

テキストマイニングの実践

—3日目—

2020/7/17

ビジネス科学研究科
経営システム科学専攻

講義スライド

- <https://github.com/haradatm/lecture/tree/master/gssm-202007>



スケジュール

- 1日目: 7/1(水)
 - 説明 — テキストマイニングの手順
 - 実習 — データをよく知る (Excel)
- 2日目: 7/10(金)
 - 説明 — テキストマイニングツールの使い方 (KHCoder)
- **3日目: 7/17(金)**
 - **説明 — データ分析の実践 (KHCoder)**
 - **実習 — データ分析の実践 (KHCoder)**
- **体育の日: 7/24(金)**
- 4日目: 7/31(金)
 - Text Mining Studio 利用体験
- 5日目: 8/7(金)
 - 発表 — データ分析の実践 (KHCoder)

グループ分け

グループ1	201940108
	202040051
	202040055
	202040074
	202040077
グループ2	201945014
	202040057
	202040063
	202040080
	MSI
グループ3	202040052
	202040062
	202040070
	202040072
	202040170

グループ4	201947529
	202040058
	202040060
	202040068
	202040069
グループ5	201940015
	201947523
	202040059
	202040066
	202040073
グループ6	201840109
	201940129
	202040064
	202040078
	202040079

グループ7	202040071
	202040076
	202040409
	202040413
グループ8	202020027
	202020051
	202040067
	202040075
グループ9	202040053
	202040054
	202040056
	202040061

KH Coder で単語登録する

- 目的
 - 複数の単語に分かれる → 1単語として抽出できるようにする
例) 「湯」「畠」の 2単語 → 「湯畠」として 1単語
- 方法
 - 「前処理の実行」前に「強制出力する語の指定」に追加する
- 手順
 1. メニューから「前処理」「語の取捨選択」を選ぶ
 - 「強制出力する語の指定」欄に抽出したい単語を登録する
 - 「OK」ボタンで画面を閉じる
 2. メニューから「前処理」「前処理の実行」を選ぶ

KH Coder で表記ゆれを吸収する (1/2)

- 目的
 - 同じ意味の単語を同一視する別の単語として扱わない
例) 「お湯」 「湯」 の 2単語 → どちらも「お湯」としてカウント

- 方法
 - 「表記揺れを吸収」 プラグインを利用する
- 手順

1. プラグインをダウンロードし, 解凍して ~~plugin_jp~~ 配下へコピー

[ダウンロード URL] http://koichi.nihon.to/psnl/tmp/z1_edit_words3.zip

[解凍後ファイル名] z1_edit_words3.zip → z1_edit_words3.pm

—— [配置後のパス] khcoder3\plugin_jp\z1_edit_words3.pm

注: 最新版ではこのプラグインが
あらかじめインストールされ
ています

(次ページにつづく)

KH Coder で表記ゆれを吸収する (2/2)

- 手順

2. プラグインファイル
`z1_edit_words3.pm` を編集する

→

```
1 package z1_edit_words3;
2 use utf8;
3
4 my $config = {
5     '友達' =>
6         [
7             '友人',
8             '旧友',
9             '親友',
10            '盟友',
11            '友',
12        ],
13        '格別' =>
14        [
15            '特別',
16            '格別', # 通常
17        ], # の
18        '偶然' =>
19        [
20            '偶然', # 形容
21        ],
22    };
23 }
```

編集前

```
1 package z1_edit_words3;
2 use utf8;
3
4 my $config = {
5     'お湯' =>
6         [
7             '湯',
8         ],
9    };
10 }
```

編集後

- ↓
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

適用後の例 →

「お湯」と「湯」が
ひとつの単語にまと
まっている

#	抽出語	品詞/活用	頻度
1	お湯	名詞	779
2	湯		426
3	お湯		353

(復習) テキストマイニングの手順

・データをよく知る

- ・データ件数や構成比を集計 → データを理解する
 - ・旅行目的別の人気エリアは?
 - ・同伴者別の人気エリアは?
 - ・数値評価による人気エリアの差異は?

・テーマを設定する

- ・解決すべき課題を決める → 分析目的を明確にする
 - ・数値評価が低い原因は?
 - ・高評価の施設に学ぶ改善点は?

・データ分析に取り組む

- ・これら課題を解決するために、テキスト分析を実施

(再掲) 数値評価の平均

エリア別

- レジャーは、風呂や食事が設備や部屋に比べて高評価

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.16	4.20	4.04	3.98	4.22	4.22	4.23
01_登別	3.88	4.13	3.83	3.78	4.16	3.93	4.02
02_草津	4.12	4.29	3.95	3.87	4.25	4.15	4.22
03_箱根	4.18	4.07	4.04	3.97	4.19	4.27	4.19
04_道後	4.01	4.21	3.97	3.90	3.96	4.14	4.15
05_湯布院	4.60	4.29	4.43	4.35	4.53	4.60	4.57
B_ビジネス	3.91	4.24	3.98	3.85	3.69	・湯布院は、レジャーの中で、軒並み高評価が多い	
06_札幌	3.95	4.20	4.01	3.81	3.66	・レジャーもビジネスも立地が評価される ・ビジネスは、立地がその他に比べて高評価	
07_名古屋	3.96	4.15	3.99	3.89	3.74		
08_東京	3.85	4.33	3.91	3.81	3.67	3.91	4.04
09_大阪	3.92	4.32	4.04	3.90	3.74	4.02	4.13
10_福岡	3.85	4.21					4.00

カテゴリー「レジャー」「ビジネス」別)

- レジャーもビジネスも立地が評価される
- ビジネスは、立地がその他に比べて高評価

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.16	4.20	4.04	3.98	4.22	4.22	4.23
B_ビジネス	3.91	4.24	3.98	3.85	3.69	3.94	4.08

実践的な分析 — 特徴語の集計

- 宿泊客は、どの項目に注目しているか？
 1. カテゴリー「レジャー」と「ビジネス」を比較する
 2. カテゴリー「レジャー」(or 「ビジネス」) の 5エリアを比較する
- 手順
 - テキスト中の特徴語を集計

「ツール」→「抽出語」→「関連語検索」→「#直接入力[and]“<>カテゴリー-->A_レ
ジャー”」「集計単位:文」→「フィルタ設定」→「品詞=名詞, 形容動詞, 未知語, タグ, 形容詞,
名詞B, 形容詞B, 名詞C」を選択→「集計」→結果を選択し「コピー」
 - エリアによって特徴語がどう異なるかを比較
 - 注目する項目の違いを考察する

