題目

第6回日本語の歴史的典籍国際研究集会（20201107）

人文学における探索的データ分析

報告者

千葉大学　人文社会科学系教育研究機構　助教

小風尚樹

今回の報告の流れ

1. 自己紹介
2. 探索的データ分析
3. テキストマイニング
4. 考察

1. 自己紹介

◆略歴

- 2014～　東京大学西洋史学修士・博士課程
- 2017～　Tokyo Digital History代表
- 2018～19 King’s College London, MA in Digital Humanities
- 2020～　千葉大学人文社会科学系教育研究機構 助教
  （人社系卓越大学院プログラムにおけるDH研究・教育）

◆主要な研究関心

- 20世紀転換期イギリスの海事史（とくに海軍と海域経済の関わり）
- 歴史研究におけるDHの実践（とくにTEIに準拠したテキストの構造化や分析）

Cf. https://researchmap.jp/naoki_kokaze/

2. 探索的データ分析

2.1. 大量のデータがもたらす可能性

- 現在、広義のデジタルアーカイブの普及に伴い、活用可能なデータは飛躍的に増加している（※とくに歴史（西洋史）研究において）
  - 史料の選択に関して、議論できることが増えてきた
  - 大量のデータのどこに注目するか、その妥当性をどう説明すればいいのか

cf. Lemercier & Zalc, 2019, *Quantitative Methods in the Humanities: An Introduction*, pp. 16-47

