量的テキスト分析を用いた通時的な社会科学研究の方法

早稲田大学

渡辺耕平

自己紹介

- 所属
  - 早稲田大学高等研究所　招聘研究員
  - ロンドン政治経済大学（LSE）訪問研究員
  - ラザード・ジャパン・アセットマネジメント　データ科学者
- 研究
  - 政治コミュニケーション
    - 新聞やソーシャルメディアを通じて外国メディアの影響を研究
  - 量的テキスト分析方法論
    - 分析手法やRパッケージの開発

本講座の目標

- 文書データを用いた通時的な研究方法を学ぶ
  - 量的テキスト分析によってテキストデータを数値化する
    - Rパッケージの基本的な操作を習得する
    - 地理的分析、感情分析、トピック分析を文書に適用する
  - 数値データに対して回帰分析を適用し通時的および共時的な変化を発見する
    - テキスト分析の結果を変数として多変量回帰分析を行う
  - 文書データやツールに対する態度を理解する
- 来年刊行予定の教科書の内容を実践してみる
  - ICA Handbook of Computational Communication Researchが刊行予定
  - 本講座はTime-dynamic Analysisという章に基づいている

本講座のスケジュール

時限	時間	内容
1コマ目	10：35～12：30（115分）	理論的な説明 ソフトウェアの準備 データの収集と前処理
昼休み	12：30～13：30	
2コマ目	13：30～15：00（90分）	辞書分析 感情分析 トピック分析
休憩	15：00～15：15	
3コマ目	15：15～16：45（90分）	各自が作業を行う （スクレーピングの説明）
質疑応答	16：45～17：00	

概要

量的テキストとは何か？

テキスト分析による研究対象

- 政治学
  - 比較政治学（選挙マニフェスト、大統領演説）
  - 議会研究（議会・国会・国会での演説）
  - 政治コミュニケーション（新聞記事、ソーシャルメディア、オンライン掲示板）
  - 国際関係（外交電報、外交政策文書、条約）
- メディア研究
  - ジャーナリズム（新聞記事、プレスリリース）
  - 国際コミュニケーション（通信社の外国電報）
- その他
  - 心理学（自由回答式の質問、実験での会話）
  - 人文科学（小説、詩）
  - ファイナンス、経済学（上場企業の年次報告書、中央銀行の議事録）

テキスト分析を用いた研究の目的

- テキスト分析では文書そのものが研究の目的とは限らない
  1. 文書の内容
     - 文書の特定の側面を系統的に特徴づける
  2. 文書の影響
     - 文書が社会に及ぼした影響を調べる
  3. 代理としての文書
     - 文書を通じて直接観察できないもの（潜在変数）を測定する

テキスト分析の目的（１）

- 文書の内容の研究
  - 17-19世紀の米国政府とネイティブアメリカンの部族との間の条約を分析し、公平性を評価した (Spirling 2012)
  - イスラム教徒に関する英国の新聞記事を分析し、報道の客観性を分析した (Baker et al 2012)
  - 2014年のウクライナ危機についてのロシアの国営通信社の報道を分析し、政府の影響によるバイアスを測定した (Watanabe 2017)

Watanabe (2017)