直接入力: [and] の右側に入力する条件

レジャー:

<>カテゴリー-->A_レジャー

<>エリア-->01_登別

<>エリア-->02_草津

<>エリア-->03_箱根

<>エリア-->04_道後

<>エリア-->05_湯布院

ビジネス:

<>カテゴリー-->B_ビジネス

<>エリア-->06_札幌

<>エリア-->07_名古屋

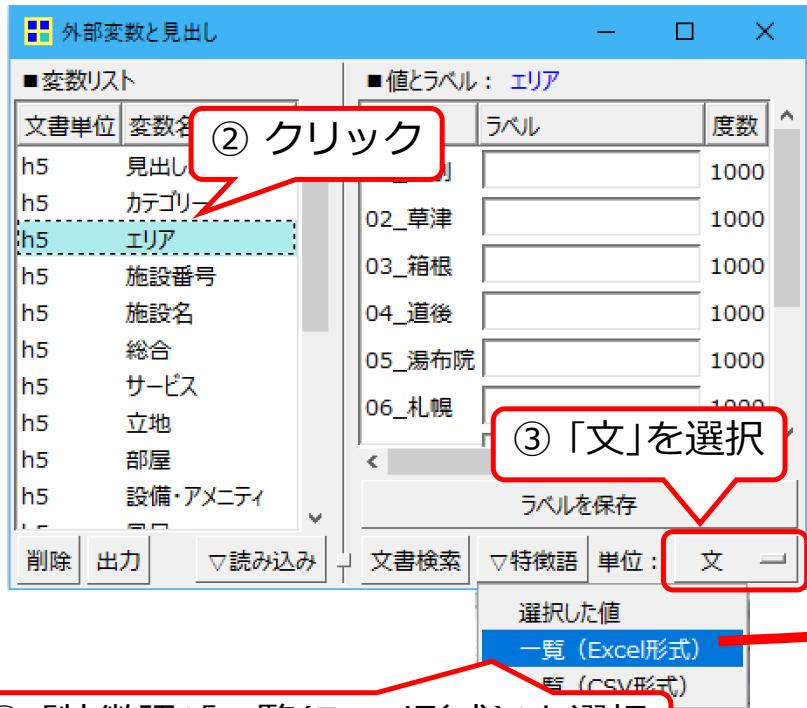
<>エリア-->08_東京

<>エリア-->09_大阪

<>エリア-->10_福岡

使い方 — 外部変数(エリア)を利用する

①メニューから「ツール」「外部変数と見出し」「リスト」を開く



	B	C	D	E	F	G	H	I	J	K
2	01_登別		02_草津		03_箱根		04_道後			
3	食事	.059	湯畑	.081	食事	.070	温泉	.058		
4	部屋	.058	草津	.066	良い	.064	良い	.053		
5	良い	.055	温泉	.066	風呂	.056	利用	.052		
6	風呂	.053	良い	.064	美味しい	.053	ホテル	.045		
7	宿泊	.045	風呂	.064	お部屋	.045	朝食	.044		
8	温泉	.043	食事	.056	満足	.044	道後	.042		
9	美味しい	.039	宿泊	.046	スタッフ	.041	宿泊	.041		
10	満足	.035	満足	.041	温泉	.040	満足	.034		
11	残念	.035	宿	.040	宿	.038	松山	.033		
12	行く	.031	美味しい	.040	露天風呂	.036	美味しい	.033		
13	05_湯布院		06_札幌		07_名古屋		08_東京			
14	食事	.070	札幌	.058	名古屋	.060	利用	.062		
15	美味しい	.067	思う	.056	利用	.056	部屋	.057		
16	宿	.064	部屋	.055	朝食	.053	ホテル	.056		
17	風呂	.063	ホテル	.053	ホテル	.052	駅	.046		
18	思う	.060	利用	.051	部屋	.048	便利	.044		
19	満足	.048	朝食	.051	駅	.038	朝食	.038		
20	料理	.045	宿泊	.046	便利	.031	立地	.035		
21	スタッフ	.043	立地	.039	立地	.029	近い	.034		
22	露天風呂	.042	広い	.031	フロント	.028	フロント	.032		
23	温泉	.042	近い	.031	近い	.027	近く	.026		
24	09_大阪		10_福岡							
25	利用	.068	利用	.059						
26	ホテル	.060	部屋	.056						
27	部屋	.054	ホテル	.049						
28	思う	.051	博多	.046						
29	大阪	.047	朝食							
30	朝食	.043	立地							
31	便利	.040								
32	立地	.040								
33	駅	.037								
34	近く	.031								

各エリアの特徴語を10件ずつ
一覧 (数値は Jaccard係数)

実践的な分析 — 特徴語の集計例

A_レジヤー	数値評価指標
良い	.094
風呂	.077
温泉	.062
食事	
美味しい	.062
お部屋	.042
宿	.042
スタッフ	.041
露天風呂	.032
残念	.030
大変	.029

B_ビジネス	数値評価指標
部屋	.102
ホテル	.085
立地	.048
ない	.047
駅	.044
便利	.043
フロント	.038
近い	.038
広い	.030
綺麗	.030

01_登別	02_草津	03_箱根	04_道後	05_湯布院					
部屋	.058	湯畠	.081	良い	.064	温泉	.058	美味しい	.067
良い	.055	温泉	.066	風呂	.056	良い	.053	宿	.064
風呂	.053	良い	.064	美味しい	.053	ホテル	.045	風呂	.063
温泉	.043	風呂	.064	お部屋	.045	ない	.035	スタッフ	.043
美味しい	.039	宿	.040	スタッフ	.041	美味しい	.033	露天風呂	.042
お部屋	.035	美味しい	.040	温泉	.040	立地	.030	温泉	.042
宿	.034	お部屋	.032	宿	.038	フロント	.025	お部屋	.041
スタッフ	.028	立地	.028	露天風呂	.036	よい	.023	家族	.035
最高	.027	スタッフ	.028	残念	.030	浴場	.023	最高	.031
バイキング	.027	残念	.027	大変	.028	便利	.023	夕食	.031