20世紀後半における数量的歴史研究の実践者による指摘：

「明確には定義されていない母集団から非明示的に選択された事例を用いて、「しばしば」や「一般に」などの副詞を使って歴史を語ることを警戒」

⇒「伝統的」な質的叙述：史料解題、文脈、研究関心、先行研究との比較…に依拠
⇒この説明を数量的な論拠で補完することができるように

- データの大まかな内容を掴み、その中から注目に値する箇所がなぜ注目に値するかを説明する良い手立てはないか？

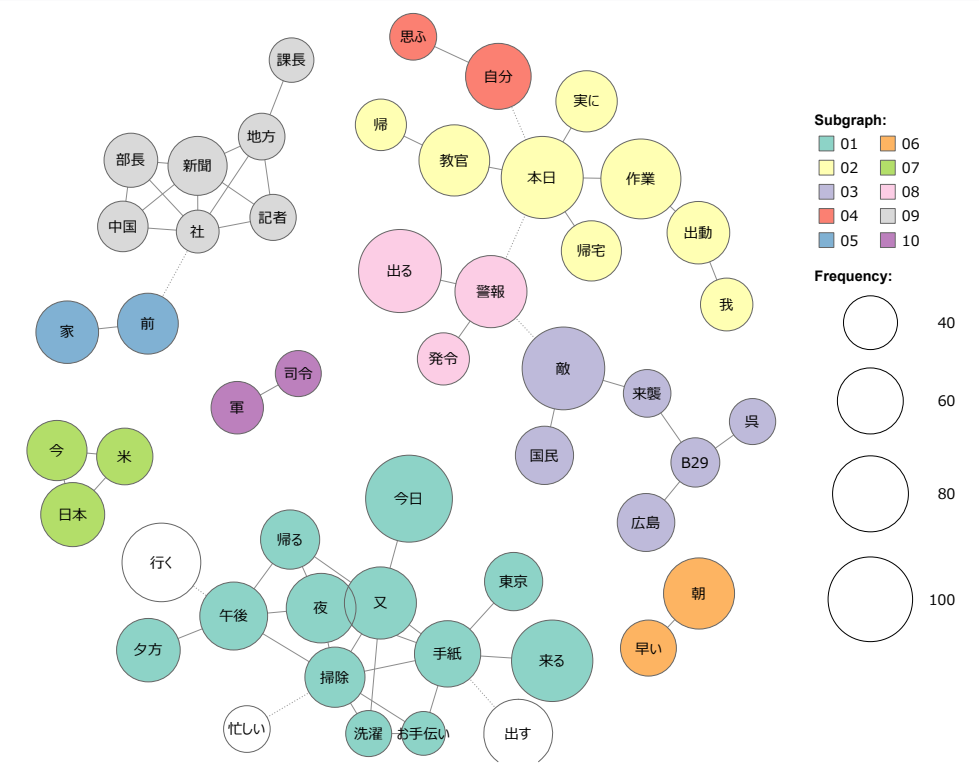
2.2. データサイエンスにおけるEDA

- Practical Data Science with R and Python「探索的データ分析」より引用

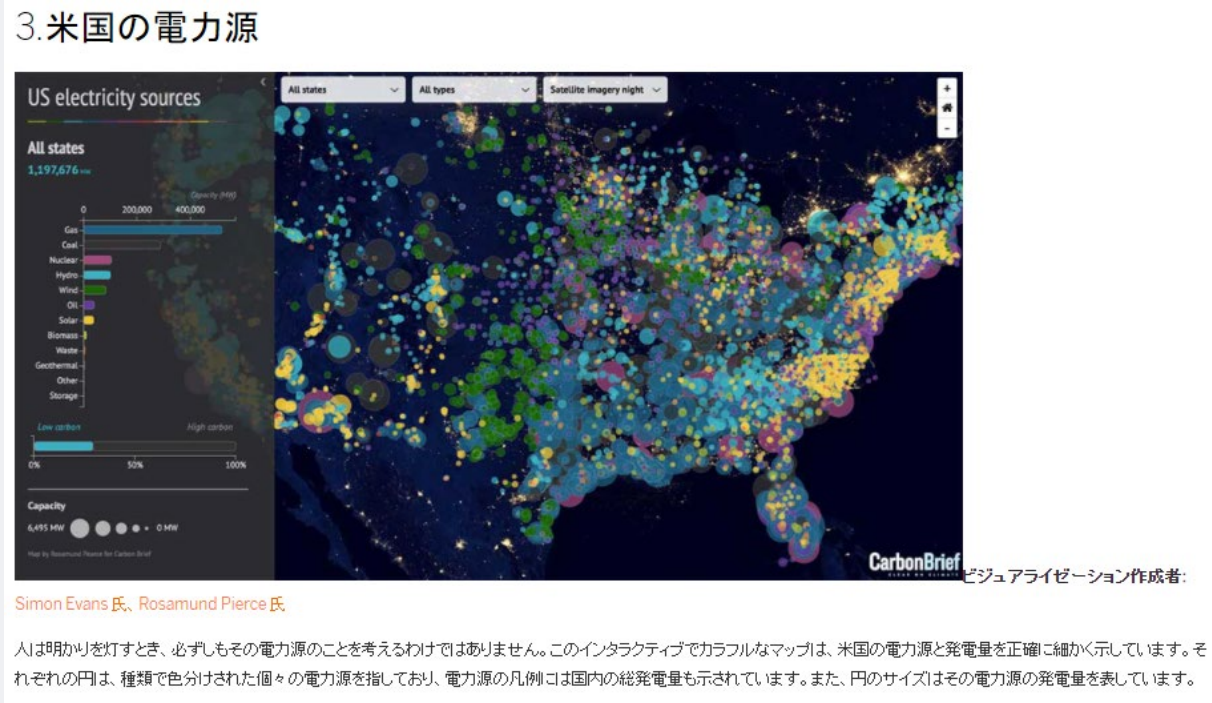
- データを料理する前に、どのようなデータが与えられているのか確認することが大切です。
  この段階を踏むことで、データに対する理解が深まり、より良いモデルの構築に繋がる可能性もあります。こうした一連の作業は探索的データ分析 (Exploratory Data Analysis: EDA)と呼ばれます。
  この作業には、データの集計、要約、可視化が含まれます。
- EDAがデータ分析の作業において早期段階で行われるのは、データの異常（思い込みとの比較を含めて）や特徴を把握するためです。これらは分析全体のアプローチや良い出発点を見つけるために有効です。出発点と表現したのは、モデルの構築や特徴量の生成によって改めてデータを見つめ直す作業が発生するためです。そのため必ずしも徹底的である必要はありません。
- まずは手元のデータを眺め、簡単な集計をしてみましょう。続いてデータをグラフによって表現してみましょう。データを要約、図示することで、個々の値からは見えなかった情報やデータ間の関係を把握できます。特に欠損値や異常値（外れ値）、データの分布などデータ全体あるいはデータ間の関係性やそのばらつきを見るのに可視化は重要です。