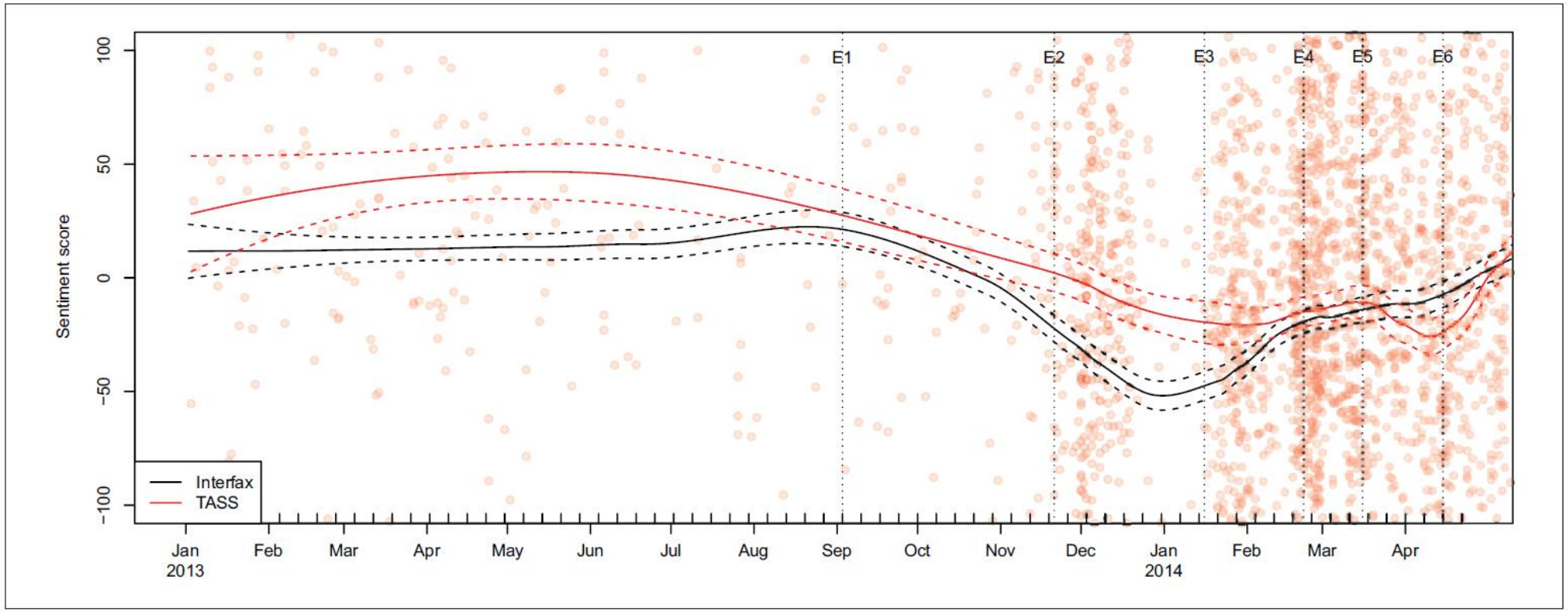
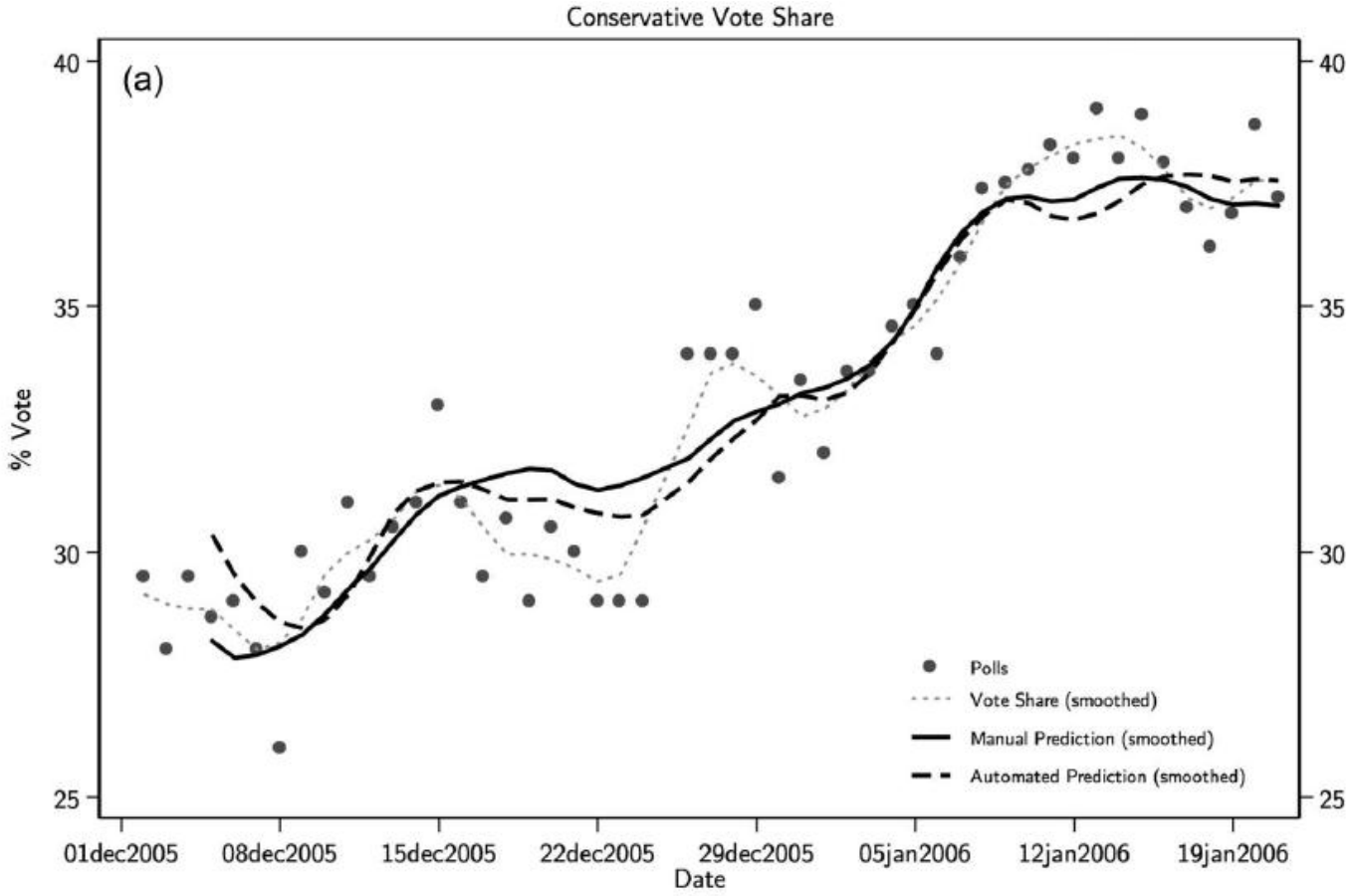


Figure 1. Framing of democracy.

