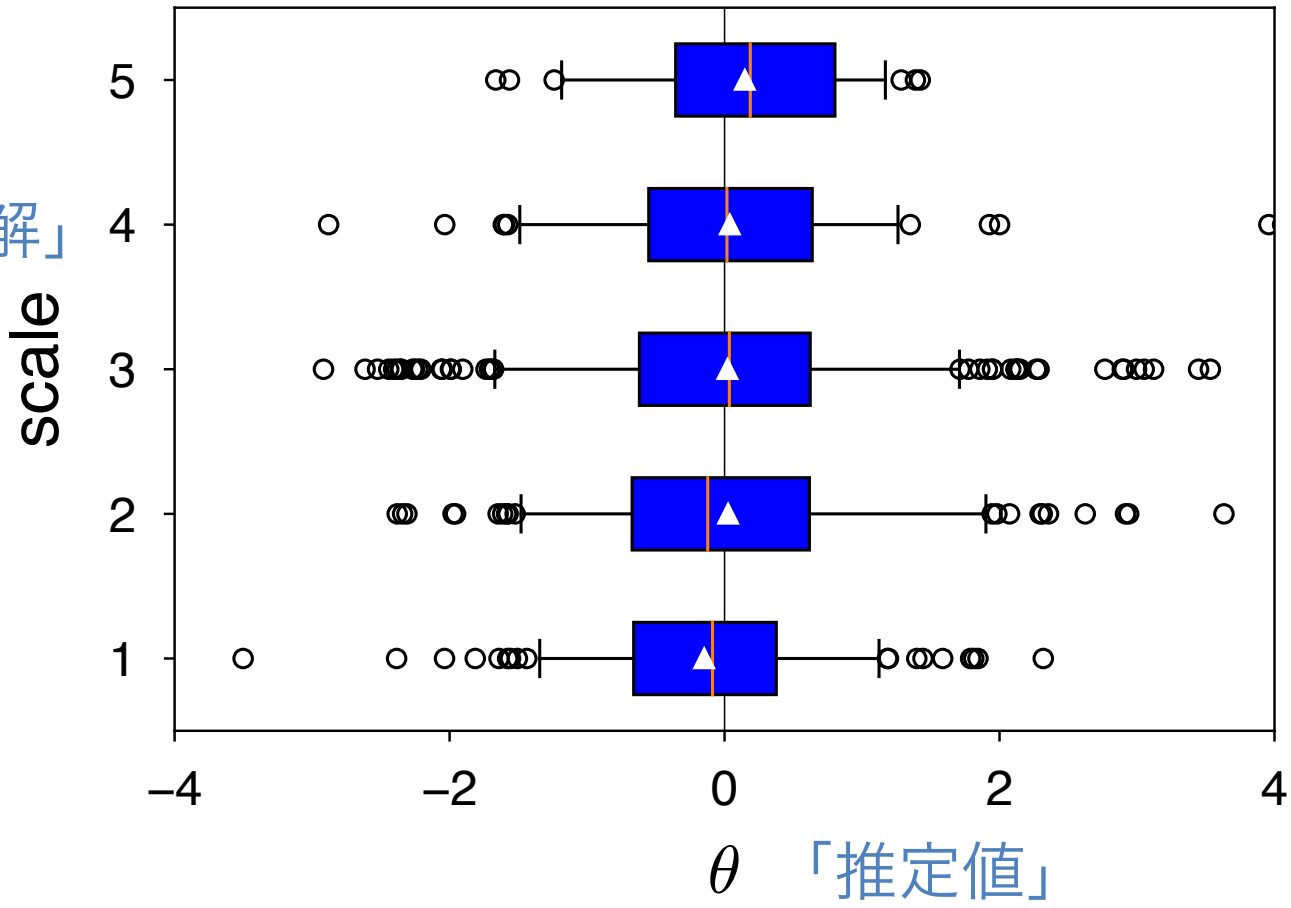
極性語

- 極性語の辞書としては、LSS (Watanabe 2020)で使われている次の標準的なpositive-negative辞書を使用

```
# LSS's data_dictionary_sentiment
positive      good nice excellent positive fortunate correct superior
negative      bad nasty poor negative unfortunate wrong inferior
```

