

@島根大学人間科学部
2019.07.13

応用心理学研究 I

データ収集後探索的解析（テキストマイニング）

明治大学 研究・知財戦略機構
佐藤浩輔

講義の前に

講義の前に①

- 講義ではRがインストールされているマシンが必要です
 - 本講義、特に2, 3講目ではRを用いて説明します。
 - Rについて基礎的な知識を持っていることが望ましいですが、講義の内容を理解する上で必須ではありません
 - 講義の特性上、大量の情報が流れますが、逐一詳細まではフォローはしないので、エッセンスをつかんで後で復習してください
 - 課題はR以外の言語を使っても構いません
 - 以下のページに講義で使う資料を置きます
 - <https://github.com/satocos135/lecture2019shimane>
 - スライド・Rコードも講義終了後公開します

講義の前に②

- 講義時間：13:00~18:30
 - 長いので適宜休憩を入れます
 - 以下は許可をとる必要はありません
 - 飲み物の補給（熱中症にならないよう）
 - トイレ
 - 質問は隨時受け付けます

講義の前に③

- 発展的な内容も扱います (★マークで表示)
 - 講義内では詳しくは扱いませんが、興味があれば調べてください
- 課題があります
 - この講義で学んだことを使ってテキストを分析する内容です
 - 課題の詳細は講義の終了時にお伝えします

講義の概要

- この授業は
 - 計量テキスト分析/テキストマイニング/自然言語処理とはなにか
 - テキストを扱う社会科学の研究をどのように計画するか
 - テキストを分析することで何がわかるか
 - テキストデータをどのように処理するか
 - テキストをどのように分析するか
 - 結果をどのように報告するか

について

300分で説明から実習までやってしまおうという
大変無謀な野心的な講義

講義の流れ

- 第1部：講義パート
 - 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
 - テキストをもちいた研究のデザイン、研究計画の立て方を学ぶ
- 第2部：実習パート
 - 実際にデータを扱いながら学ぶ
 - 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、解釈する

第1部

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

第1部のアウトライン

- 前半
 - 計量テキスト分析・テキストマイニング・自然言語処理
 - それぞれの用語の整理
 - 特に、自然言語処理とは何か
 - なぜ量的手法が必要か
 - 人文・社会科学での応用例
- 後半
 - 計量テキスト分析・テキストマイニングの研究デザイン
 - 研究の立案から
 - 研究に関わる誤差
 - データの収集法
 - 分析手法について

計量テキスト分析・テキストマイニング・
自然言語処理

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

自然言語処理の技術

- 基礎技術
 - 形態素解析
 - 構文解析
 - 意味解析
 - 固有表現抽出
- 応用技術
 - 文書分類
 - 自然言語理解
 - 自然言語生成
- 実社会への応用例
 - 検索エンジン
 - 自動翻訳
 - 質問応答・チャットボット

Google

Google 検索

I'm Feeling Lucky

<https://www.google.co.jp/>

文 テキスト

ドキュメント

言語を検出する

日本語

英語

韓国語



吾輩は猫である。名前はまだ無い。



どこで生れたかとんと見当がつかぬ。何でも薄暗いじめじめした
所でニヤーニヤー泣いた。声がけはなかった。アハフ。五年目だ。
始めて人間というも

Wagahaihanekodearu. N
Nani demo usugurai jim
iru. Wagahai wa koko de



英語

日本語

韓国語



I am a cat. There is no name yet.



I have no idea where I was born. I remember only that I was crying
in a place where it was dull bullying anything. For the first time here
I saw a human being.





●●●● au

9:53

97%

“面白い話”

タップすると編集できます

わかりました...

昔々、遙か彼方の仮想銀河
に、Siriという若くて知的な工
ージェントが住んでいまし
た。

ある晴れた日、Siriはパーソナ
ルアシスタントとしてAppleに

既読
12:25

探偵ごっこ

郊外の道路でオートバイ
が横転し、運転者がケガ
をした。捜査を進めた結
果、被疑者AとBが浮かび
上がったが、どちらが犯
人かはわからない。2人は
事件に対し、こう供述し
た。



Amazon.co.jp : Echo 第2...
amazon.co.jp



スマートスピーカー(AIスピーカー)徹...
yuki-no-yabo.com



価格.com - スマートスピーカー...
kakaku.com



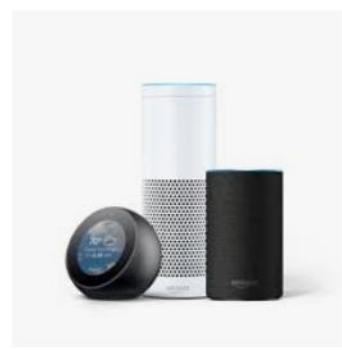
スマートスピーカー(AIスピ...
watch.impress.co.jp



価格.com - スマートスピーカ...
kakaku.com



スマートスピーカー・AIスピーカーでできること・選び方...
e-earphone.jp



スマートスピーカー(AIスピ...
watch.impress.co.jp



あらゆるスマートスピーカーを徹...
moov.ooo



最新スマートスピーカー徹底比較】「Amazon Ech...
robotstart.info

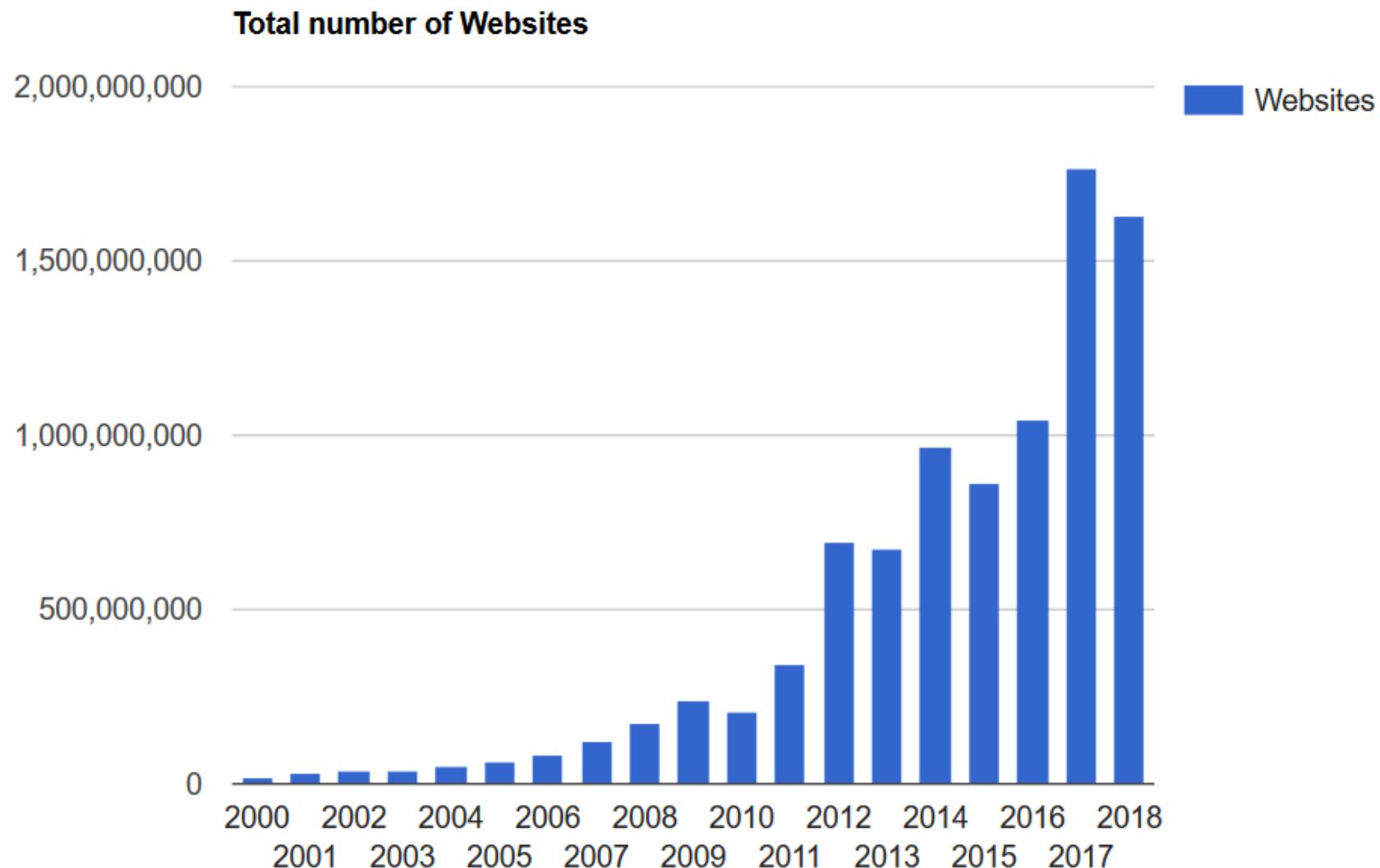
なぜテキストマイニングか

- 言語データの蓄積 + データ処理環境の整備
- →言語データを手軽に大規模に利用できるようになった

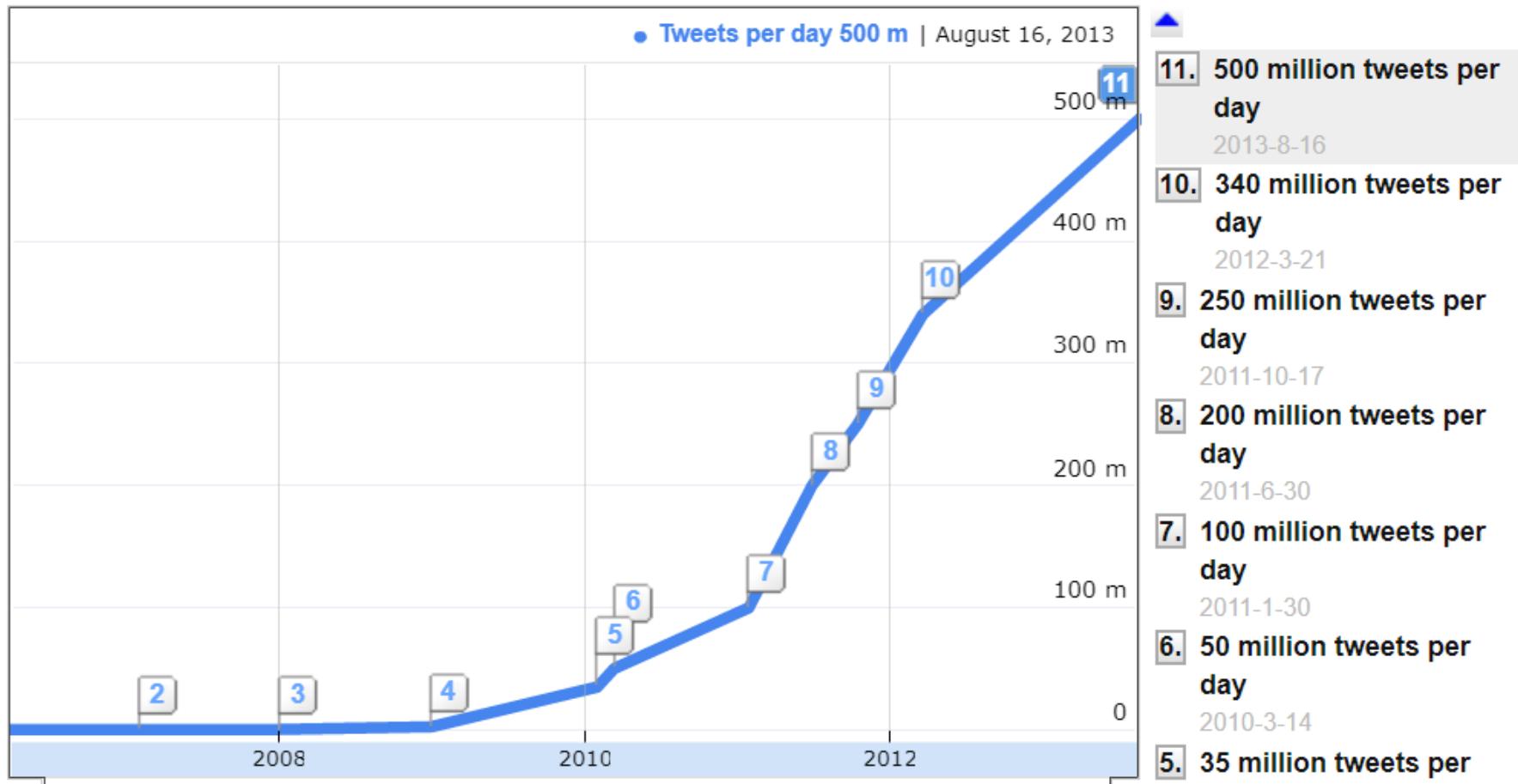
言語データの蓄積

- 人間同士のやりとり：言語情報
 - 文字情報
 - 文書・書籍
 - SNS・チャット
 - 電子メール
 - Webサイト
 - 音声情報
 - 会話

インターネット上のWebサイトの数



Twitterにおける一日あたりのtweetの数



データ処理環境の整備

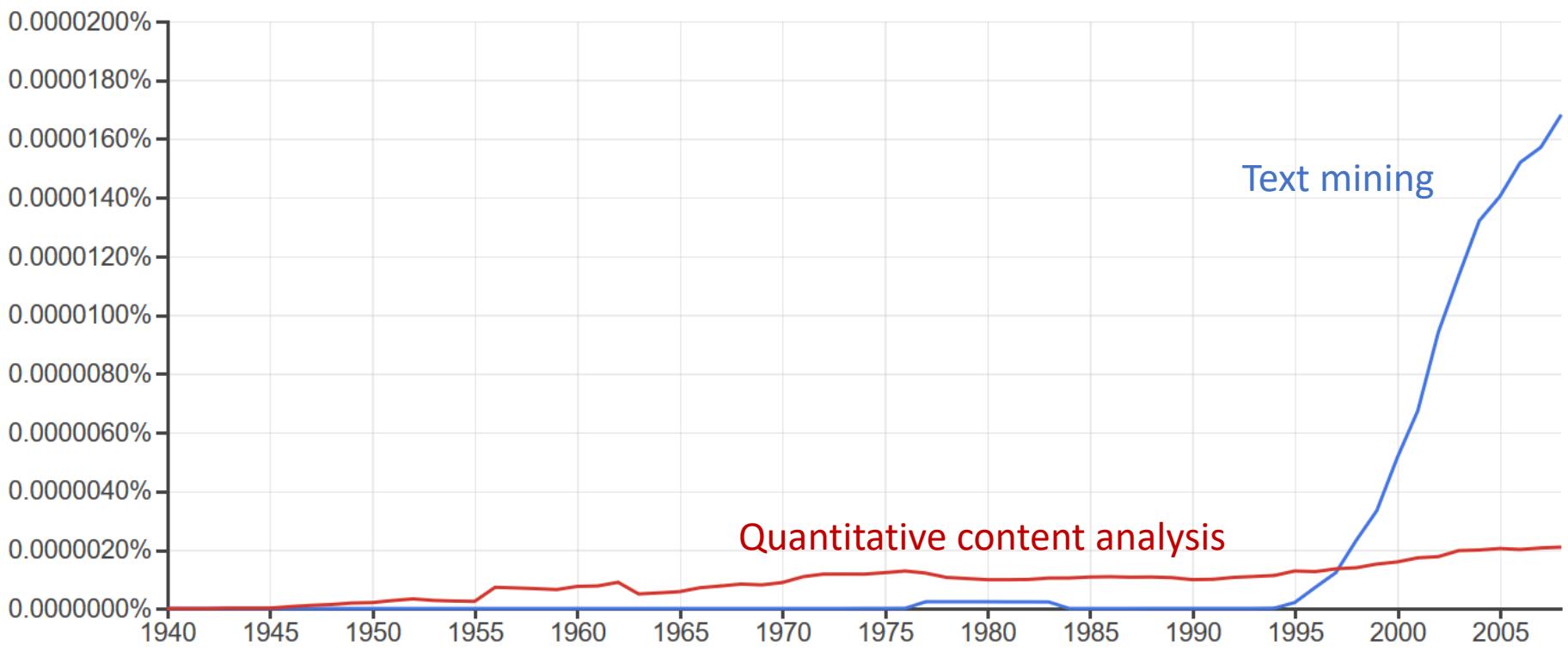
- ハードウェア能力の向上
 - 保存容量の増加：大量のデータを蓄積できる
 - 計算機の処理速度：大量のデータを処理できる
 - 通信速度の向上：大量のデータをやりとりできる
 - 通信インフラの整備：誰でもインターネットにアクセスできる
- データ処理技術の発展
 - データマイニング/統計的手法（特に機械学習）の発展
 - データセットの整備
 - 扱えるデータの増加
 - 自然言語処理技術の発展
 - 言語データを（一定の精度で）大量に、自動的に処理できる

Google Books Ngram Viewer

Graph these comma-separated phrases: case-insensitive

between and from the corpus with smoothing of .

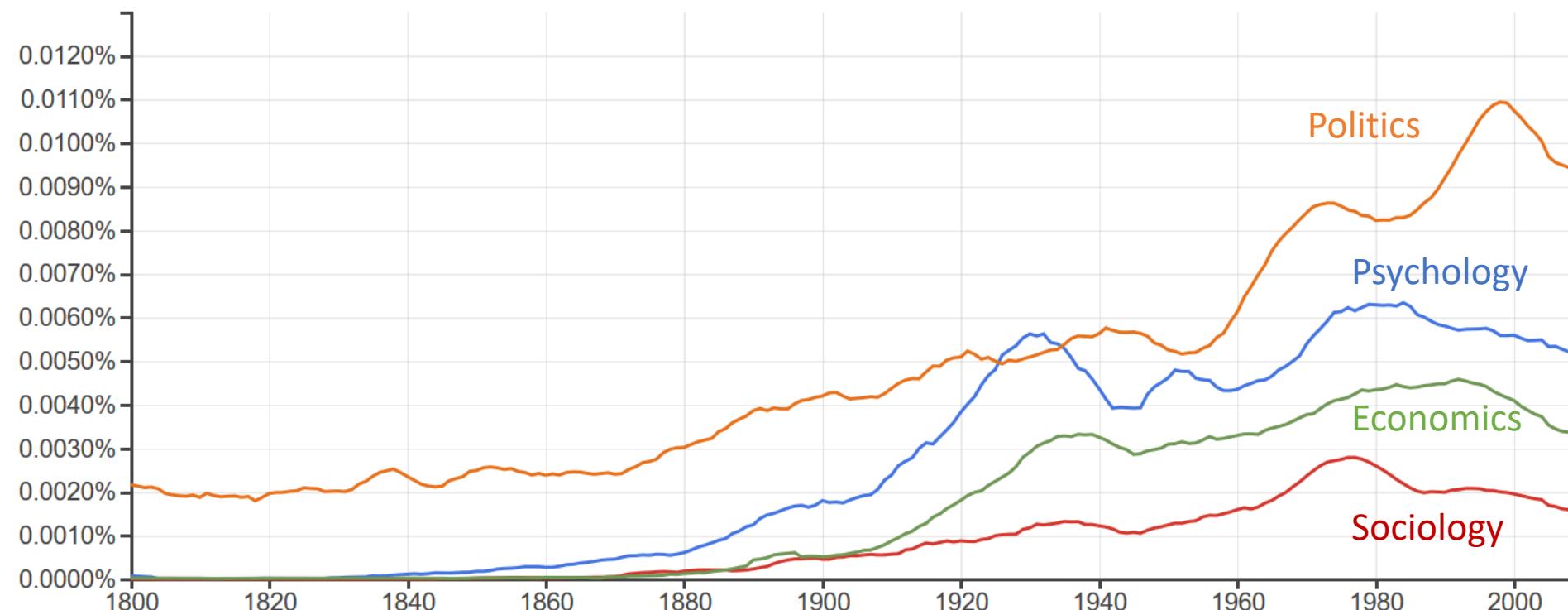
Search lots of books



Google Books Ngram Viewer

Graph these comma-separated phrases: case-insensitive

between and from the corpus with smoothing of



計量できると何が嬉しいか

- 質的なものを量的に扱える
 - 処理の効率性
 - 機械的かつ大量に処理できる
 - 手続きの明瞭性
 - 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
- 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

人文学分野における応用例

- デジタル人文学 digital humanities :
情報処理の技術を人文学の研究に応用
 - 文学
 - Distant reading (Moretti, 2013)
 - 精読 close reading に対して、情報処理技術で大量の文献を扱う
 - 計量文体学 stylometry / stylometrics (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

社会科学分野における応用例

- 計算社会科学: computational social sciences

Webのソーシャル化や実空間での様々な行動センシングが進行している現在、人々の自発的な情報行動やコミュニケーションなどの詳細はデジタルに記録・蓄積されるようになりました。このような大規模社会データを情報技術によって取得・処理し、分析・モデル化して、人間行動や社会現象を定量的・理論的に理解しようとする学問が「計算社会科学」(Computational Social Science)です。

計算社会科学はその目的の達成の方法論として、大規模社会データ分析研究、社会シミュレーションによる理論的研究、バーチャルラボによる実験的研究などを用いています。

<https://css-japan.com/about/>

Figure 3. “Censorship Magnitude,” The Percent of Posts Censored Inside a Volume Burst Minus Outside Volume Bursts.

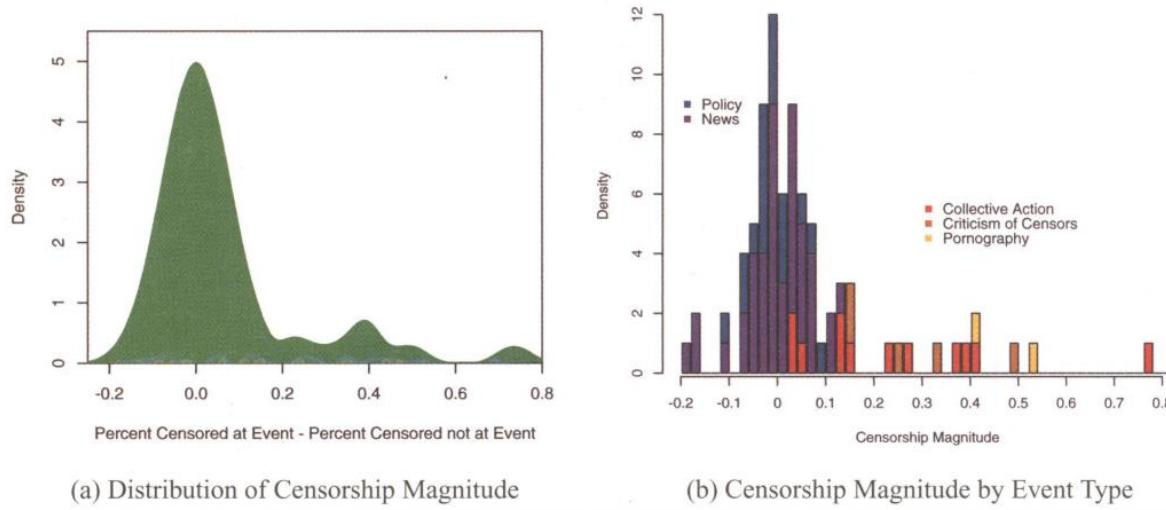
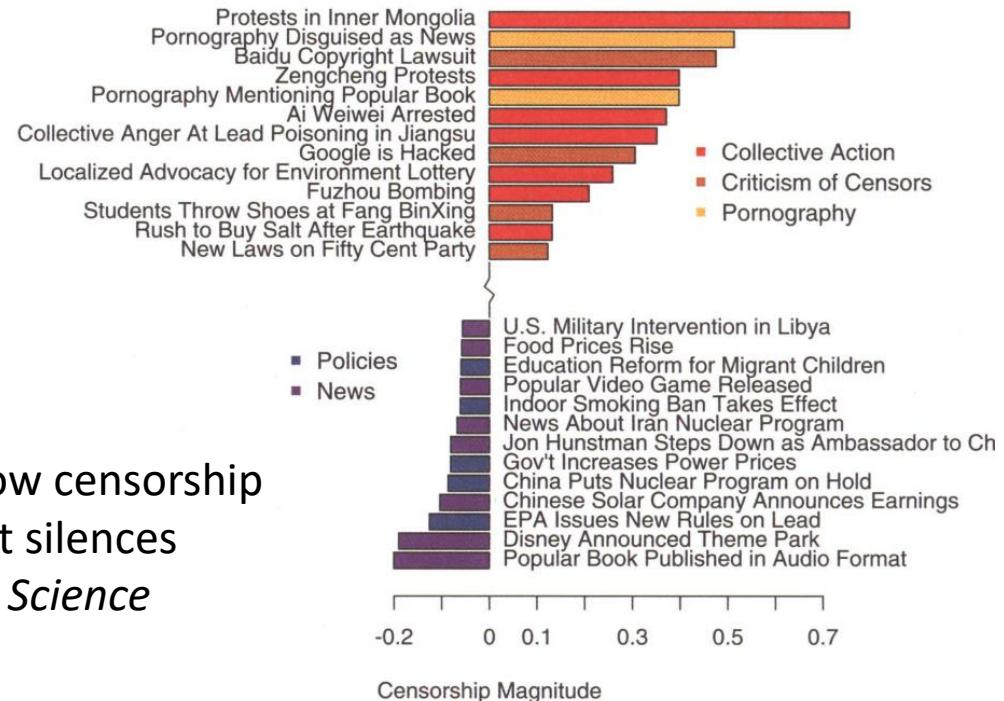


Figure 4. Events with Highest and Lowest Censorship Magnitude



King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, 107(2), 326–343.