テキスト分析の目的（２）

- 文書の影響の研究
  - 米州議会に提出された法案を分析することにより、立法過程において富裕州が貧困州に与える影響を推定する (Jansa et al. 2018)
  - 人種問題に関するニュース記事を調査データと併せて分析し、メディア報道が世論に与えた影響を推定する (Kellstedt 2010)
  - カナダの新聞と世論調査を用いて、政党に関する記事の内容と政党支持率が強く相関していることを示した (Young & Soroka 2012)

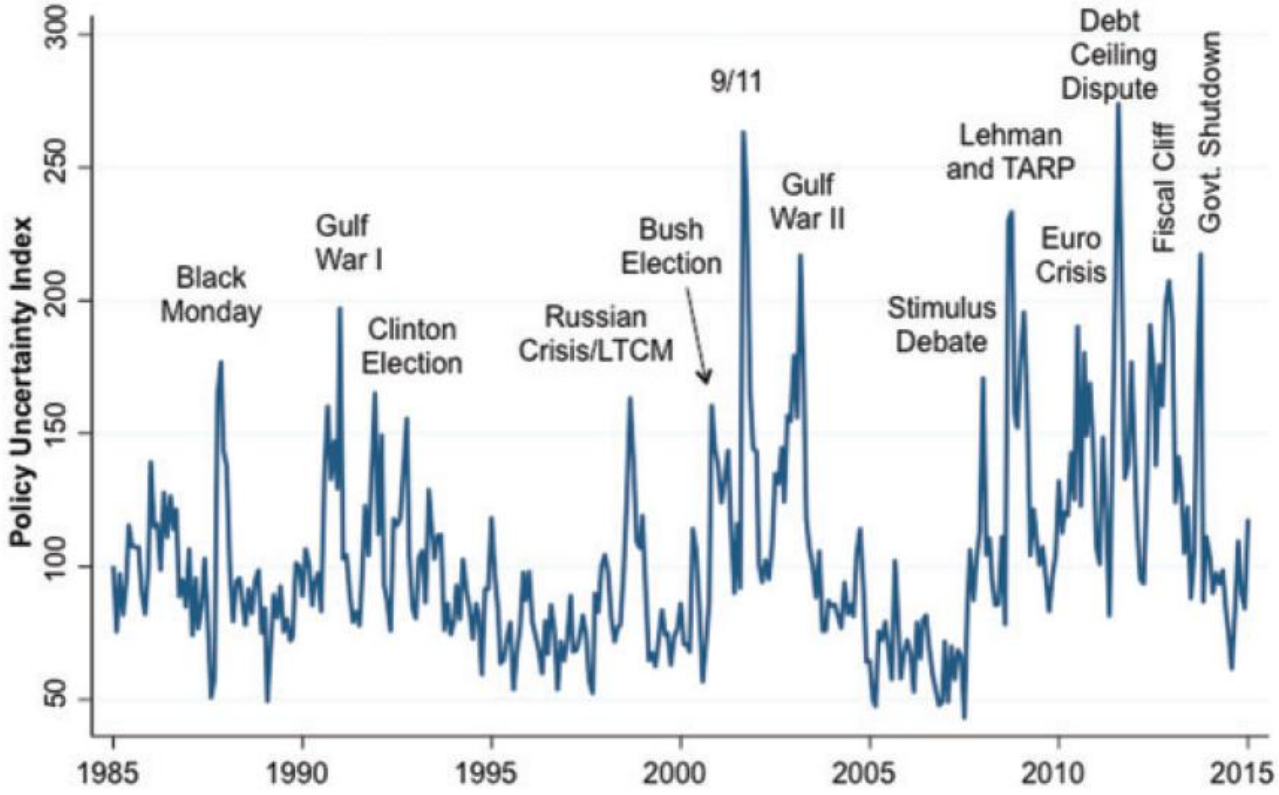
Young & Soroka (2012)