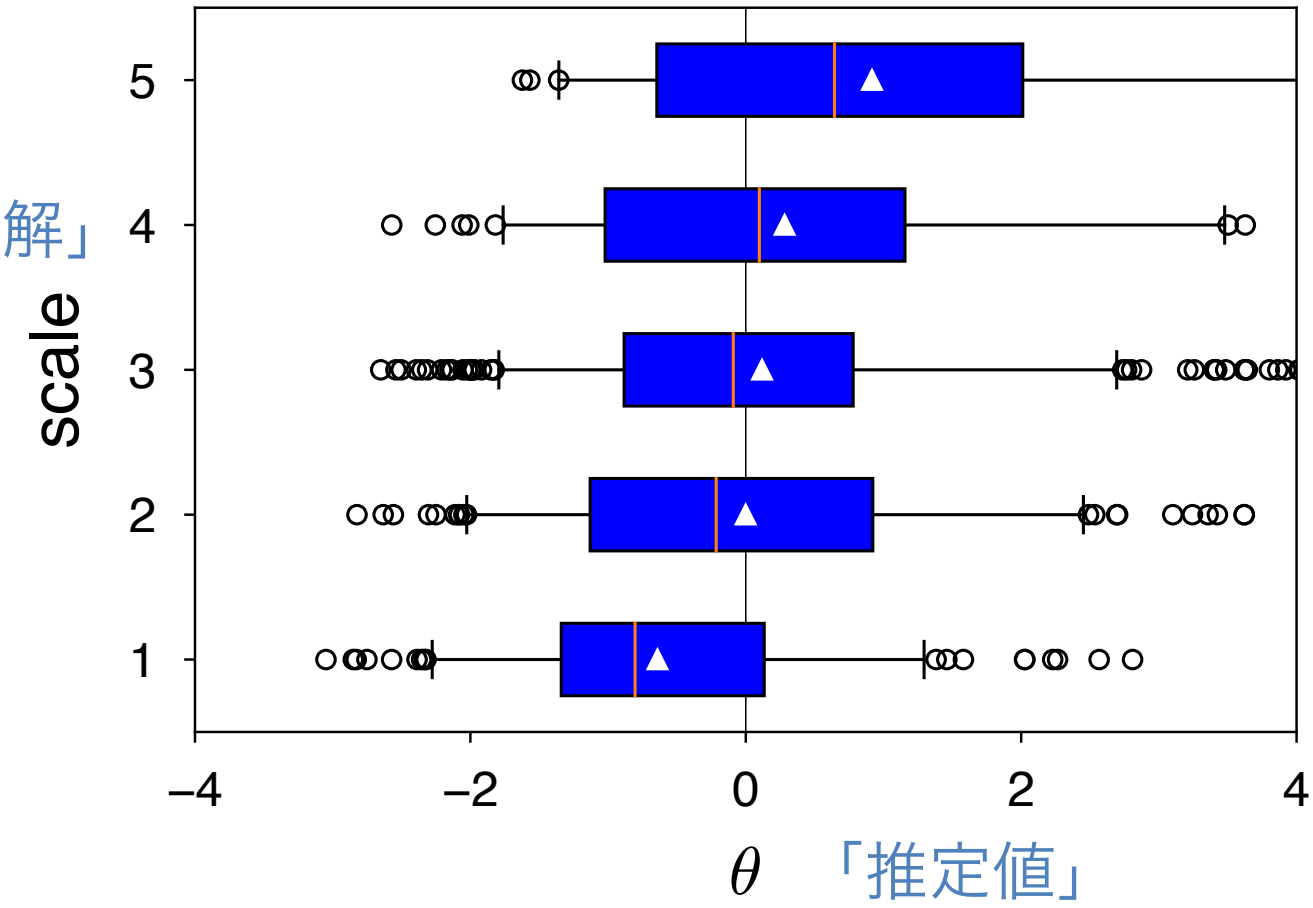
- 注: 提案法はPositive-Negativeのような自明な軸だけでなく、任意の複雑な意味的な軸を扱える

LSSと人手のスケールとの相関



- 平均値(△)は微妙にスケールに相関しているが、非常にノイズが多い状態 (相関係数 ρ=0.065、平均だと0.901)