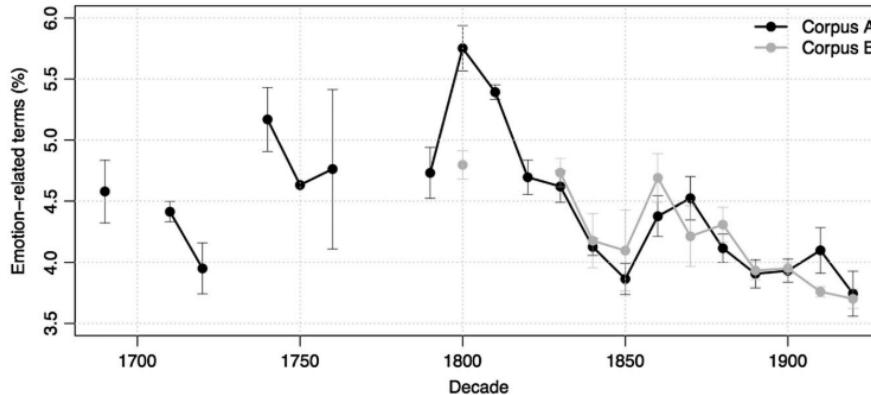


Figure 2. Emotionality changes in Anglophone literature, for the two “small data” corpora. Error bars represent 95% confidence intervals.

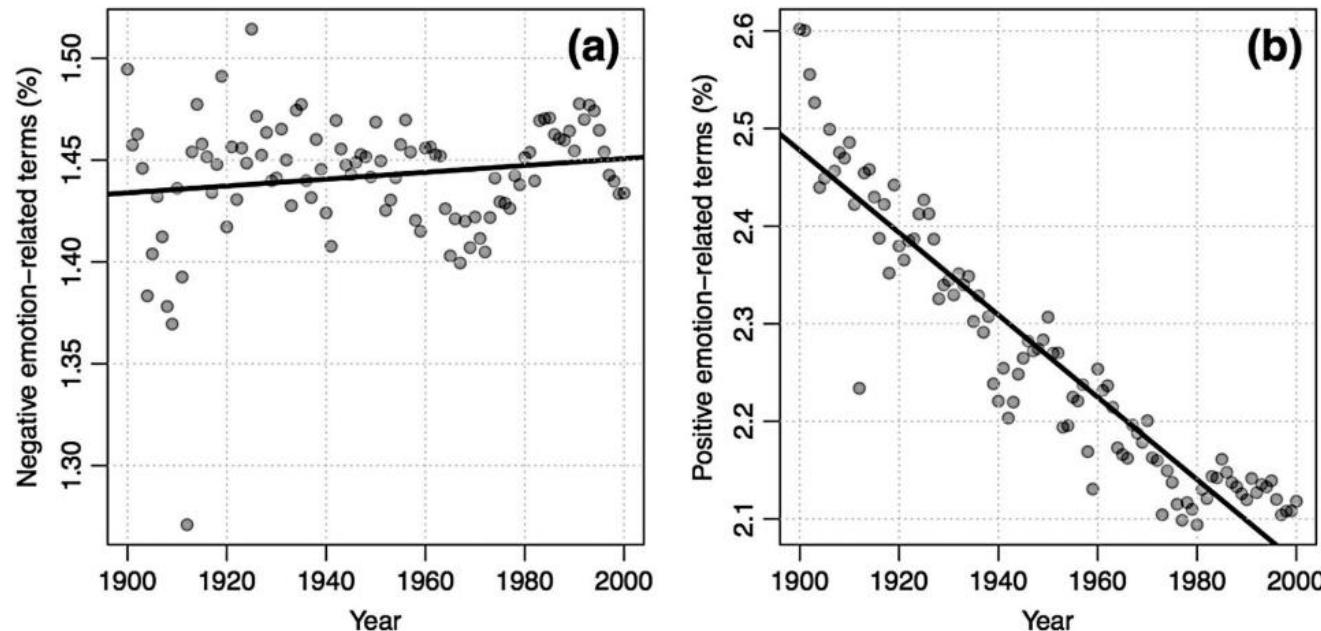


Figure 3. Emotionality changes in Anglophone literature, for the Google Books corpus. (a): negative emotions-related terms. (b): positive emotions-related terms. Solid lines represent linear regressions of the data.

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.

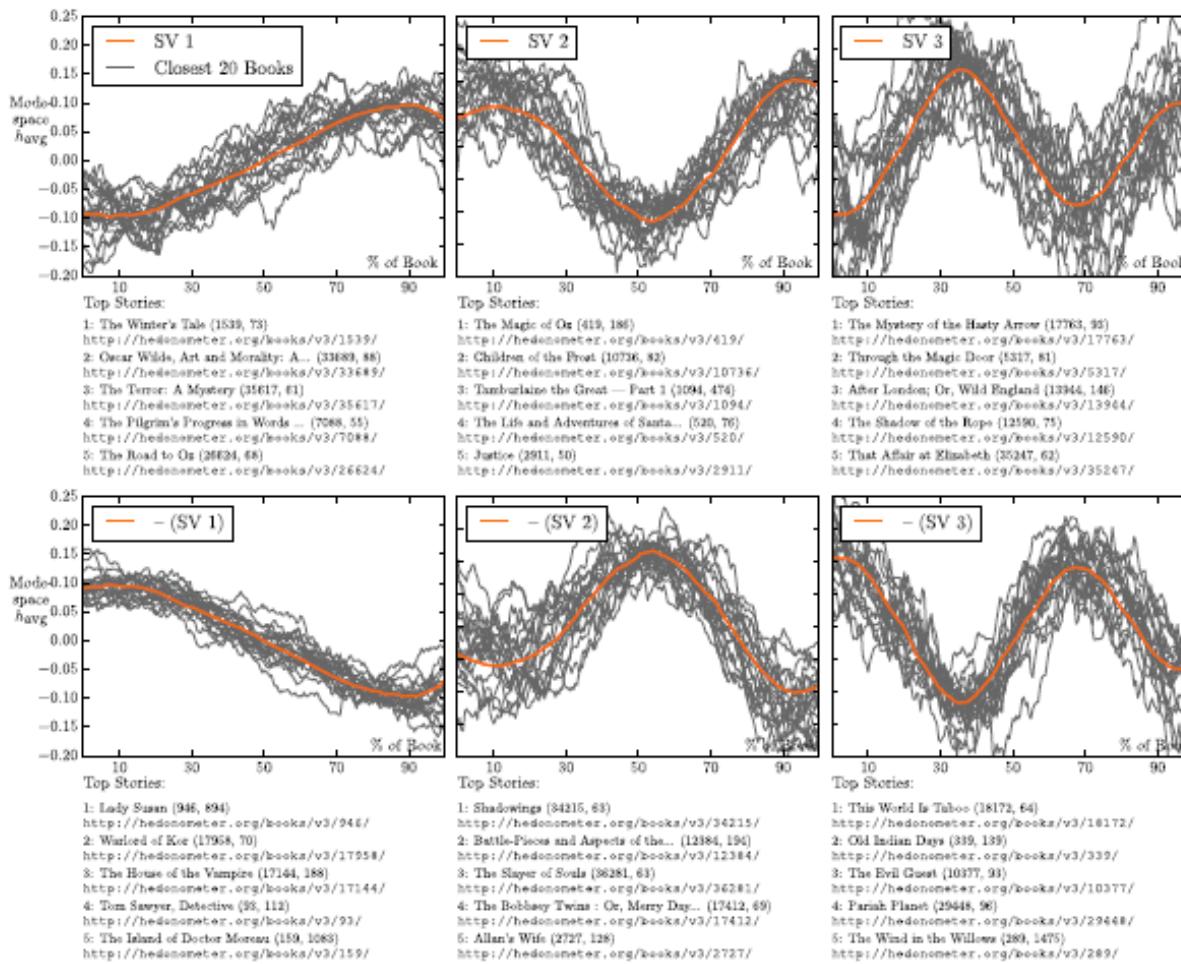


Figure 4 First 3 SVD modes and their negation with the closest stories to each. To locate the emotional arcs on the same scale as the modes, we show the modes directly from the rows of V^T and weight the emotional arcs by the inverse of their coefficient in W for the particular mode. The closest stories 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://hedonometer.org> which enables detailed exploration of the emotional arc for the story.

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
 - 人文学・社会科学の様々な領域に広がっている

計量テキスト分析・テキストマイニングの 研究デザイン

研究の流れ

- 研究計画の立案
 - 研究目的の設定
 - データ収集手法の決定
- データ収集
- 分析
 - 前処理・クリーニング
 - 分析
 - 検証
- アウトプット
 - 報告・発表
 - レポート・論文・学会発表, etc.

←研究デザイン

研究計画の立案

“Garbage in, garbage out”

→ゴミを入れればゴミが出てくる

- 勝負はデータを取る前にほとんど決まっている
 - 段取りが9割
- 「So What?」な研究にならないために
 - その研究のオーディエンスは誰か
 - その研究がうまくいくとなぜ嬉しいのか
 - 理論的な価値：理論的な意味がある
 - 応用的な価値：何かの役に立つ
 - 資料的な価値：そのデータを取ること自体に価値がある
 - 分野によって関心が異なる
 - →色々な分野の研究を知ってセンスを磨く

研究目的の設定

- 仮説**検証**型研究
 - 特定の仮説を検証するための研究
 - 妥当性と信頼性が求められる
 - 妥当性：測りたいものが測れているか
 - 信頼性：測りたいものが精度よく測れているか
 - ※この辺の検討が雑だと総じて無意味な研究になりがち
 - 検証的な手法と相性がよい
- 仮説**生成**型研究
 - (意味のある) 仮説を生成するための研究
 - 事前の仮説はないものの、関心のある変数の分布や変数間の関連を調べる
 - 探索的な手法と相性がよい
 - 探索的データ解析(EDA: Exploratory Data Analysis; Tukey, 1977)
 - データマイニング

仮説検証と仮説生成

- ひとつの研究の中で組み合わせてもよい
 - e.g. 仮説検証パートと、主要な変数との関連を探索的に調べるための質問群
 - 重要なのは、それぞれこの項目はどのような目的のためにとるのか、を意識すること
 - 研究に使えるリソースは有限
 - 何が重要で何が重要でないか、優先順位を明らかにする
 - 探索的な分析で出た結果
 - 議論するに十分な測定精度がないかもしれない
 - 統計的なアーティファクトかもしれない
 - 多重検定の問題

データ収集方法の決定

- 考慮すべき事項
 - 研究における誤差
 - 各手法の特性
 - 実験
 - 調査
 - ビッグデータ
 - データの入手方法
 - 収集後の分析手法
 - 各種制約
 - 倫理・コスト（予算・時間）・その他ロジスティクス上の問題

研究における誤差

- 偶然誤差 random error
 - ランダムに生じる真の値からのずれ（ばらつき）
 - サンプルサイズを増やせば減らすことができる
- 系統誤差 systematic error
 - ランダムでない真の値からのずれ
 - サンプルサイズを増やしても減らすことができない
 - 系統誤差の 3 つのカテゴリ
 - 選択バイアス selection bias
 - 情報バイアス information bias
 - 交絡 confounding

研究における誤差

選択バイアス

- 研究対象者を選定する際に生じるバイアス
(=標本の性質に関わるバイアス)
 - 標本抽出バイアス **sampling bias**
 - 抽出した参加者は母集団を代表していないかもしれない
 - 自己選択バイアス **self-selection bias** /
参加バイアス **participation bias**
 - 研究に参加してくれるのは特殊な人かもしれない
 - 特定の人々は研究に参加してくれないかもしれない

研究における誤差

情報バイアス

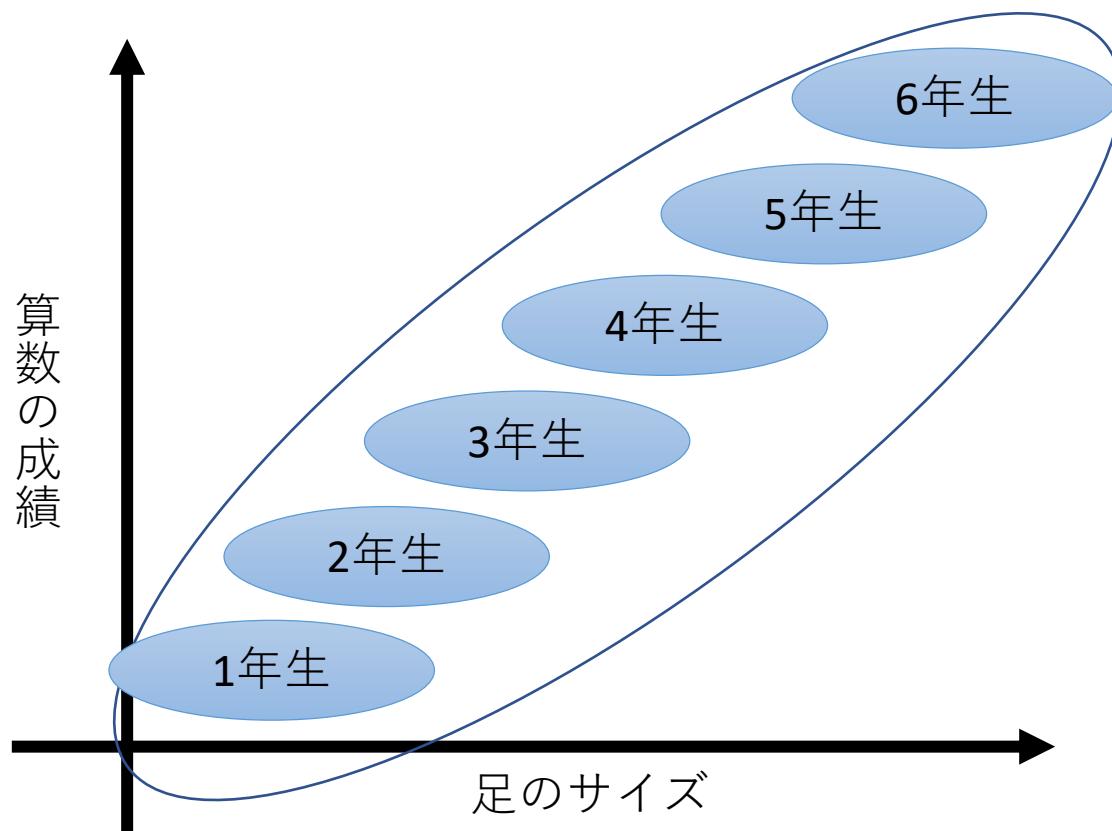
- 研究対象者からデータを得る際に生じるバイアス
（=測定に関わるバイアス）
 - 実験者効果 experimenter effect / 要求特性 demand characteristics
 - 実験者が仮説を知っていると、仮説を支持する方向に（意図せずに）参加者を誘導してしまう (Rosenthal, 1966)
 - 実験参加者は、実験者の期待する回答をしてしまう傾向がある (Orne, 1962)
 - 社会的望ましさ social desirability / 報告バイアス reporting bias
 - 望ましくない回答は抑制される
 - 想起バイアス recall bias
 - 群によって記憶の正確さが違うかもしれない
 - 誤分類 misclassification
 - 測定の精度や方向性が群間で異なると問題になることがある

研究における誤差

交絡

- ・他の変数の効果が混ざってしまうために生じるバイアス
- ・第三の変数の影響を受けること
- ・例：足の大きさと成績
 - ・「ある小学校の児童全体を対象に、算数の能力テストと足のサイズの計測を行った。その結果、強い有意な正の相関がみられた。ゆえに足の大きさと算数の能力は関係しているといえる」
⇒ ? ? ? ?

考えられる可能性



⇒ 「学年」が「成績」と「足のサイズ」両方に影響を与えている

研究における誤差

交絡への対応

- 研究デザインで対応
 - 無作為割り当てる (実験)
 - 交絡要因が実験群と対象群で異なるよう群を無作為に割り当てる
 - 交絡要因が同じと思われる集団を対象に分析する
 - e.g. 職業コホート
- 分析で対応
 - 層化 stratification して分析する
 - 交絡をもたらすと思われる変数(e.g. 年齢)によって層に分け、グループごとに分析する

実験の特徴

- 実験 experiment
 - pros
 - 自由度が高い
 - 条件の無作為割り当てができる
 - 交絡を抑えられる
 - 因果関係を議論できる
 - 様々な外的要因を統制できる
 - cons
 - コストが高い
 - 参加者を確保する
 - 謝金・謝礼が必要
 - 多人数を対象にするのには向かない

- 生理指標・行動指標
 - 信頼のおける指標
 - コストが高い
 - 一度にたくさんの変数を測るのが難しい
- 質問紙実験
 - 容易に実施可能
 - 一度にたくさんの変数が測れる
 - 結果が細かいワーディングに左右される
 - 態度と行動は一貫しない(LaPiere, 1934)
→行動データではない
 - 実際の人間の振舞いを表しているとは限らない
→妥当性を担保する工夫が必要

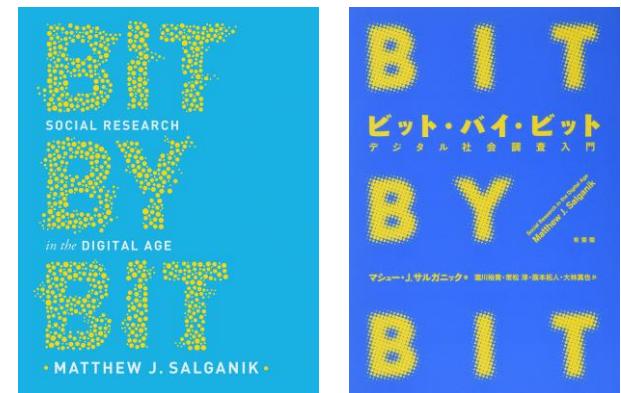
調査の特徴

- 調査 survey
 - pros
 - 標本が特定の母集団を代表するように設計できる
 - "一般の"人々の回答を得られる
 - 多人数を対象にできる
 - cons
 - 交絡がある
 - コストが高い
 - 回答率が低い場合には代表性が失われる
 - 基本的には一時点の調査では因果関係を議論できない

- 社会調査
 - 何らかの抽出台帳(e.g. 住民基本台帳)から無作為抽出した人々を対象に調査票に回答してもらう
 - 訪問留置法・郵送法など様々な手段があるが一般に高コスト
 - 回答率を確保することが難しい
- Web調査
 - 調査会社に依頼し、企業の確保・整備しているアンケートモニターを対象に収集してもらう
 - 性別・地域等が柔軟に設計できる
 - 回答率が高い
 - 調査の中では比較的低コスト
 - 選択バイアスを考慮すべき
 - 質問紙よりも回答者が適当に答えがちな傾向がある
- その他
 - パネル調査
 - 同じ回答者に期間をおいて何度も調査に参加してもらう
 - コホート研究
 - 特定の集団を長期間追跡

ビッグデータの特徴

- ビッグデータの10の特徴(Salganik, 2017)
 - 研究にとって有益な特徴
 - 巨大さ
 - 常時オン
 - 非反応性
 - 問題となる特徴
 - 不完全性
 - アクセス不能性
 - 非代表性
 - ドリフト
 - アルゴリズムによる交絡
 - 汚染
 - センシティブ



Salganik (2017) *Bit by bit*

ビッグデータの特徴①

巨大さ

- データセットが巨大であること有益な研究
 - まれなできごとの研究
 - 不均質性(**heterogeneity**)の研究
 - 実験の処理の効果の違い
 - 地域の特性の違い(e.g. Chetty et al., 2014)
 - 微小な差異の検出
 - 1%の差異が意味を持つ分野もある
- 落とし穴
 - 系統誤差に注意しなければならない
 - 偶然誤差は減っても系統誤差は減らない

ビッグデータの特徴②

常時オン

- 絶えずデータを収集
 - 時系列データを取ることができる
 - 予期せぬ出来事の研究
 - 歴史的なできごと・事件
 - リアルタイム推定が可能になる

ビッグデータの特徴③

非反応性

- 社会科学における「反応性 reactivity」
 - 人は観察されると行動を変える(Webb, 1966)
 - 実験者効果
- オンラインのデータ
 - データをとられることを人々が通常意識していないという意味で、非反応的
 - 落とし穴
 - 非反応的であるからといって、そのままの態度や行動を表しているわけではない
 - 社会的望ましさなどといった要因の影響はなお残る

ビッグデータの特徴④

不完全性

- 欲しい情報が入っていない
- 研究上の構成概念と対応するか
 - 構成概念妥当性

ビッグデータの特徴⑤

アクセス不能性

- データが存在しても研究者がアクセスできるとは限らない
 - 政府や自治体、企業の中にあるデータ

ビッグデータの特徴⑥

非代表性

- ビッグデータの多くは非代表的
→母集団を代表してはいない
 - 研究結果を一般化できるか？

ビッグデータの特徴⑦

ドリフト

- ドリフト(浮動)
 - 時間にともなうシステムの変化
 - どのようなシステムか
 - 誰が使うのか
 - どのように使うのか

ビッグデータの特徴⑧

アルゴリズムによる交絡

- システム上の行動：人間のありのままの行動ではない
 - システム設計者の企図によって人工的な結果 (artifact)が生じる
 - Ugander(2011): Facebookにおけるネットワーク
 - 友達の人数は「20」が突出して多い
 - 友人を20人になるまで増やすようシステムがうながす仕組みがある
 - 「友達の友達」同士は友達になりやすい
 - 社会ネットワークにおいては推移性 **transitivity** として知られる現象
 - 社会理論を知っている設計者がシステムに理論を組み込んでいる(遂行性 **performativity**)

ビッグデータの特徴⑨

汚染

- スパムやボットなど、人間の行動を反映しないデータが紛れ込んでいる
 - Back, Kühner, & 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の投稿を取り除くとBackらの結果は再現されない
→人工的結果(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, Kühner, & Egloff, 2011)