テキスト分析の目的（３）

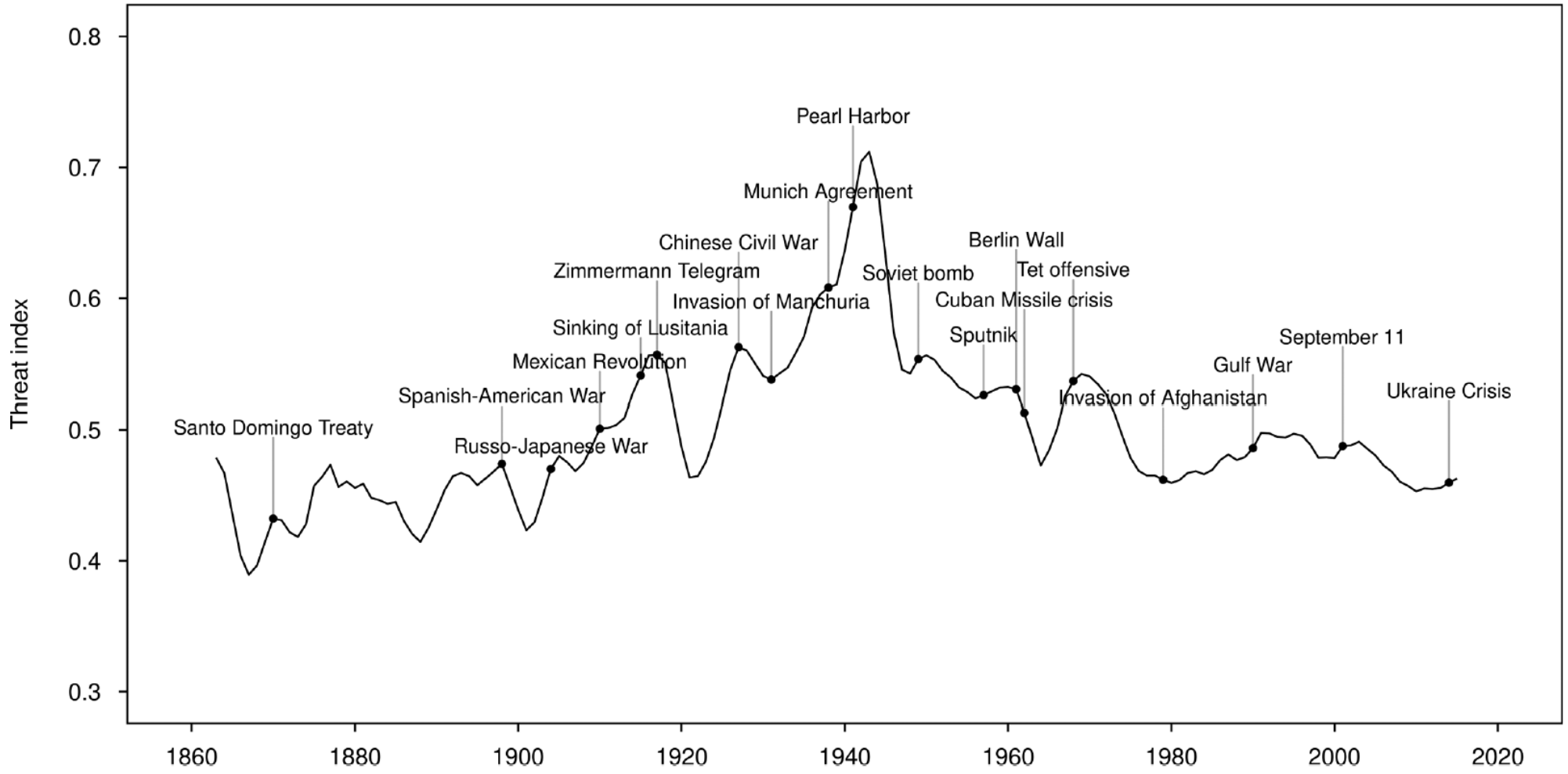
- 代理としての文書
  - 二酸化炭素の排出規制に関するEUと利益団体の文書を数値的に比較することで、政策決定過程における利益団体の影響を測定した (Klüver 2009)
  - 選挙のマニフェストやプレスリリースを分析することで、政治家や政党の政策の優先順位を推定した (Slapin & Proksch 2008; Catalinac 2018; Grimmer 2010)
  - 米国の新聞を横断的に分析することで、30年間の経済政策に対する不確実性を推定した (Baker et al. 2016)
  - New York Timesの記事を分析し、米国に対する地政学的脅威を160年間に渡って測定した (Trubowitz & Watanabe 2021)

Baker et al. (2016)