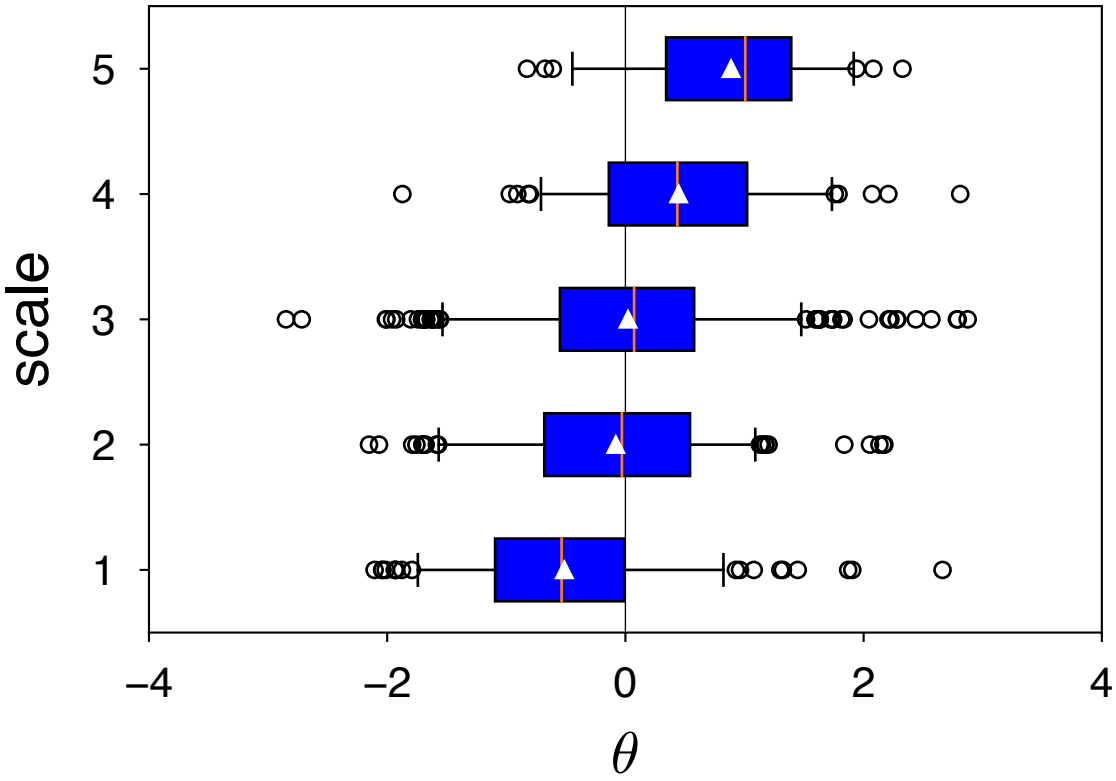
PSSと人手のスケールとの相関



- LSSよりかなり高い相関が出る (ρ =0.240, 平均だと0.965)
- モデルが Bag of words なので、さらに改善の余地あり

31

極性辞書＋半教師あり学習



- 標準的な極性辞書に正例/負例を15件加えて学習
- 相関係数 ρ=0.348, 平均(△)を使えば0.985に上昇

32

国会議事録の分析

- 2021年春の通常国会の衆議院・農林水産委員会の議事録を使用 (Webからスクレーピング)
  - 議員の論点の違いが比較的明らかなため
  - 12回の開催、1,324個の発言
- 例として、玉木雄一郎議員(国民民主党)の発言を+、田村貴昭議員(日本共産党)の発言を-として、それぞれランダムに20発言を抽出して半教師ありテキストとする
- β を最適化して求め、他の発言をPLSSで分析して潜在的な極性 θ を計算
- 極性辞書は使わない (準備できない)

国会議事録の分析 (結果)