06_札幌	07_名古屋	08_東京	09_大阪	10_福岡					
部屋	.055	ホテル	.052	部屋	.057	ホテル	.060	部屋	.056
ホテル	.053	部屋	.048	ホテル	.056	部屋	.054	ホテル	.049
立地	.039	駅	.038	駅	.046	便利	.040	立地	.040
ない	.031	便利	.031	便利	.044	立地	.040	フロント	.037
駅	.031	立地	.029	ない	.039	駅	.037	近い	.033
便利	.031	フロント	.028	立地	.035	ない	.034	ない	.032
フロント	.029	近い	.027	近い	.034	近い	.031	駅	.032
近い	.028	綺麗	.026	フロント	.032	フロント	.029	便利	.031
広い	.027	快適	.024	綺麗	.026	綺麗	.028	快適	.025
綺麗	.026	広い	.022	コンビニ	.022	広い	.025	広い	.023

Tips: 「ツール」→「外部変数と見出し」→「リスト」→「変数リスト=カテゴリー」を選択→「▽特徴語」→「選択した値」→「関連語検索画面」→「フィルタ設定」→「品詞=名詞,形容動詞,未知語,タグ,形容詞,名詞B,形容詞B,名詞C」を選択→「▽特徴語」→「一覧(EXCEL形式)」で連続実行

実践的な分析 — 特徴語の共起ネット

- 宿泊客は、どの項目のどこに注目しているか？
 1. カテゴリー「レジャー」と「ビジネス」を比較する
 2. カテゴリー「レジャー」(or 「ビジネス」) の 5エリアを比較する
- 手順
 - 特徴語の共起ネットワーク図を作成

「ツール」→「抽出語」→「関連語検索」→「#直接入力[and]“<>エリア-->01_登別”」「集計単位:文」→「フィルタ設定」→「品詞=名詞,未知語,タグ,形容詞,名詞B,形容詞B,名詞C」を選択→「集計」→「共起ネット」→「調整:上位120,共起関係ほど濃い線に」
 - エリアによって特徴語(とその背景)がどう異なるかを比較
 - 注目する項目の違いを考察する

直接入力: [and] の右側に入力する条件

レジャー:

<>カテゴリー-->A_レジャー

<>エリア-->01_登別

<>エリア-->02_草津

<>エリア-->03_箱根

<>エリア-->04_道後

<>エリア-->05_湯布院

ビジネス:

