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)

前回の課題 — 正解例

- 手順
  1. ファイル「khcoder3¥plugin_jp¥z1_edit_words3.pm」を編集する
  2. ↓
  3. KH Coder を再起動する
  4. プロジェクトファイルを開く
  5. メニューから「ツール」「プラグイン」「表記ゆれの吸収」を選ぶ
  6. 分析を続ける

```
package z1_edit_words3;
use utf8;

my $config = {
    '友達' =>
        [
            '友人',
            '旧友',
            '親友',
            '盟友',
            '友',
        ],
    '格別' =>
        [
            '特別',
            '格別', # 通常
        ],          # の
    '偶然' =>
        [
            '偶然', # 形容
        ],
};
```

→

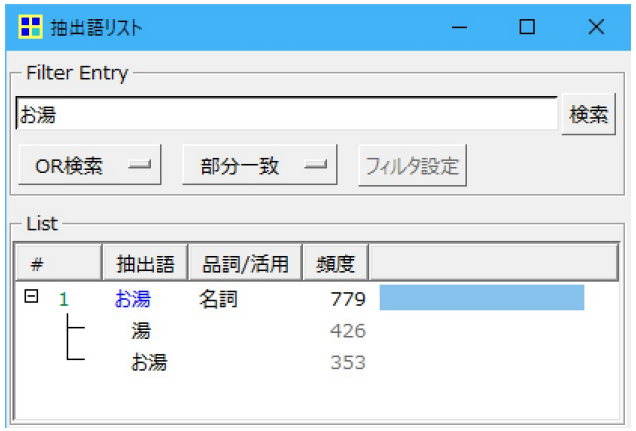
```
package z1_edit_words3;
use utf8;

my $config = {
    'お湯' =>
        [
            '湯',
        ],
};
```

編集前　　編集後

適用後の例 →

「お湯」と「湯」がひとつの単語にまとまっている



前回 講義後のQ&A

Q: 樋口先生のチュートリアル,夏目漱石の小説を一体どうやって読み込んだの?

A: 青空文庫 (著作権が消滅した作品や著者が許諾した作品のテキストを公開している電子図書館) で公開されています

- https://www.aozora.gr.jp/

01_登別	
風呂	.058
温泉	.045
美味しい	.043
残念	.034
お部屋	.032
最高	.030
バイキング	.030
露天風呂	.029
大変	.028
夕食	.027

Q: この最下位の0.027は非常に数値が低いため特徴とは言い難い…という解釈ですか?

A: データは,その収集方法や時期,掲載元の特性などの影響を受けるため, 単独で数値の高(頻度や共起率)を議論するのでなく,エリア内/外で比較するなど, 常に**相対的に見る**ことが重要です

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

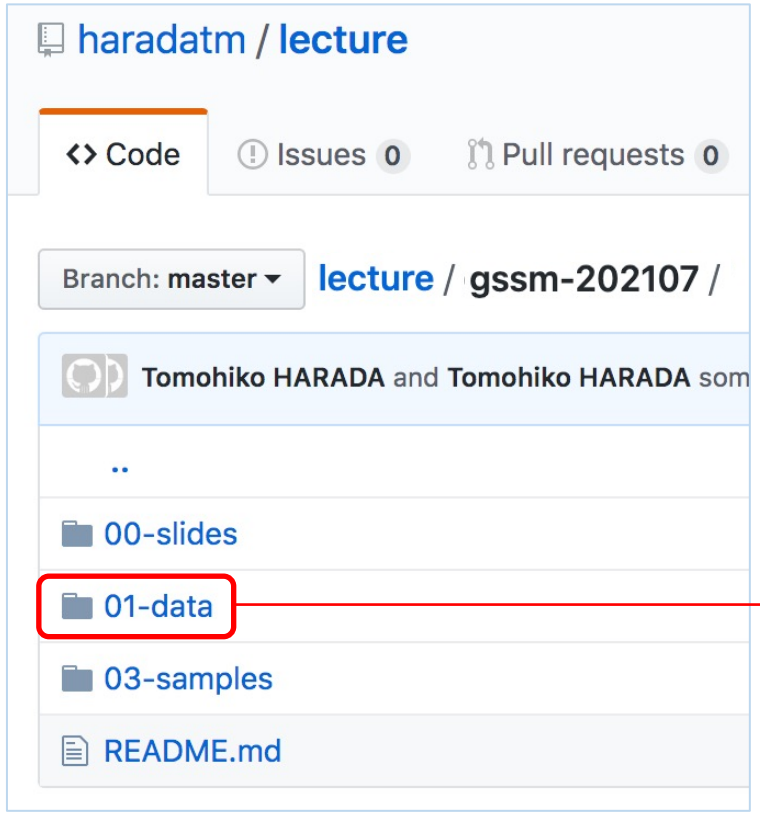
2. 分析結果 (プロットおよび考察)
   - テーマや仮説にもとづくストーリで,分析を進めるのがベター
   - 支持する**プロット**とユーザーの**生の声(原文)**を使って主張する

