

@関西学院大学神学部
2021.01.18

講演会／ワークショップ
「テキストを対象とした認知人類学・宗教学研究」

Rを用いたテキスト分析 実習

①計量テキスト分析/テキストマイニング入門

明治大学 研究・知財戦略機構
佐藤浩輔

位置づけ

- 2020/12/28
「テキスト分析のためのR入門」
 - 2021/1/18
「テキストを対象とした認知人類学・宗教研究の実例」*
 - 研究発表：民話の分析
 - 認知人類学・宗教研究におけるテキスト分析の動向
- 「Rを用いたテキスト分析実習」←このワークショップ
- 計量テキスト分析/テキストマイニング 入門
 - RとMeCabを用いた日本語テキスト解析実習

*中分通先生（高知工科大学）担当

構成

- 計量テキスト分析/テキストマイニング入門
 - テキストマイニング(計量テキスト分析)と自然言語処理の概要を学び、何ができるかを知る
- RとMeCabを用いた日本語テキスト解析実習
 - データ分析の実演：前処理から分析まで
 - ①前処理を行い、分析ができるようデータを加工する
 - ②分析を行い、結果を解釈する

計量テキスト分析/テキストマイニングとは何か
それを用いていったい何ができるか

計量テキスト分析とテキストマイニング

- 計量テキスト分析 quantitative text analysis
 - テキストデータを量的 quantitative な手法を用いて分析すること
 - 社会科学の内容分析 content analysisの流れをくむ(樋口, 2006, 2014)
- テキストマイニング text mining
 - 大量のテキストデータから（機械を用いて）価値のある情報を取り出すこと
 - 工学、マーケティングの流れをくむ
 - データマイニング:
 - mining: *Mining is the industry and activities connected with getting valuable or useful minerals from the ground, for example coal, diamonds, or gold.* -Collins COBUILD English dictionary
 - 大量のデータの中から価値のある情報を取り出す技術
 - cf. Webマイニング: 大量のWebデータの中から
 - 探索的な手法 というニュアンス
 - 価値のある情報が埋まっているとは限らない

自然言語処理

- テキストマイニング/計量テキスト分析
→自然言語処理技術を用いてテキストデータから情報を抽出
- 自然言語処理(Natural Language Processing: NLP)
 - 構造化されていない自然言語を扱うための技術
 - 自然言語：普通の人が使うような言葉や文章
 - vs. 形式言語：人工的に作られた言葉(e.g. プログラム言語)
 - 自然言語を処理して様々な情報を抜き出したり生成したりする
→処理の自動化・効率化

機械を使うと何が嬉しいか

- 動物民話論文の査読コメント（一部）：
 - "Since the corpus is relatively **small** (with less than 400 very short 'stories'), I honestly believe it would be more effective to manually extract all animal references from the summaries."
 - 「本研究のコーパスは比較的小さいので、動物の名前については要約の文章から手動で抽出した方が効果的だと思います」

機械を使うと何が嬉しいか

- 処理が効率的になる
 - 大量に処理できる
- 手続きが明瞭になる
 - 機械は言われたことしかやらないので明示的に指示を与える必要がある→**ブラックボックスな領域が減る**
 - 同じデータに同じ手続きを適用すれば、同じ結果が得られるはず→**検証可能である**
- 質的なものを量的に扱える
 - 量的に分析することで、質的な分析では見えてこないものを発見できる→**質的な分析と相互に補完しあえる**
- 分析の幅が広がる
 - 知りたいことはひとつの事柄だけではない
 - 動物の共起頻度「だけ」が知りたいのであれば、確かに手動でもOK

人文学分野における応用例

- デジタル人文学 digital humanities: 情報処理の技術を入文学の研究に応用
 - 文学
 - Distant reading (Moretti, 2013)
 - 精読 close reading に対して、(情報処理技術を使って)大量の文献を扱う
 - 計量文体学 stylometry / stylistics (村上, 2002)
 - 文体を量的に扱う
 - 言語学
 - 計算言語学 computational linguistics
 - 歴史学
 - Digital history
 - 民俗学/民話学
 - 計算民話学 computational folkloristics (Abello et al. 2012; Tangherlini, 2016)
 - 民話の自動タグ付けやデータベースに活用