<>カテゴリー-->B_ビジネス

c>エリア-->06_札幌

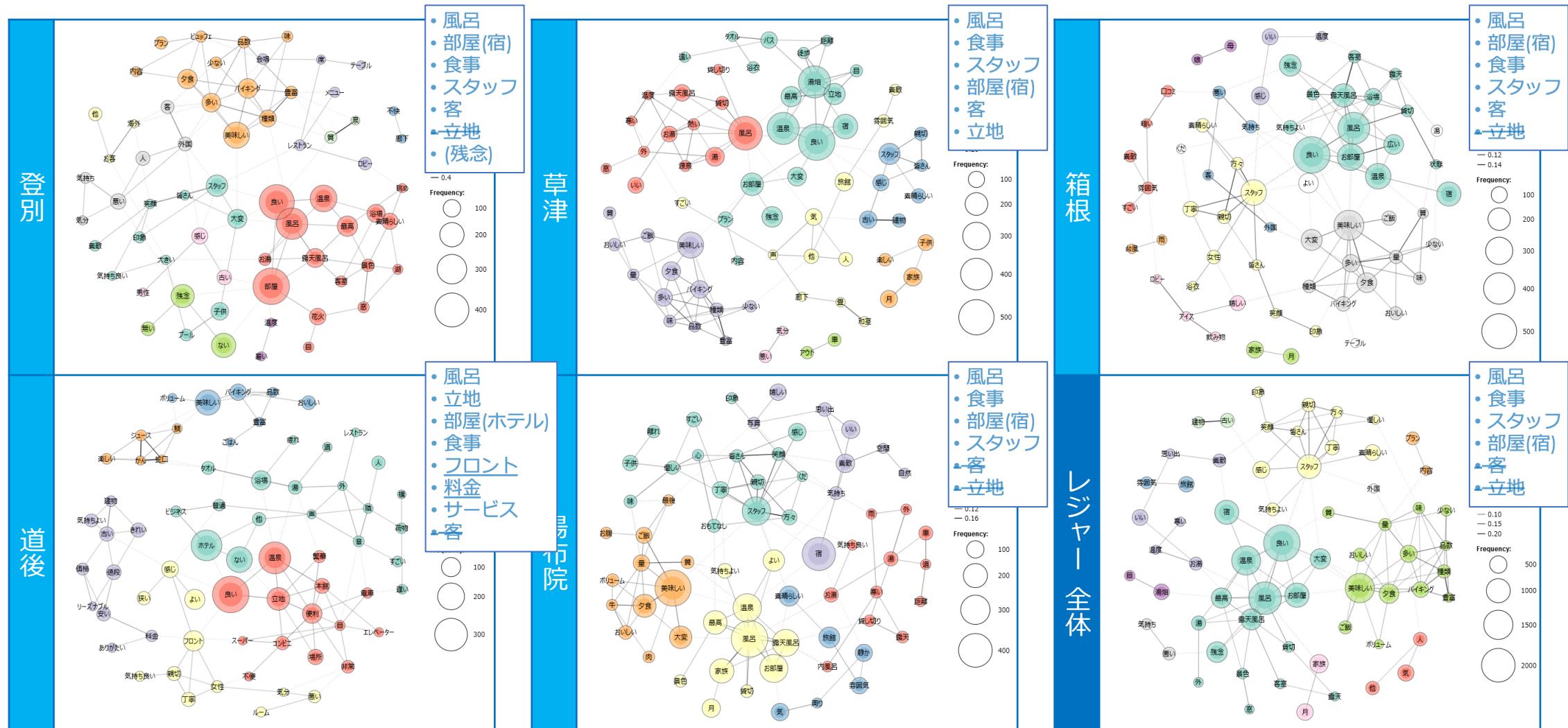
<>エリア-->07_名古屋

<>エリア-->08_東京

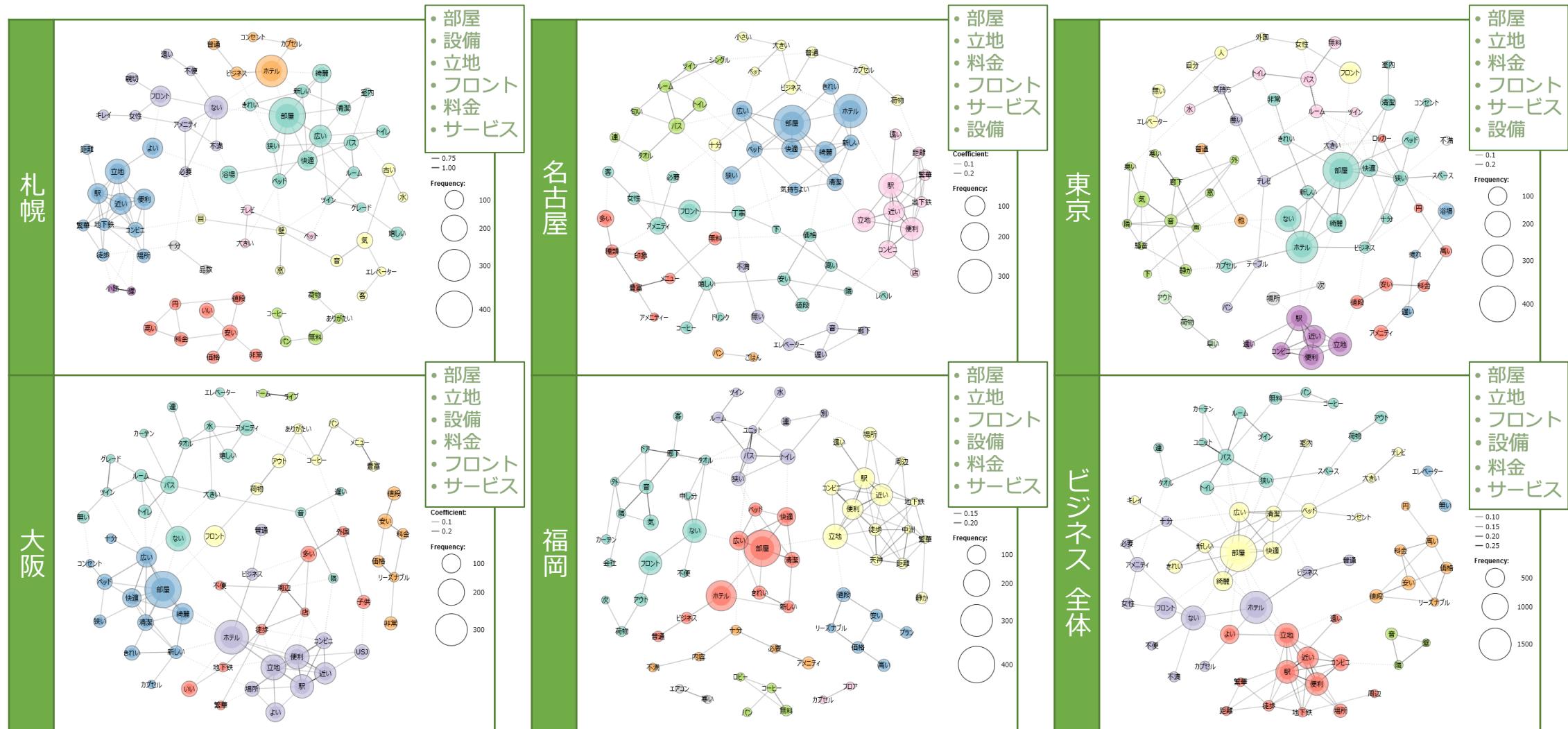
<>エリア-->09_大阪

<>エリア-->10_福岡

実践的な分析 — 共起ネットの出力例(1)



実践的な分析 — 共起ネットの出力例(2)



まとめ方の例

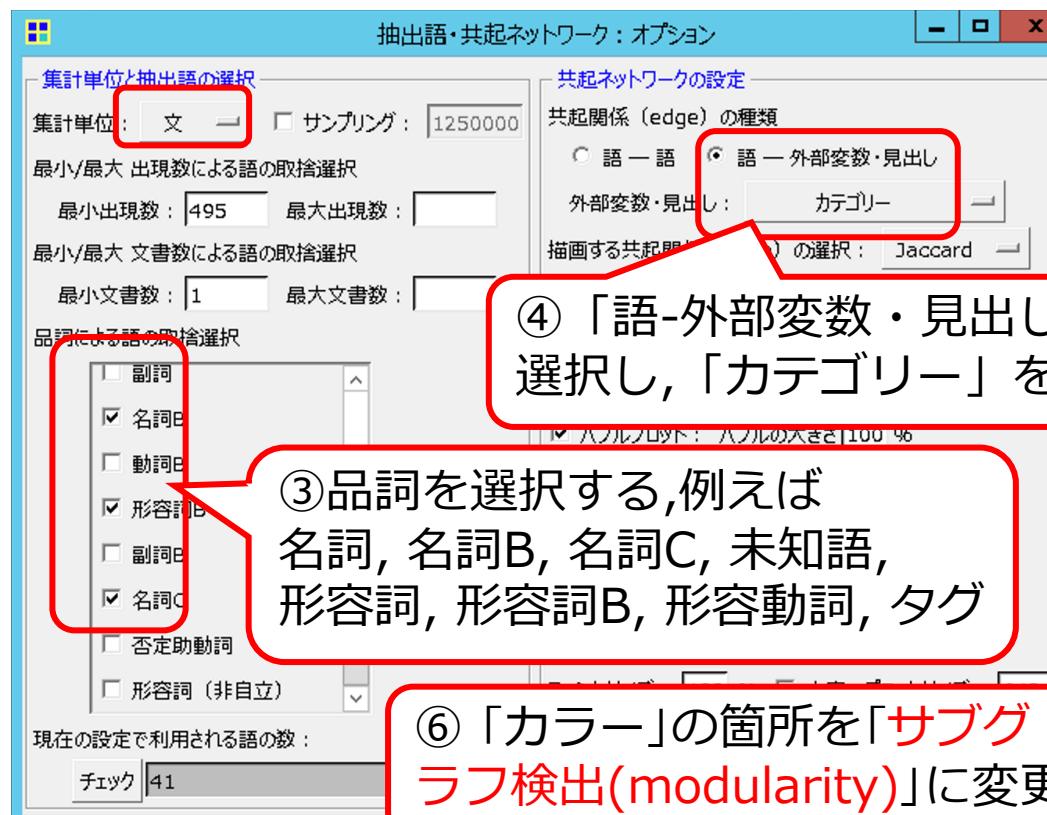
- ・宿泊客が、どの項目のどこに注目しているかを列挙する
 - ・エリアごとに、注目ポイントを列挙
 - ・エリアごとで、注目ポイントを「好評」と「不評」に分類

カテゴリー	エリア	好評	不評
レジャー	XXX	<ul style="list-style-type: none">・風呂が広い・...	<ul style="list-style-type: none">・エアコンが臭い・...