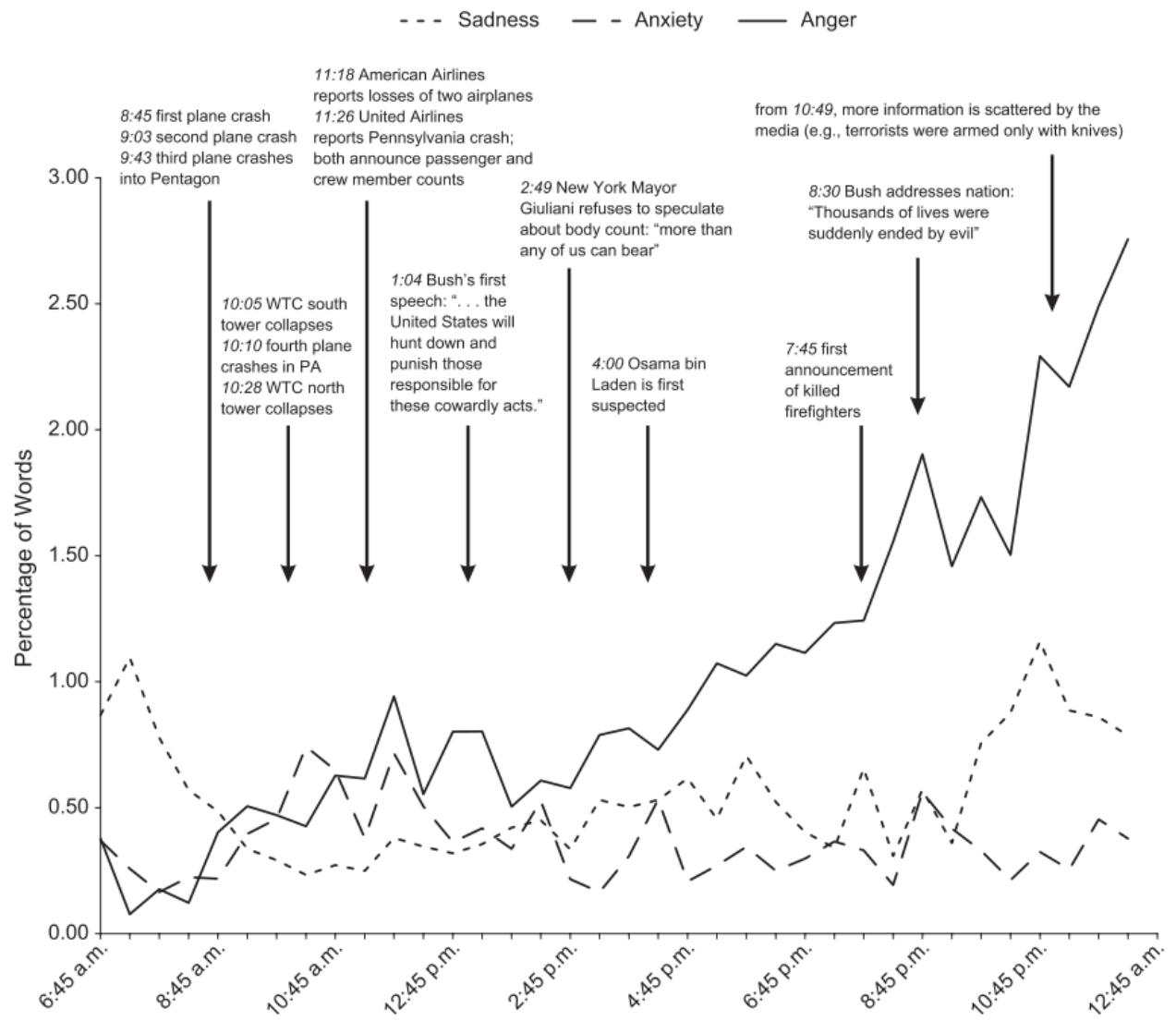
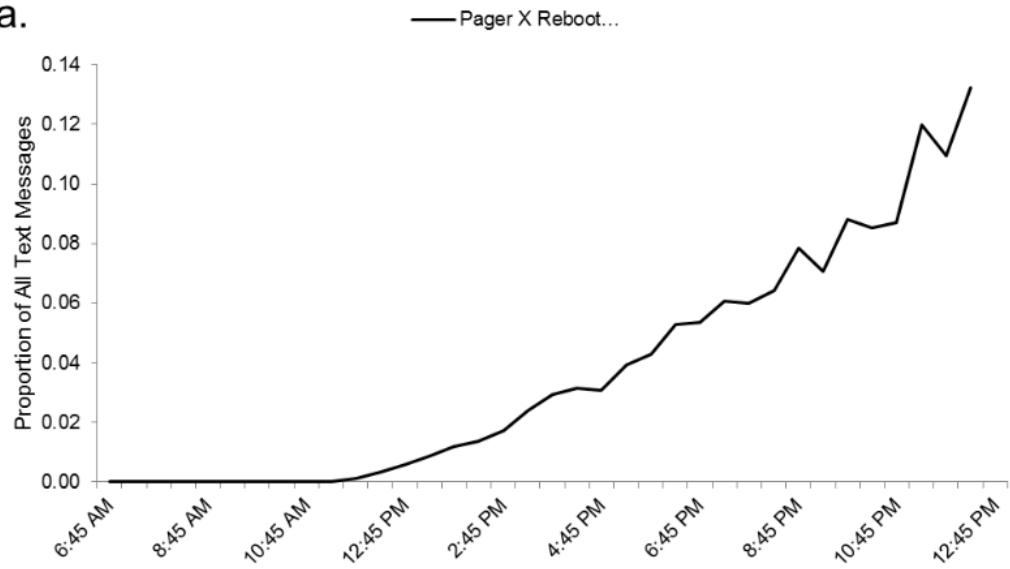


Fig. 1. The timeline of sadness, anxiety, and anger on September 11 as expressed in messages sent to text pagers. Each data point represents the mean percentage of words related to the specific negative emotion, averaged across 30 min. The time slots start at 6:45 a.m. to 7:14 a.m. on September 11, 2001, 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 timelines. WTC = World Trade Center

a.



b.

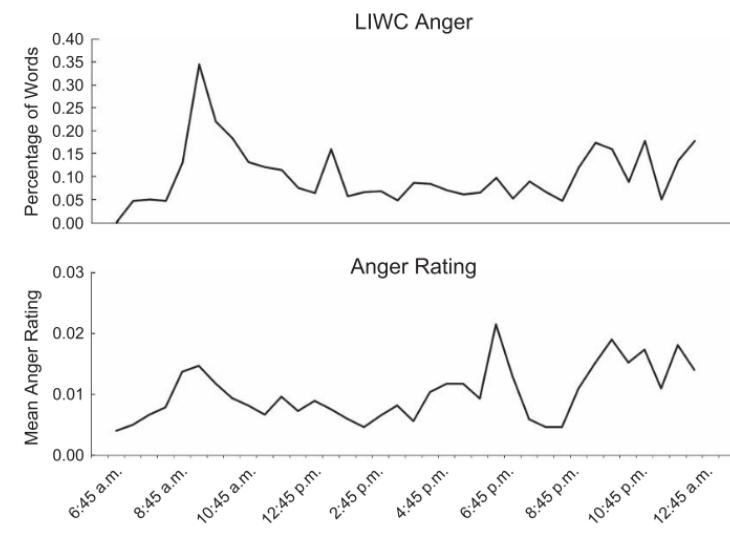
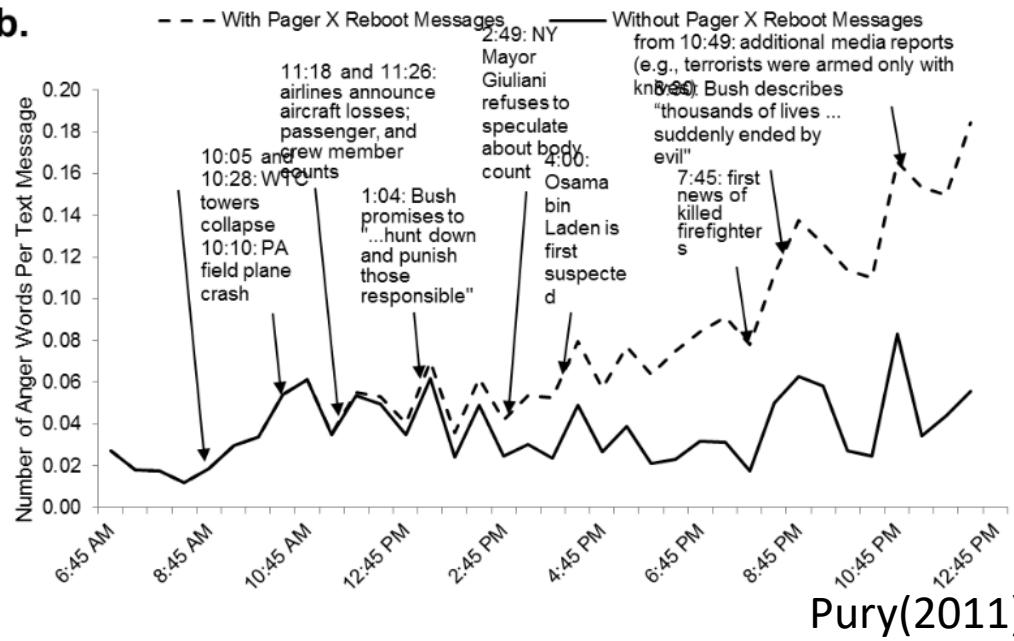


Fig. 1. A revised timeline of anger as expressed in 37,606 social messages sent to text pagers on September 11, 2001. The graphs show (a) the mean percentage of words related to anger (as classified by Linguistic Inquiry and Word Count; Pennebaker, Francis, & Booth, 2001) and (b) the mean anger rating (0 = no anger, 1 = some anger, 2 = strong anger; averaged across three raters for each message) across time slots starting at 6:45 a.m. to 7:14 a.m. on September 11, 2001, and ending at 12:15 a.m. to 12:44 a.m. on September 12, 2001.

Back, Kühner, & Egloff (2011)

ビッグデータの特徴⑩

センシティブ

- 個人のセンシティブな情報が含まれる
 - 複数のデータをつなげることで個人が特定できてしまうかもしれない

テキストデータの収集法

- 新たに入手する
 - 実験や調査の実施
- 既存のデータの活用・Web上での収集
 - 文書・書籍
 - 公開データ
 - データを持っている団体との接触
 - Web API
 - スクレイピング

テキストデータの収集法

実験・調査による収集

- 実験・調査
 - 文章データ
 - 記述式・自由回答
 - 日記法
 - 音声・映像データ
 - 面接（インタビュー）
 - 討議などの録音
- 注意点
 - 事前に電子化する方法を考えておく
 - 必ずしも取りたいデータが取れるとは限らないので、あくまで副次的な方法として考えておく

テキストデータの収集法

文書・書籍からの収集

- 手動入力
 - 手間がかかる
- OCRによるスキャン
 - 光学的自動文字認識 Optical Character Recognition
 - 日本語の文章は弱い
- 注意点
 - 版による違いがある
 - 入力チェックがある分、データクリーニングに時間がかかる

テキストデータの収集法

公開データの利用

- 公開されているデータセットの例
 - 各種オープンデータ
 - 国・地方公共団体・官公庁のオープンデータ
 - 研究用データセット
 - 各種言語資料（コーパス）
 - 情報学研究データリポジトリ
<https://www.nii.ac.jp/dsc/idr/datalist.html>
 - パブリックドメインの文学作品
 - Project Gutenberg
 - 青空文庫
 - その他
 - Wikipedia
 - Kaggle

テキストデータの収集法

データを持っている団体との接触

- 国や地方公共団体・企業
 - 様々なデータを持っている
 - その多くは通常アクセスできない
- アクセスできる可能性：ゼロではない
 - お願いしてみる
 - 共同研究

テキストデータの収集法

Web APIを用いた収集

- API: Application Programmable Interface
 - 他のプログラムからアクセスするために提供されているツール群
- Webサービスの中にはAPIを通じて様々な情報を取得できるものがある
 - Twitter
 - Instagram
 - Facebook

テキストデータの収集法

スクレイピング

- スクレイピング(scraping: こそげ落とす)
 - Webページを取得し、意味のある情報を抽出する
 - すべてのWebページにAPIが用意されているわけではない
→ダウンロード・加工してデータに
 - Webページ：HTMLで記述されている
 - 様々なツールがある

テキストデータの収集法

データ取得時の注意

- 公式の取得法があればそれを使う

クローラを使わない [編集]

記事を大量にダウンロードするためにクローラを使わないで下さい。強引なクローリングは、ウィキペディアが劇的に遅くなる原因となります。

ウィキペディアのデータベースから自動的にデータの収集がなされた場合、システム管理者によってあなたのサイトからウィキペディアへのアクセスを禁止する措置が取られることもあります。またWikimedia Foundationが法的措置を検討することもあります。

<https://ja.wikipedia.org/wiki/Wikipedia:データベースのダウンロード>

- 取得先に過度の負荷をかけないようにする
 - 短期間・高頻度にアクセスすると攻撃とみなされるかもしない

どういう分析をするか

- 質的なデータ（自然言語）を量的なデータに変換する
 - 頻度
 - 分布
 - 各種指標・統計量

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

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

どういう分析をするか

指標・統計量の例

- TF-IDF (term-frequency / inverse document frequency)
 - 文書における単語の重要度
- 類似度
 - 特徴ベクトル間の類似度
 - Pearsonの積率相関・Spearmanの順位相関・コサイン類似度
 - 集合同士の類似度
 - Jaccard係数
 - 文字列同士の類似度
 - Levenshtein距離（編集距離）
- 相互情報量 mutual information
 - 共起の重要度
- TTR (token type ratio)
 - 延べ語数・異なり度数。語彙の多様性
- Simpson's D
 - 繰り返し表現の多さ
- 各種スコア (e.g. 感情分析)

どういう分析をするか

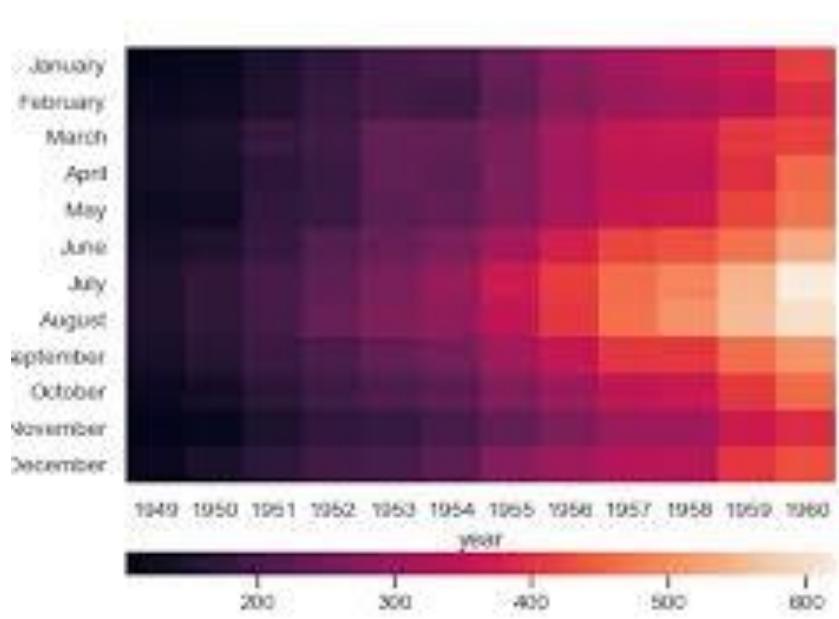
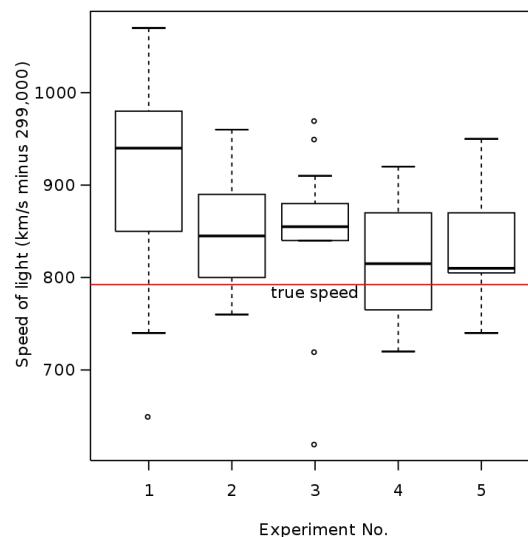
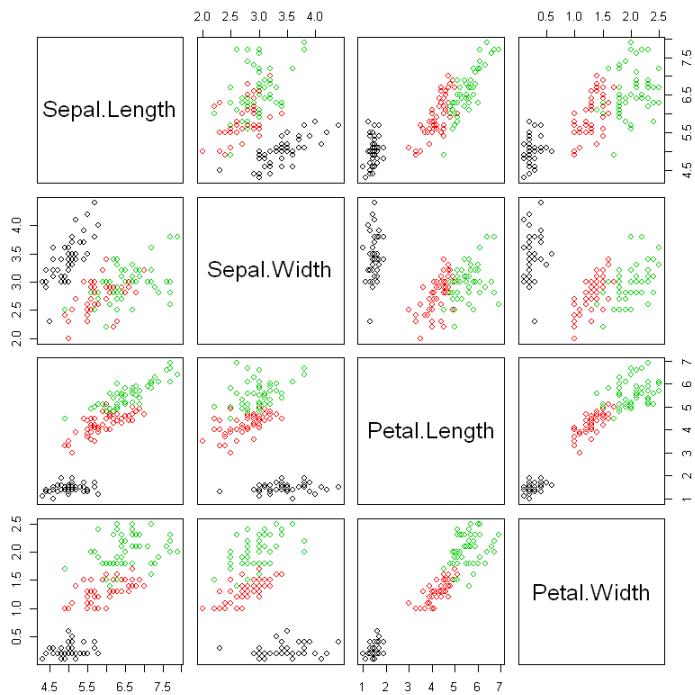
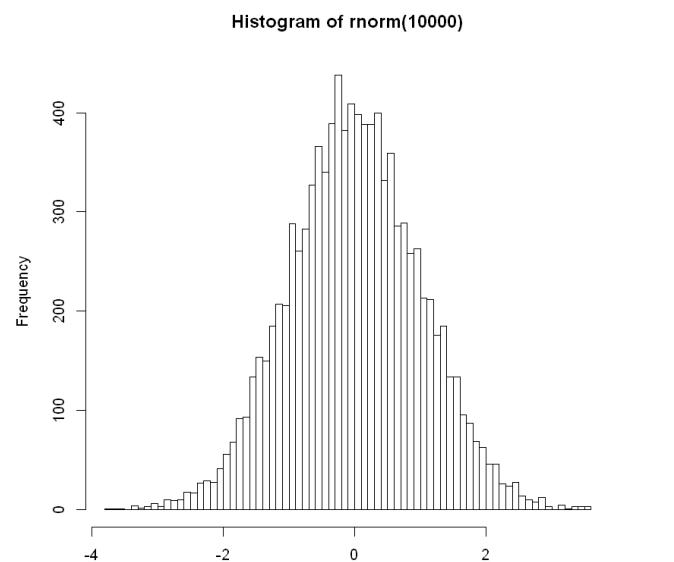
分析手法

- ①図示する
 - 各種グラフ
 - ネットワークグラフ
- ②比べる
 - カイ二乗検定：分布間の比較
 - 尤度比検定：頻度の比較
- ③まとめる
 - クラスター分析
 - 次元削減：主成分分析/因子分析
- ④分類する
 - 潜在意味解析/トピックモデル
 - 感情分析
 - ニューラルネット

どういう分析をするか

①図示する

- 可視化することで全体のパターンを把握する
 - 各種グラフ
 - ヒストグラム
 - 箱ひげ図
 - 散布図行列
 - ヒートマップ
 - ネットワークグラフ



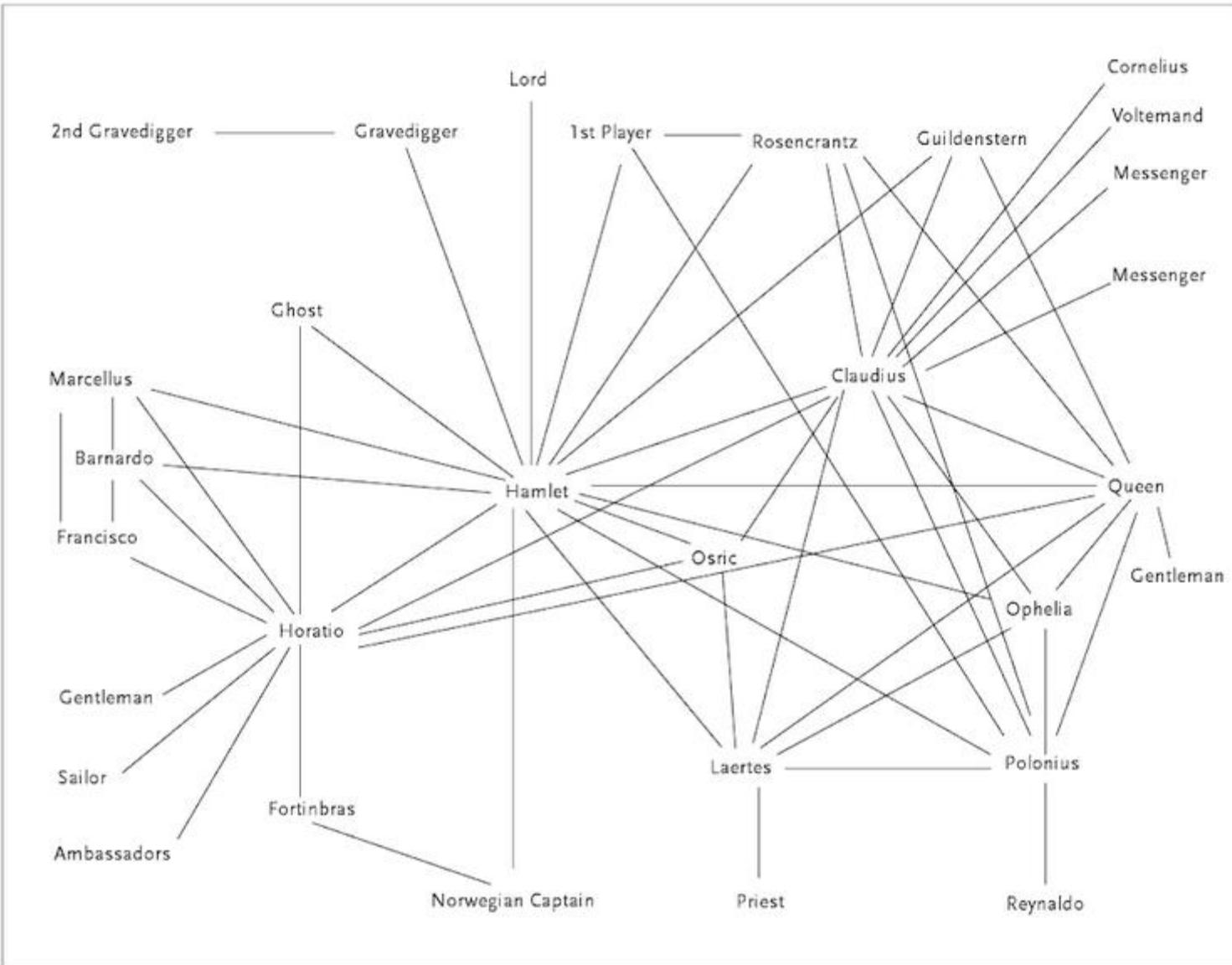


Figure 1: The Hamlet network

Moretti, F. (2013). *Distant reading*.