θ	発言者	発言内容
1.6410	大串（博）委員	立憲民主党・無所属の大串です。　早速質疑に入ります。　貯保法ですけれども、私は、
1.4988	矢上委員	時間の関係で次の質問に移らせてもらいますけれども、低コスト化対策ですね。二問あっ
1.4838	重徳委員	だから、農水省に何の非もないのかと言っているんですよ。　例えば、大臣、富山県の御
1.3139	大串（博）委員	全くちぐはぐですね。一時的な要因で余っているんだったら、一時的に市場から切り離さ
1.1895	本郷政府参考人	木材流通に関してでございます。需給のミスマッチを起こさないように、生産、加工の事
1.0874	玉木委員	国民民主党の玉木雄一郎です。　本法案についてまず質問いたします。　先ほどから、規
1.0784	玉木委員	コロナにはいろいろなことを教えてもらったなと思ったんですが、例えば、マスク一つ国
1.0716	近藤（和）委員	石川県能登半島の近藤和也でございます。よろしくお願いいたします。　ＣＯＶＩＤ―1
1.0316	金子（恵）委員	今、イノベーションの話もされたので、済みません、順番を変えて、林業労働力の育成、
0.8315	神谷（裕）委員	そうしますと、遡れる限り遡るということだと思うんですが、そこで、先ほど議論になっ
⋮		
−0.5228	野上国務大臣	御指摘のございました主要農作物種子法につきましては、昭和二十七年に、戦後の食料増
−0.7074	野上国務大臣	間伐等特措法によりまして、平成二十年の法律制定後、一定以上の森林面積を有します市
−0.8432	葉梨副大臣	お答えいたします。　佐々木先生の資料の二の品目横断的経営安定対策、これが導入され
−1.0359	水田政府参考人	お答えいたします。　委員御指摘の冊子の二ページのところに「ＥＵやアメリカの現状」
−1.5408	高鳥委員長	お諮りいたします。　ただいま議決いたしました法律案に関する委員会報告書の作成につ
−1.5771	高鳥委員長	起立少数。よって、本修正案は否決されました。　次に、原案について採決いたします。
−1.9059	大串（博）委員	今回の農中さんの議論を契機に是非いい議論をしていただきたいと思いますし、間違って
−2.2223	田村（貴）委員	私は、日本共産党を代表して、本法案に反対の立場から討論を行います。　第一に、改正
−2.5166	田村（貴）委員	私は、日本共産党を代表して、畜舎等の建築及び利用の特例に関する法律案に反対の討論
−2.6210	重徳委員	立憲民主党の重徳和彦です。　今日は矢上筆頭、先輩、同僚議員の御了解をいただきまし
−3.5215	新井政府参考人	お答えいたします。　ＯＩＥ連絡協議会は、産業界及び学界における技術者又は学識経験

- 玉木ー田村の極性軸上に、連続的に発言を並べられる

国会議事録の分析 (結果)

- 玉木ー田村議員を極性軸とした際の、各単語の極性

$$\phi_v = \boldsymbol{\beta}^T \vec{v}$$

(玉木側)

v	ϕ_v	v	ϕ_v
まずは	0.5167	訴訟	−0.5769
なので	0.4560	傍聴	−0.5646
一つ	0.4364	毀損	−0.5519
ミリ	0.4246	原告	−0.5481
増やす	0.4178	敗訴	−0.5147
整い	0.4109	控訴	−0.5010
もっと	0.4014	係争	−0.4963
もう少し	0.4012	裁判所	−0.4962
植え付ける	0.3995	判決	−0.4808
切り替える	0.3985	審	−0.4727
一番	0.3982	シベリア	−0.4599
どうにか	0.3940	退け	−0.4491
しっかり	0.3903	高裁	−0.4462
同時に	0.3897	弁護	−0.4431
早く	0.3890	最高裁	−0.4410
戦える	0.3886	紛争	−0.4303

(田村側)

テキストの統計的抽出

- 分析に用いるテキストの選択を、客観的に行う必要
  - Wordfishなど教師なし手法では、結果に直結
  - キーワード検索では漏れが大きい
  - Polmeth分野ではこれまで、トピックの選択などが人手で行われてきた (トピック数>1000だと破綻)
- このために、
  - 背景分布付き潜在トピックモデルを使う方法

