Token 化

テキスト分析

川田恵介

テキストデータ

実例

- 日本 & 経済学におけるコロナ研究リスト

```
# A tibble: 120 x 2
   Title                                                                Type
   <chr>                                                                <chr>
 1 新型コロナ対策としてのマスク着用義務化――アメリカの政策評価と日本への示~ 一般
 2 現状放置なら５月下旬に緊急事態宣言の恐れ                            一般
 3 接触確認アプリ COCOA を行動経済学で読み解く                         一般
 4 新型コロナ対策がもたらす効果の定量的分析〜緊急事態宣言解除後のシナリオ~ 一般
 5 新型コロナ危機と経済政策                                            一般
 6 感染症データにおける「従属性」と因果推論                            一般
 7「疫学」と「マクロ経済学」の視点から ―最新論文に見る感染症対策と経済活~ 一般
 8 コロナ危機は需要ショックなのか供給ショックなのか？                  一般
 9 コロナ危機の経済学 ： 提言と分析                                    一般
10 新型コロナウイルスのマクロ経済学（1）感染症拡大防止政策のトレードオフ 一般
# i 110 more rows
```

テキスト分析

- 事例の多くは、”テキスト”で記録されてきた
  - 膨大な電子化されたテキストデータがより容易に入手可能（SNS, オンラインアンケート、口コミ評価）
- テキスト分析への需要拡大
  - 統計・機械学習的手法の応用が急速に進む

例：自由記述欄

- 多くの調査には、自由記述欄が含まれる
  - 例：商品への感想を書いてください
- Open end question（回答結果を制約しない）
  - いろんな回答結果が記録できる
  - 分析者が想定していない情報を得られる可能性

例: 武蔵大学の印象

```
# A tibble: 3 x 2
     ID Text
  <dbl> <chr>
1     1 キャンパスが綺麗
2     2 池袋から近い
3     3 キャンパスがおしゃれ
```

本講義での応用

- 特定のグループが使いがちな単語とは？
- 予測モデル $E_P[Y|\text{テキスト}] \sim g(\text{テキスト})$ の構築
  - 機械学習に比較優位

前処理

復習: 母平均の推定問題

- 基本: “似ている” 複数の事例を集計して、$E_P[Y|X]$ を推定したい
- 予測変数 X の役割 $=$ 似ている事例かどうかの判断基準
  - 統計・機械学習 $=$ データから似ている度を判定する
- 伝統的な X : カテゴリー/連続変数
  - テキストは大きく異なる特徴を持つ

VS カテゴリー変数

- 性別、国籍、学部など
- カテゴリー変数 = “少数” の値しかとらない
  - 同じ値をとるサンプルが複数存在
  - “値が同じであれば似ている”
- テキスト変数 = 無限大の種類がある

VS 連続変数

- 年齢、身長など
- 連続変数 = 同じ値をとるサンプルは極めて少数
  - 値が近いかどうかは自明
- テキスト変数
  - テキストが近いかどうかは自明ではない

テキスト変数の難しさ

- テキストは情報が” 豊富” すぎる
- 変数の値間の距離が定義できない
  - 似ている文章とは？
- $\rightarrow$ 何らかの単純化が必要

テキストデータを使った予測モデル構築

- 予測モデル $g(X, Text)$ を構築
- 合わせ技でモデル構築
  - 事前に $Text$ を単純化する (前処理)
  - 大量の X に対応した手法で推計

前処理: Token

- テキストを単語の羅列（Token 化）
- 日本語は単語化が難しい
  - 分かち書きをしない
  - quanteda パッケージを用いれば解決可能

例: 武蔵大学の印象

```
Tokens consisting of 3 documents and 1 docvar.
text1 :
[1] "キャンパス" "が"          "綺麗"

text2 :
[1] "池袋" "から" "近い"

text3 :
[1] "キャンパス" "が"          "おしゃれ"
```