どういう分析をするか

② 比べる

- 頻度データ
 - 分布の比較： χ^2 二乗検定
 - 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
 - 「母集団」に相当するものはなにか
 - 記述統計で十分な場合もある
 - 検定の多重性

どういう分析をするか

★多重検定 multiple-test

- 検定を繰り返すことによって、研究全体の第Ⅰ種の過誤の確率が増大してしまうこと
 - 第Ⅰ種の過誤 Type I error / α error
 - 本当は有意な差がないのに有意な差があると判断してしまう
 - 第Ⅱ種の過誤 Type II error / β error
 - 本当は有意な差があるのに有意な差がないと判断してしまう
- 真の差がなくても、たくさん検定をすれば「どこか」では統計的に有意な結果ができる
 - 擬陽性 false positive な結果
 - 再現性のない結果を量産する一因
 - p-hacking: 納得のいく結果が出るまで「試行錯誤」を続ける
 - cherry-picking: 都合のいい結果だけを報告する
 - いずれも科学的には無意味な結果

- 有意確率
 - e.g. 「有意水準0.05で有意な差がみられた」
 - ○帰無仮説のもとでそのデータが得られる確率
 - ×帰無仮説が正しい確率
- 有意水準 α の検定を n 回繰り返した場合に、真の差がなくても有意だと判定される結果を含む確率

$$\alpha_{total} = 1 - (1 - \alpha)^n$$

p	n	total alpha	p	n	total alpha
0.05	1	0.050	0.01	1	0.010
	2	0.098		2	0.020
	5	0.226		5	0.049
	10	0.401		10	0.096
	100	0.994		100	0.634

どういう分析をするか

★多重検定の対策

- 個別の分析では
 - 多重性を考慮した分析をする
 - 多重比較 multiple comparison (e.g., Tukey's HSD)
 - Bonferroniの補正
- 研究全体では
 - 事前にどういう分析をするか決めておく
 - 研究をpreregisterする
 - 行った分析をすべて記述する
 - 様々な角度から結果の妥当性を検証する
 - その差は実質科学的に意味のある差なのかをチェックする
 - e.g. 効果量のチェック
 - 当然あるべき関連をチェックする
 - 「AということがいえるならBという結果も得られるはずだ」
⇒基準関連妥当性
 - 交差検証をする：分析用と検証用にデータに分割する
 - 追試をする：再現できることを確認する

どういう分析をするか

③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
- 似たような性質を持つ変数 (=列) をまとめたい
 - 次元削減
 - 主成分分析
 - 因子分析

どういう分析をするか

クラスター分析

- クラスター分析 cluster analysis
 - データ間の「距離」または「類似度」をもとに、データの集まり（クラスタ）を抽出する分析手法
 - 階層的手法
 - 距離の近いデータ同士からボトムアップにクラスタを統合していく
 - 欠点：データが多いと計算時間が膨大になる
 - 非階層的手法
 - K-means法
 - 計算時間が比較的少なくて済む
 - 欠点：一意に定まらない
 - 変数選択の問題：どの変数を使うか
 - みにくいアヒルの子の定理(Watanabe, 1969)：変数を増やすとどれも同じ程度似てしまう

どういう分析をするか

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 **principal component analysis; PCA**
 - 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
 - データの分散をもっともよく説明する軸（第1主成分）から順に直行するように軸を抜き出していく
 - 因子分析 **factor analysis**
 - 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」)
 - 全体に共通する因子 + 誤差、というモデル
 - 主成分分析とは想定するモデルが異なる
 - 因子回転の手法・解釈に任意性がある

どういう分析をするか

④分類する

- **機械学習 machine learning**
 - 機械が分類や予測などのタスクの成績をデータをもとに（自動的に）改善していく技術
- 教師あり学習：データとともに分類ラベルを与えて学習
 - 感情分析
 - 決定木分析
 - ナイーブベイズ
 - サポートベクターマシン(SVM)
- 教師なし学習：データのみから学習
 - 主成分分析
 - クラスター分析
 - 潜在意味解析

どういう分析をするか

感情分析

- 感情分析 **sentiment analysis**

- 極性語と呼ばれる単語をもとにスコアを算出
- 単純に計算する場合と、少数のデータに学習させてタグ付けする場合がある。
- 心理学的な妥当性は疑問(Basely and Mason, 2005; Panger, 2016)

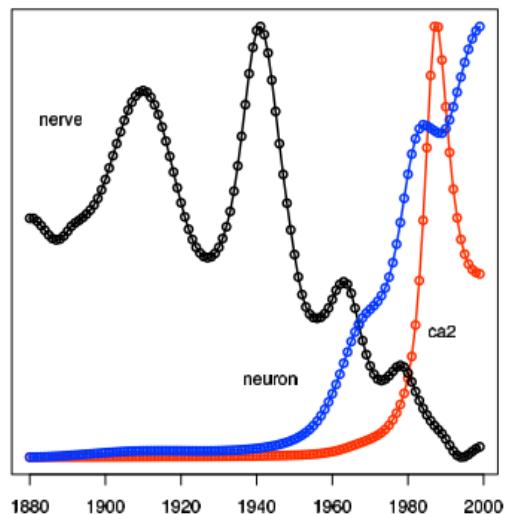
どういう分析をするか

★潜在意味解析

- Latent Semantic Analysis(LSA) / Latent Semantic Indexing(LSI)
 - 単語の生起頻度をもとに似た意味を持つ単語のグループを抽出
 - ざっくりいえば単語を主成分分析にかけるようなもの
- トピックモデル topic model
 - より統計的に洗練された手法
 - 文書群の背景にある「トピック」と、各文書がどれくらいそのトピックに該当するかを同時に推定

1881 brain movement action right eye hand left muscle nerve sound	1890 movement eye right hand brain left action muscle sound muscle sound experiment	1900 brain eye movement sound nerve active muscle left hand nerve vision sound	1910 movement brain sound nerve active nerve stimulate muscle left eye right nervous	1920 movement sound muscle active nerve stimulate fiber reaction brain response	1930 stimulate muscle sound movement response nerve frequency fiber active brain	1940 record nerve stimulate response muscle electrode active brain fiber potential	1950 respons record stimulate nerve muscle active frequency electrode potential study	1960 response stimulate record condition nerve muscle active potential stimulus nerve subject eye	1970 respons cell potential stimul neuron active nerve eye record abstract	1980 cell channel neuron response active brain stimul muscle system nerve receptor	1990 cell channel neuron ca2 active brain receptor muscle respons current	2000 neuron active brain cell fig response channel receptor synapse signal
---	---	--	--	---	--	--	---	---	--	---	---	--

"Neuroscience"



- 1887 Mental Science
 1900 Hemianopsia in Migraine
 1912 A Defence of the ``New Phrenology''
 1921 The Synchronal Flashing of Fireflies
 1932 Myoesthesia and Imageless Thought
 1943 Acetylcholine and the Physiology of the Nervous System
 1952 Brain Waves and Unit Discharge in Cerebral Cortex
 1963 Errorless Discrimination Learning in the Pigeon
 1974 Temporal Summation of Light by a Vertebrate Visual Receptor
 1983 Hysteresis in the Force-Calcium Relation in Muscle
 1993 GABA-Activated Chloride Channels in Secretory Nerve Endings

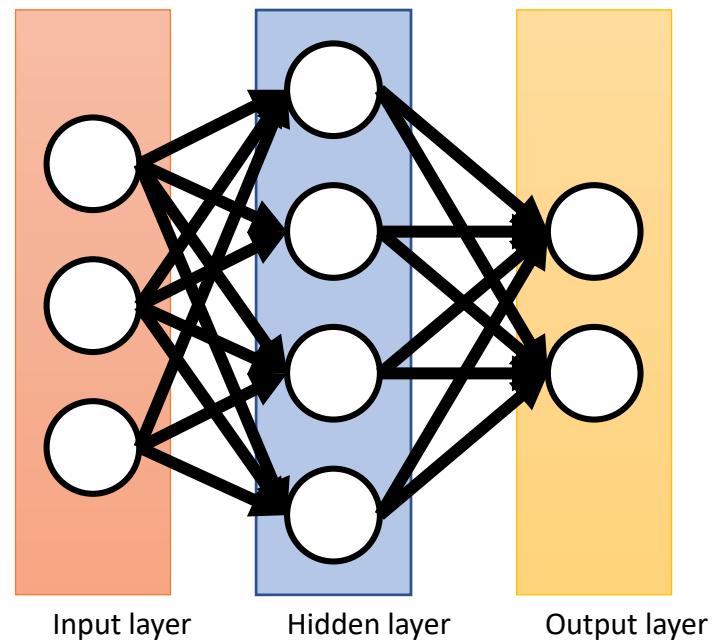
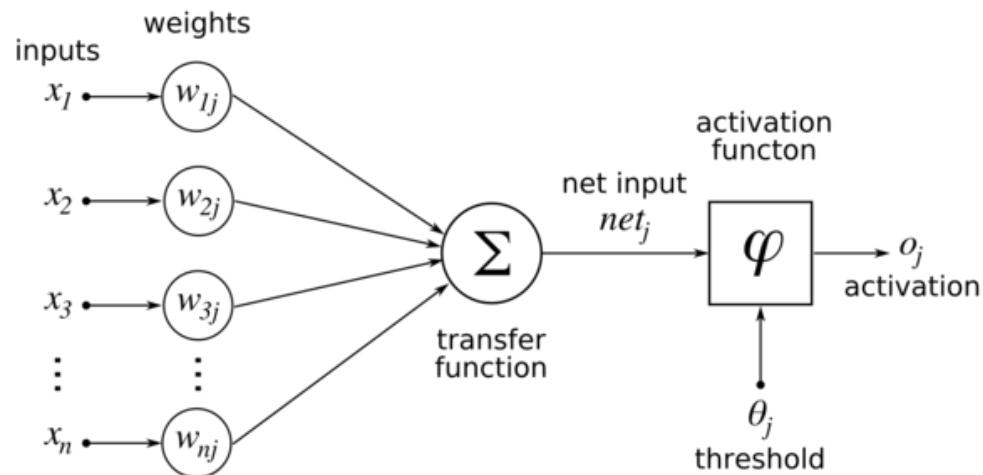
Figure 4. Examples from the posterior analysis of a 20-topic dynamic model estimated from the *Science* corpus. For two topics, we illustrate: (a) the top ten words from the inferred posterior distribution at ten year lags (b) the posterior estimate of the frequency as a function of year of several words from the same two topics (c) example articles throughout the collection which exhibit these topics. Note that the plots are scaled to give an idea of the shape of the trajectory of the words' posterior probability (i.e., comparisons across words are not meaningful).

Blei, D. M., & Lafferty, J. D. (2006). Dynamic Topic Models. In *Proceedings of the 23rd international Conference on Machine Learning*, 113–120)

どういう分析をするか

ニューラルネット

- ニューロンを模した学習器を多数組み合わせて学習させる
 - 分類や生成、様々なタスクに応用できる



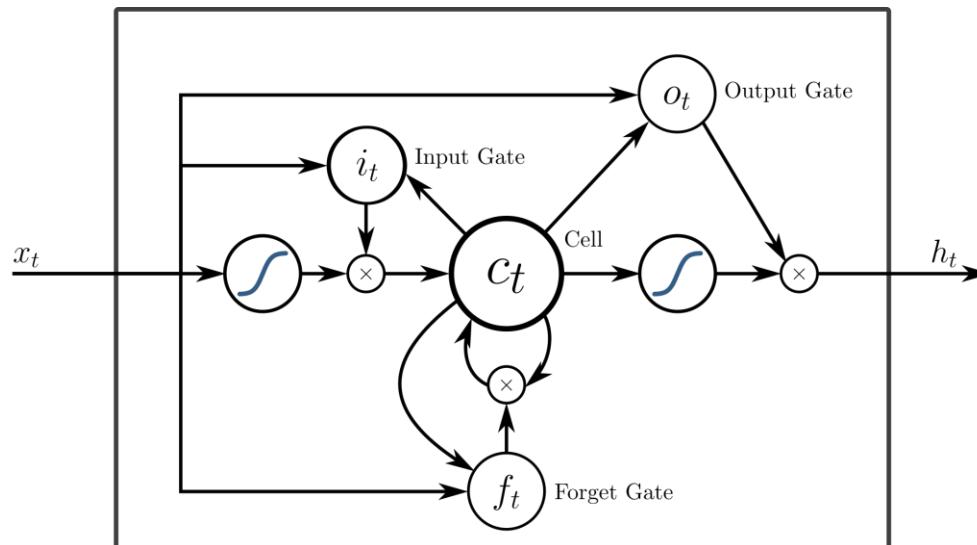
★ニューラルネット小史

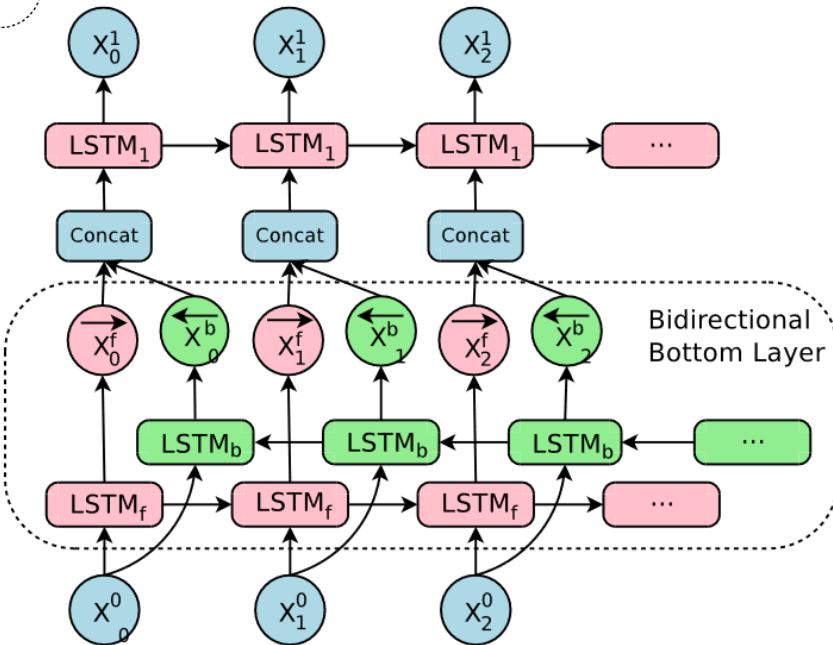
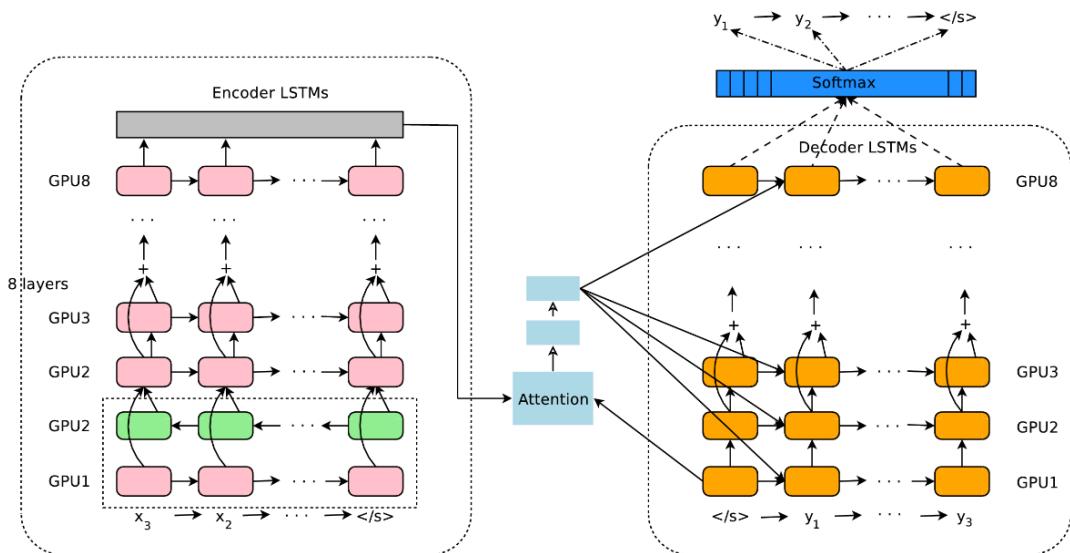
- McCulloch & Pitts(1943): ニューロンのモデル
- Hebb(1949): シナプス結合の変化則
- Rosenblatt(1958): パーセプトロン perceptron
- Minsky & Papert(1963): パーセプトロンの限界
 - AI 冬の時代 第1期(1974-1980)-
- Hopfield(1984): Hopfield net 相互結合型ネットワーク
- Hinton & Sejnowski(1985): 確率的相互結合型ネットワーク
- Rumelhart, Hinton & Williams(1986): 誤差逆伝播法
 - パーセプトロンの多層化
 - AI 冬の時代 第2期 (1987-1993)-
- Hinton & Salakhudinov(2006): AutoEncoder
- Krizhevsky, Sutskever & Hinton(2012): ImageNet
 - 深層学習の時代へ

★LSTM

- LSTM: Long-Short Term Memory

- 可変長の引数を扱える→文章などの データを扱える
- 複数のLSTMを組み合わせて自動翻訳を強化





Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. <http://arxiv.org/abs/1609.08144>

★word2vec

- 単語の分散表現（単語ベクトル）を学習
 - 分散仮説(Firth, 1957)
 - 「言葉の意味は周辺の語彙によって決まる」
 - 学習モデル
 - CBoW
 - Skip-gram
- 単語の加減算ができる

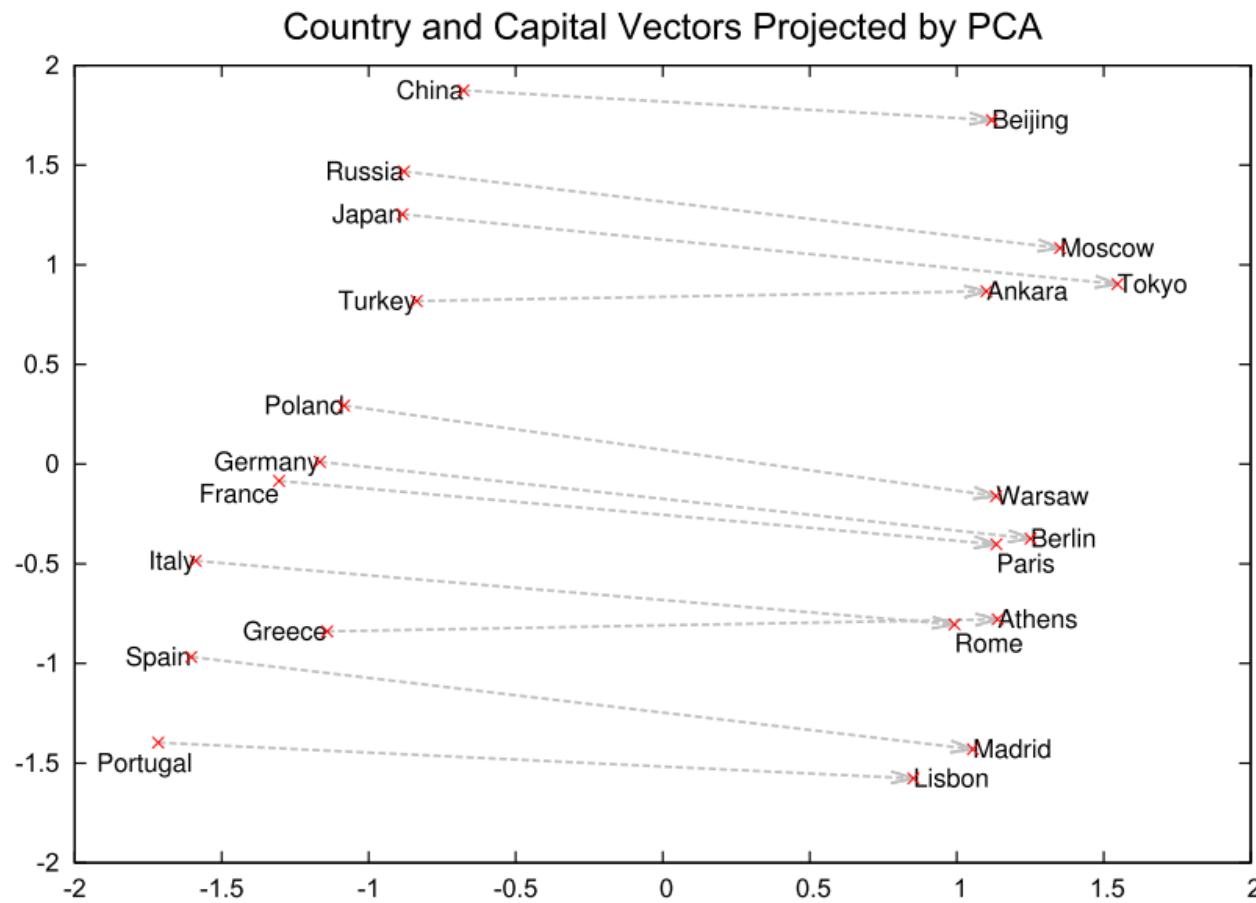


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

どういう分析をするか

分析のまとめ

- テキストを分析する選択肢は多い
 - すべてを理解する必要はない
 - 「こういう技術がある」として知っておくとよい
- 大量の変数, 大量の分析
 - 自由度が高すぎるために何をしていいかわからない
 - たくさん検定をやればどこかには統計的に「有意な」結果が存在してしまう → 多重検定
 - 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
 - 何が知りたいことなのか、研究デザインの段階でよく考えておく

研究の各種制約

- 倫理
 - 通常の心理学研究の倫理綱領にもとづく(cf. 日本心理学会, 2011)
 - ビッグデータの場合はさらに注意が必要
 - 個人情報の管理は適正に
- コスト
 - 予算
 - インセンティブを使えるか（実験・調査）
 - 時間
 - 収集にかかる時間
 - 収集後の処理にかかる時間
 - データ入力
 - クリーニング・前処理
 - PCの計算時間
- その他ロジスティクス上の問題
 - 使用できるPCの性能
 - CPU
 - メモリ
 - 利用できるデータ容量 etc.

小まとめ

- 実験・調査に限らず様々な方法でテキストデータを収集することができる
 - オンラインで得られるデータは通常の実験・調査とは異なる種類の性質がある
 - 落とし穴にはまらないよう、研究デザインをしっかり立てる
- 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
 - 行動データではない→何を測っているのか常に意識する
 - どうすれば測りたいことを測れるのかを考える

References

- Abello, J., Broadwell, P., & Tangherlini, T. R. (2012). Computational folkloristics. *Communications of the ACM*, 55(7), 60–70. <https://doi.org/10.1145/2209249.2209267>
- Back, M. D., Küfner, A. C. P., & Egloff, B. (2011). "Automatic or the people?" Anger on september 11, 2001, and lessons learned for the analysis of large digital data sets. *Psychological Science*, 22(6), 837–838. <https://doi.org/10.1177/0956797611409592>
- Back, M. D., Küfner, A. C. P., & Egloff, B. (2010). The Emotional timeline of September 11, 2001. *Psychological Science*, 21(10), 1417–1419. <https://doi.org/10.1177/0956797610382124>
- Beasley, A., & Mason, W. (2015). Emotional states vs. emotional words in social media. In *Proceedings of the ACM Web Science Conference on ZZZ - WebSci '15* (pp. 1–10). New York, New York, USA: ACM Press. <https://doi.org/10.1145/2786451.2786473>
- Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international Conference on Machine Learning* (pp. 113–120). <https://doi.org/10.1145/1143844.1143859>
- Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? the geography of intergenerational mobility in the United States, 129(November), 1553–1623.
- Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955" in studies in linguistic analysis. *The Philological Society*.
- Grimes, D. A., & Schulz, K. F. (2002). Bias and causal associations in observational research. *The Lancet*, 359, 248–252.

- Healy, K. (2015). The performativity of networks. *Archives Européennes de Sociologie*, 56(2), 175–205. <https://doi.org/10.1017/S0003975615000107>
- Hebb, D. (1949). *The organization of behavior: : A neuropsychological theory*. New York: Wiley & Sons.
- King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, 107(2), 326–343. <https://doi.org/10.1086/667000>
- LaPiere, R. (1934). Attitudes vs. Actions. *Social Forces*, 13(2), 230–237.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115–133.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176–182. <https://doi.org/10.1126/science.1199644>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*
- Minsky, M., & Papert, S. (1969). Perceptron: an introduction to computational geometry. *The MIT Press, Cambridge, Expanded Edition*, 19(88), 2.
- Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17(11), 776.
- Panger, G. (2016). Reassessing the Facebook experiment: critical thinking about the validity of Big Data research. *Information Communication and Society*, 19(8), 1108–1126. <https://doi.org/10.1080/1369118X.2015.1093525>
- Pury, C. L. S. (2011). Automation can lead to confounds in text analysis: Back, Küfner, and Egloff (2010) and the not-so-angry Americans. *Psychological Science*, 22(6), 835–836. <https://doi.org/10.1177/0956797611408735>
- 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. *EPJ Data Science*, 5(1), 31. <https://doi.org/10.1140/epjds/s13688-016-0093-1>

- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408.
<https://doi.org/10.1037/h0042519>
- Rosenthal, R. (1966). Experimenter effects in behavioral research.
- Rothman, K. J. (2012). *Epidemiology: an introduction*. Oxford university press.
- Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*, 32(1–2), 51–63. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/447779>
- Salganik, M. J. (2017). *Bit by bit: social research in the digital age*. Princeton University Press.
- Tangherlini, T. R. (2016). Big folklore: A Special issue on computational folkloristics. *Journal of American Folklore*, 129(511), 5–14. Retrieved from
<http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=113224879&site=ehost-live>
- Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.
- Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The anatomy of the Facebook Social Graph, 1–17. Retrieved from <http://arxiv.org/abs/1111.4503>
- Watanabe, S. (1969). Knowing and guessing a quantitative study of inference and information.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). Unobtrusive measures: Nonreactive research in the social sciences.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). Google's Neural Machine Translation System: Bridging the gap between human and machine translation. Retrieved from <http://arxiv.org/abs/1609.08144>
- 村上征勝. (2002). 文化を計る 文化計量学序説. 朝倉書店.
- 橋口耕一. (2006). 内容分析から計量テキスト分析へ--継承と発展をめざして. 大阪大学大学院人間科学研究科紀要, 32, 1–27. <https://doi.org/info:doi/10.18910/11920>
- 橋口耕一. (2014). 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. ナカニシヤ出版.
- 橋口耕一. (2018). 計量テキスト分析およびKH Coderの利用状況と展望. 社会学評論, 68(3), 334–350.