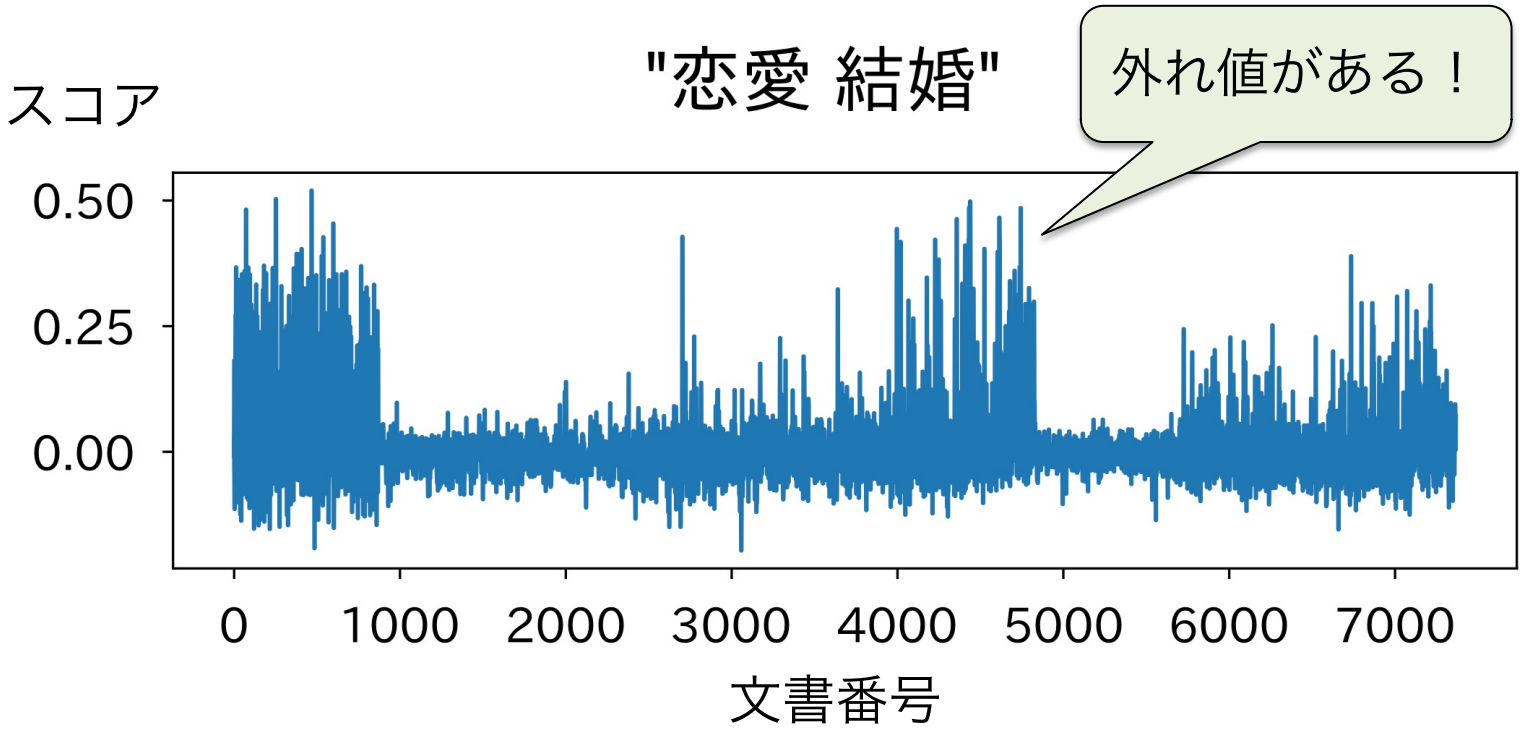
  ✓ニューラル文書ベクトルを使用する方法

  を提案した (詳細は論文参照)

テキストの統計的抽出 (2)

