前処理

テキスト分析

川田恵介

前処理: 発展

- tokens 化しただけでは、一般に不十分
  - 新語への対応
  - 大量のあまり意味のない単語 (助詞や助動詞) が含まれる
- 文章の特徴把握や予測モデル構築を困難にする
- 対応策
  - 辞書の更新
  - 統計的基準の導入

新語への対応

- 常に新しい単語が出現する
  - 辞書だけに頼った前処理では、対応が不十分になってしまう
  - 分かち書きをしていない日本語では特に深刻
- 例: 新型コロナ感染症 → 新型/ コロナ/ 感染症

共起語の接続

- 文章中に連続して出現しやすい単語
  - 一つの単語として接続する
  - 統計的に判定可能

	collocation	count	count_nested	length	lambda	z
1	ウイルス 感染	12	0	2	4.195249	8.775305
2	購買 動向	5	0	2	7.840313	6.367479
3	新型 コロナ	55	0	2	9.155246	6.308722
4	ウィルス 感染	6	0	2	4.766117	6.156754
5	対応 緊急	3	0	2	5.691806	6.144400
6	症 拡大	5	0	2	3.542957	5.883316
7	感染 症	27	0	2	8.396832	5.770567
8	禍 購買	5	0	2	5.368066	5.637789
9	データ 用	3	0	2	5.775035	5.539923
10	事態 宣言	7	0	2	9.248118	5.525857
11	緊急 事態	8	0	2	8.523090	5.496678
12	2 〜	2	0	2	5.693732	5.461940
13	コロナ ショック	9	0	2	3.166398	5.427353
14	勤務 生産	2	0	2	6.542952	5.400897
15	用 実証	2	0	2	6.542952	5.400897
16	サービス 業	2	0	2	7.054737	5.356976
17	ワクチン 接種	6	0	2	8.256756	5.285278
18	感染 拡大	5	0	2	3.057141	5.203793
19	緊急 調査	3	0	2	3.602310	5.139375
20	マクロ 経済	4	0	2	4.867698	5.041003

Stop words の除去

- 文章に大量に出現する、「意味のない」、単語
  - 一般に除去することが望ましい
- 方法
  - 辞書の活用
  - 統計的基準を導入 (1 文字言葉の除去など)

Rare words の除去

- 文章に一度しか出てこない単語は、予測に活用不可能
  - 除去すべき
- 非常にまれ/ほとんど出てくる単語についても?

例: WordCloud

コロナ	新型	感染	経済	症	:	・	ウイルス
96	55	39	29	27	27	25	24
—	分析	禍	影響	企業	調査	ショック	拡大
23	22	22	21	15	14	13	13
行動	緊急	的	危機				
12	11	11	11				

例: WordCloud

新型コロナ	コロナ	経済	分析	影響
53	34	25	22	21
感染	企業	行動	ウイルス	危機
15	15	12	12	11
データ	covid-19	調査	政策	ウイルス感染
11	11	11	9	9
日本	考察	雇用	緊急事態宣言	コロナショック
8	8	8	7	7

まとめ

- テキスト変数の前処理は、まだまだ研究が続いている
  - 例えば表記ブレをどのように対処するか?
- 推定方法の面での対応も必須