テキストマイニング

石田 基広

徳島大学総合科学部

テキストマイニングとは何か

従来の構造化されたデータ

- 質問例: 自民党をどう思いますか
  - 1) 好き 2) 嫌い 3) 無関心
- 回答例 :
  - 選択 3. (30 代，男，徳島市)
  - 選択 2. (20 代，女，徳島市)
  - 選択 2. (20 代，男，徳島市)
  - … 以下同様に続く…

定量的データ

さまざまな集計化

- クロス表（分割表）
- カイ二乗検定へ

年代	支持する	支持しない	どちらでもない
10 代	3	10	20
20 代	5	8	16
30 代	11	10	20
40 代	18	9	12

定量的なデータ

テキストマイニングとは何か

構造化されていない大量のデータの処理

- 質問例:自民党をどう思いますか．自由に書いてください．
- 回答例:「自民党というのは何となく企業からお金をもらうために議員をしているような政治家ばかりのイメージがある．だからといって民主党の方がましとも思えない．」(20 代 女 徳島市)

定性的データ

年代	自由記述欄
10 代	企業の都合を優先する政治家が多い気がする.
20 代	企業からお金をもらう政治家ばかり.
30 代	官僚の都合を優先するな.
40 代	大衆に迎合する政治家が多い.

定性的なデータ

- このままでは解析データとして扱えない
- クロス表に変換できないか？

テキストマイニングとは何か

大量に蓄積された非構造化データ

- 顧客からのアンケート
- サポートセンターへ寄せられたクレーム
- 医療カルテ
- ホームページやブログ

データを構造化してデータ・マイニングの対象とする

- 構造化する（形態素解析）
- 定量化する（ターム・文書行列）

文書を構造化する

単語単位に分割する

- 形態素解析
- 文を単語に分ける
  - スペースで分かち書きされない日本語
  - 曖昧性（恣意性）
  - 統計解析 = 統計 + 解析 ？
  - データ解析 = データ + 解析 ？
- ChaSen や MeCab といったフリーウェア
  - すももももももものうち
  - すもも も もも も もも の うち

文書を構造化する

意味関係を考慮しながら形態素解析

被験者	回答
番号１	その本は面白くないよ.
番号２	その本は面白かった.

- 係り受け解析
- CaBoCha といったフリーウェア
  - 面白い ＝ 面白い
  - 面白くない ＝ 面白い＋ ない

形態素解析とターム・文書行列

企業	名詞, 一般,*,*,*,*, 企業, キギョウ, キギョー
の	助詞, 連体化,*,*,*,*, の, ノ, ノ
都合	名詞, 一般,*,*,*,*, 都合, ツゴウ, ツゴー
を	助詞, 格助詞, 一般,*,*,*, を, ヲ, ヲ
優先	名詞, サ変接続,*,*,*,*, 優先, ユウセン, ユーセン
...	(以下続く)

形態素解析とターム・文書行列

企業	名詞, 一般,*,*,*,*, 企業, キギョウ, キギョー
の	助詞, 連体化,*,*,*,*, の, ノ, ノ
都合	名詞, 一般,*,*,*,*, 都合, ツゴウ, ツゴー
を	助詞, 格助詞, 一般,*,*,*, を, ヲ, ヲ
優先	名詞, サ変接続,*,*,*,*, 優先, ユウセン, ユーセン
...	(以下続く)

ターム・文書行列の作成

terms	文書 1	文書 2	文書 3	文書 4
企業	1	1	0	0
の	1	0	0	0
都合	1	0	1	0
を	1	1	1	0
優先	1	0	1	0
...	(以下続く)			