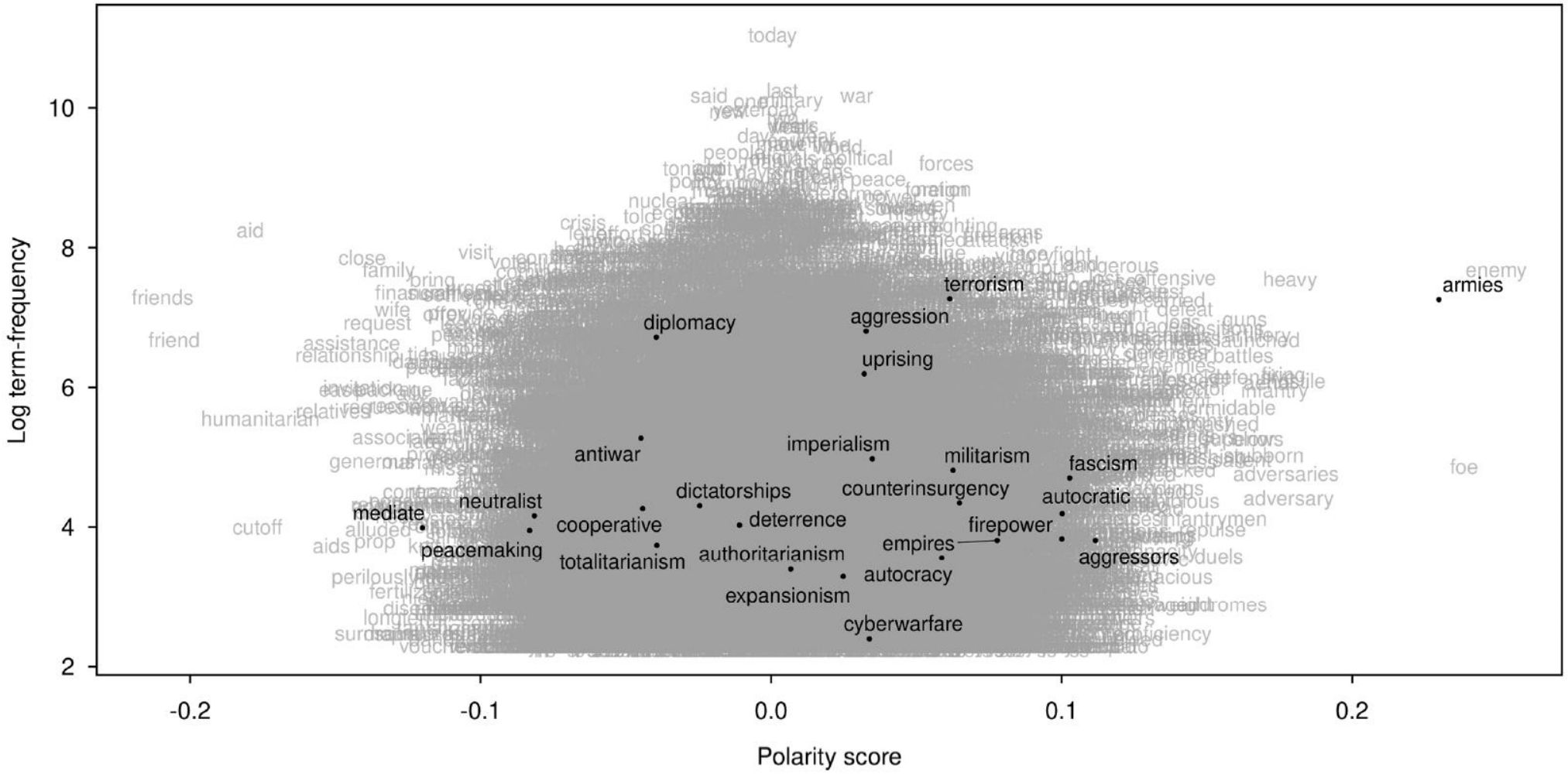


Index reflects scaled monthly counts of articles containing 'uncertain' or 'uncertainty', 'economic' or 'economy', and one or more policy relevant terms: 'regulation', 'federal reserve', 'deficit', 'congress', 'legislation', or 'white house'. The series is normalized to mean 100 from 1985-2009 and based on queries run on 2 February, 2015 for the USA Today, Miami Herald, Chicago Tribune, Washington Post, LA Times, Boston Globe, SF Chronicle, Dallas Morning News, NY Times, and the Wall Street Journal.

Trubowitz & Watanabe (2021)



Trubowitz & Watanabe (2021)



テキストマイニングなのか?

- テキスト分析では社会科学的に意味のあるパターンを探す
  - 文書データは変数が多いため多くの無意味なパターンを含んでいる
  - コンピュータの計算能力に頼りすぎると、偽の相関を見つけることになりやすい
- テキスト分析は理論に導かれる
  - 分析をはじめる前に理論的な期待（仮説）を形成する必要がある
  - 背景知識に基づいて見つかったパターンを説明する必要がある

内容分析なのか?

- テキスト分析は内容分析の一種
  - テキスト分析は内容分析の自動化ではなく「コンピュータに支援された」内容分析である (Krippendorff 2004)
  - テキスト分析の結果の人間の解釈によってはじめて意味が与えられる
- コンピュータは信頼性 (reliability) が高いが、妥当性 (validity) が低い
  - 人間とコンピューターの間の不一致が問題になる
- テキスト分析は、あくまで文書内の概念を数値化する方法
  - 感情、フレーミング、トピックなど
  - 文書を数値化することで、さらなる統計分析を行える

「ビッグデータ」なのか?

- テキスト分析は「ビッグデータ」を調査する方法
  - 「ビッグデータ」とは、実社会で生成された大規模なデータ（観察データ）
  - 文書データは、個人や組織によって毎日大量に生成されている
- 文書データは大きくなる傾向がある
  - 微妙なパターンを見つけるには、大規模なデータが必要
  - 大規模なデータにはノイズが多く分析が大変
- テキスト分析はデータ科学の一分野でもある
  - データ科学（計算社会科学）
    - ネットワーク分析
    - 地理空間分析
    - 画像分析
    - テキスト分析

「AI」なのか?

- テキスト分析は「AI」を応用した研究の方法
  - いわゆる「AI」は、機械学習を用いた自動化ソフトウェア
  - テキスト分析では、比較的単純な機械学習モデルを用いる
- 文書をコンピューターで処理するのは難しい
  - 文書の集合（コーパスは）数万種類の語によって構成される
  - 語の意味は文脈によって変化する
- テキスト分析で使う機械学習モデル
  - 単純ベイズ：モデルに与えた例に基づく文書の分類
  - 共起分析：語の連続に基づく複合語の特定
  - トピックモデル：語の共起関係に基づくトピックの特定
  - Word2vec：語の出現確率に基づく意味の推定