2.3. DHにおけるさまざまな「探索」

テキストマイニング/ネットワーク分析



GIS分析



ビジュアライゼーション作成者:
Simon Evans 氏、Rosamund Pierce 氏

人は明かりを灯すとき、必ずしもその電力源のことを考えるわけではありません。このインタラクティブでカラフルなマップは、米国の電力源と発電量を正確に細かく示しています。それぞれの円は、種類で色分けされた個々の電力源を指しており、電力源の凡例には国内の総発電量も示されています。また、円のサイズはその電力源の発電量を表しています。

2.4. 探索的データ分析の意義

- 先行研究との差異、自分の問題関心といった「バイアス/予断」から距離を置き、扱おうとする事例の一般性・例外性を把握/説明する際に数量的手法が役立つ
  - テキストマイニング・GIS・統計学的アプローチ・ネットワーク分析はその一部

- 本日は、テキストマイニングに注目

3. テキストマイニング

3.1. テキストマイニングの概略

(cf. Stefan Sinclair and Geoffrey Rockwell, 'Text Analysis and Visualization: Making Meaning Count', in *A New Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 2nd ed., 274–90. Chichester, West Sussex, UK: Wiley-Blackwell, 2016)

- テキストコーパスを詳しく読む前に、コーパスの特徴をつかもうとする手法
  - 統計分析の結果やそのデータを可視化した図表の力を借りて
  - つまり、情報を圧縮した上で、何らかの観点で注目に値する点を浮き彫りにする

- 向き・不向きはあれど、テキストを分析するきっかけを増やす
  - 語りのニュアンス・事実の詳細など、ナラティブに関心がある場合には不向き
  - 言語学的・意味論的な特徴を計量的に知ろうとするのに向いている

- 二つの原則を抑えておくと良い
  - ツールに期待しすぎない⇒計算(ツール)と解釈(人間)の分業
  - 様々な点から試行錯誤を繰り返す⇒結果を得るためではなく、問いを立てるため

3.2. イギリス史の事例(Hitchcock & Turkel, 2016)

- イギリス法制史における法廷実務の変遷の画期について
- 【先行研究】
  先行研究における統計的サンプリングと「選別された」精読に基づいて、18世紀最後の四半世紀と結論

- 【Hitchcock & Turkel, 2016】
  包括的な単語の集計による、テキストが持つ属性と相対頻度に着目したテキストマイニング研究
  ∵答弁取引の増加と裁判数の減少
  →1800～1860年に再考