文書行列とデータ・マイニング

定量化されたテキスト・データに各種データマイニング技法を適用

- 形態素(係り受け)解析後に ターム・文書行列 作成
- 行列に変換して 定量化 する

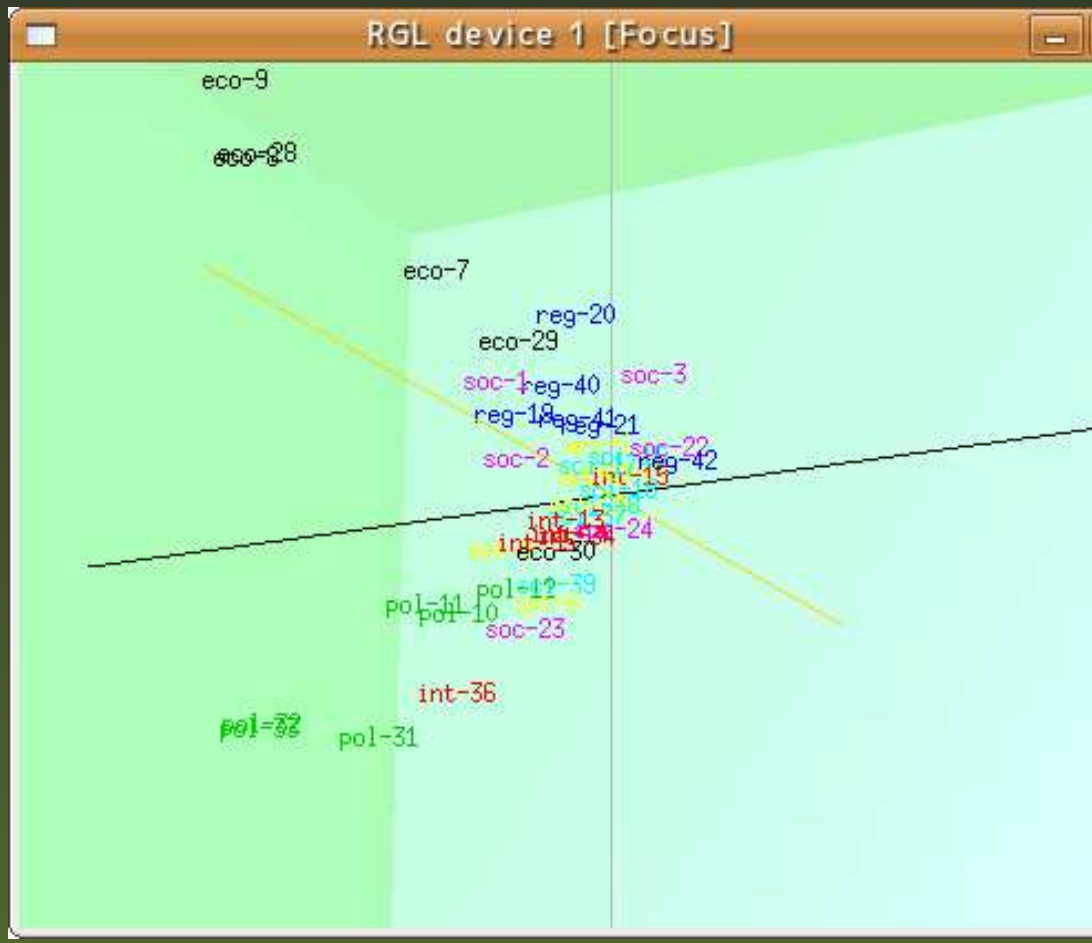
データを 構造化 してデータ・マイニングの対象とする

- テーマごとに文書を類別
- 文書のテーマを推定
- 書き手の判別

新聞記事分類

2008 年 3 月某日の読売新聞の主な見出し

term	経済欄 1	経済欄 2	経済欄 3	経済欄 4
金融	0	2	0	3
金利	0	2	0	0
見る	0	1	0	0
厳格	0	1	0	0
減る	0	2	0	0

