アウトライン

- テキストマイニングについて
- 分析
- 課題

アウトライン

- テキストマイニングについて
- 分析
- 課題

テキストマイニング

テキスト　→　数値データ　→　グラフ等

計量言語学

テキストの特徴を示す数値を抽出して分析

・シェークスピア・ベーコン論争

単語の長さの平均値

・ジップの法則

出現頻度と出現順位との間に相関

形態素解析

- 言語を単語単位に分割、品詞情報なども加える
- 形態素：意味を持つ最小の単位

一つ一つの品詞のこと

例）すももももももものうち

名詞 ："すもも" 助詞 ："も" 名詞 ："もも" 助詞 ："も”
名詞 ："もも" 助詞 ："の" 名詞 ："うち"

Rを用いた解析

- RMeCabパッケージ

```
> library(RMeCab)
> RMeCabC("すももももももものうち")
[[1]]
    名詞
"すもも"

[[2]]
助詞
"も"

[[3]]
  名詞
"もも"

[[4]]
助詞
"も"

[[5]]
  名詞
"もも"

[[6]]
助詞
"の"

[[7]]
  名詞
"うち"
```

RMeCab

- Ngram関数

抽出する単位を単語や形態素、品詞のいずれかに指定できる

n-gram:

連続する文字ないし形態素、品詞をペアとした頻度情報

２つ連続→bigram　３つ連続→trigram

```
> tail(matsumoto.bi2)
     Ngram1 Ngram2 Freq
1934   誕生     日    4
2000     達     酒    3
2008     酒   芸人    2
2067   飲食     店    2
2081      %     人    2
2091      3     0    2
```

アウトライン

- テキストマイニングについて
- 分析
- 課題

アウトライン

- テキストマイニングについて
- 分析
- 課題

分析

調査目標：
　Twitterのつぶやきから個人の嗜好を分析できるのか

対象：
　有名人（分析と実際の情報を照らし合わせやすい）

キーワード：
　マイクロブログ、ネットワーク分析

手順

1. TwitterAPIを通して過去の投稿を抽出

   TwitteRパッケージを使用

2. Perlにて文字データの加工

   http~（リンク）や@~（アカウント名）などの削除

3. 形態素解析（bigram）

4. ネットワーク分析

Twitterからの抽出

- 分析対象：松本人志　３００ツイート

三村のふわっふわ感。
楽しい楽しい さまぁ～ずライブでした。。。
月給1億なんてまさかです～

わしゃ浜田か！
え？ 9割以上が賛成してんのに？
これって炎上？　東スポweb
たまにあるワイドナショーの感想。。。

松本嫌いだけどこの意見は同意。

あ、ありがとう…
マツナンデス！
耳クソは人のグチを聞きすぎた日よくたまる。。。
君は芸人が壊れる瞬間を見たくはないか？！

ドキュメンタル2
最近ウチの鳩が鳴いた後も引っ込まなくなった。
おそらく向かいの置き時計をカンニングして鳴いとるな。
私のコレクションです。
聴覚障害の人に握手を求められた。マスク着けたまま ありがとうって言ってしまった。
オレあほやな。
僕の妹達です