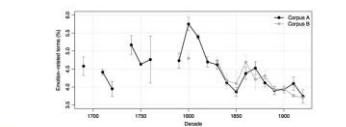
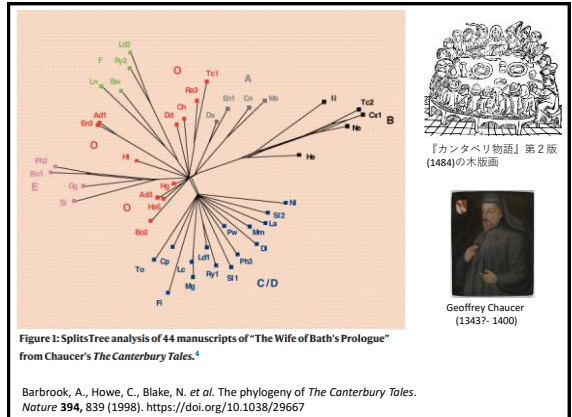


Figure 2. Emotionality changes in Anglophone literature for the two "small data" corpora. Error bars represent 95% confidence intervals.

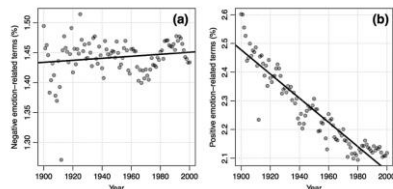


Figure 3. Emotionality changes in Anglophone literature for the Google Books corpus. (a) negative emotion-related terms; (b) positive emotion-related terms. Solid lines represent linear regressions of the data.

Morin, O., & Acerbi, A. (2017). Birth of the cool: a two-centuries decline in emotional expression in Anglophone fiction. *Cognition and Emotion*, 31(8), 1663–1675.

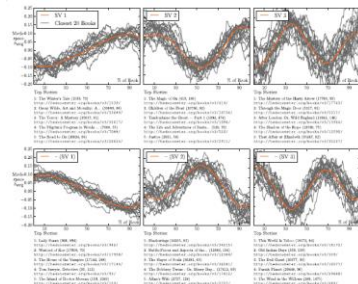


Figure 4 First 3 SVD modes and their negation with the closest stories to each. To locate the emotional arcs on the same scale as the modes, we show the modes directly from the rows of V^T and weight the emotional arcs by the inverse of their coefficient in W for the particular mode. The closest stories shown for each mode are those stories with emotional arcs which have the greatest coefficient in W . In parentheses for each story is the Project Gutenberg ID and the number of downloads from the Project Gutenberg website, respectively. Links below each story point to an interactive visualization on <http://phenomenonem.org> which enables detailed exploration of the emotional arc for the story.

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPI Data Science*, 5(1), 31.

小まとめ

- 計量テキスト分析・テキストマイニング
 - 自然言語処理技術を用いて、テキストから価値のある情報を抽出することができる
 - 質的なものを量的に扱える
 - コンピュータを用いて、高速に大量に情報を処理することができる
 - 人文・社会科学の様々な領域に広がっており、新たな研究の可能性を示すものである

テキストをどう分析するか

分析までの概観

- 分析対象となるコーパスを選ぶ
 - コーパス(corpus; pl. corpora)
 - 「言語資源」
 - あるテーマに沿って集められた文書の集合
- 自然言語（質的データ）を量的なデータに変換する
 - 頻度
 - 分布
 - 各種指標・統計量

基本は「数える」こと

どういう分析をするか

頻度

- 文書や文を単位に頻度を算出する
 - 文字
 - 単語
 - トークン token : ひとつひとつの単語の出現「延べ語数」
 - タイプ type : 単語の種類「異なり語数」
 - 共起
 - 単語同士が文や文書に同時に登場する回数
 - n-gram
 - 連続するn個の単語
 - 機械学習によるタグ付け
 - 感情分析などによる「感情」の判定
 - 各種分類器による判定