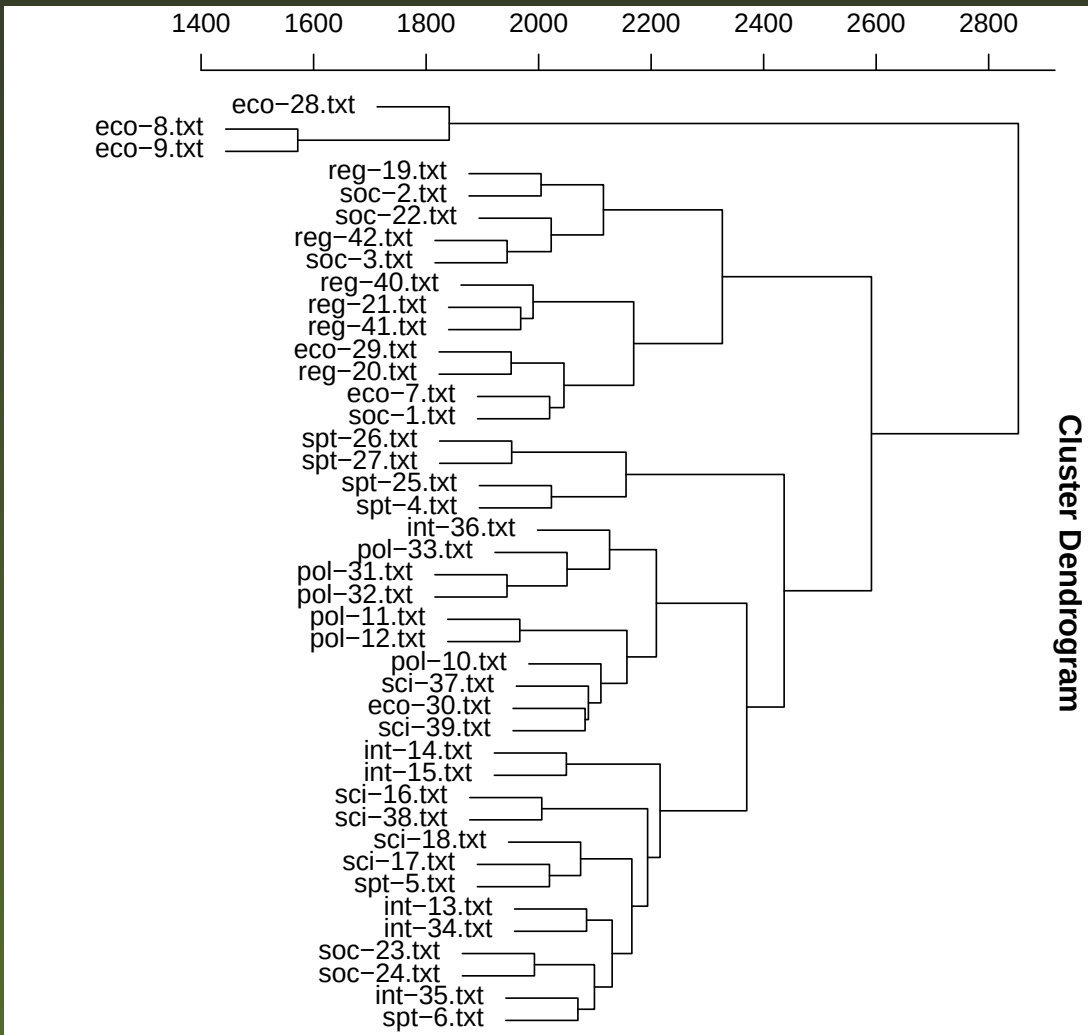


三次元散布図を描いてみる

新聞記事分類

クラスター分析

term	経済欄１	経済欄２	経済欄３	経済欄４
金融	0	2	0	3
金利	0	2	0	0
見る	0	1	0	0
厳格	0	1	0	0
減る	0	2	0	0

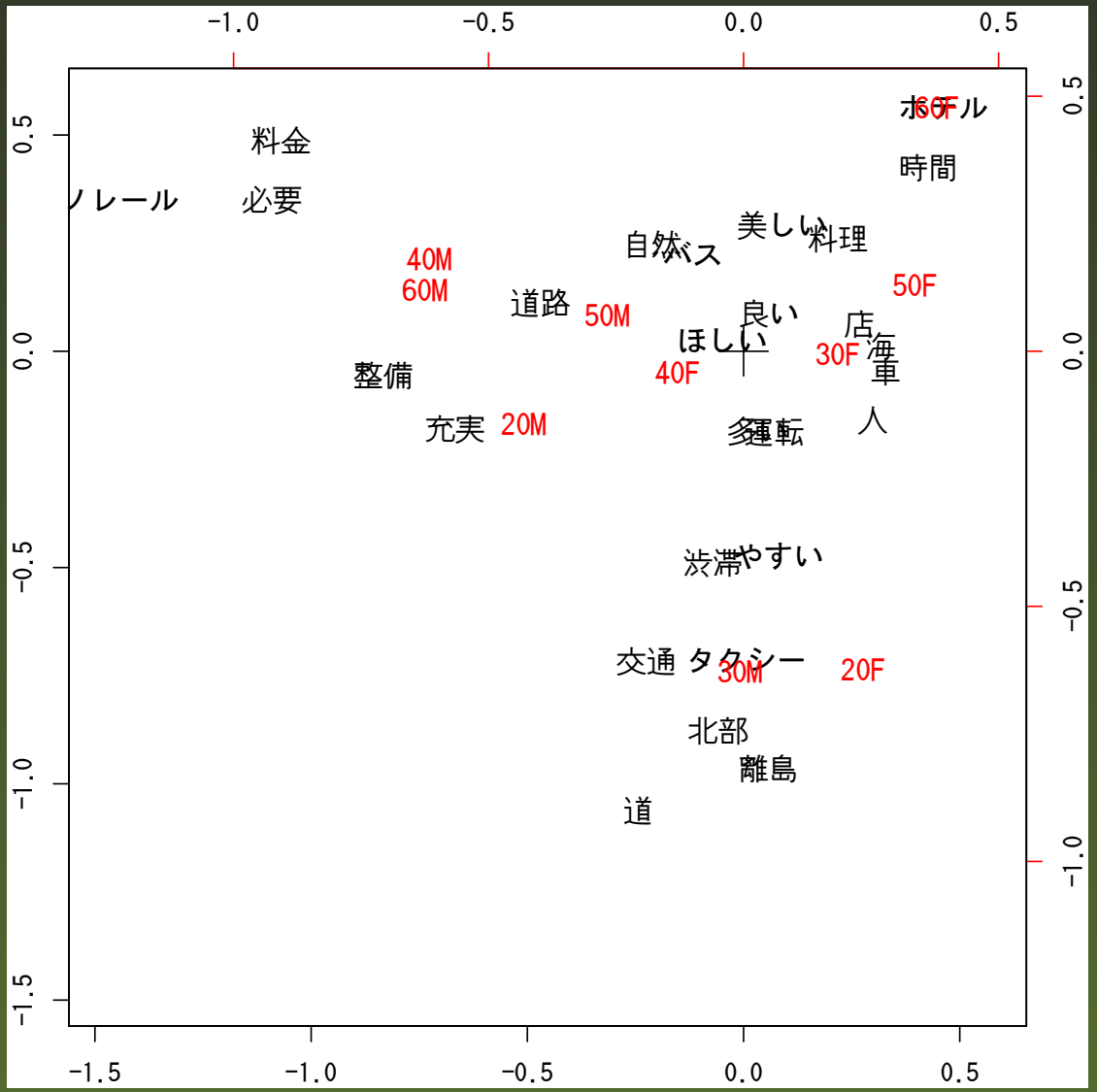


沖縄観光分析

性別	年代	意見
女性	60 代	花粉で悩んでいますので、今回過去に訪れて思い切り美味しい
		クシャミの出ない空気を吸えて幸せでした。
		今後花粉の時季に 3ヶ月くらいを目途に滞在を考えております。
		そういう宿泊施設の宣伝等ありましたら、観光情報とともに
		インターネットで流していただくと利用しやすく有難く思っています。
男性	50 代	空港→国際通りまでタクシーを利用。
		乗車したらどちらまで、下車したらありがとうございますの基本動作が全く無い。
		沖縄の観光のためにもお客様に気持ちよく帰ってもらう為にも教育をした方が良い。
女性	20 代	気候が暖かくて住みやすく、沖縄の人たちは名古屋人と違ってのんびりしていて楽しい。
		車もせかす人もいなく運転しやすかった。
		今回友達と来たけど、次回は母と一緒に遊びに行きたいです。