第2部 実習：Rを使ったテキスト分析

第2部のアウトライン

- 前半
 - 実習：データ前処理
 - 前処理の流れ
 - Rによる処理の基礎
 - クリーニングの実際
 - データハンドリング入門
 - 単語の頻度
 - キーワード抽出
- 後半
 - 実習：データ分析
 - 主成分分析
 - ネットワークグラフによる共起の可視化
 - クラスター分析（時間があれば）
 - 結果の報告と解釈
 - 結果の書き方に関するヒント
 - クロージング
 - 講義全体のまとめ
 - 課題について

第2部 実習①データ前処理編

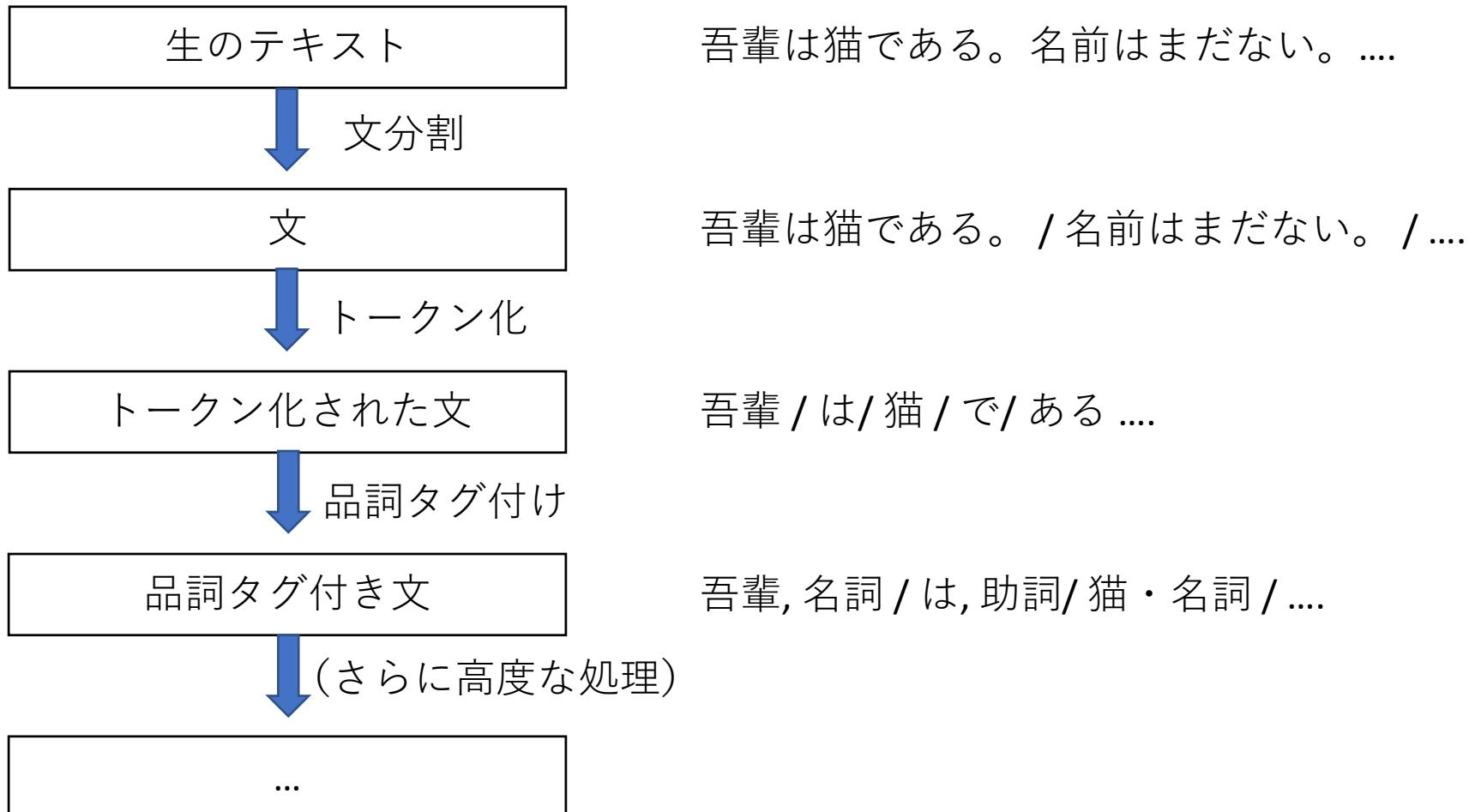
なぜ前処理が必要か

- 通例心理学で扱うようなデータ
 - →きれいに整形されている（べき）
 - 表形式・構造化データ
 - ノイズは少ない
 - 変数は少ない
 - すぐに分析できる
- テキストデータ
 - →dirty
 - 整形されていない・非構造化データ
 - ノイズがたくさん
 - 変数が膨大
 - そのままでは扱えない

★構造化データ

- 整然データ **tidy data** (Wickham, 2014)
 - 個々の変数を1つの列
 - 個々の観測を1つの行
 - 個々の観測の構成単位の類型が1つの表
 - 整然ではないデータ
 - 列見出しが、値であって変数名でない
 - 複数の変数が、1つの列に格納されている
 - 変数が、行と列の両方に格納されている
 - 観測の構成単位の類型が、同じ表に複数格納されている
 - 1つの観測の構成単位が、複数の表に格納されている
- DBデータ
 - Relational Database (RDB)

情報抽出アーキテクチャの例



生のテキスト（平テキスト）

吾輩は猫である。名前はまだ無い。

どこで生れたかとんと見当がつかぬ。何でも薄暗いじめじめした所でニヤーニヤー泣いていた事だけは記憶している。吾輩はここで始めて人間というもののを見た。しかもあとで聞くとそれは書生という人間中で一番獰惡な種族であったそうだ。

…

——夏目漱石『吾輩は猫である』

- 何のタグもつけられていないテキスト
- このままでは分析できない

クリーニングと前処理

文に分割

吾輩は猫である。 / 名前はまだ無い。 / どこで生れたかとんと見当がつかぬ。 / 何でも薄暗いじめじめした所でニヤーニヤー泣いていた事だけは記憶している。 / 吾輩はここで始めて人間というものを見た。 / しかもあとで聞くとそれは書生という人間中で一番獰惡な種族であったそうだ。 /

...

- 文ごとに分割されたテキスト

クリーニングと前処理

単語(トークン)に分割

吾輩/は/猫/で/ある/。

名前/は/まだ/無い/。

どこ/で/生れ/た/か/とんと/見当/が/つか/ぬ/。

...

- 文を単語ごとに分割
→こうやって分析できる単位に分割していく

クリーニングと前処理

データ前処理

- クリーニング
 - ノイズを取り除く
 - 辞書を整備する
- 前処理
 - テキストの分割
 - 文書→文→単語に分割
 - 単語の処理
 - 様々なタグをつける
- データハンドリング・構造化
 - 様々な分析手法が可能なようにデータを整える

クリーニングと前処理

クリーニング

- 分析にかけられるよう、データを整備する
 - 誤字脱字のチェック
 - 辞書の整備
 - 専門用語
 - 新語・死語
 - 方言
- ノイズが多く混じっていると結果が歪む

クリーニングと前処理

形態素解析

- 英語などの言語：
 - 単語と単語の間に切れ目がある
→スペースで分割できる
- 日本語のような言語
 - 単語と単語との間に切れ目がない
 - 分割できるように「分かち書き」する必要がある
→「形態素解析」という技術を使う
- 形態素解析ソフトを使うことで、
 - 分かち書き
 - 品詞タグ付けを、一定の精度でまとめて処理することができる

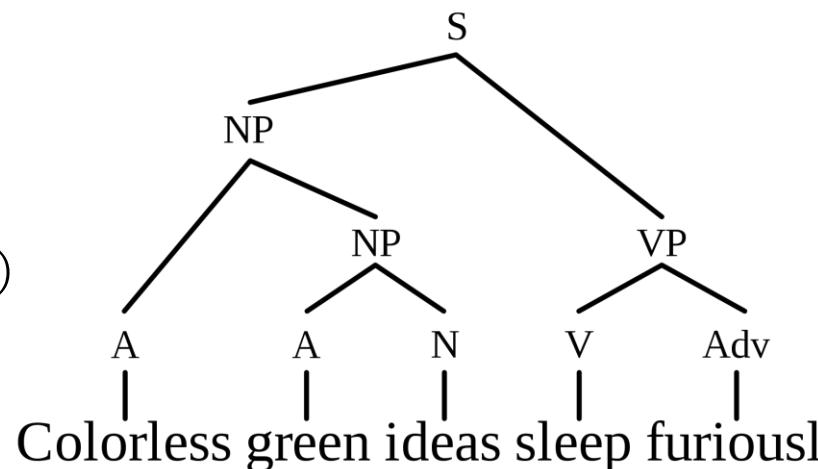
単語の処理

- ストップワード **stopword**
 - 話題の種類と関連を持たない語(e.g. a, 冠詞、助詞)
 - 分析に必要なければ除外する
- ステミング **stemming**
 - 派生語などを含めて同じ素性とみなす処理
 - e.g. operate→oper, operational→oper
 - 単語の形状をみて一律に処理する
 - e.g. Porter's stemmer
- 見出し語化 **lemmatization**
 - 単語を基本形に戻す処理
 - 文脈を考慮して処理

クリーニングと前処理

タグ付け

- 付加情報をつける作業
 - 品詞 part of speech (POS)
 - 名詞
 - 動詞
 - 形容詞
 - 副詞
 - 助詞
 - 構文情報
 - 文法的な構造 (e.g. 主語, 述語)



クリーニングと前処理

構造化

- 構造化の例

- bag of words
 - 単語の頻度
 - 語順の情報は無視される
- n-gram
 - 連続するn語の組み合わせ
 - 連続する2語→bigram
 - 連続する3語→trigram
- 共起
 - 同じ文内に出現した語の組み合わせ

bag of words

this	1
is	1
a	1
pen	1

bigram

this-is	1
is-a	1
a-pen	1

共起

(this, is)	1
(this, a)	1
(this, pen)	1
(is, a)	1
(is, pen)	1
(a, pen)	1

→構造化されたデータを分析・可視化する

実習

事前準備

- ソフトウェア
 - R version 3.4.3以上
 - MeCab
 - Windowsの場合はShift-JIS版がインストールされていること
 - IPA辞書がインストールされていること
- Rパッケージ
 - RMeCab
 - dplyr
 - stringr

ソースコードについての一般的な注意

- 元データを直接いじらない
 - どこを変更したかわからないので
 - 元データとクリーニング後のデータは別にする
 - データの変更 (e.g. 参加者の除外) は分析プログラム上で行い、生データから変更手続きが再現できるようにする
- 必ずソースコードを保存する
 - 元のデータから結果を再現するための証拠

パイプ処理入門①

- パイプ処理：データを効率的に処理するための枠組み
 - 通常の処理：ひとつの処理が完全に終わったら次の処理、というように逐次的に処理する
 - 処理n(処理n-1(...(処理2(処理1)))
 - パイプ処理：終わった部分から徐々に次の処理に流していき、並列的に処理する
 - 処理1 | 処理2 | ... | 処理n
 - Linuxシェルなどで処理するときに重宝する

パイプ処理入門②

- “dplyr”パッケージを使う
 - パイプ記号 : %>%
 - 通常の関数の適用 : `head(df)`
 - パイプ記号を使った関数の適用 : `df %>% head()`
 - 関数の第一引数に放り込まれる
 - データの選択
 - `filter()`: 条件で行を選択
 - `select()`: 列を選択、名前の表示法も変えられる
- RStudioのページで公開されているチートシートが参考になる
 - <https://www.rstudio.com/resources/cheatsheets/>

テキスト処理入門

- “stringr”パッケージを使う
 - 検索
 - `str_subset()`: 指定されたキーワードが含まれている文字列を返す
 - `str_detect()`: 指定されたキーワードが含まれていればTRUE, そうでなければFALSEを返す
 - `grep`: Rの標準の関数。ヒットしたベクトルの位置を返す
 - 置換
 - `str_replace()`: 指定された検索語を置換語で置き換える
 - `sub`: Rの標準の関数。検索語を置換語で置き換える
 - 「正規表現」が使える

Rによる処理の基礎

正規表現

- 文字列の集合を表現するための記法
- 文字列の検索に使うことができる
- 基本
 - ^ 行頭
 - \$ 行頭
 - . 任意の1文字
 - | 「または」
- エスケープシーケンス
 - ¥¥d 数字(digits)
 - ¥¥w 単語に使われる文字・数字・アンダーバー(words)
 - ¥¥s 空白(spaces)
 - ¥n 改行(newline)
 - ¥t タブ(tab)

- 文字クラスとグループ化
 - [] 括弧内に含まれる文字にマッチ
 - [^] 括弧内に含まれない文字にマッチ
 - () グループ化
- 記号
 - * 0回以上の繰り返し
 - + 1回以上の繰り返し
 - ? 0または1回
- その他文字クラス
 - [:alpha:] アルファベット
 - [:lower:] 小文字アルファベット
 - [:upper:] 大文字アルファベット
 - [:digit:] 数字
 - [:alnum:] アルファベット + 数字
 - [:punct:] 記号
 - [:graph:] 数字 + 記号 + アルファベット
 - [:blank:] 空白、タブ

RMeCabの使い方

- RMeCab

- MeCab: 日本語の形態素解析エンジン
- RMeCab: R上でMeCabを使うためのパッケージ(石田, 2017)
 - 作者のページに各関数の詳しい解説がある
 - <http://rmecab.jp/wiki/index.php?RMeCabFunctions>
 - RMeCabC(): 文字列を形態素解析して返す
 - docMatrixDF(): フォルダ内のファイルごとに解析する
 - docDF(): データフレームの行ごとに解析する
 - 解析時に辞書ファイルを指定できる

クリーニングの実際

- 例文：第一班のインタビューより
 - Q: 「やまいり？やまはいって木を丸める？」
 - A: 「切ってきて、丸めて、家の屋根にしげておくん。
それはてんかごめんけどどこのやまいってきってきてもよいということになっちゃった。」
- これを形態素解析にかけてみる

- 動詞: '切つ'
- 助詞: 'て'
- 動詞: 'き'
- 助詞: 'て'
- 記号: '、'
- 動詞: '丸め'
- 助詞: 'て'
- 記号: '、'
- 名詞: '家'
- 助詞: 'の'
- 名詞: '屋根'
- 助詞: 'に'
- 形容詞: 'しげ'
- 助詞: 'て'
- 動詞: 'おく'
- 助動詞: 'ん'
- 記号: '。'
- 名詞: 'それ'
- 助詞: 'は'

方言または
誤字・聞き取りミス

- 名詞: 'てんか'
- 感動詞: 'ごめん'
- 接続詞: 'けど'
- 名詞: 'どこ'
- 助詞: 'の'
- 助詞: 'や'
- 動詞: 'まいっ'
- 助詞: 'て'
- 動詞: 'きっ'
- 助詞: 'て'
- 動詞: 'き'
- 助詞: 'て'
- 助詞: 'も'
- 形容詞: 'よい'
- 助詞: 'という'
- 名詞: 'こと'
- 助詞: 'に'
- 動詞: 'なっ'
- 名詞: 'ちょ'
- 動詞: 'つ'
- 助動詞: 'た'
- 記号: '。'

「てんかごめん」
(天下御免) の誤認識

誤字・聞き取りミス？

「やま いって」が
「や まいって」に

方言

クリーニングの実際

クリーニング方略

- ノイズを減らす
 - 誤字・脱字をチェックする
 - 句読点を入れる
 - 漢字に変換する
 - 表記揺れを減らす
- 固有名詞への対応
 - 辞書を整備する
- 方言への対応
 - 辞書を整備する
 - 標準語に変換する
 - 方言に関する知識を持つ

クリーニングの実際 修正の例

「切ってきて、丸めて、家の屋根にしげておくん。
それはてんかごめんけどどこのやまいってきって
きてもよいということになっちゃった。」

「切ってきて、丸めて、家の屋根にしげておくん。
それは**天下御免で**どこの山に行って**切**ってきても
よいということになつていた。」

Before

1. 動詞: '切つ'
2. 助詞: 'て'
3. 動詞: 'き'
4. 助詞: 'て'
5. 記号: '、'
6. 動詞: '丸め'
7. 助詞: 'て'
8. 記号: '、'
9. 名詞: '家'
10. 助詞: 'の'
11. 名詞: '屋根'
12. 助詞: 'に'
13. 形容詞: 'しげ'
14. 助詞: 'て'
15. 動詞: 'おく'
16. 助動詞: 'ん'
17. 記号: '。'
18. 名詞: 'それ'
19. 助詞: 'は'

After

1. 動詞: '切つ'
2. 助詞: 'て'
3. 動詞: 'き'
4. 助詞: 'て'
5. 記号: '、'
6. 動詞: '丸め'
7. 助詞: 'て'
8. 記号: '、'
9. 名詞: '家'
10. 助詞: 'の'
11. 名詞: '屋根'
12. 助詞: 'に'
13. 動詞: 'すげ'
14. 助詞: 'て'
15. 動詞: 'おく'
16. 助動詞: 'ん'
17. 記号: '。'
18. 名詞: 'それ'
19. 助詞: 'は'

**Before**

20. 名詞: '天下'
21. 感動詞: 'ごめん'
22. 接続詞: 'けど'
23. 名詞: 'どこ'
24. 助詞: 'の'
25. 助詞: 'や'
26. 動詞: 'まいっ'
27. 助詞: 'て'
28. 動詞: 'きっ'
29. 助詞: 'て'
30. 動詞: 'き'
31. 助詞: 'て'
32. 助詞: 'も'
33. 形容詞: 'よい'
34. 助詞: 'という'
35. 名詞: 'こと'
36. 助詞: 'に'
37. 動詞: 'なっ'
38. 名詞: 'ちょ'
39. 動詞: 'つ'
40. 助動詞: 'た'
41. 記号: '。'

After

20. 名詞: '天下'
21. 名詞: '御免'
22. 助詞: 'で'
23. 名詞: 'どこ'
24. 助詞: 'の'
25. 名詞: '山'
26. 助詞: 'に'
27. 動詞: '行っ'
28. 助詞: 'て'
29. 動詞: '切っ'
30. 助詞: 'て'
31. 動詞: 'き'
32. 助詞: 'て'
33. 助詞: 'も'
34. 形容詞: 'よい'
35. 助詞: 'という'
36. 名詞: 'こと'
37. 助詞: 'に'
38. 動詞: 'なっ'
39. 助詞: 'て'
40. 動詞: 'い'
41. 助動詞: 'た'



クリーニングの実際

MeCabの辞書の設定

- 追加辞書用csvを作る
 - MeCab/dic/ipadic配下に辞書のcsvがたくさん入っているので参考にする
- コンパイルする
 - MeCab/bin配下のmecab-dict-index.exeを使う
- RMeCabの関数実行時に辞書ファイルを指定

クリーニングの実際

辞書CSVの作成

指定不要

IPA 品詞体系を参考にする

表層形	左文脈ID	右文脈ID	コスト	品詞	品詞 細分類1	品詞 細分類2	品詞 細分類3	活用型	活用形	原形	読み	発音
けん	*	*	1000	助詞	接続助詞	*	*	*	*	けん	ケン	ケン

同じカテゴリの単語を参考にする



けん,*,*,1000,助詞,接続助詞,*,*,*,*,けん,ケン,ケン

クリーニングの実際

辞書のコンパイル

MeCabフォルダに移動し、

```
¥bin¥mecab-dict-index.exe -d デフォルト辞書フォルダ  
-u 出力ファイル名  
-f 文字エンコーディング（入力ファイル）  
-t 文字エンコーディング（出力ファイル）  
入力ファイル名
```

を実行（行を分けずに入力する）

- 実行例

```
¥bin¥mecab-dict-index.exe -d dic¥ipadic -u  
c:¥Users¥satoc¥projects¥nlp2019¥example.dic -f shift-jis -t  
shift-jis c:¥Users¥satoc¥projects¥nlp2019¥shimane.csv
```

```
RMeCabC('まあそういうことで、余分な話はいいけん。')
```

1. **副詞:** 'まあ'
2. **連体詞:** 'そういう'
3. **名詞:** 'こと'
4. **助動詞:** 'で'
5. **記号:** '、'
6. **名詞:** '余分'
7. **助動詞:** 'な'
8. **名詞:** '話'
9. **助詞:** 'は'
10. **形容詞:** 'いい'
11. **助詞:** 'けん'
12. **助詞:** 'ん'
13. **記号:** '。'

```
RMeCabC('まあそういうことで、余分な話はいいけん。', dic='example.dic')
```

1. **副詞:** 'まあ'
2. **連体詞:** 'そういう'
3. **名詞:** 'こと'
4. **助動詞:** 'で'
5. **記号:** '、'
6. **名詞:** '余分'
7. **助動詞:** 'な'
8. **名詞:** '話'
9. **助詞:** 'は'
10. **形容詞:** 'いい'
11. **助詞:** 'けん'
12. **記号:** '。'

クリーニングの実際

どれくらいクリーニングすればよいか？

- テキストデータは一般に膨大
 - 人手で完璧にチェックすることは不可能
 - 自動で大量に処理できることのメリットが失われる
 - 事前に何が問題になるかはわかりづらい
 - 探索的なプロセス / データセットそれぞれの固有の問題
- 無難なアプローチ：分析しながら、漸進的に改善
 - ノイズになっている個所を見つけたら対応する
 - 固有名詞や方言 / 形態素解析の誤認識
 - 単語の頻度表をチェックする
 - 当然多くなるべき単語が多くなっているか
 - 不可解な言葉が多くなっていないか
 - 分析
 - いつでも元データから最新のデータを作れるようコードを保存する
 - コアとなる分析に関する部分は入念にチェックする
 - 日ごろから様々なエラーの可能性を検討しておく

クリーニングの実際

クリーニングのヒント

- 辞書に載っている表記に変える
 - ひらがな・カタカナを漢字にする
 - 伸ばし棒なのか
- 方言に対応する
 - MeCabの辞書に追加する
 - 追加する際の情報は標準語の対応する単語を参考にする
 - 辞書で対応できそうになければ、一括で置換する
 - Googleスプレッドシートは正規表現での検索・置換に対応している

データハンドリング入門：「夢十夜」の分析

- 夏目漱石 「夢十夜」

- 1908年に朝日新聞紙上で連載
- 夢を題材にした10話からなる小説
- パブリックドメイン



- データ

夏目漱石
(1867-1916)

- 青空文庫で公開されているデータ(新字新仮名)を用いた
- ルビの削除等の処理はAozora()関数(石田, 2017) を用いた
 - 著者のホームページでも公開されている
 - URL
- 今回はtsv形式に加工したものを使う