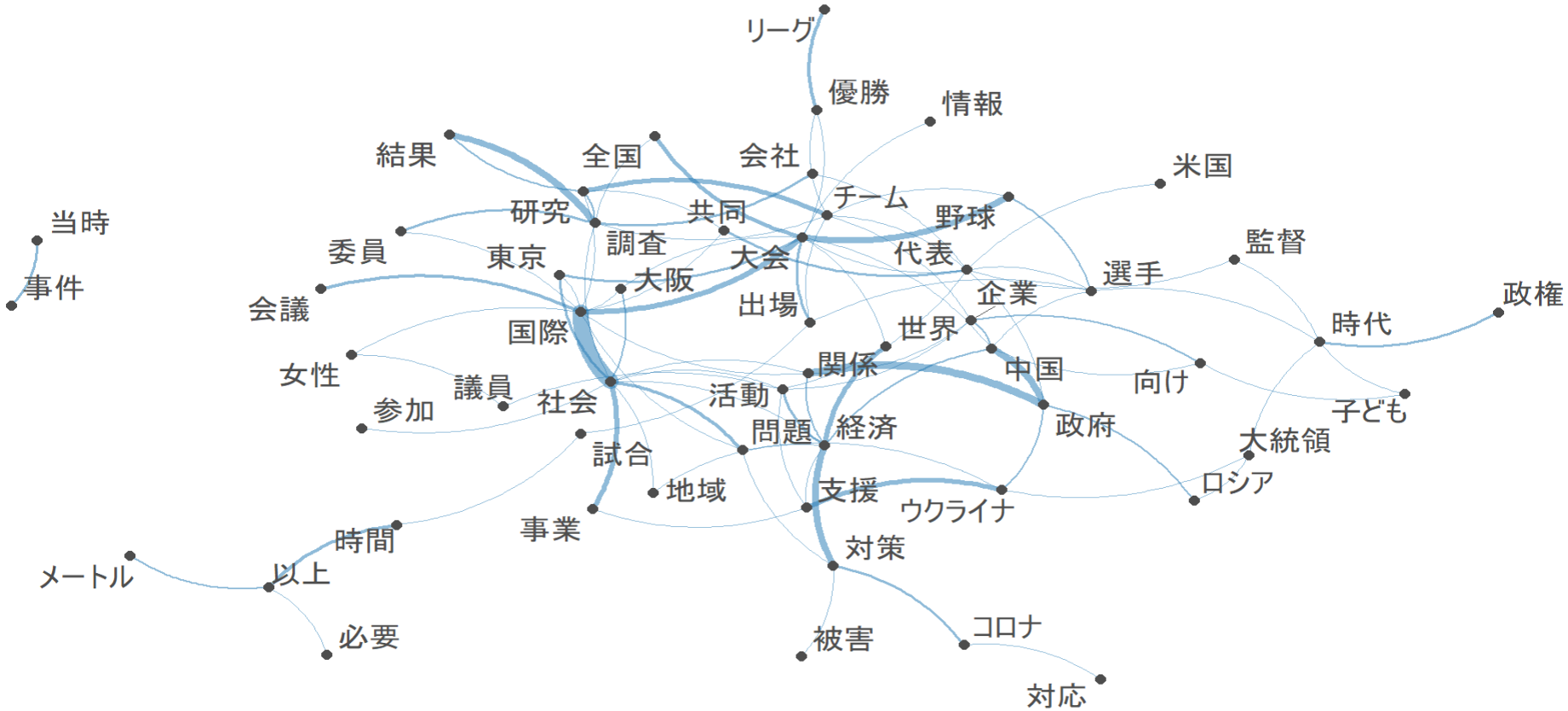
量的テキスト分析とは

- 系統的な文書の分析を通じて、社会科学的に意味のあるパターンを見つける
- 社会科学者は高次の言語現象に注目する
  - 低次の言語現象 （自然言語処理）
    - 形態素：意味の最小単位
      - 例：「日本＋政府」「女性＋議員」「リーグ＋優勝」「大＋統領」
    - 語彙：個々の単語の意味
    - 構文：単語間の関係（文法）
  - 高次の言語現象（テキスト解析）
    - ディスコース（言説）：文書の集合（コーパス）としての意味
    - プラグマティクス（語用）：社会的文脈における文書の意味