実践的な分析 — 共起ネットの出力例(3)

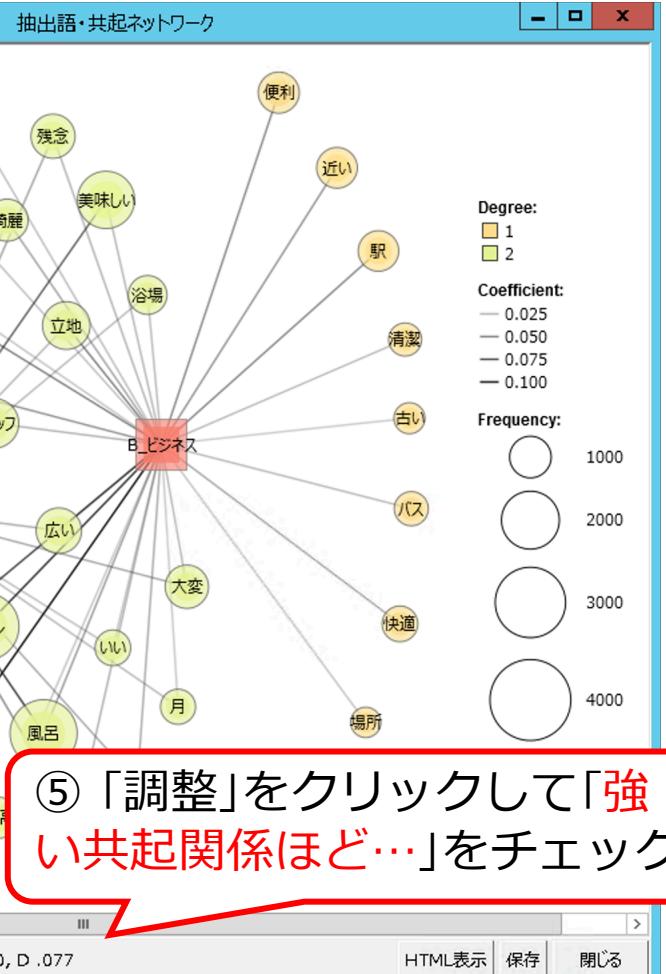
①メニューから「ツール」「抽出語」「共起ネットワーク」を選ぶ

②「集計単位」として「文」を選んで「OK」をクリック



③品詞を選択する, 例えば
名詞, 名詞B, 名詞C, 未知語,
形容詞, 形容詞B, 形容動詞, タグ

⑥「カラー」の箇所を「サブグ
ラフ検出(modularity)」に変更



⑤「調整」をクリックして「強
い共起関係ほど…」をチェック

実践的な分析 — 改善案を提案する(1/2)

- ・ユーザーは何をどう高評価しているか?
 1. カテゴリー「レジャー」と「ビジネス」を比較する
 2. 対照的な2エリアを比較する
- ・手順
 - ・特徴語とポジティブ意見の共起ネットワーク図を作成

「ツール」→「抽出語」→「関連語検索」→「#直接入力(and)“<>エリア-->01_登別”」「Search Entry:*ポジ」「AND検索」「集計単位:文」→「フィルタ設定」→「品詞=名詞,未知語,形容詞,名詞B,形容詞B,名詞C」を選択→「集計」→「共起ネット」→「調整:上位=120,共起関係ほど濃い線に」
 - ・エリアによってポジティブ意見(とその背景)どう異なるかを比較
 - ・何がどう評価されているかを考察する

実践的な分析 — 改善案を提案する(2/2)

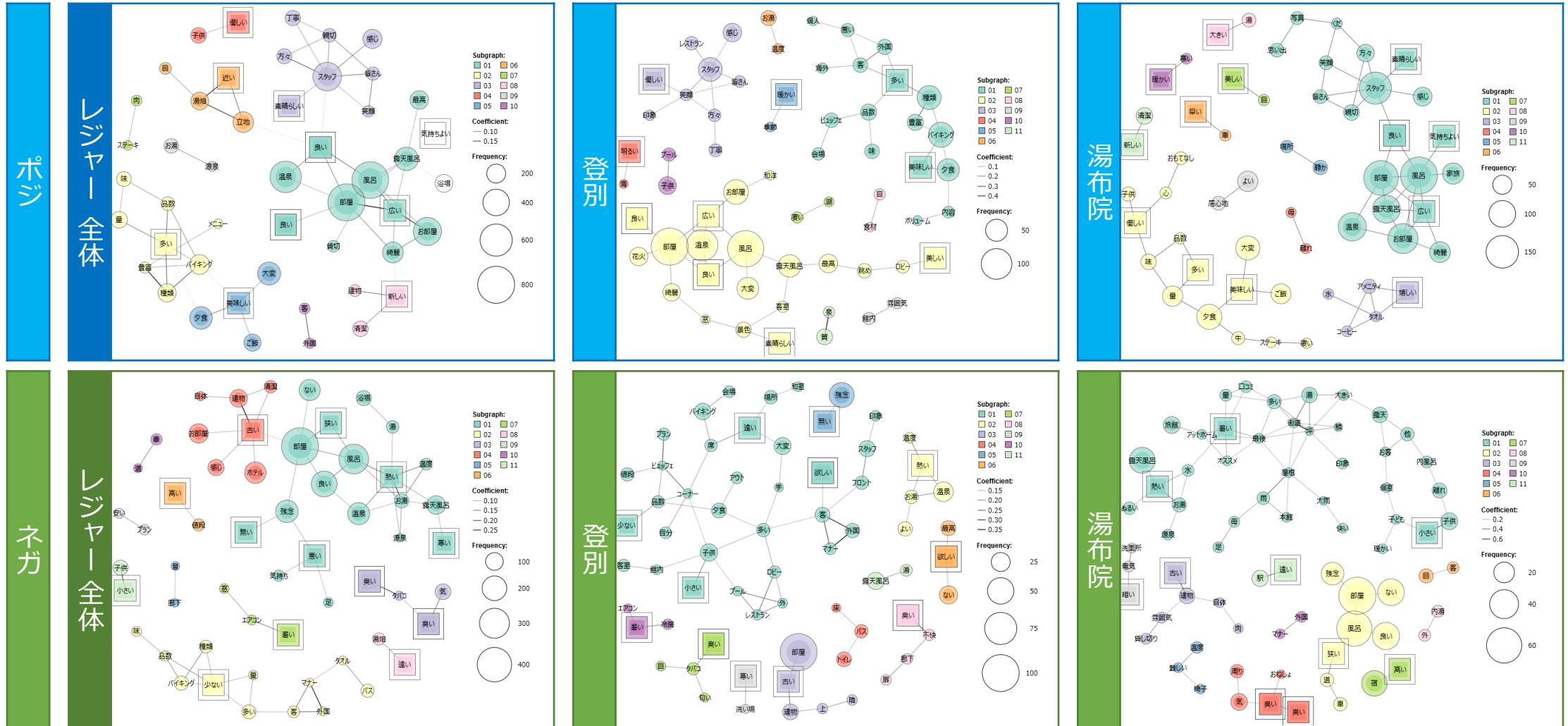
- ・ユーザーは何をどう**低評価**しているか?
 1. カテゴリー「レジャー」と「ビジネス」を比較する
 2. 対照的な2エリアを比較する