分析①：「夢十夜」

データセットの構造

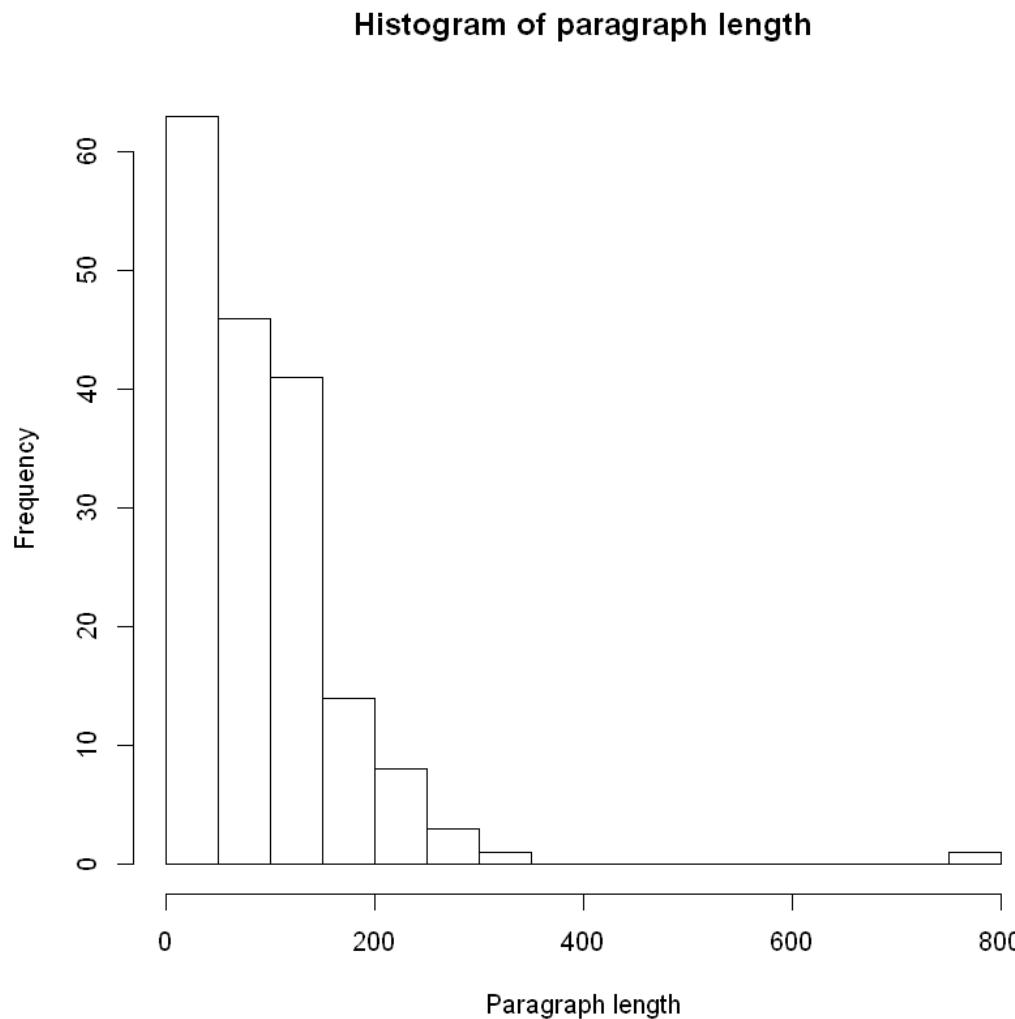
話ID	段落ID	本文
section_id	paragraph_id	content
1	1	こんな夢を見た。
1	2	腕組をして枕元に坐っていると、仰向に寝た女が、静かな声でもう死にますと云う。女は長い髪を枕に敷いて、輪郭の柔らかな瓜実顔をその中に横たえている。真白な頬の底に温かい血の色がほどよく差して、唇の色は無論赤い。とうてい死にそうには見えない。しかし女は静かな声で、もう死にますと判然云った。自分も確にこれは死ぬなと思った。そこで、そうかね、もう死ぬのかね、と上から覗き込むようにして聞いて見た。死にますとも、と云いながら、女はぱちりと眼を開けた。大きな潤のある眼で、長い睫に包まれた中は、ただ一面に真黒であった。その真黒な眸の奥に、自分の姿が鮮に浮かんでいる。
1	3	自分は透き徹るほど深く見えるこの黒眼の色沢を眺めて、これでも死ぬのかと思った。それで、ねんごろに枕の傍へ口を付けて、死ぬんじゃなかろうね、大丈夫だろうね、とまた聞き返した。すると女は黒い眼を眠そうに※たまま、やっぱり静かな声で、でも、死ぬんですもの、仕方がないわと云った。
1	4	じゃ、私の顔が見えるかいと一心に聞くと、見えるかいって、そら、そこに、写ってるじゃありませんかと、にこりと笑って見せた。自分は黙って、顔を枕から離した。腕組をしながら、どうしても死ぬのかなと思った。
1	5	しばらくして、女がまたこう云った。
1	6	「死んだら、埋めて下さい。大きな真珠貝で穴を掘って。そして天から落ちて来る星の破片を墓標に置いて下さい。そして墓の傍に待っていて下さい。また逢いに来ますから」

分析①：「夢十夜」

データハンドリング

- テキスト：様々な分析単位がありえる
 - 文字・単語・文・段落・章・文書全体・文書の集合...
→様々な分析単位を自由に行き来できる形式が望ましい
- インタビューデータ：話者の発話ごと
 - 特定の話者の発言のみを抜き出す
 - 特定の話題を含む発言のみ抜き出す

分析①：「夢十夜」 段落の長さの分布

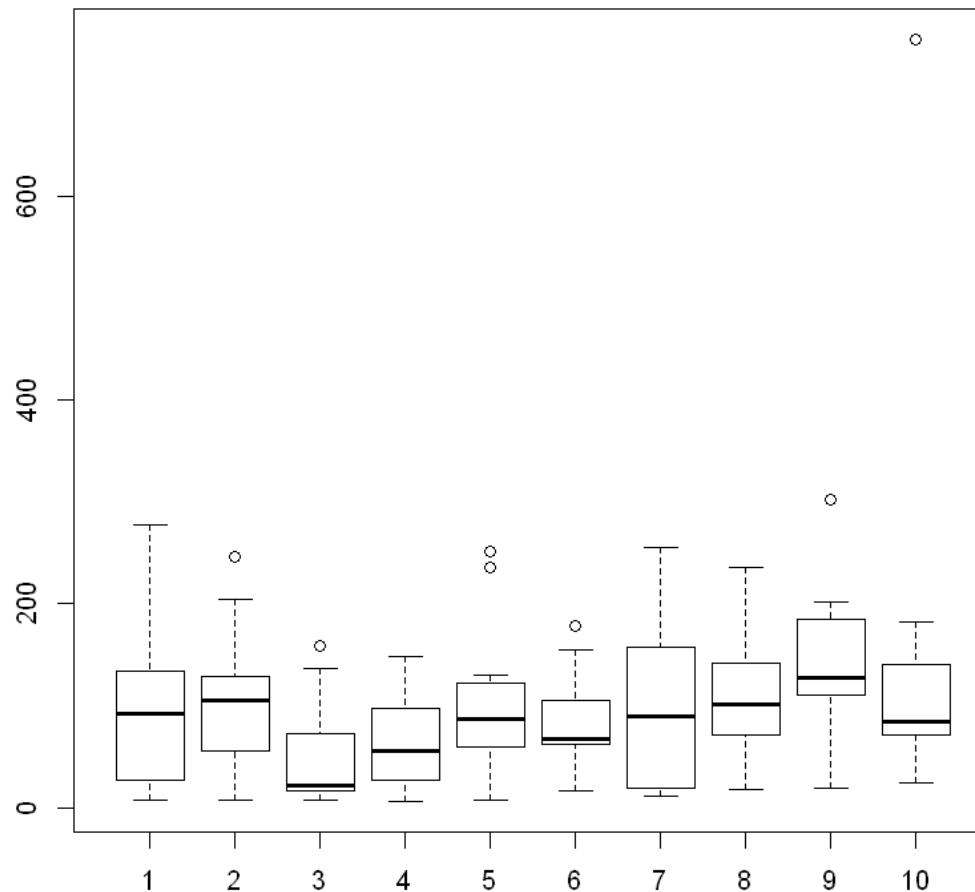


- ほとんどが400字以下だが、非常に長い段落がごく少数ある

分析①：「夢十夜」

段落の長さ

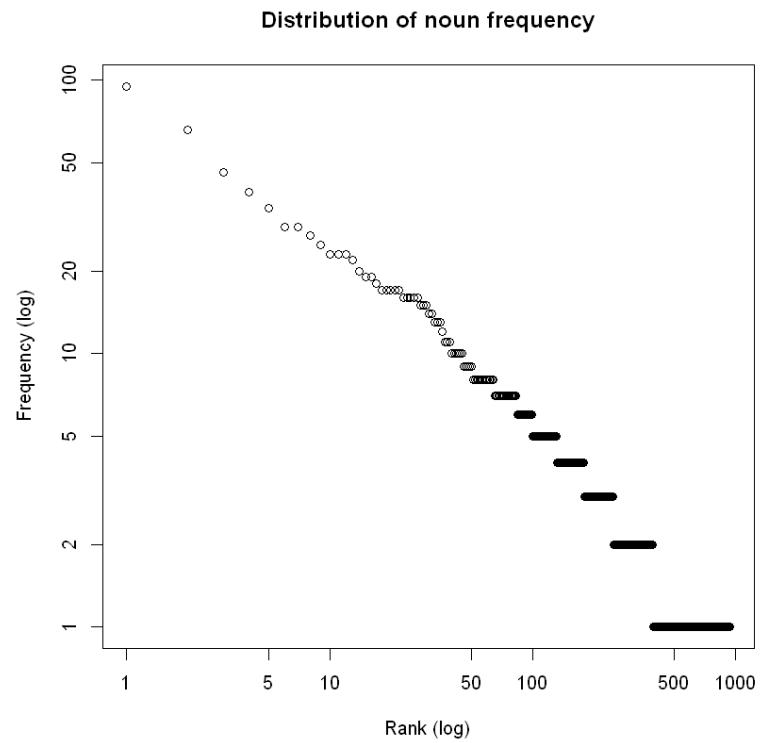
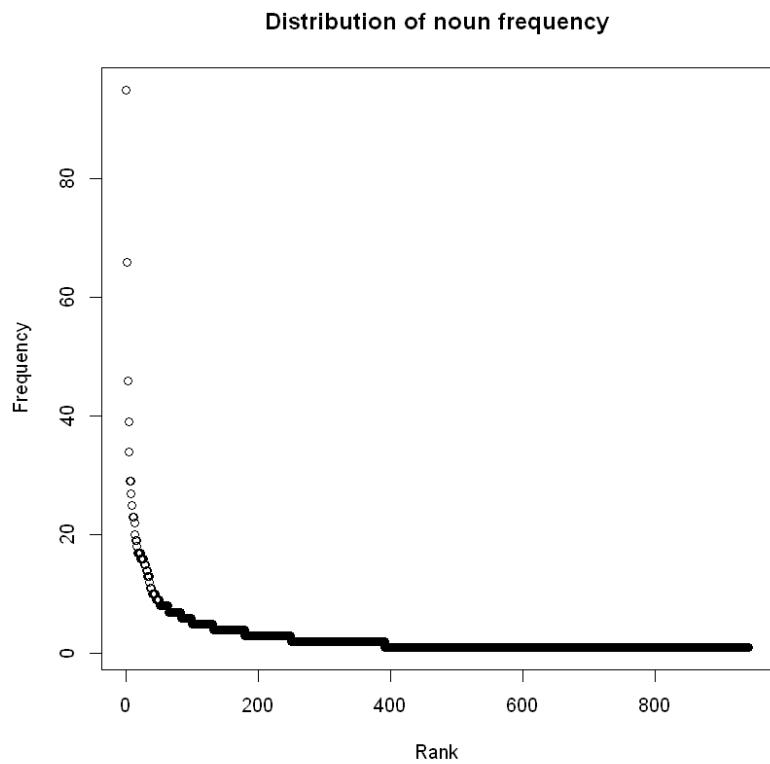
Paragraph length of each section



- 第十夜に非常に長い段落があるのがわかる
- 他の話でも、1,2段落ほど外れ値的に長い段落がある

分析①：「夢十夜」

名詞の分布



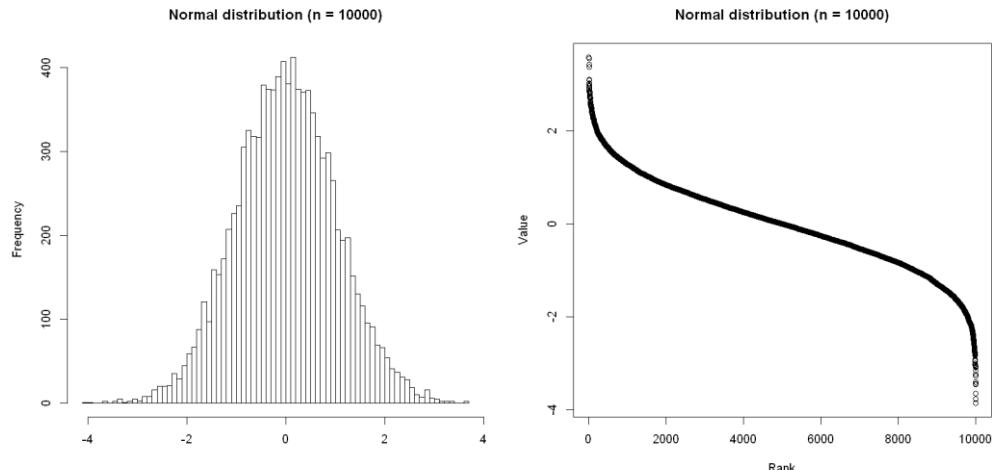
- ごく少数の単語が多く出現するが、大多数の単語はほとんど出現しない
- 両対数グラフにとると、直線のようにみえる
→ 単語の分布一般の特徴（Zipfの法則）

★頻度データの特徴

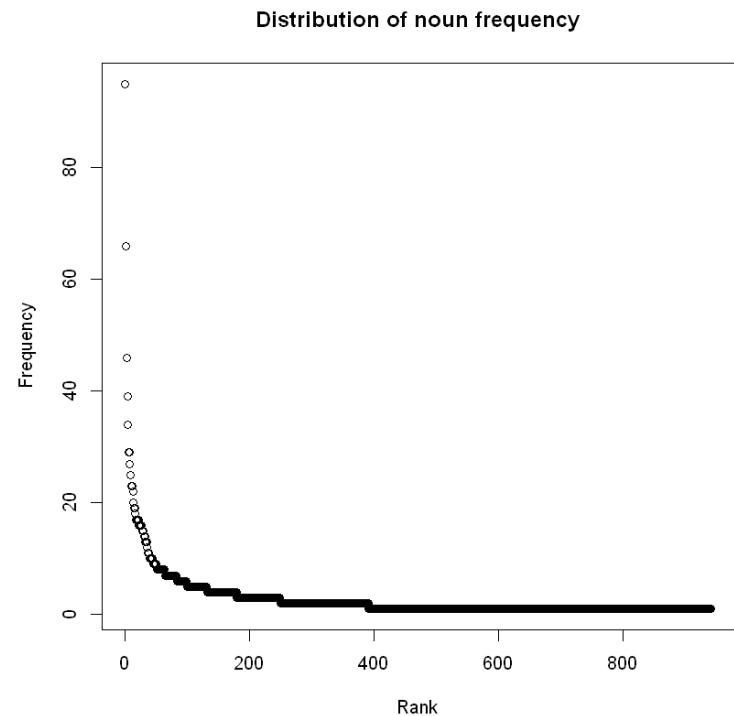
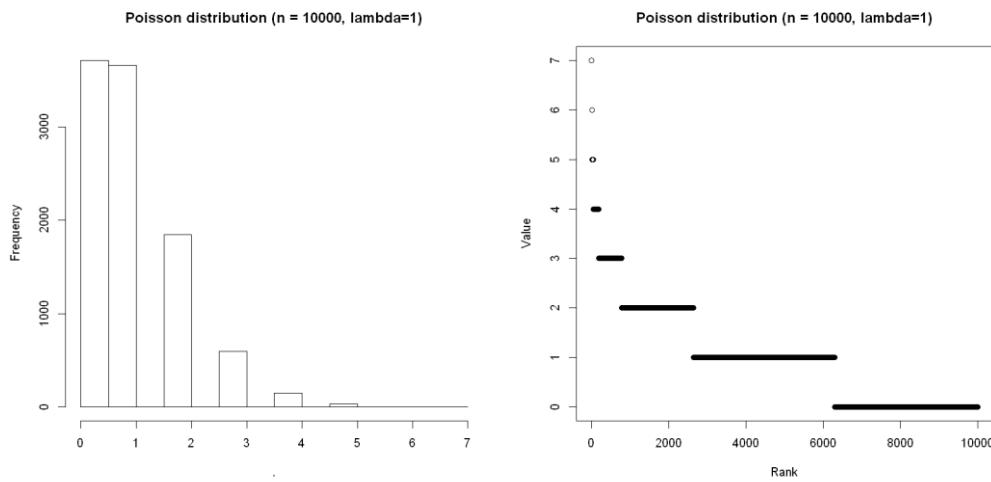
- 頻度：0以上の値をとる
 - 正規分布とは違う確率分布に従う
 - e.g. Poisson分布
- べき分布 power distribution
 - べき則 $f(x) = ax^k$ の形に従う分布
 - 突出した頻度をもつ少数のアイテムと、頻度が小さい膨大な数のアイテムからなる (long-tailな分布)
 - 自然現象や社会現象の一部で観察される
 - Paretoの法則(Pareto, 1896)
 - 所得分布は上位20%が全体の80%を占める
 - Zipfの法則(Zipf, 1949)
 - 文書内の単語の頻度は順位に反比例する
 - Gutenberg-Richter則(Gutenberg & Richter, 1941)
 - 地震の頻度は規模に反比例する

★べき分布：他の分布との比較

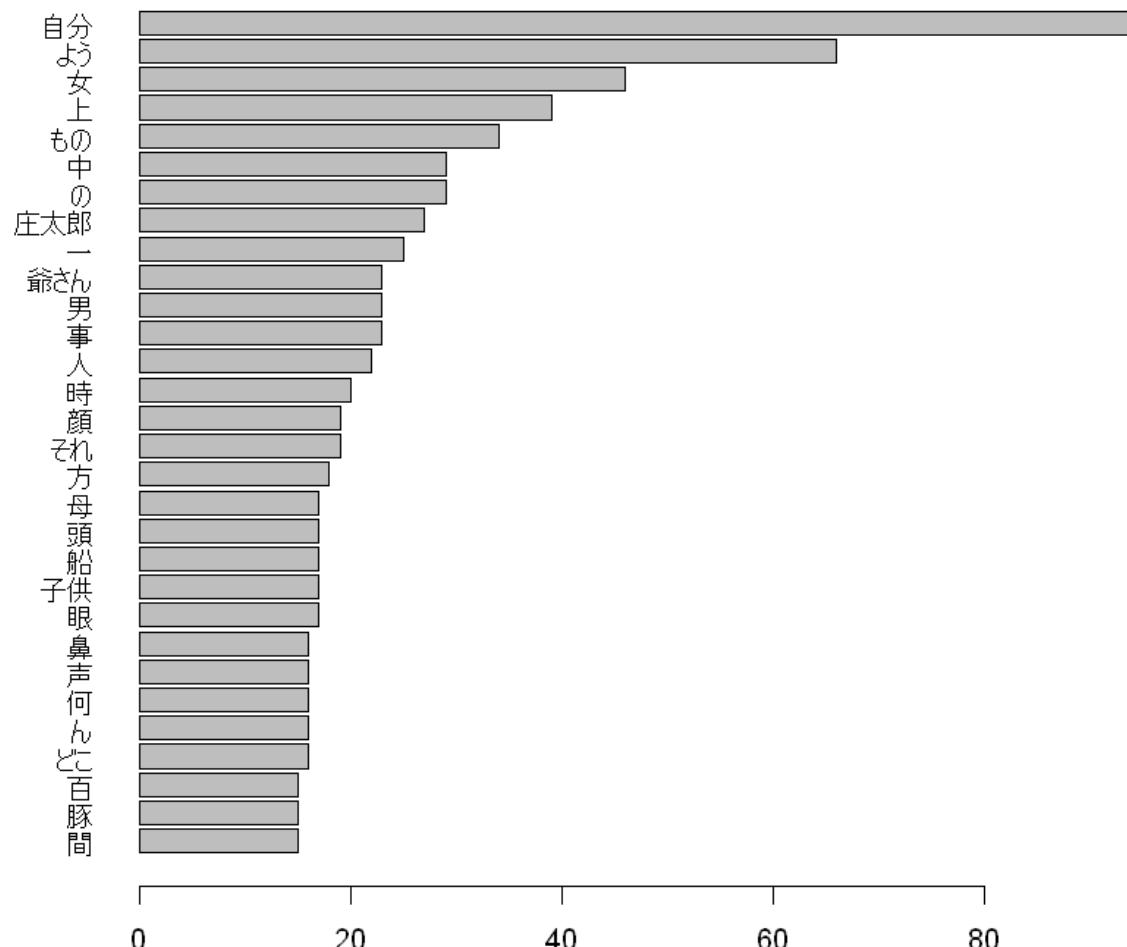
正規分布



Poisson分布



分析①：「夢十夜」 頻度の多い単語



- 一部の単語は内容を理解するのに寄与しない
- 例：よう・上・もの・中・の・一・それ・事・それ・etc.
→ストップワードとして分析から除外する

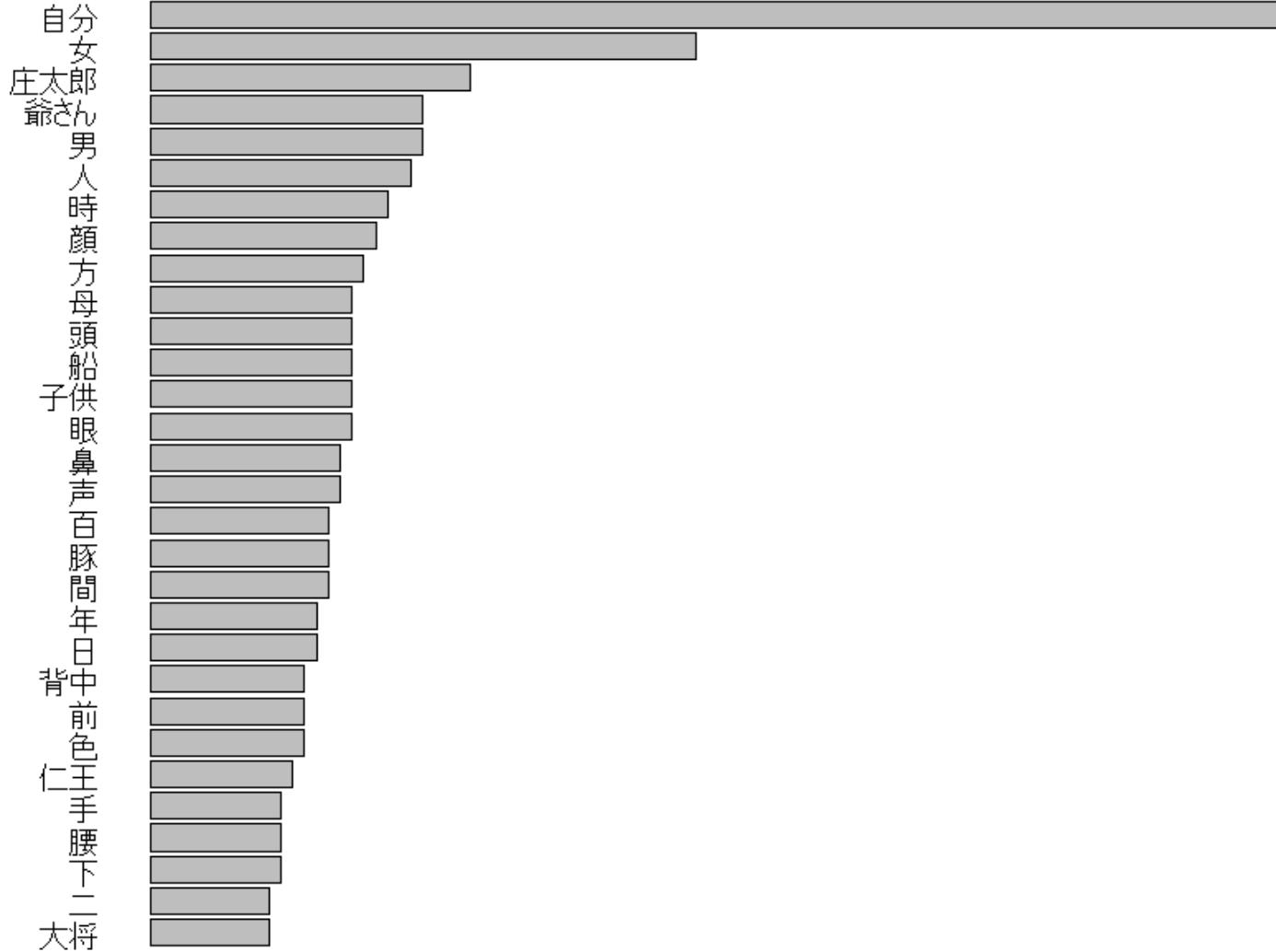
分析①：「夢十夜」ストップワードの除外

Before

自分
よ
女
上
も
中の
庄太郎
一
爺さん
男
事
人
時
顔
それ
方
母
頭
船
供
眼
鼻
声
ほん
ど
百
豚
間



After



分析①：「夢十夜」

TF-IDF

- **TF-IDF: term frequency – inverse document frequency**
 - 文書群について、単語がどれくらい特徴的かを表す指標
→文書のキーワードを抜き出すために使える
 - TF:Term Frequency 単語頻度
 - それぞれの文書について、その単語が出てくる程度
 - IDF:Inverse Document Frequency 逆文書頻度
 - 全体の文書のうち、その単語を含む文書の程度（の逆数）
 - 複数の文書に出現する単語ほど特徴的でない
 - TF-IDF
 - TFとIDFの積
 - 少数の文書に頻出する単語ほど強く重みづける
- 経験的な指標：理論的な基礎ははっきりしないが、有用なためテキストマイニングや検索エンジンなどで幅広く使われている
- 共通する単語は低く重みづけられるので、同系統の文書を分析するときには注意が必要