トークン vs. タイプ

Sanctus, Sanctus, Sanctus
Dominus Deus Sabaoth.
(from Sanctus)

単語"sanctus"(羅:「聖なる」)はこの文に、

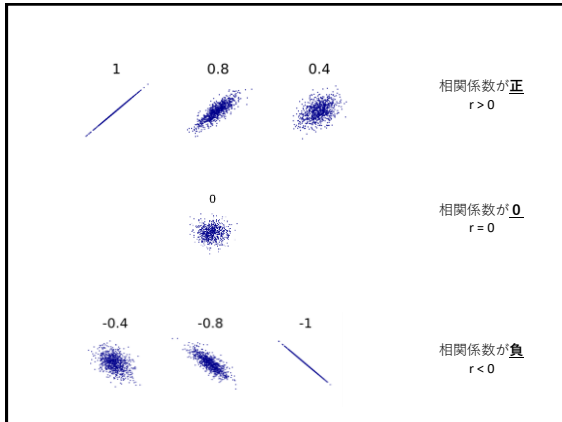
トークンとして: 3回出現
タイプとして: 1回出現

どちらを使うかは分析の目的による

どういう分析をするか

分布

- 各種要素の頻度の分布
 - 長さ
 - 単語の長さ
 - 文の長さ
 - 単語の種類
 - 品詞
 - 識別語
 - 機能語
 - その他
 - 語彙・漢字・仮名・読点・文節・音韻・文頭文字



どう分析をするか

③ 比べる

- 頻度データ
 - 分布の比較： χ^2 乗検定
 - 頻度の比較：尤度比検定
- 数値データ：
 - 各種パラメトリック・ノンパラメトリック検定
- 注意すべきこと
 - 検定を行う必要/必然性があるか
 - 何と何を比較しているか（意味のある比較なのか）
 - 記述統計で十分な場合もある

比べるのは難しい

“Why is a raven like a writing-desk?”

「カラスと書き物机が似ているのはなぜ？」





Lewis Carroll (1876)
Alice's Adventures in Wonderland

似ていないものを比較する？

e.g. ナイフ・フォーク文化と箸文化：ある不適切な比較の例

「食器の違いが社会制度にもたらす影響を調べるため、ナイフ・フォークを使う文化、箸を使う文化の国をそれぞれ一国ずつ選び出し、比較した」

	ナイフ・フォーク文化	箸文化
行政府の長	大統領	首相
政党	二大政党制	多党制
公的医療保障制度	一部のみ	手厚い支援
雇用	ジョブ型	メンバーシップ型
価値観	個人主義的	集団主義的
...
		

比較の難しさ

- 比較しているのは代表的な例ではないかもしれない
 - 適切な比較にならない
 - 対策
 - 何らかの外的な基準に基づいて代表的な例を選ぶ
 - 複数の例をサンプリングする
 - e.g. 箸文化である国を複数持ってきて比較する
- 違いをもたらししている原因は比較の軸ではない
第三の変数かもしれない（**交絡 confounding**という）
 - 相関関係は因果関係を意味しない
 - 対策
 - 比較したい変数以外の要因をなるべく揃える
 - e.g. 同じコーパスを分割する, 同じ位置づけのコーパスを比較する

どう分析をするか

④ まとめる

- 似たような性質を持つデータ（＝行）をまとめたい
 - クラスター分析
 - データ間の「距離」または「類似度」をもとにデータの塊（クラスター）を抽出する
- 似たような性質を持つ変数（＝列）をまとめたい
 - 次元削減：複数の変数（次元）をデータの性質を保ったまま少ない変数で表現する
 - 主成分分析
 - 複数の変数を数個の「主成分」に合成する
 - 主成分：データをよく説明する合成スコア
 - 因子分析
 - 複数の変数をいくつかの「因子」に分解する
 - 因子：観測変数の背後にある潜在的な変数(e.g. 「知能」)