- ・手順
 - ・特徴語と**ネガティブ意見**の共起ネットワーク図を作成

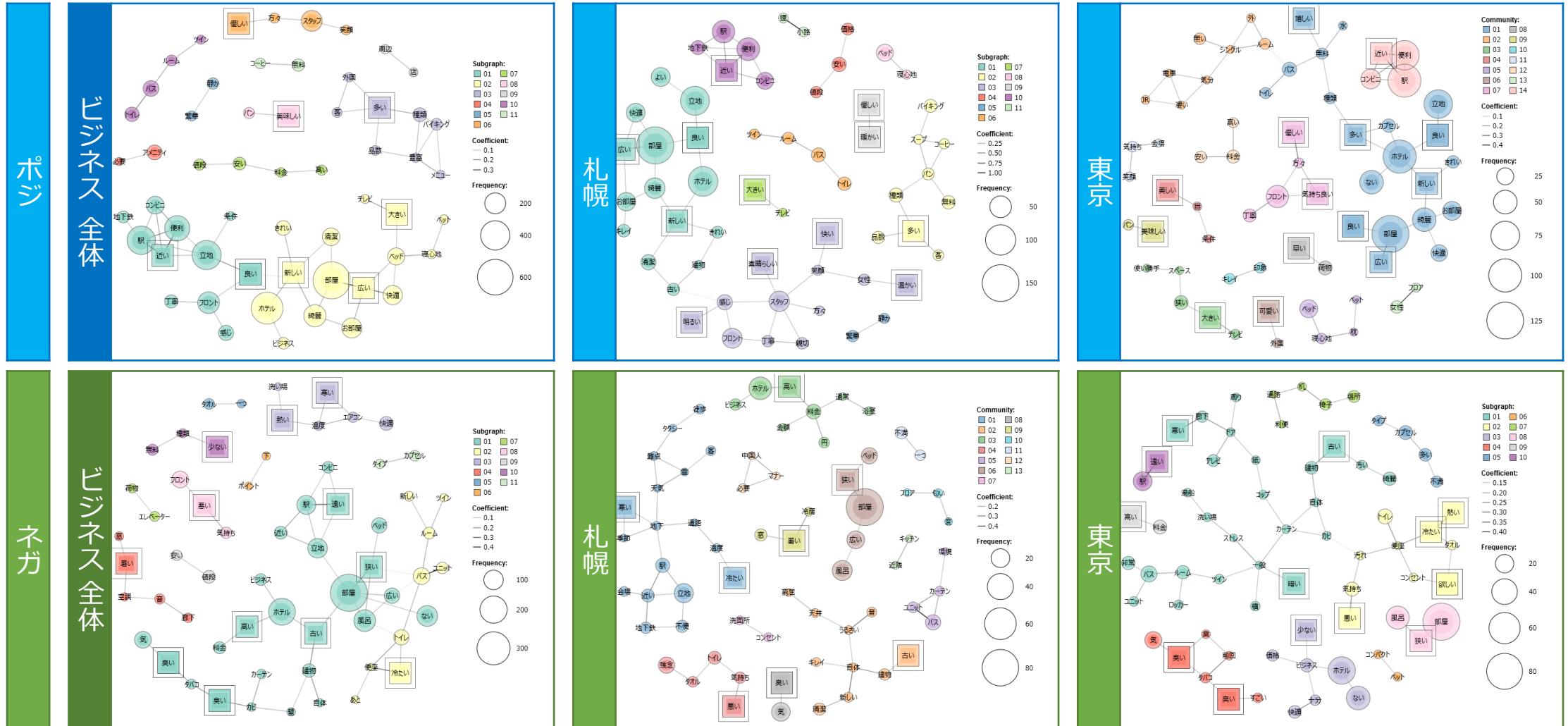
「ツール」→「抽出語」→「関連語検索」→「#直接入力(and)"<>エリア-->01_登別"」「Search Entry:*ポジ」「AND検索」「集計単位:文」→「フィルタ設定」→「品詞=名詞,未知語,形容詞,名詞B,形容詞B,名詞C」を選択→「集計」→「共起ネット」→「調整:**上位=120,共起関係ほど濃い線に**」

- ・エリアによって**ネガティブ意見**(とその背景)どう異なるかを比較
- ・エリアの課題を考察する

実践的な分析 — 登別と湯布院のポジネガ比較



実践的な分析 — 東京と札幌のポジネガ比較



まとめ方の例

- ・主張を支持する図とユーザーの生の声(原文)を使って議論する
 - ・エリア X が評価されている点は何か?
 - ・エリア Y の課題は何か?
 - ・エリア Y の改善に向けた提案?

対象エリア	エリアX の評価ポイント	エリアY の課題	エリアYの改善案
エリアX: XXX エリアY: XXX	・風呂が広い 根拠原文: ... ・...	・エアコンが臭い 根拠原文: ... ・...	・... ・...

