

テキストマイニングの実践 — 4日目 —

2021/7/16

人文社会ビジネス科学学術院
ビジネス科学研究群

スケジュール

- 1日目: 6/25(金)
 - 説明 — テキストマイニングの手順
 - 説明 — データをよく知る (Excel)
- 2日目: 7/2(金)
 - 説明 — テキストマイニングツールの使い方 (KHCoder)
- 3日目: 7/9(金)
 - 説明 — データ分析の実践 (KHCoder)
 - 実習 — データ分析の実践 (KHCoder)
- 4日目: 7/16(金)
 - Text Mining Studio 利用体験
 - 実習 — データ分析の実践 (KHCoder)
- 体育の日: 7/23(金)
- 5日目: 7/30(金)
 - 発表 — データ分析の実践 (KHCoder)

※ お知らせ ※

3日目, 4日目後半, 5日目 は, Zoom のブレイクアウトルーム機能を使ったグループワークになります。

本日の内容

- 前半 18:20 ~ 20:10
 - Text Mining Studio 紹介
- 後半 20:20 ~ 21:00
 - 実習 グループワーク (議論や資料作成)

予告: 発表会(7/30)の予定

- Q&A (Max ~18:35)
- 発表会 (各グループ 説明10分, 質疑5分)
 - 前半 18:35 ~ 19:35
 - グループ1 ~ グループ4
 - 後半 19:45 ~ 21:00
 - グループ5 ~ グループ9

発表内容

1. テーマ設定

例) 「A. クチコミデータ」であれば, 好評価のエリアに倣って, 低評価のエリアを改善するプランを提案する

例) 「B. Twitterデータ」であれば, シーズン(期間1, 期間2)ごとの国民の心情の変化を分析し, 新たな施策を提案する

2. 分析結果 (プロットおよび考察)

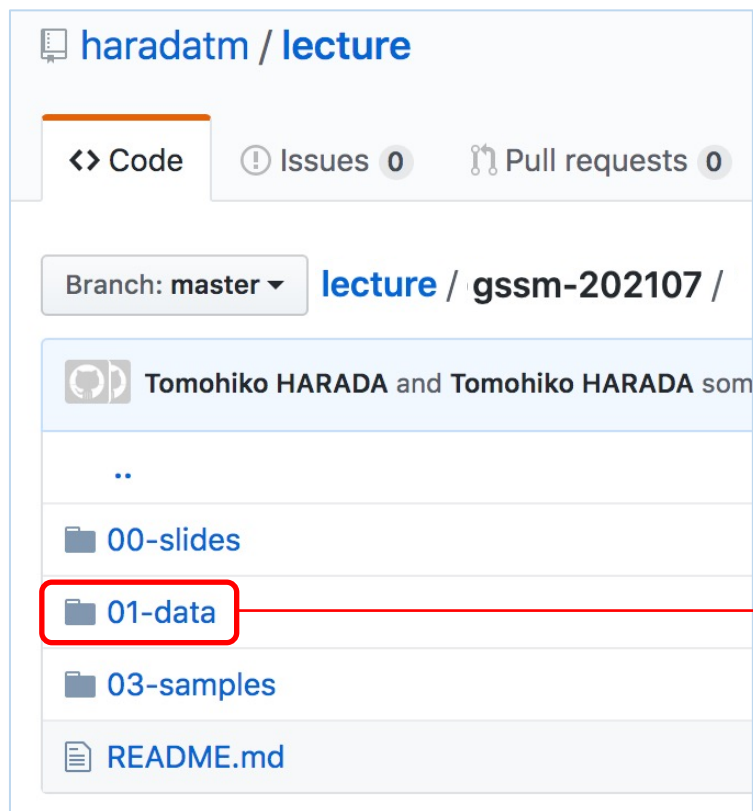
- テーマや仮説にもとづくストーリーで, 分析を進めるのがベター
- 支持する**プロット**とユーザーの**生の声(原文)**を使って主張する