どういふ分析をするか

小まとめ

- テキストを分析する選択肢は多い
 - 何を指標にするか、どう比較するか
 - 事前にすべてを理解しておく必要はない
 - 「こういう技術がある」として知っておくとい
- テキスト：大量の変数, 大量の分析
 - 自由度が高すぎるために何をしたいかわからない
 - 統計的な検定だけにこだわらず、データ可視化手法などといった記述的な手法も視野に入れる
 - 比較の際はデータの持つ制約を理解する

一般的な注意点

- ・ 計量テキスト分析はあくまで「テキスト」を分析しているに過ぎない
 - ・ 「このテキストはこれこれこういう性質を持っている」ということしか基本的にはいえない
- ・ テキストが、その背後の「心理」や「行動」、「実態」のようなものをストレートにとらえている保証はない
 - ・ ランダムに生成した文章に対しても同じ分析ができる
 - ・ 編纂者のバイアス/サンプリングのバイアス
 - ・ 態度の行動を予測しない(LaPiere, 1934)
- ・ 分析した結果がどのような意味を持っているかは他の証拠も踏まえた上で慎重に検討する必要がある

テキストの罫

- 人間の行動を反映しないデータが紛れ込んでいる場合がある
 - Back, Küfner, & Egloff(2010): 9.11後のSNS上のメッセージを分析
 - 「9.11後に怒りの感情がSNS上で増加している」
 - Pury(2011):「Backらの結果は誤り」
 - Backらの結果はBotの仕業
 - Botの投稿を除き取くとBackらの結果は再現されない
 - 人工の結果(artifact)
 - その後、Backら自身の再集計後の分析でも結果は再現されず(Back, Küfner, & Egloff, 2011)

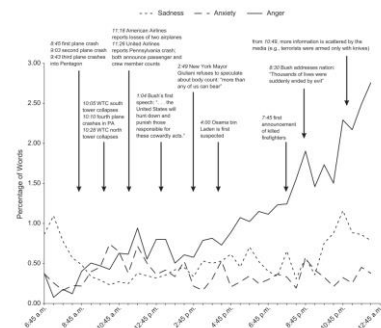
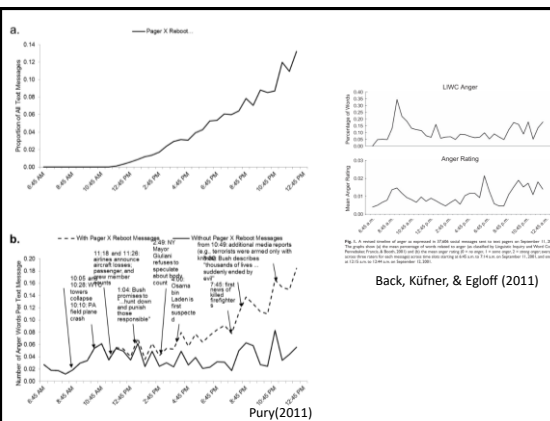


Fig. 1. The timeline of sadness, anxiety, and anger on September 11 as expressed in messages sent to text pagers. Each data point represents the mean percentage of words related to the specific negative emotion, averaged across 30 min. The time does start at 6:45 a.m. to 7:14 a.m. on September 11, 2001, and end at 12:15 a.m. to 12:44 a.m. on September 12, 2001. Exact times and brief descriptions of the most important events of September 11 are included above the timeline. WTC = World Trade Center.

Back, Küfner, & Egloff(2010)



前半まとめ

- ・ 計量テキスト分析/テキストマイニングは、機械（コンピュータ）を用いてテキストデータを処理し量的に解析する
 - ・ テキストデータの分析手法は多様である
 - ・ 自由度が高い分、定型的な手法というものがない
 - ・ 数え方や指標の選び方に任意性がある
 - ・ 都度、正当化が必要
 - ・ テキストデータには固有の特徴がある
 - ・ 必ずしも現実世界のことを反映しているとは限らない
 - ・ 落とし穴にはまらないよう注意する