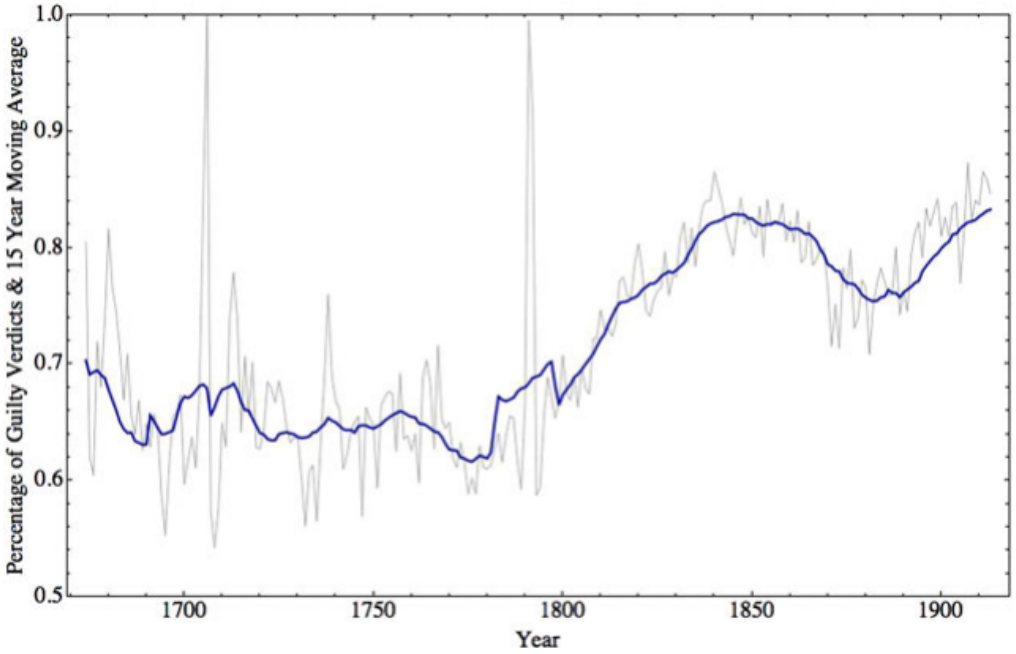


Figure 8. Percentage of trials resulting in a "guilty" verdict. Between October 1792 and December 1793, trials resulting in an acquittal on all charges were excluded from the *Proceedings*. This exclusion has a marked and misleading impact on the moving average between 1777 and 1808; the apparently similar spike in convictions in 1706 results from a whole issue of the *Proceedings* being given over to a single trial, which was judged "guilty." See s17061206.

Cf.

- **Tim Hitchcock and William J. Turkel**, 2016, 'The Old Bailey Proceedings, 1674–1913: Text Mining for Evidence of Court Behavior', *Law and History Review* 34 (4): 929–55. https://doi.org/10.1017/S0738248016000304

3.3. KH Coder概説

- 2001年に樋口耕一が開発した計量テキスト分析ツール
  - 日本語分析に力を入れている
- 充実した解説書（cf. 樋口, 2020）
- 2017年時点で、日本国内における研究導入実績は1500件以上

- KH Coderを導入する2つの目的
  - **データ探索**：機械的に言葉を数えることによるセレンディピティ（偶然の発見）
  - **分析の信頼性向上**：第三者に分析の過程や妥当性を示し、検証・再現可能性を担保

- 従来の「精読」を否定しない、むしろ相互補完的に読解を深めるツール
  - 欧米圏のDHでは、Voyant Toolsがテキスト解析ツールとしては有名

3.4. KH Coderの解説

- 樋口耕一『社会調査のための計量テキスト分析—内容分析の継承と発展を目指して【第2版】KH Coder オフィシャルブック』ナカニシヤ出版、2020年

- 河野武司「計量テキスト分析のすすめ」2014年12月
  https://www.youtube.com/watch?v=kGpimrzIILA

- KH Coderチュートリアル
  https://khcoder.net/tutorial.html

4. 考察

4.1. 分析過程にひそむバイアス

- 情報検索過程
  - 電子図書館やデジタルアーカイブの検索アルゴリズム
  - Googleなどウェブ検索エンジンのヒットしやすさ（SEO）という商業的理由
  - そもそも、なぜそのコンテンツがデジタル化される対象として選ばれたか
  - デジタル化の過程で、何かが代替されていたり、抜け落ちたりしていないか

- 情報分析過程
  - データや処理ソフトウェアの選択
  - 分析結果を表示するブラウザやインタフェース
  - ⇒これらは、成果物に影響を及ぼす要素。複雑さが過度に単純化されていないか

- Cf. Romein, et al., 'State of the Field: Digital History', 2020, pp. 17-20, https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-229X.12969