分析①：「夢十夜」

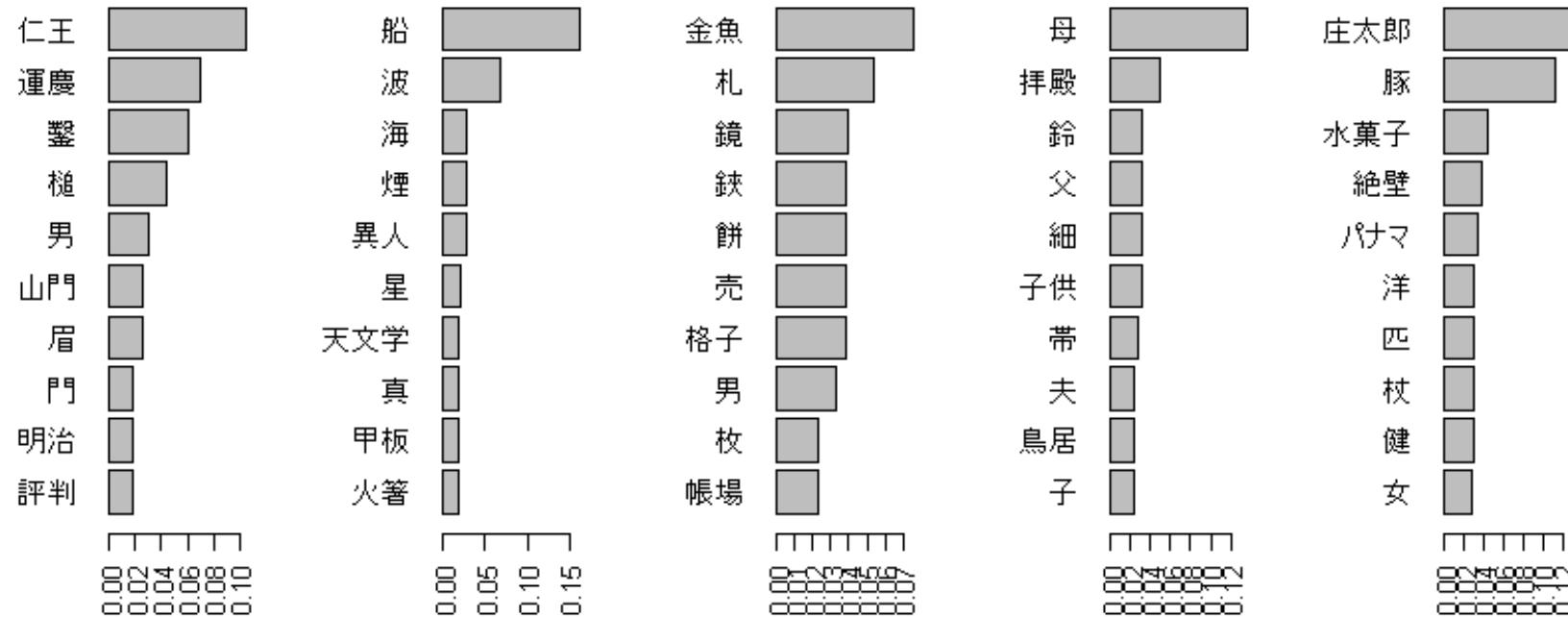
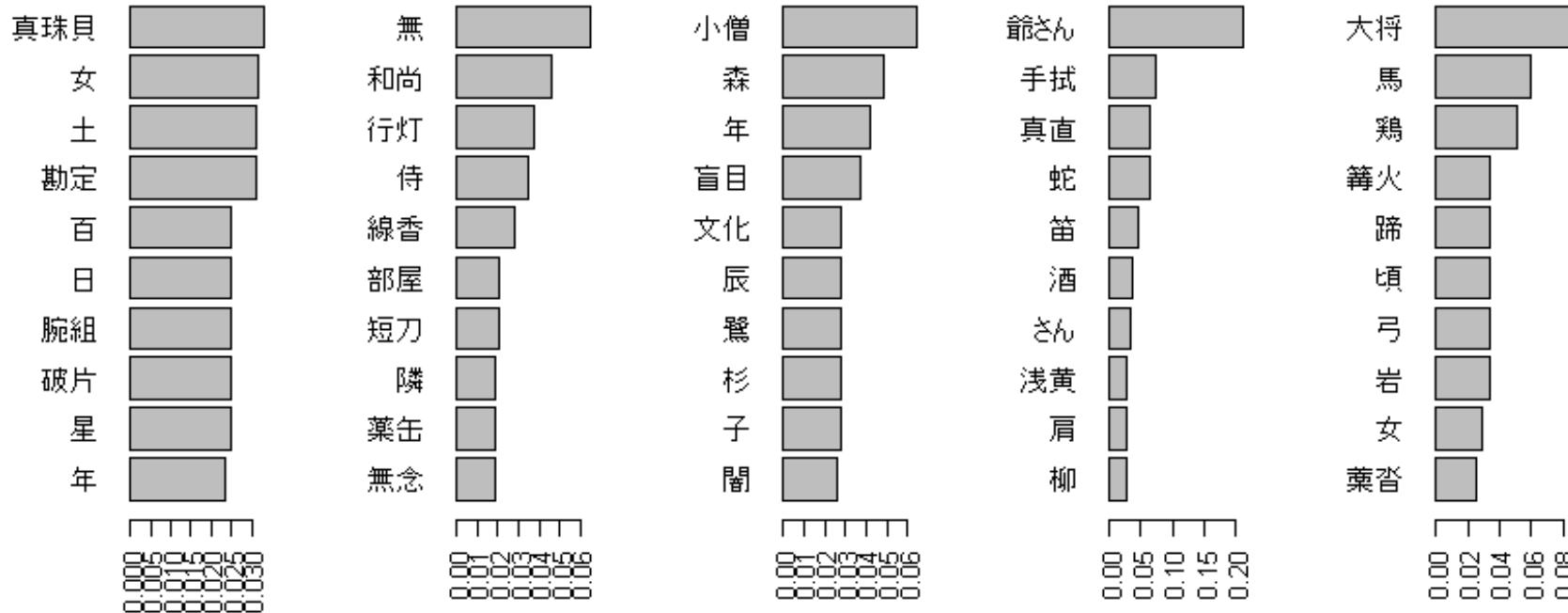
TF-IDFの計算方法

$$TF = \frac{\text{当該の単語の出現回数}}{\text{文書内の総単語数}}$$

$$IDF = \log \frac{\text{総文書数}}{\text{当該の単語を含む文書の数}}$$

$$\begin{aligned} &= \log \left(1 / \frac{\text{当該の単語を含む文書の数}}{\text{総文書数}} \right) \\ &= \log \left(\frac{1}{\text{文書頻度}} \right) \end{aligned}$$

- すべての文書に当該の語が含まれる場合、IDFはゼロになる
- 「当該の単語を含む文書の数」がゼロになると計算できないので、実用上 1 を足して計算することがある



小まとめ

- データのクリーニング
 - ノイズを減らす、辞書を整備する
 - 分析しながら適宜データを洗練させていく
 - データ固有の知識 (e.g.方言) が役に立つ
- 前処理/データハンドリング
 - テキスト：様々な単位で分析したい
→柔軟に扱えるデータ構造にする
- 簡単な分析でも色々なことがわかる
 - 頻度の分析
 - TF-IDF

第2部 実習②テキスト分析編

テキスト分析：『こころ』の分析

- 夏目漱石『こころ』

- 1914年に朝日新聞で連載
- 三部構成
 - 「先生と私」「両親と私」「先生と遺書」
- 新潮文庫版：発行718万部
 - 新潮文庫で最も売れている小説(日本経済新聞, 2016)
- パブリックドメイン



- データ

- 青空文庫で公開されているデータ(新字新仮名)を用いた

分析②：『こころ』

分析の方略

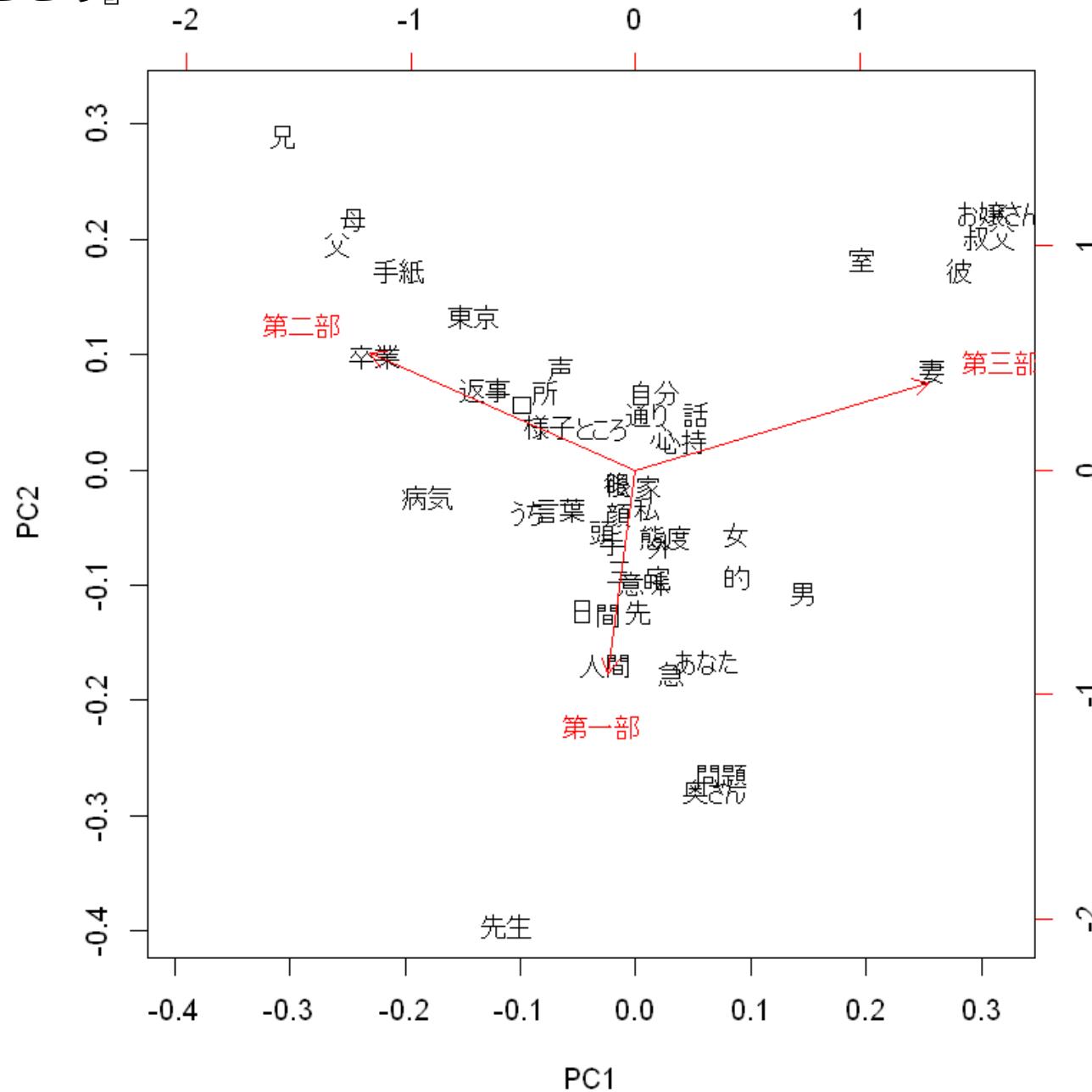
- 『こころ』：三部構成
 - 各部によって登場人物・主題が異なる
→各部の特性の違いを可視化する
- 単語の分布の違いの分析
 - 主成分分析を用いて、各部の単語（名詞）の分布を可視化する
- 単語同士の共起関係を分析する
 - 全体の単語の共起関係を分析
 - 登場人物による共起する単語の違いを可視化する
- ~~クラスタ分析（時間があれば）~~

分析②：『こころ』

主成分分析

- データのばらつきをよく説明する「主成分」を大きいものから順に抜き出していく
- 各軸は直交する（相関がない）ように選ばれる
- バイプロット biplot
 - 多次元データを二次元空間に射影したものが得られる
 - 合成スコアなので軸の解釈は特にする必要はない
 - 多次元のデータを図示するのに使える

分析②：『こころ』



『こころ』各部と名詞同士の共起関係を用いた主成分分析

分析②：『こころ』

共起分析 co-occurrence analysis

- 共起をカウントする範囲

- 文書
- 章・節
- 文
- 窓関数
 - 前後n単位

※範囲が広くなるほどデータサイズが膨大になる

- カウントの仕方

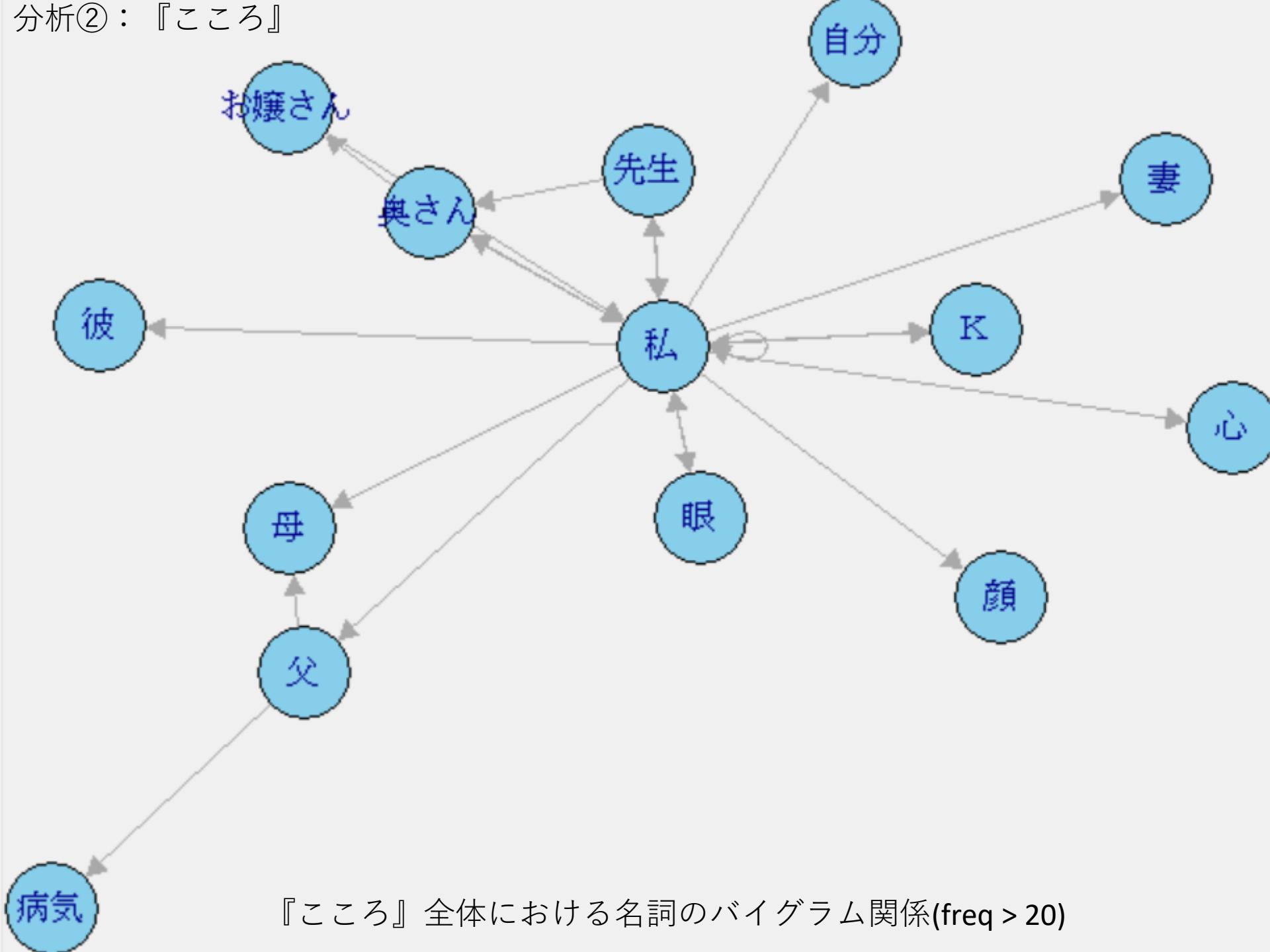
- Bag of words: 出現のみを考慮する
- N-gram: 語順も考慮に入れる

分析②：『こころ』

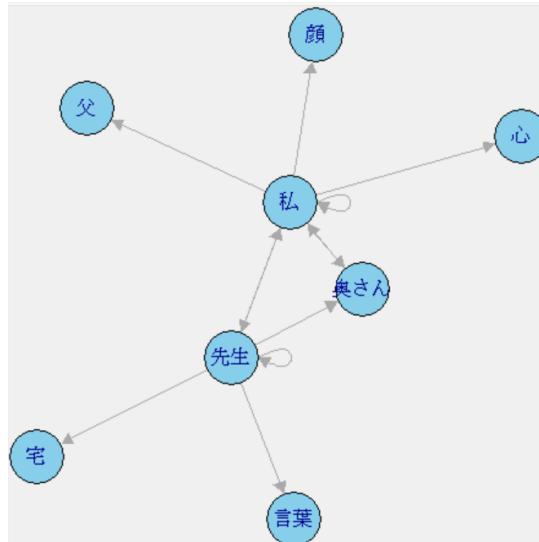
N-gramを用いた共起分析の例

- 手続き
 - 対象とする単語群を決める (e.g. 品詞)
 - N-gramを抽出する
 - カウントする
 - 分析・可視化する
- 特徴
 - 順序・距離が保持される
 - 文章の空間的構造が比較的反映されている
→ 距離的に近い組み合わせを抽出できる
 - 数が少なくすむ
 - 直近の共起関係しかわからない

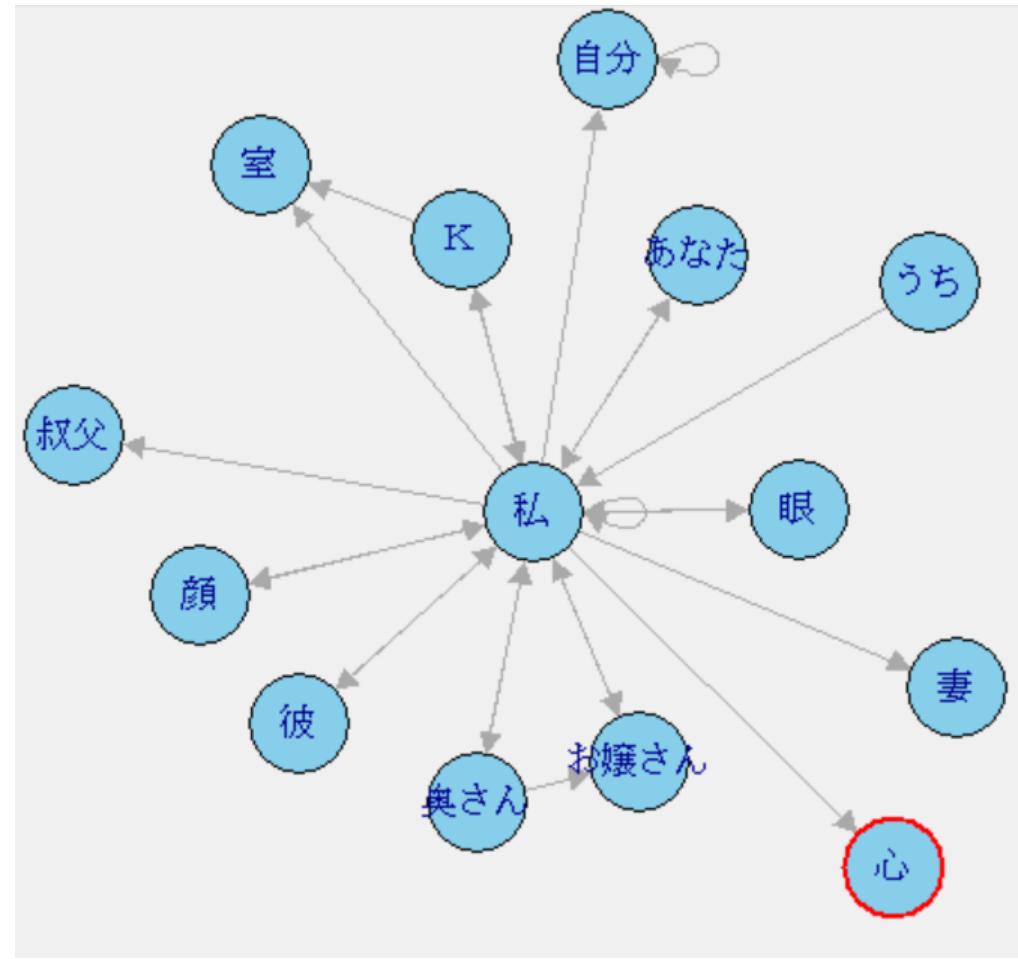
分析②：『こころ』



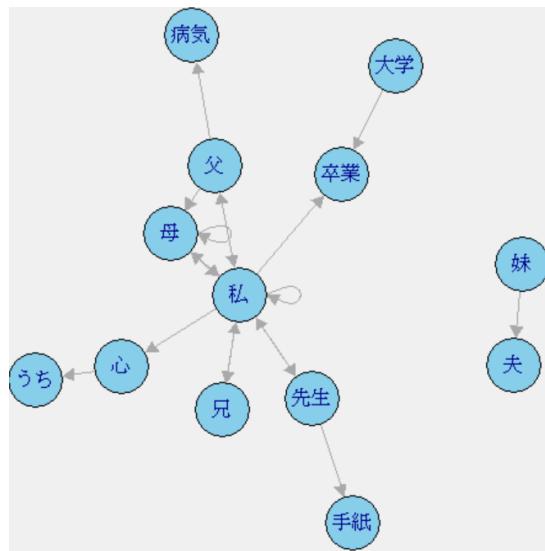
分析②：『こころ』



第一部(freq > 10)



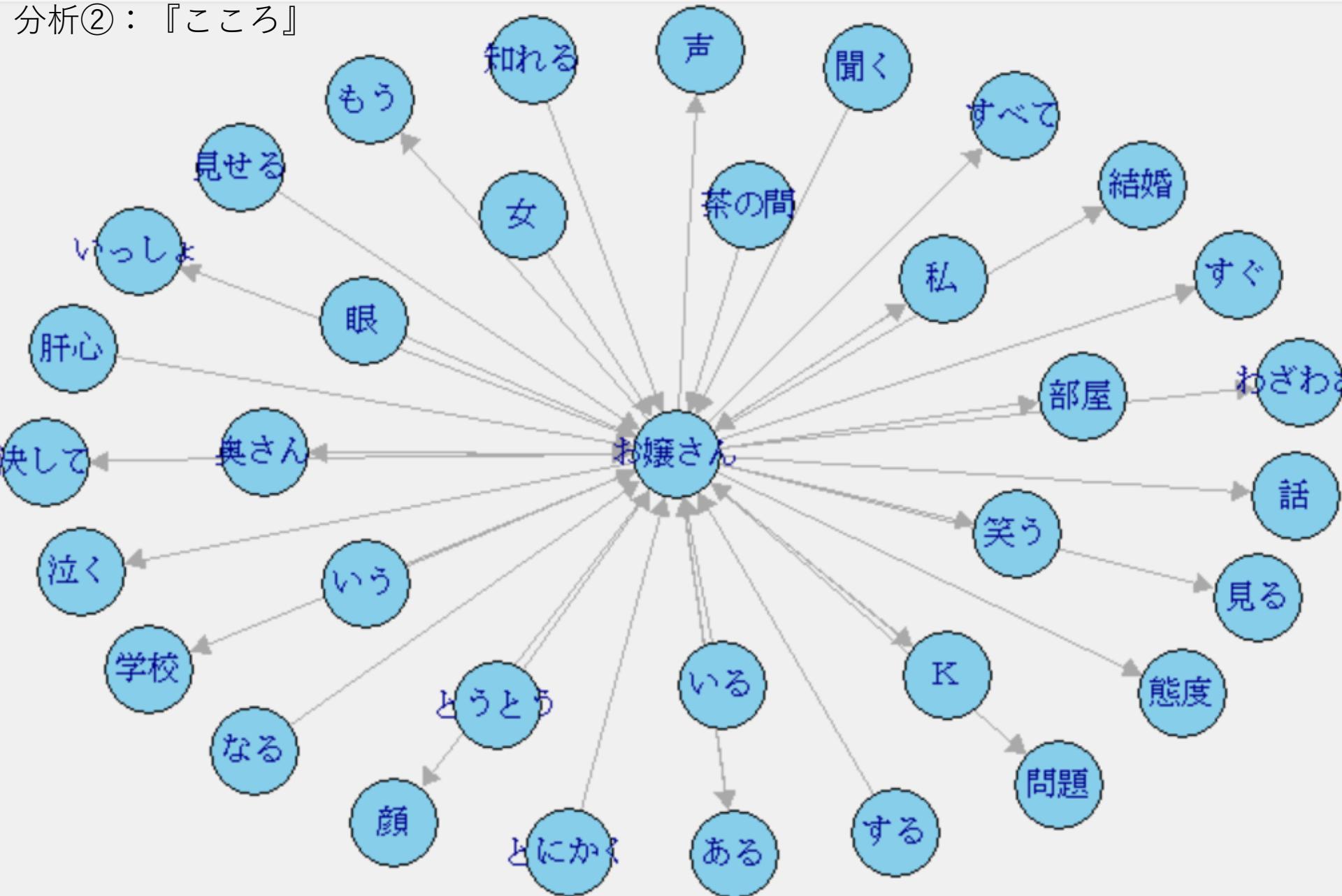
第三部(freq > 10)