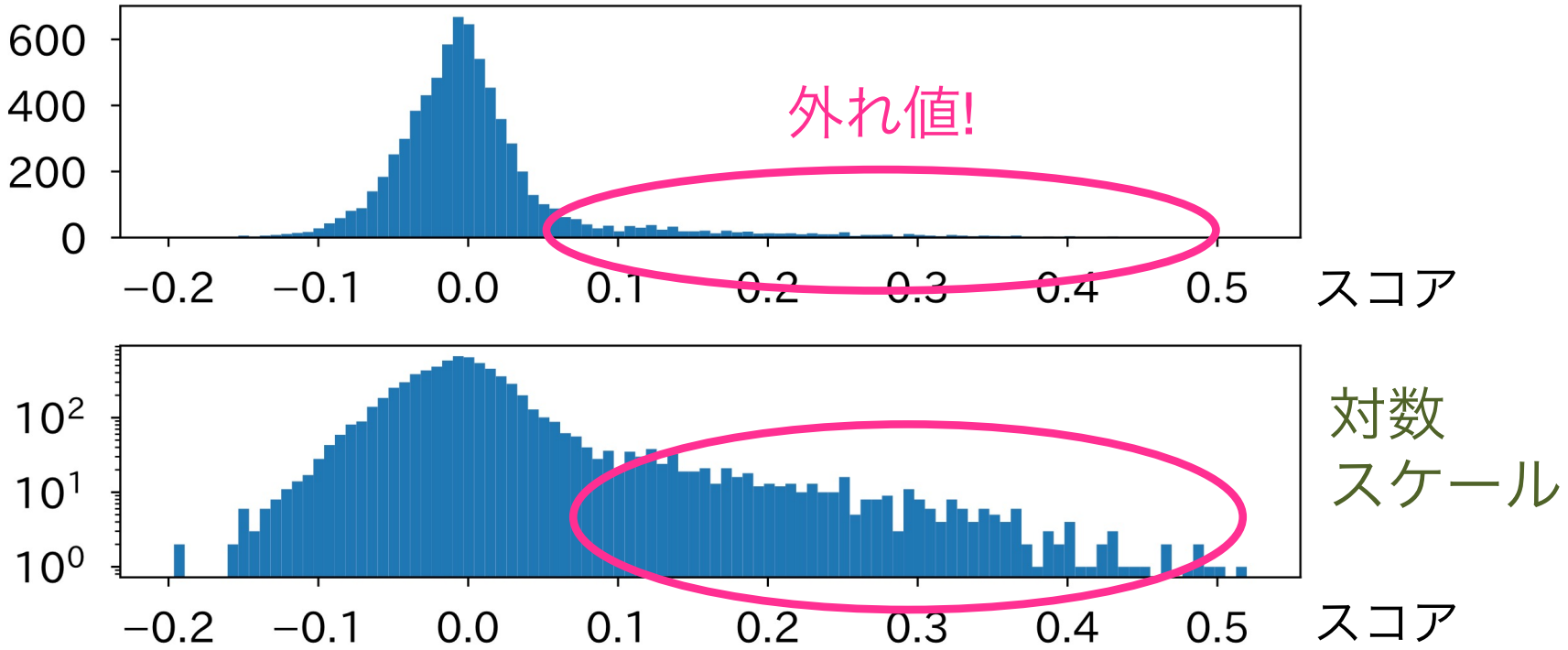
Doc2Vecより高性能

- 線形代数で解析的に計算できるニューラル文書ベクトルから、検索語と各文書のコサイン類似度を計算
- 例: Livedoorニュースコーパスを、“恋愛 結婚”で各文書のコサイン類似度を計算