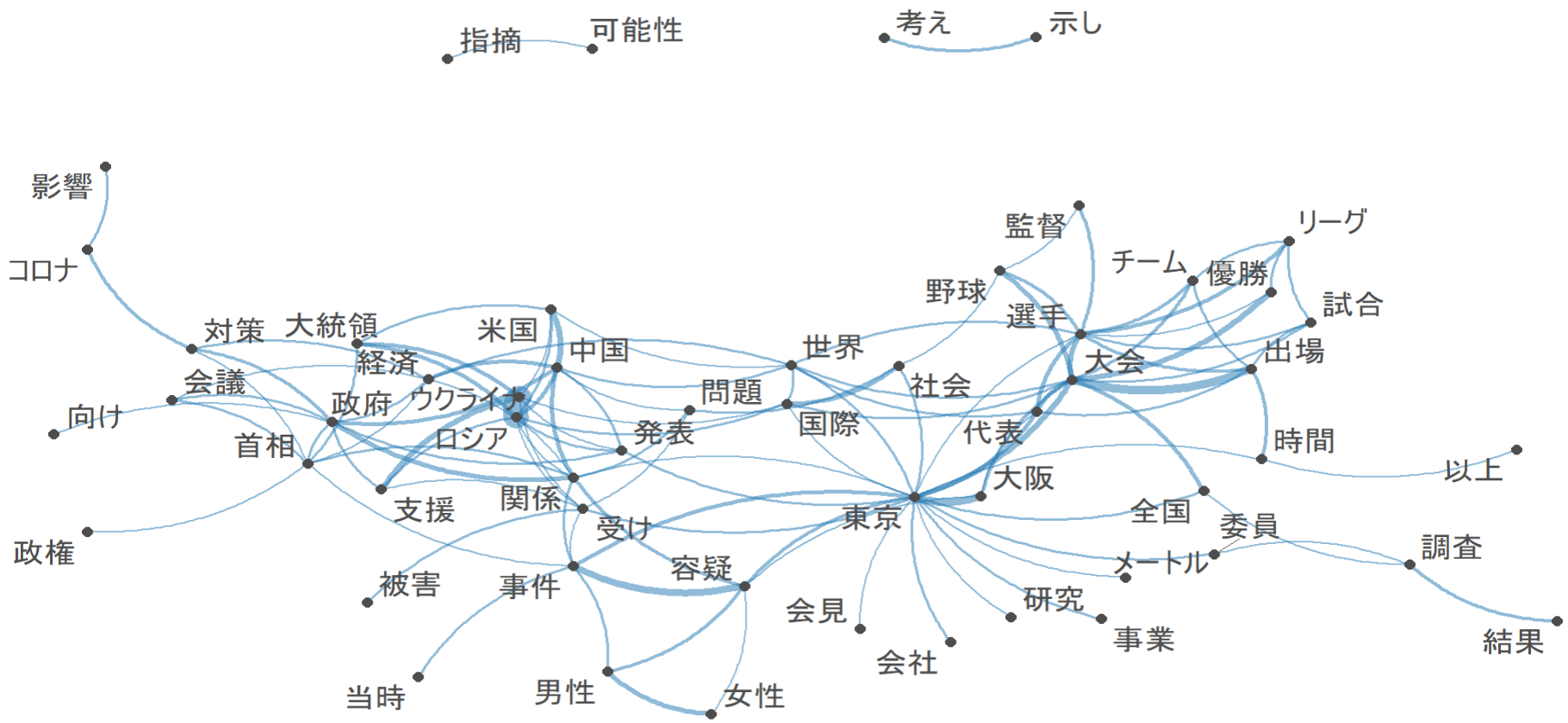
高次の現象を見つける方法

- 言説分析
  - 深い読み (deep reading)
    - 少数の文書を人間が読む（質的テキスト分析）
  - 浅い読み (shallow reading)
    - 多数の文書を機械で分析する（量的テキスト分析）
- 語用論的な分析
  - 言説の社会的意味を理解する
    - いつ、だれが文書が作成（または公表）したのか?
    - なぜ、その時、その者によって文書が作成（または公表）されたのか?

低次の言語現象


リーグ
優勝
情報
結果
全国
会社
米国
当時
研究
共同
チーム
野球
委員
調査
東京
大会
代表
監督
事件
大阪
選手
会議
出場
企業
政権
国際
世界
時代
女性
関係
中国
議員
社会
活動
向け
参加
問題
経済
政府
子ども
試合
大統領
地域
支援
ロシア
ウクライナ
事業
時間
対策
メートル
以上
必要
被害
コロナ
対応

高次の言語現象


指摘
可能性
考え
示し
影響
コロナ
対策
大統領
米国
中国
経済
会議
政府
ウクライナ
向け
首相
ロシア
発表
問題
世界
国際
社会
監督
野球
選手
チーム
優勝
リーグ
試合
大会
出場
代表
時間
以上
大阪
支援
関係
受け
東京
全国
委員
政権
容疑
メートル
調査
被害
事件
研究
会見
事業
結果
会社
当時
男性
女性

テキスト分析の歴史（１）

- 1950-1960年代：メインフレームによる辞書分析
  - General Inquirer システムと辞書 (Stone et al. 1966)
    - ハーバードIII心理社会学辞書
    - サンタフェ人類学辞書
    - スタンフォード政治辞書
  - 辞書分析は非力なコンピューターでデータの次元を縮小する方法

テキスト分析の歴史（２）

- 1970-1980年代：パソコンによる辞書分析
  - The Regressive Imagery Dictionary (Martindale 1975)
  - DICTION (North et al. 1984)
- 1990年代：統計的な次元縮減
  - Latent Semantic Analysis (Deerwester et al. 1990)
- 2000-2010年代：さまざまな機械学習モデル
  - Wordscores (Laver et al. 2003)
  - Latent Dirichlet allocation (Blei 2003)
  - Wordfish (Slapin & Proksch, 2008)
  - Word2vec (Mikolov et al. 2013)
- 2010年代後半：深層学習モデル

基本的な手順

量的テキスト分析の一般的な作業の流れ

テキスト分析の用語

- 文書（documents）
  - データの単位（ニュース記事、スピーチの写し）
  - 文書が段落や文節の場合もある
- コーパス（corpora/corpuses）
  - 文書の集合
- トークン（tokens）
  - 文書内の個々のシンボル (単語、数字、記号など)
  - 文書をコンピューターで処理するために分割した単位
- 特徴（features）
  - 文書内のシンボルの集合
- 共起（collocations）
  - 複数のトークンが同じ文書に現れること
- 文書行列（document-feature matrix; DFM）
  - 文書に現れる特徴の頻度を記録した行列

テキスト分析の流れ

1. 研究課題の特定
2. データの収集
   - 資料収集してコーパス構築
     - 多くの場合、資料収集が一番大変
3. データの前処理
   - 文書をトークンに変換
     - 自然言語処理ツールを用いて文書をトークン化する
4. 文書行列の構築
   - 特徴を選択してデータを単純化する
5. 統計分析の実施
6. 結果の解釈

研究課題の特定

- 理論に従って研究課題を見つける
  - 理論は研究者の間で共有された考え方のこと
  - 研究では理論を発展、強化、訂正することが重要
    - 文献レビューを通じて理論を明確にし、課題を特定する必要がある
  - 理論に基づき仮説を設定し、テキスト分析で検証する
- 研究課題によって分析するデータを選ぶ
  - テキスト分析が最適な研究手法である場合とは限らない
    - 文書の内容や影響を調べたいのか?
    - 直接観察できないものを測定したいのか?