前処理：Bag of words

- Token 化しただけでは、依然として、全ての事例が異なる値を有する
  - さらなる単純化が必須
- 代表的な手法は、Bag of words
  - 単語の出現頻度を数える
- 文脈や語順は捨象
  - 発展: N-gram, embedding

例: 武蔵大学の印象

```
Document-feature matrix of: 3 documents, 7 features (57.14% sparse) and 1 docvar.
       features
docs    キャンパス が 綺麗 池袋 から 近い おしゃれ
  text1          1  1    1    0    0    0        0
```

```
text2          0  0    0    1    1    1        0
text3          1  1    0    0    0    0        1
```

頻度分析

- どのような単語が使われているか
  - グループごとに集計も可能
  - テキスト分析版、記述統計分析

全体

```
    の  コロナ      と    新型    感染      に      を    経済
   167      96      64      55      39      31      29      29
    症      :   ・ウイルス       —    分析    禍    影響
    27      27      25      24      23      22      22      21
における      へ      で    企業
    19      16      15      15
```

- よくわからない

グループの特徴づけ

- 単純に集計すると助詞や助動詞など、“文章を特徴づける上で、そこまで重要ではない単語” が上位に来る
- グループ (例: 満足度が高い VS 低い, **一般むけ** VS **専門むけ**) を特徴づける単語の探索
- chi2 指標: 単語がグループ間で偏りなく使用される場合に比べて、分布がどの程度偏っているのか？

一般

```
   feature     chi2           p n_target n_reference
1       「 7.318962 0.006823080        8           1
2       」7.318962 0.006823080        8           1
3       学 7.318962 0.006823080        8           1
4       で 7.302326 0.006886539       11           4
5        2 5.381745 0.020348513        5           0
6      pos 5.381745 0.020348513        5           0
7     みる 5.381745 0.020348513        5           0
8     経済 4.567783 0.032578695       17          12
```

9	**格差**	3.888633	0.048613979	4	0
10	**編**	3.888633	0.048613979	4	0

専門

	feature	chi2	p	n_target	n_reference
1	**における**	6.690888	0.009690694	17	2
2	**調査**	3.719224	0.053789379	12	2
3	**た**	3.692958	0.054642633	8	0
4	**データ**	3.070639	0.079718138	10	1
5	**感染**	2.081123	0.149130465	28	11
6	**い**	1.812277	0.178235280	5	0
7	**による**	1.812277	0.178235280	5	0
8	**関係**	1.812277	0.178235280	5	0
9	**行動**	1.739370	0.187218660	10	2
10	**ウィルス**	1.431780	0.231474161	7	1

OLS による予測モデル

- $E_P[Y|X, \text{テキスト}] \sim g(X, \text{テキスト})$ を推定する
- 代表的な推定方法が、一般に機能しない
  - テキスト分析が難しかった理由

OLS の前提条件

- $Y_i \sim \beta_0 + .. + \beta_L X_L$
  - 経験則として、事例数 $> 3\times$ 変数数であれば、ある程度の推定精度を期待できる
- 変数数 $>$ 事例数であれば、原理的に推定できない
  - テキスト分析ではしばしば発生する

実例

- $Y = 1$ 一般むけの文章, $Y = 0$ 専門家向けの文章
- $X =$ 論文のタイトル

実例

```
Call:
SuperLearner(Y = Y, X = X, SL.library = list("SL.lm", "SL.mean", "SL.ranger",
    "SL.glmnet"))

                     Risk       Coef
SL.lm_All     209.6583759 0.00315369
SL.mean_All     0.2482596 0.00000000
SL.ranger_All   0.1943378 0.99684631
SL.glmnet_All   0.2266339 0.00000000
```

まとめ

- カジュアルなグループの特徴づけの手法がたくさんある
  - 単純な集計に比べれば、有意な情報が多い
- OLS は予測モデル構築の役に立たない