自己紹介 10分

グループ1	201940108
	202040051
	202040055
	202040074
	202040077
グループ2	201945014
	202040057
	202040063
	202040080
	MSI
グループ3	202040052
	202040062
	202040070
	202040072
	202040170

グループ4	201947529
	202040058
	202040060
	202040068
	202040069
グループ5	201940015
	201947523
	202040059
	202040066
	202040073
グループ6	201840109
	201940129
	202040064
	202040078
	202040079

グループ7	202040071
	202040076
	202040409
	202040413
グループ8	202020027
	202020051
	202040067
	202040075
グループ9	202040053
	202040054
	202040056
	202040061

演習 — グループワーク

- テキストマイニングの手順(1日目)に倣って、テキストデータの分析を進めてください
 - データによく知る → テーマを設定する → テキスト分析に取り組む
- この演習はグループワークです。グループに分かれて、グループ単位で分析や議論を進めてください
- なお、次回(8/7)は発表会です。グループごとに発表準備をお願いします
 - グループごとに、説明10分、質疑5分

演習用のデータ

データファイル名	件数	データセット	備考
rakuten_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件をサンプリングEXCEL 形式 (シート名「2019」)	<ul style="list-style-type: none">本講義の全体を通して利用する
rakuten_2020.xlsx	8,518	<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 (登別,草津,由布院は 1,000件以下のため全数, それ以外はランダムサンプリング)EXCEL 形式 (シート名「2020」)	<ul style="list-style-type: none">演習用 (3~4日目)
covid_2020.xlsx	30,000	<ul style="list-style-type: none">Search API で取得した 2020/4/24~6/30 のハッシュタグ「#新型コロナ」のツイートデータ分布を保持して,3万件をサンプリングEXCEL 形式 (シート名「30000」)	<ul style="list-style-type: none">演習用 (3~4日目)

※ 自身の業務課題を題材にしても構いませんが、個人情報や機密情報など**守秘義務に特に留意**してください。

rakuten_2019.xlsx

- ・楽天トラベルから収集した「お客様の声」のデータ
 - ・宿泊日が2019年、下記の10エリアが対象

レジャー	5エリア	登別, 草津, 箱根, 道後, 湯布院	1,000件×10エリア = 計10,000件
ビジネス	5エリア	札幌, 名古屋, 東京, 大阪, 福岡	

- ・データ項目

施設情報	4項目	カテゴリ, エリア, 施設番号, 施設名
口コミ	1項目	コメント
ユーザー評価	7項目	総合, サービス, 立地, 部屋, 設備・アメニティ, 風呂, 食事
その他の分類	2項目	旅行の目的, 同伴者
宿泊日	1項目	宿泊年月
ユーザー情報	3項目	ユーザー, 年代, 性別

rakuten_2020.xlsx

- ・楽天トラベルから収集した「お客様の声」のデータ
 - ・宿泊日が2020年1~4月(5月は一部), 下記の10エリアが対象

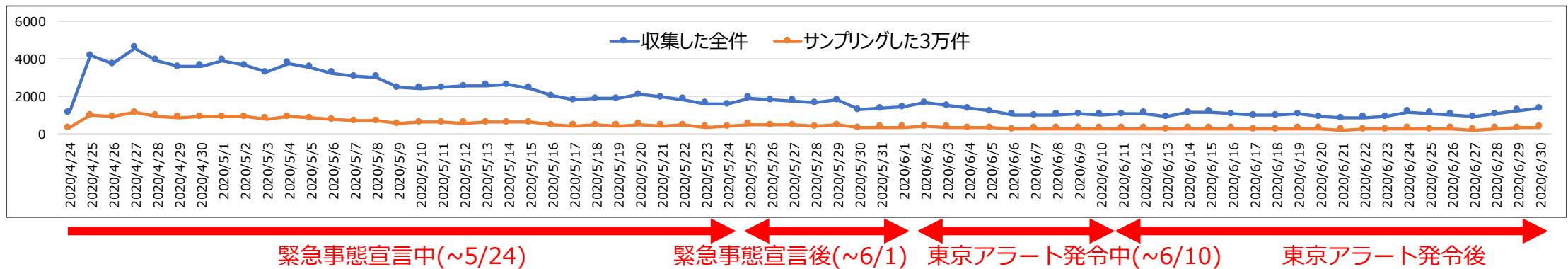
レジャー	5エリア	登別, 草津, 箱根, 道後, 湯布院	登別, 草津, 由布院は全件, それ以外は1,000件 = 計 8,518件
ビジネス	5エリア	札幌, 名古屋, 東京, 大阪, 福岡	

- ・データ項目

施設情報	4項目	カテゴリ, エリア, 施設番号, 施設名
口コミ	1項目	コメント
ユーザー評価	7項目	総合, サービス, 立地, 部屋, 設備・アメニティ, 風呂, 食事
その他の分類	2項目	旅行の目的, 同伴者
宿泊日	1項目	宿泊年月
ユーザー情報	3項目	ユーザー, 年代, 性別