データの収集（１）

- 文書はインターネット上に多数ある場合
  - Word、PDF、HTMLなどのファイル形式
  - 手動もしくは自動的でファイルをダウンロードする
  - ダウンロードの自動化はスクレイピングと呼ばれる
- 文書がデジタル化されていない場合
  - デジタル化するには、現物をスキャンし、OCR（光学式文字認識）を適用する
  - データに誤りがないか人間が確認する必要がある

データの収集（２）

- ニュース記事は商用データベースから入手できる
  - NexisやFactivaなどの横断的データベース
  - 朝日、読売、毎日新聞などは独自のデータベース
- 政治文書はコーパスとして公開されている
  - 日本の国会議事録
  - 欧州の選挙マニフェスト
  - 欧州の議会議事録
  - 国連総会演説
  - 米国大統領演説

データの収集（３）

- 代表性のあるサンプルを採取するのが難しい
  - データ生成過程の知識に基づいてソースを選択する
    - 文書がいつ、どこで、なぜ、そして誰のために/誰によって作成されのか？
    - 文書の内容の変化が、データ生成過程の変化を表すと考える
  - 異なったデータ生成過程を持つソースからデータを収集する
    - テキスト分析の結果を解釈するためには比較対象が必要
    - 異なったソースからの文書を比較すると分析がやりやすい
      - 商業メディアと国営メディアのニュース
      - 右翼政党と左翼政党のソーシャルメディア

テキスト分析用のソフトウェア

- 本講座で使うRパッケージ
  - Quanteda：Rで文書データを扱うための基盤
  - LSX：種語を用いて文書の極性の測定
  - SeededIda：LDAを応用し文書を指定したトピックに分類
  - Newsmap：地名辞書を自動で拡張し文書を地理的に分類
- その他のパッケージ
  - TidyText： Rで文書データを手軽に操作するめのツール
  - STM：文書のトピックを文書変数と合わせて特定
  - Wordvector：Word2vecを実装し容易に単語ベクトル作成
  - proxyC：並列処理によってベクトル間の距離を効率的に計算

実習

量的テキスト分析の実践

実習用のデータ

- コーパス
  - 「世界と日本」（政策研究大学院大学　田中明彦研究室）
  - 「帝国議会・国会内の総理大臣演説」からダウンロード
    - https://worldjpn.net/documents/indices/pm/index.html
- Rスクリプトとデータ
  - Google Drive：事務局からのメールを参照
  - Github： https://github.com/koheiw/workshop-ISS
- 参考文献
  - アルゴリズムや研究事例を必要に応じて参照

実習の内容

1. 首相演説コーパスから変数を作成する
   - 辞書分析：どの国が言及されているかを検出する
   - 感情分析：どれぐらい肯定的（否定的）かを測定する
   - トピック分析：何について論じているかを特定する
2. 作成した変数を用いて回帰分析を行う
   - 辞書分析：ダミー変数
   - 感情分析：連続変数
   - トピック分析：カテゴリー変数

実習の進め方

1. Rスクリプトをそのまま実行してみる
2. 重要な引数を理解する
   - 講義の中で説明する
   - Rパッケージのマニュアルも参照する
3. 引数を変更して実行してみる
   - 結果がどれくらい変わるか確認する
   - どのように最適な値を設定するかを理解する
4. 種語を変えてみる
   - 種語によって結果がどのように変化するか理解する
5. コーパスを変えてみる
   - データの収集方法を理解する

講座の最後に

- 多言語テキスト分析研究会
  - https://groups.google.com/g/multa-rg
    - 年に４回ほど若手を中心に研究会を開催している
    - 現在進行中や立案中の研究について議論する
- 今後の連絡先
  - watanabe.kohei@gmail.com
    - 研究に関する相談を歓迎します

参考文献（１）

- Baker, P., Gabrielatos, C., & McEnery, T. (2012). Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word “Muslim” in the British Press 1998-2009. Applied Linguistics, 34(3), 255–278. https://doi.org/10.1093/applin/ams048
- Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. The Quarterly Journal of Economics, 131(4), 1593–1636. https://doi.org/10.1093/qje/qjw024
- Catalinac, A. (2018). Positioning under Alternative Electoral Systems: Evidence from Japanese Candidate Election Manifestos. American Political Science Review, 112(1), 31–48. https://doi.org/10.1017/S0003055417000399
- Grimmer, J. (2010). A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. Political Analysis, 18(1), 1–35. https://doi.org/10.1093/pan/mpp034
- Jansa, J. M., Hansen, E. R., & Gray, V. H. (2018). Copy and Paste Lawmaking: Legislative Professionalism and Policy Reinvention in the States. American Politics Research, 1532673X18776628. https://doi.org/10.1177/1532673X18776628
- Kellstedt, P. M. (2000). Media Framing and the Dynamics of Racial Policy Preferences. American Journal of Political Science, 44(2), 245. https://doi.org/10.2307/2669308

参考文献（２）

- Klüver, H. (2009). Measuring Interest Group Influence Using Quantitative Text Analysis. European Union Politics, 10(4), 535–549. https://doi.org/10.1177/1465116509346782
- Slapin, J. B., & Proksch, S.-O. (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. American Journal of Political Science, 52(3), 705–722. https://doi.org/10.1111/j.1540-5907.2008.00338.x
- Spirling, A. (2012). U.S. Treaty Making with American Indians: Institutional Change and Relative Power, 1784–1911. American Journal of Political Science, 56(1), 84–97. https://doi.org/10.1111/j.1540-5907.2011.00558.x
- Trubowitz, P., & Watanabe, K. (2021). The Geopolitical Threat Index: A Text-Based Computational Approach to Identifying Foreign Threats. International Studies Quarterly, sqab029. https://doi.org/10.1093/isq/sqab029
- Watanabe, K. (2017). Measuring news bias: Russia's official news agency ITAR-TASS' coverage of the Ukraine crisis. European Journal of Communication, 32(3), 224–241. https://doi.org/10.1177/0267323117695735
- Young, L., & Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. Political Communication, 29(2), 205–231. https://doi.org/10.1080/10584609.2012.671234

補足

トピックの収束を検知する仕組み