第二部(freq > 5)

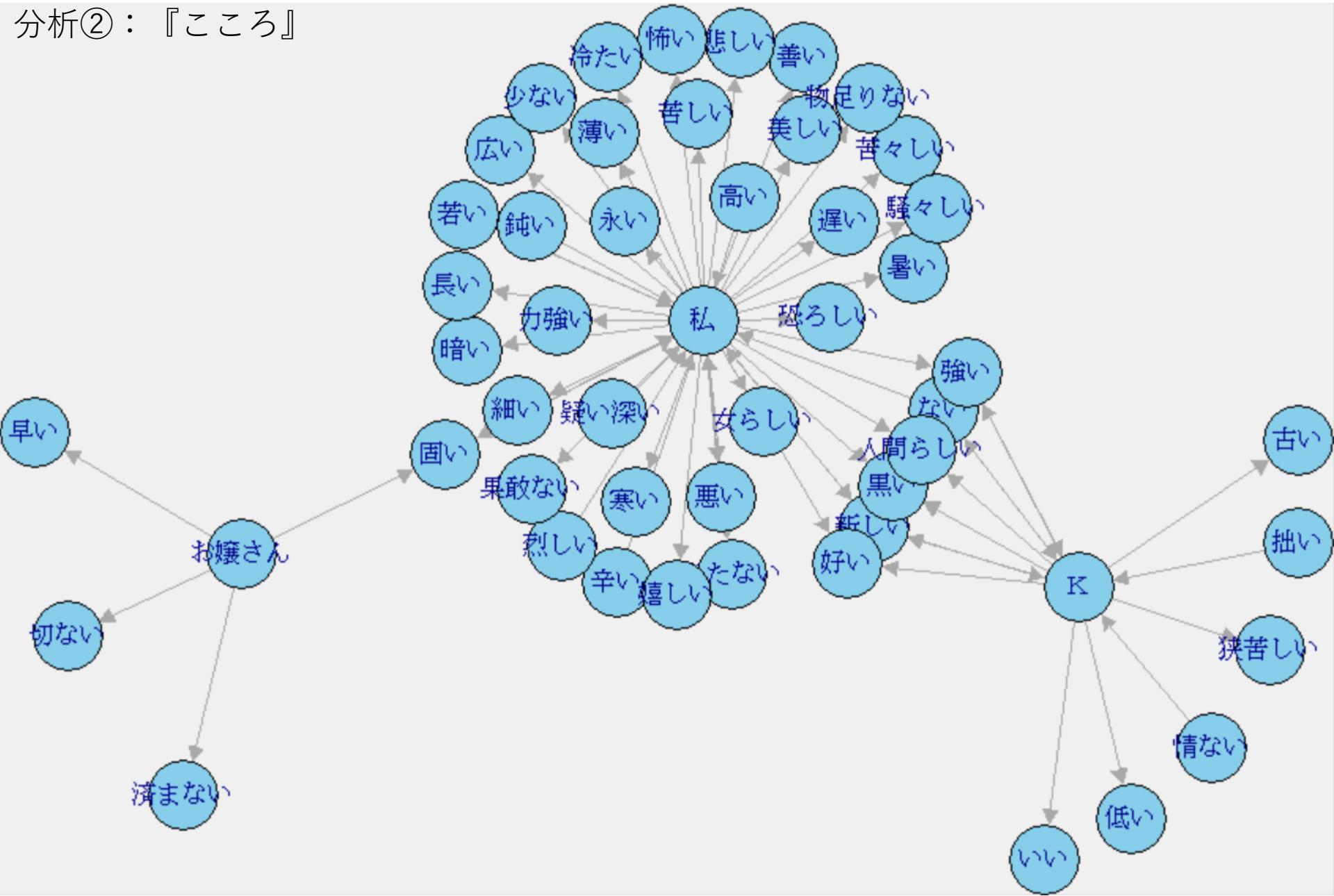
『こころ』各部における名詞のバイグラム関係

分析②：『こころ』



第三部における単語「お嬢さん」のバイグラム関係
(動詞・名詞・形容詞・副詞, freq > 1)

分析②：『こころ』



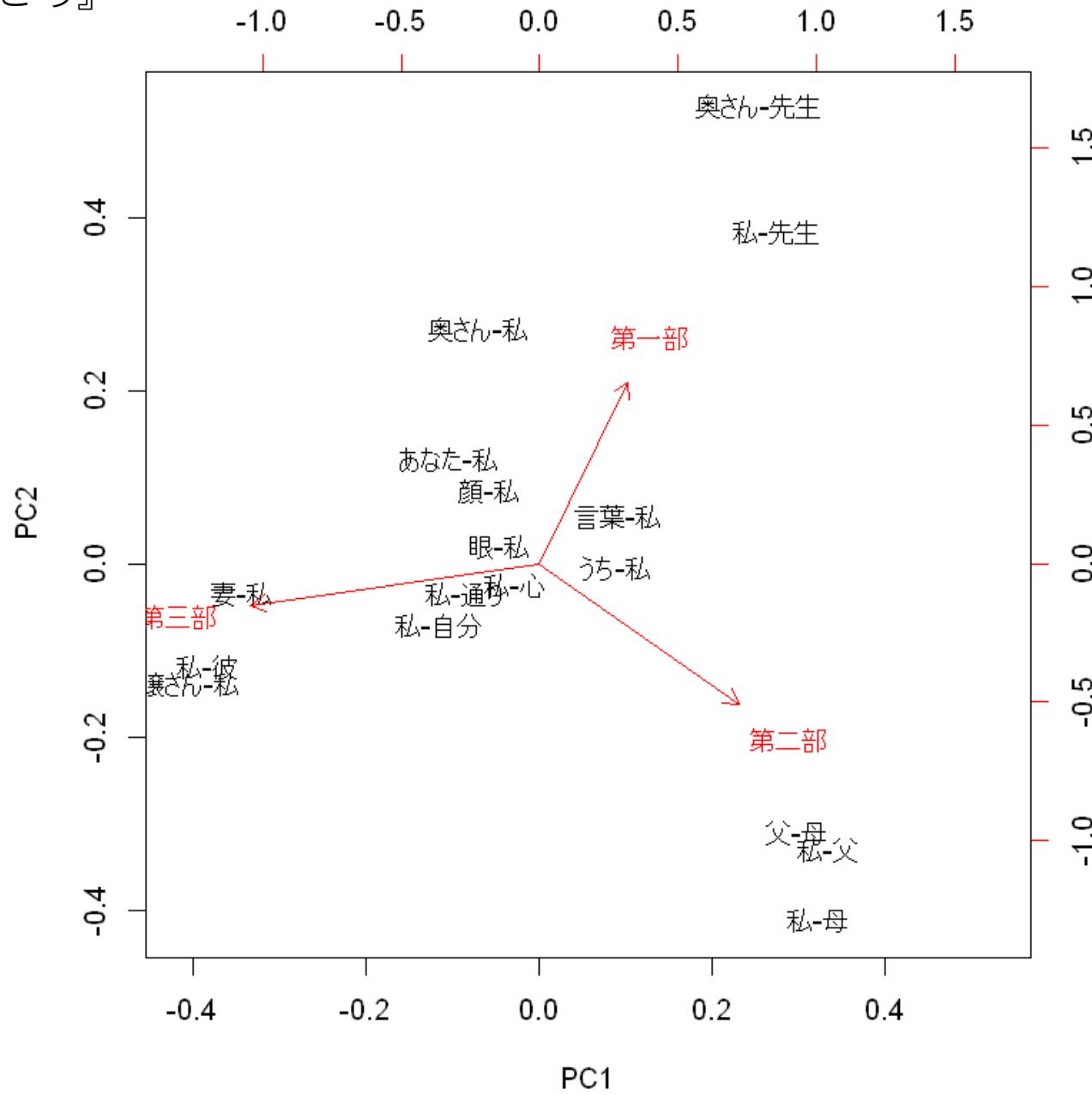
第三部における単語「私」「お嬢さん」「K」と形容詞とのバイグラム関係 ($\text{freq} > 0$)

分析②：『こころ』

Bag of wordsを用いた共起分析

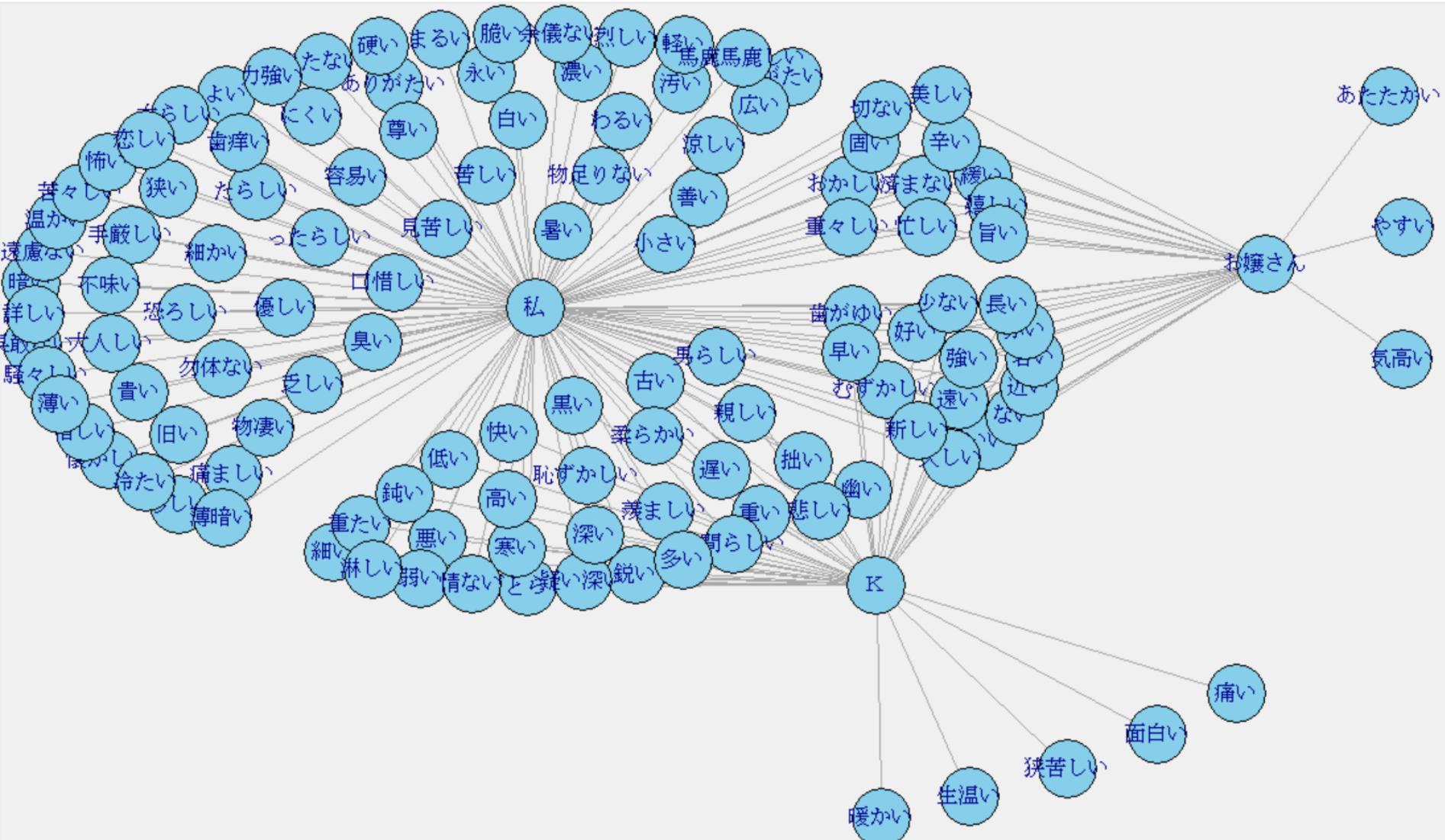
- 手続き
 - 対象とする単語群を決める(e.g. 品詞)
 - 文ごとに単語の共起を抽出する
 - Bag of wordsに基づいて、総組み合わせを抽出する
 - 集計する
 - 分析・可視化
- 特徴
 - 文内の距離・順序を保持しない
 - 空間的構造は反映されない
→文内の距離に影響されない
 - 組み合わせが膨大になる
 - Bag of wordsの長さの二乗に比例する
 - 意味的なつながりの薄い組み合わせも拾ってしまう

分析②：『こころ』



『こころ』各部と名詞同士の共起関係を用いた主成分分析

分析②：『こころ』



第三部における単語「私」「お嬢さん」「K」の形容詞との共起関係 ($\text{freq} > 0$)

分析②：『こころ』

分析の落とし穴

- 人物の名前
 - 「K」が固有名詞でなく記号として分類
 - 第一部・第二部に登場する「私」と第三部に登場する「私」は違う人物
 - 会話文の中も考えれば他の人物も「私」と発言している可能性がある
 - 第一部・第二部と、第三部の「奥さん」は違う人物
- ⇒領域/データ固有の知識が必要
- 否定表現
 - 単純に共起語を抜き出しただけではそれがどのように使われているかまではわからない
 - 「Aをした/Aである」ではなく「Aをしなかった/Aでない」かも知れない
 - 否定表現まで含めて分析したいなら構文解析などといった処理を使う必要がある
- ⇒データの性質/分析の限界を理解する

小まとめ

- データを可視化する様々な手法がある
 - 主成分分析：少ない次元に可視化する
 - 共起ネットワーク：結びつきを可視化する
 - 統計的検定に載せられないような「質的な」性質も表現できる可能性がある
- ⇒ どうすればうまくデータの性質を可視化できるかを考える
- テキスト分析の罠
 - 否定表現
 - 単語の誤認識
 - アーティファクトではないか(cf. Back et al.)
- ⇒ データの内容および背景知識、手法に関する知識を持つことで、罠にはまる可能性を減らせる

第2部 結果の報告と解釈

論文の形式

- 心理学論文はおおむねIMRAD形式(APA, 2008, 2013;日本心理学会, 2015)
 - **Introduction**
 - 緒言または序論。研究の意義と目的について説明する
 - **Materials and Methods**
 - 方法。研究の方法について説明する
 - **Results**
 - 結果。方法により得られた結果について説明する
 - **(And)**
 - **Discussion**
 - 議論と考察。得られた結果がどのような意味を持つかについて議論し、結論を述べる。

序論

- なぜその研究を行う必要があったのか
 - 先行研究のレビュー
 - 新たにその研究を行う意義
- どのような目的でその研究を行うのか
 - どのようなデータが得られるのか
 - なぜその分析をするのか
- 何を明らかにするのか
 - 理論仮説
 - 作業仮説
 - 予測

目的があるということ

- 例①
 - 「本研究では夏目漱石の『こころ』を分析した。」
- 例②
 - 「本研究では、夏目漱石の『こころ』について、各部における表現の違いを調べるために、計量テキスト分析の手法を用いた。」
- 例③
 - 「本研究では、計量的な手法でテキストの内容を可視化できるか検討するために、夏目漱石の『こころ』を題材に分析を行った。」

方法

- どのようにその研究を行ったのか
 - 研究に用いた材料と方法論について述べる
 - 研究の妥当性の根幹にかかわる部分
- 何をどのように測定したか
 - 特に、理論的な概念と材料（データ）がどのように対応するのか→構成概念妥当性
 - どんなに立派な序論を書いてもここがダメだとダメ
 - 対応が取れていないと総じて無価値な研究になりがち
 - 他の人がそれを読んで研究結果を再現できるように

報告すべき内容

- 分析対象（データセット）
 - データセットの概要
 - データセットの入手方法
 - データセットの構造
- データ処理
 - クリーニング
 - 前処理
 - 除外データの有無
 - 分析に用いる変数・尺度
- 使用ツール
 - プログラム言語
 - 形態素解析エンジン
- その他特記事項
 - 研究倫理（個人情報の保護など）

方法の書き方の例

- データセット
 - 分析には『こころ』本文を用いた。
 - テキストは青空文庫(<https://www.aozora.gr.jp/>)にて公開されているものを用いた。データの取得には石田(2017)のAozora関数を用い、同時に元のHTMLにあるルビ等のタグは削除した。
 - テキストは一文を一行のレコードとし、部・節の番号、並びに登場した順に段落・文の番号を1から連番のIDとして付した。このIDは、第二部第一節第三段落第四文であれば2,1,3,4となる。このように一文を単位にしたデータをタブ区切りのファイルとし、分析に用いた。

方法の書き方の例

- クリーニング・前処理
 - クリーニング作業として、誤字や脱字が含まれていないかどうかをチェックした。具体的には形態素解析の結果分割した内容を集計した頻度表を作成し、登場人物などの固有名詞が認識されているか、また、誤認識の結果が集計されていないかを確認した。
 - 確認作業の結果、うまく認識されていなかった単語については形態素解析ソフトの辞書に追加した。
 - 漢字の送り仮名などの表記ゆれについては同じ単語として集計されるよう統一した
- 除外データの有無
 - 今回の分析にはすべての文を分析の対象とした
 - 頻度の高い語のうち、「もの」「こと」などのように、内容の理解に寄与しない単語群をストップワードとして集計から除外した

方法の書き方の例

- 分析に用いる尺度・変数
 - 分析に用いる変数は以下のように集計した。
 - 頻度：出現した単語をそのまま集計した。
 - バイグラム：文から抽出した対象の品詞の連續をバイグラムとしてカウントした。すなわち、「吾輩は猫である」から、名詞のみを抽出した場合、「吾輩-猫」がバイグラムとなる
 - 共起：文から単語をbag of wordsとして抽出し、そのすべての単語の組み合わせを共起としてカウントした。ただし、同じ単語が一文に2回以上出現しても1単語として組み合わせを求めた。

方法の書き方の例

- 使用ツール
 - 分析にはR(3.4.4)を用いた。
 - 日本語の形態素解析には工藤ら(2004)のMeCab(ver. 0.996)を用いた。Rからの操作にはRMeCabパッケージ(石田, 2017)を用いた。

結果

- 何がデータから得られたか
 - データの特徴に関する記述
 - 基本的な記述統計→基本的な事実の確認
 - 実験であれば操作の妥当性に関するチェックなど
 - 主要な分析結果とその説明

研究の結果を、内容の重要度に従って事実に即して忠実に述べる。自分の予期に反した事実も省略しない。

心理学における研究では統計的仮説検定が分析にしばしば用いられるが、仮説検定はデータ分析の一側面に限られる。必要に応じて仮説検定に限らず適切な分析手法を用いるのが望ましい。特に、仮説検定の適用にあたっては、前提とするデータの性質（データの分布の正規性や、標本相互の独立性など）が成立していることを確認する。

分析結果の記述においては、研究結果の重要性を評価できるよう効果量とその信頼区間も示す。元来の測定単位・尺度によって表された効果量は理解が容易であるが、必要に応じて尺度に依存しない標準化された効果量の指標（Cohen の d や標準化回帰係数等）を示す。

データの欠測は分析の結果に大きな影響をしばしば与える。欠測を伴うデータを分析する場合には、欠測の頻度や件数を示すとともに、欠測の発生について経験的あるいは理論的な説明を記述する。分析において採用した欠測モデルの性質（MCAR, MAR, NMAR の区分）や、欠測に対応するために採用した方法（多重埋め合わせなど）について記述することが望ましい。

結果の報告と解釈

- 報告：事実 = 得られたデータについて述べる
 - ○ 「データは～である」 → 事実
 - × 「人々は～である」 → 推論
- 解釈：事実 = 得られたデータについての説明
 - ○ 「実験の結果～であった。そのため、参加者は～で
あったと考えられる」
 - × 「実験の結果～であった。そのため、人は～であると
考えられる」
- 一般化可能性については議論のセクションで述べる

議論と考察

- 何が明らかになったのか
 - 結果のまとめ
 - そこから何が言えるのか
- どのような意味を持つか
 - 結果が仮に正しく、一般化したとすれば、どのような意味があるか
 - 先行研究との比較
- どこまで一般化できるか
 - 研究の適用範囲・限界
 - 結果の妥当性・信頼性
- 結び
 - 今後の展望など

小まとめ

- 序論
 - なぜその研究や分析をしたのかをわかるように書く
 - 特に目的は明瞭に書く
- 方法
 - 何をどうやって扱ったのかを明瞭に書く
 - 読んだ人が同じ手続きを取れるように書く
- 結果
 - データから事実としていえることを書く
 - そのことをわかりやすく言い換える
 - データは概念そのものでないことに注意する
- 考察
 - 結果から導き出される結論を書く
 - 序論で言及したことに対して結果がどういう意味を持つか議論する

References

- American Psychological Association. (2013). *Publication manual of the American Psychological Association, Sixth Edition*. American Psychological Association.
- APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, 63(9), 839–851. <https://doi.org/10.1037/0003-066X.63.9.839>.Reporting
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Sebastopol, CA: O'Reilly Media.
- Gutenberg, B., & Richter, C. F. (1941). Seismicity of the Earth. *Geological Society of America Special Paper*, 34, 1–131.
- Pareto, V. (1982). First course in applied political economy, academic year 1893--1894. *Premier Cours d'économie Appliquée, Année Académique 1893--1894*.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.
- Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, 59(10), 1–23.
<https://doi.org/10.18637/jss.v059.i10> (Wickham, H. (n.d.) 西原史暁(訳) 【翻訳】整然データ
<https://id.fnsr.info/2017/01/09/trans-tidy-data/> (2019年7月1日アクセス))
- Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.
- 奥村学. (2010). 自然言語処理の基礎. コロナ社.
- 高村大也. (2010). 言語処理のための機械学習入門. コロナ社.
- 石田基広, & 金明哲(編). 2012コーパスとテキストマイニング. 共立出版.
- 石田基広. (2017). *R*によるテキストマイニング入門 第2版. 森北出版.
- 工藤拓, 山本薰, & 松本裕治. (2004). Conditional Random Fieldsを用いた日本語形態素解析. 情報処理学会研究報告自然言語処理 (NL) , 2004(47), 89–96. Retrieved from <https://ci.nii.ac.jp/naid/110002911717/>
- 日本経済新聞. (2016). 漱石没後100年、人気衰えず書店で文庫フェア:日本経済新聞. Retrieved July 9, 2019, from https://www.nikkei.com/article/DGXLASDG08H0C_U6A211C1CR0000/

全体のまとめ

- テキストマイニング/計量テキスト分析
 - 手近なものが分析対象になる
 - インタビュー
 - 文学作品
 - スピーチ・演説
 - SNS
- ⇒研究の幅が広がる
- 研究デザインは大事
 - テキスト分析固有の罠もある
 - 手法の可能性と限界を理解しつつ活用するのがよい