沖縄観光の分析結果

対応分析



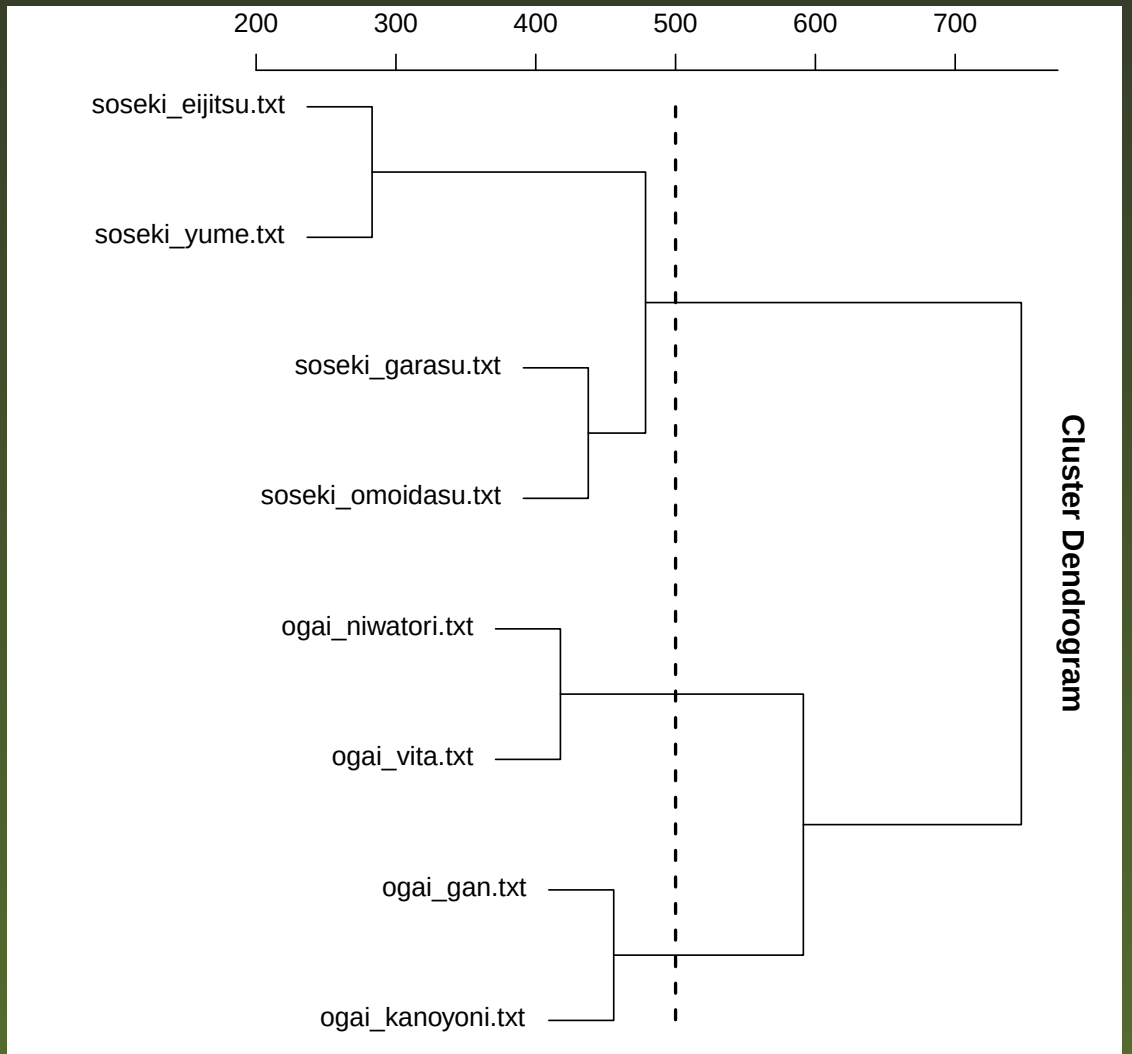
森鴎外と夏目漱石

作家	作品	冒頭部分
鴎外	鶏	石田小介が少佐参謀になって小倉に着任したのは六月二十四日であった。 徳山と門司との間を交通している蒸汽船から上がったのが午前三時である。
	雁	古い話である。僕は偶然それが明治十三年の出来事だと云うことを記憶している。 どうして年をはっきり覚えているかと云うと、
	かのように	朝小間使の雪が火鉢に火を入れに来た時、奥さんが不安らしい顔をして、 「秀麿の部屋にはゆうべも又電気が附いていたね」と云った。
	ヴィタ・セクスアリス	金井湛君は哲学が職業である。 哲学者という概念には、何か書物を書いているということが伴う。
漱石	夢十夜	こんな夢を見た。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。
	硝子戸の中	硝子戸の中から外を見渡すと、霜除をした芭蕉だの、 赤い実の結った梅もどきの枝だの、無遠慮に直立した電信柱だのがすぐ眼に着くが、
	思い出す事	ようやくの事でまた病院まで帰って来た。 思い出すとここで暑い朝夕を送ったのももう三カ月の昔になる。
	永日小品	雑煮を食って、書斎に引き取ると、しばらくして三四人来た。 いずれも若い男である。そのうちの一人がフロックを着ている。

書き手の分類

クラスター分析

term	鴎外１	鴎外２	漱石３	漱石４
が、	107	79	34	31
て、	251	245	155	81
で、	86	69	41	40
と、	63	42	99	27
に、	76	101	45	45



書き手の分類

主成分分析

term	鴎外１	鴎外２	漱石３	漱石４
が、	107	79	34	31
て、	251	245	155	81
で、	86	69	41	40
と、	63	42	99	27
に、	76	101	45	45