covid_2020.xlsx

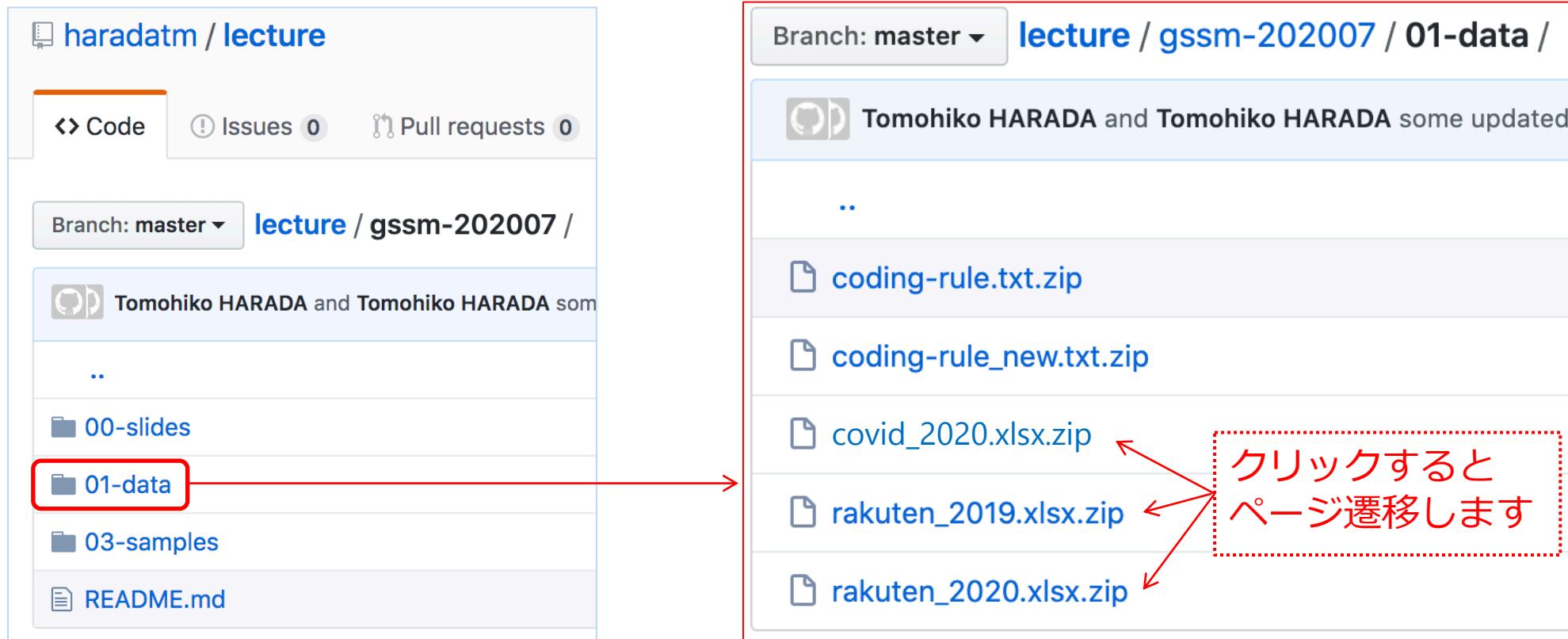
- ・ハッシュタグ「#新型コロナ」を付けて Tweet されたデータ
 - ・2020/4/24~6/30 に Tweet されたデータを Search API (1%)で収集
 - ・データ件数の日別の分布を保持して、30,000 件をサンプリング



Tweet情報	6項目	ID, 投稿日, 投稿時刻, お気に入り数, retweet数, 言語
Tweet本文	1項目	コメント
ユーザー情報	4項目	ユーザーID, フォロワー数, フォロー数, ユーザーのこれまでの記事投稿数
その他の属性	1項目	シーズン {緊急事態宣言中, 東京アラート発令中, 東京アラート解除後}

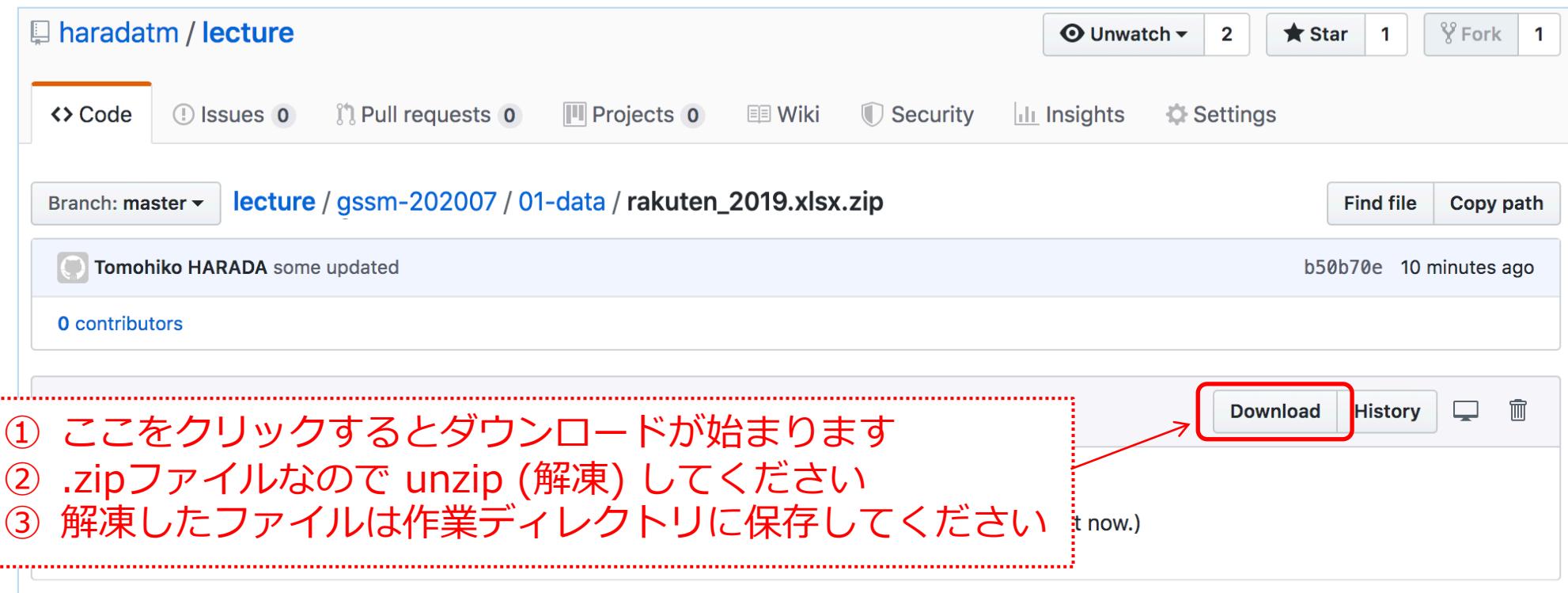
データの取得方法

- <https://github.com/haradatm/lecture/tree/master/gssm-202007>



ダウンロード方法

- Download ボタンをクリックするとダウンロードを開始



発表内容

1. 利用データ

例) 「covid_2020.xlsx」であれば、日/月/シーズン別の投稿件数など

2. テーマ設定

例) 「rakuten_2019.xlsx」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

3. 分析結果 (プロットおよび考察)

- ・設定したテーマにもとづき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

課題 (3日目)

- ・「KH Coder で表記ゆれを吸収する」を参考に, 任意の表記ゆれをまとめて, その結果を「抽出語リスト」で確認してください.
- ・グループワークについて, 所属するグループ名とで取り上げる分析テーマを記載して提出してください.
- ・形式: PPT(PDF), 提出先: manaba, 期限: 次回 7/24 18:20

参考書

(KH Coder)

- [1] 横口耕一. 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] 横口耕一. 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.