(再掲) 実習用のデータ

データファイル名	件数	データセット	備考
rakuten-1000-2020-2021.xlsx	10,000	<ul style="list-style-type: none">• レジャー+ビジネスの 10エリア• エリアごと 1,000件 (ランダムサンプリング)• 期間: 2020/1/1~2021/5/12	<ul style="list-style-type: none">• 本講義の全体を通して利用する
rakuten-1000-2018-2019.xlsx	10,000	<ul style="list-style-type: none">• レジャー+ビジネスの 10エリア• エリアごと 1,000件 (ランダムサンプリング)• 期間: 2018/1/1~2019/12/31	<ul style="list-style-type: none">• 実習用 (3~4日目)
covid19-10000.xlsx	10,000	<ul style="list-style-type: none">• ハッシュタグ「#新型コロナ」がついたツイート• Search API (1%サンプリング) で取得した 32万 →10,000件 (ランダムサンプリング)• 期間: 2020/4/24~2021/5/31	<ul style="list-style-type: none">• 実習用 (3~4日目)

(再掲) データの取得方法

- <https://github.com/haradatm/lecture/tree/master/gssm-202107>



master lecture / gssm-202107 / 01-data / Go to file

Tomohiko HARADA and Tomohiko HARADA some updated ... 3 h

..

README.md some updated

README.md

Download data (to be used in the exercise)

file name	# records	size (zipped)	period
rakuten-1000-2020-2021.xlsx.zip	10,000	2.4 MB	2020/1/1~2021/5/12
rakuten-1000-2018-2019.xlsx.zip	10,000	2.4 MB	2018/1/1~2019/12/31
covid19-10000.xlsx.zip	10,000	2.4 MB	2020/1/1~2021/5/31

本講義で主として使用

課題 (4日目)

※ TMS は, 全学の RDP にはインストールできません!
各自の Windows PC にインストールしてください

- 講義で利用した「**rakuten-1000-2020-2021.xlsx**」を使って,
TMS で「ことばネットワーク」を作成し, KHcoder (共起ネットワーク)で見えない気づきを述べてください
 - Windows PC がない方の救済: 受講したTMSセミナーを通して気づいた KHcoder との違い(メリデメ等)を述べてください
- グループワークについて,以下を記載して提出してください
 - 所属する **グループ名**、グループで取り上げる **分析テーマ**
 - **メンバ全員の名前** および **発表者名**
- 形式: PPT(PDF), 提出先: manaba, 期限: 7/23(祝) 21:00

参考書

(KH Coder)

- [1] 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.
- [2] 樋口耕一. テキスト型データの計量的分析—2つのアプローチの峻別と統合—. 理論と方法, 数理社会学会, 2004, 19(1): 101-115.
- [3] 牛澤賢二. やってみよう テキストマイニング—自由回答アンケートの分析に挑戦!. 朝倉書店, 2019

(Windows環境によるデータ収集方法の参考に)

- [4] テキストマイニングソフトを利用した新未来洞察手法の研究. 第10分科会, (財)市場創造研究会. http://www.shijo-sozo.org/news/第10分科会_1.pdf

参考書

(R を使った参考書)

- [5] 金明哲. "テキストデータの統計科学入門." 岩波書店, 2009.
- [6] 石田基広. "RMeCab によるテキスト解析. R によるテキストマイニング入門." 森北出版, 2008, 51-82.

(他のツールを使った参考書)

- [7] 那須川哲哉. "テキストマイニングを使う技術/作る技術: 基礎技術と適用事例から導く本質と活用法." 東京電機大学出版局, 2006.
- [8] 上田隆穂, 黒岩祥太, 戸谷圭子. "テキストマイニングによるマーケティング調査." 講談社, 2005.

(統計解析を中心とした参考書)

- [9] 前田忠彦; 山崎誠. 言語研究のための統計入門. くろしお出版株式会社, 東京, 2013.