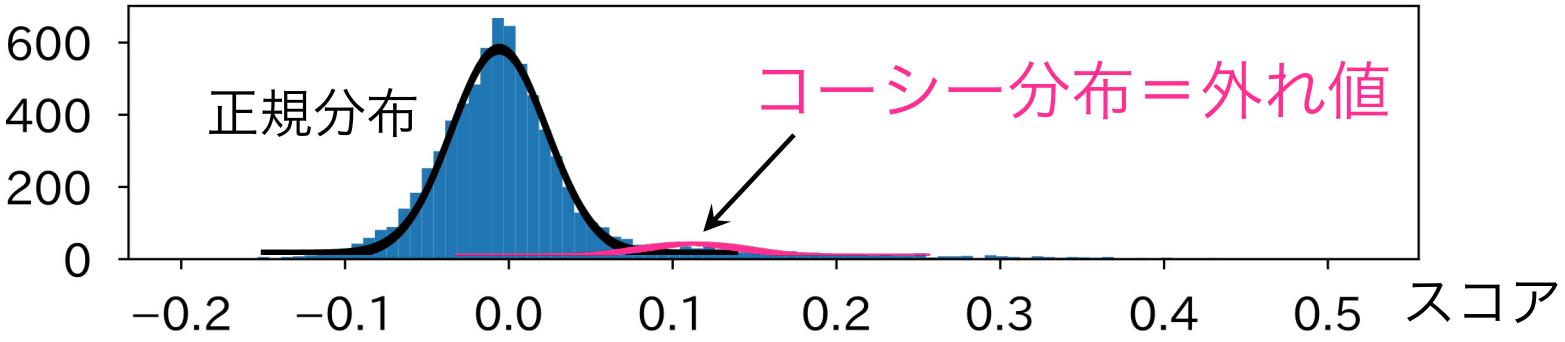


テキストの統計的抽出 (4)

- スコアをプロットすると、右端に外れ値が存在
  →対数スケールだと明らか



テキストの統計的抽出 (5)



- 外れ値を、正規-コーシー混合分布としてモデル化
  → コーシー分布の確率が高いテキストが外れ値
  → 検索したキーワードに関連するテキスト
- 正規-コーシー混合モデルは、EMアルゴリズムで高速に推定できる
  - コーシー分布は解析解がないため、BFGSで最適化
  - 詳細は論文を参照のこと

テキストの統計的抽出 (6)

- この基準で抽出すると、閾値を人間が設定する必要がなく、高精度に関連するテキストだけを抽出できる

文書番号	スニペット
48	既婚女性の話に恋愛を学ぶ独身女性ばかりの職場に勤務するサオリさん（28 歳
50	独身男性は独女より人妻と遊びたいってホント？「結婚をしたら独身男性から
86	モテる男が選ぶ女子の条件「素敵だと思う人には、もうすでに奥さんがいる」
99	人に聞いてはいけないことお盆に帰省した沙織さん（30 歳）に「実家は楽しか
116	新しい結婚の形？「事実婚」とは実際どんな制度なのか最近メディア等で「事
156	恋の駆け引きができる女、できない女「追いかけられると逃げたくなり、逃げ
181	2011 年こそ結婚したい！独女・独男の婚活事情「婚活」という言葉が流行語大
186	仕事、遠距離恋愛……。会えない時間で愛は育つか？俳優の向井理さんとモデル
239	アラフォーだって結婚式したい！ Presentedby ゆるっと cafe 独女の皆さま、はじ
298	意外に大変?!モテる定説“マメな男”との交際事情メールや電話で連絡を怠らず
309	独女的映画レビュー vol.7『』“人を好きになる気持ち”ってどんな気持ちだっけ
314	あなたはいくつまで恋ができると思いますか？今年 4 月、2 週に渡って朝日新聞
316	今さらながら、運命の出会いについて考えてみる「もしあの時……」と選ばなか
395	ケンカをしても気持ちが冷めている…。「長すぎた春」の予感先日、28 歳の誕生
448	独女通信が見た「独女たちの 5 年の軌跡」Vol.5〜婚活とは、人生最大の営業活
468	恋愛感情がイマイチでも結婚はできるのか？婚活ブームの影響で、寸前と言わ

Livedoorから抽出された262件のテキストから、ランダムに選んだもの

まとめ

- 確率的潜在意味スケーリング(PLSS):
  テキストを任意の潜在的な軸で連続的に測定する
  - 項目反応理論をベースにした、LSSの確率拡張
  - LSIからPLSI (Hofmann 1999)への拡張と同じ方向
- 極性辞書または少数の代表テキストから、測定軸を推定
- 公開されている尺度データでLSSより高性能、
  国会議事録など、様々な生テキストで実験
- 関連して、キーワードに関係する文書を統計的に選択するための正規-コーシー混合モデルを提案

Future Work

- 今回はユニグラムの、非常に限定された特徴しか使っていない
  - バイグラムなど、単語の組み合わせの利用
  - 品詞などの利用 (形容詞+名詞など)
  - 反転表現への対応 (not good など)
  - 有効な特徴を自動的に抜き出す方法はあるか?
- Benoit+ (2016)でのクラウドソーシングされた評価データでの実験
- Rパッケージは広い方々に使っていただけるよう、github で公開予定です