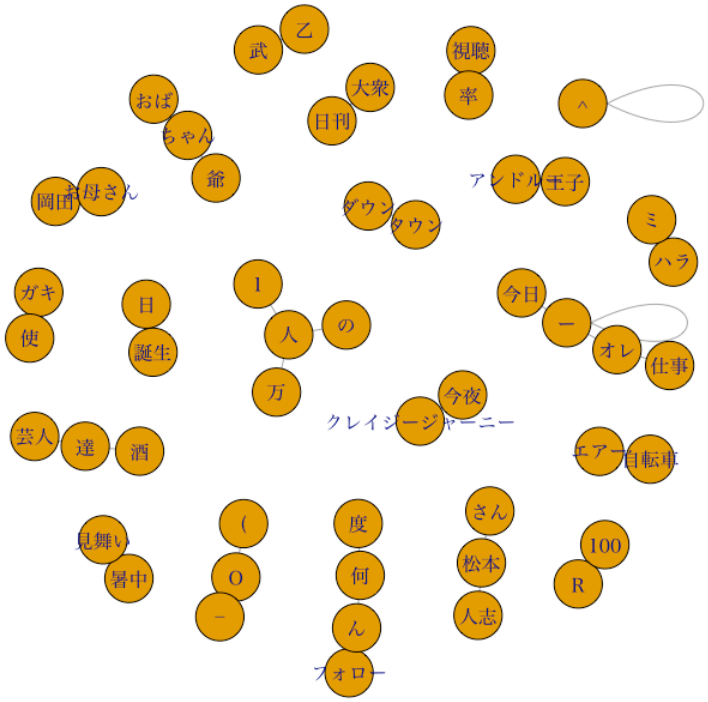
（加工済み）

- bigramで名詞のみを抽出

```
> tail(matsumoto.bi2)
     Ngram1 Ngram2 Freq
1934   誕生     日    4
2000     達     酒    3
2008     酒   芸人    2
2067   飲食     店    2
2081      %     人    2
2091      3     0    2
```

ネットワーク分析

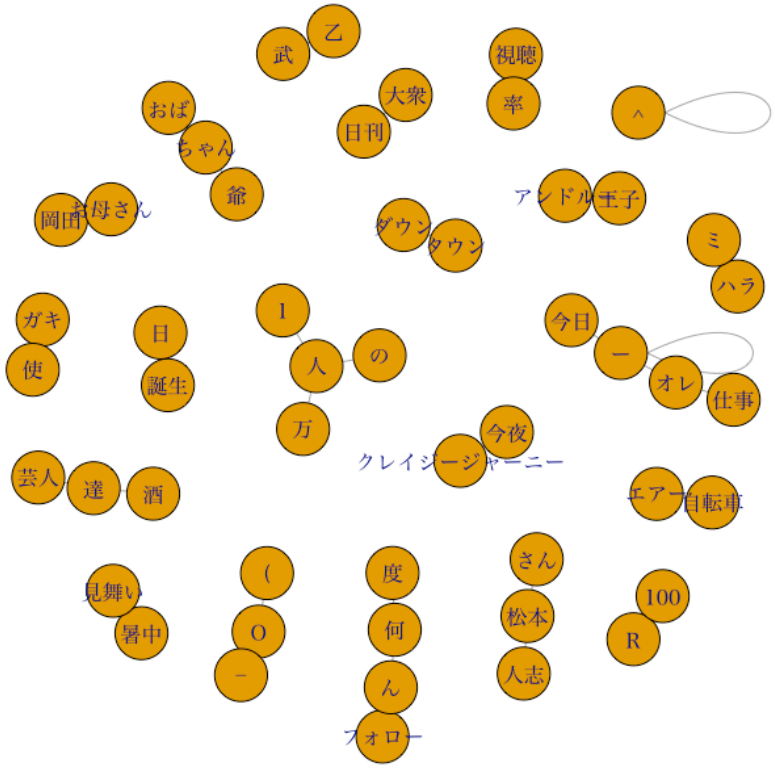
- 頻度：３回以上



ネットワーク分析

- 番組、仕事のことが多い

- 好きなもの
  芸人との酒
  ハラミ
  女好き
    面白いやつの条件の一つ
    乙武、アンドルー王子



結論

調査目標：

Twitterのつぶやきから個人の嗜好を分析できるのか

仕事関連のツイートが多い（有名人）
ごくわずかの情報について照合性あり
単語と単語との関連性が薄い

アウトライン

- テキストマイニングについて
- 分析
- 課題

アウトライン

- テキストマイニングについて
- 分析
- 課題

課題

- 投稿数、単語数の不足

  もっとノード同士が繋がるはず

- ネットワーク分析　中心性

  それぞれの要素の関係性、距離　数学的計算

- 友人の投稿、リツイート

参考文献

- 石田基広、小林雄一郎（2013）、株式会社ひつじ書房、「Rで学ぶ日本語テキストマイニング」、第１章・第３章
- 中野光一（2011）、「マイクロブログにおける個人の嗜好解析に関する研究」