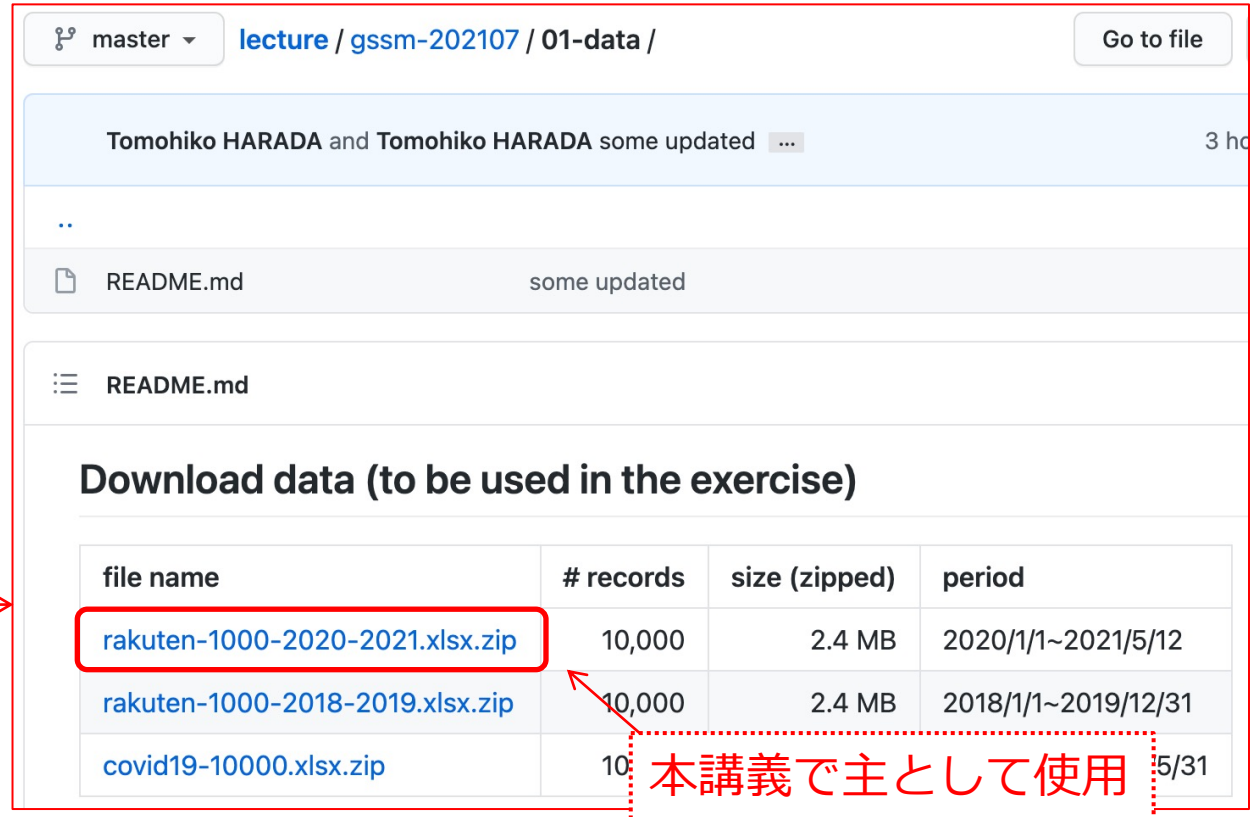
(再掲) 実習用のデータ

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	• 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	• 本講義の全体を通して利用する
rakuten-1000-2018-2019.xlsx	10,000	• 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	• 実習用 (3~4日目)
covid19-10000.xlsx	10,000	• 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	• 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**





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

A) 講義で利用した「**rakuten-1000-2020-2021.xlsx**」を使って, **TMS** で「ことばネッネットワーク」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
- 形式: PPT(PDF)，提出先: manaba，期限: 7/30(金) 21:00

B) 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.