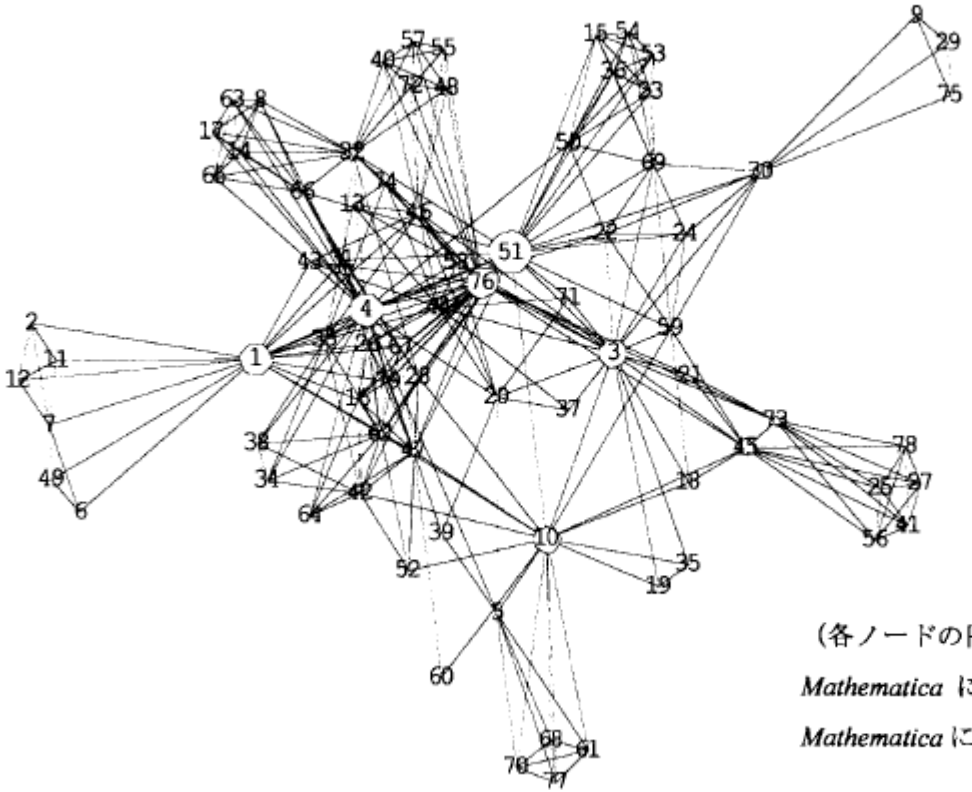
2012年7月1日
社会学研究のためのR勉強会

twitteRパッケージとRMeCabパッケージによるソーシャルメディア分析

鈴木　努
snatool@gmail.com

研究の経緯

- 政治コミュニケーション研究の中で、政治家の演説や新聞記事中の単語の共起ネットワーク分析
- 鈴木努，2005，「東京ゴミ戦争における都知事演説のテクスト分析―中心化共鳴性分析による」『社会学論 考』26: 1-24
- 鈴木努，2006，「二〇〇五年衆議院選挙における三大紙の社説比較―概念ネットワーク分析の適用」『マス・コミュニケーション研究』69: 2-21.



（各ノードの円の大きさは媒介中心性の大きさを表している。ネットワーク描画は *Mathematica* によるバネ電気モデル。本稿におけるその他の数値計算および描画も *Mathematica* による。）

図 1　ゴミ戦争宣言の概念ネットワーク　(鈴木 2005)

①いま②おくれ③ごみ④ごみ処理⑤ごみ焼却炉⑥ごみ戦争⑦ごみ対策⑧シアン⑨ひずみ⑩プラスチック⑪ヨーロッパ⑫下水道⑬家庭⑭家庭ごみ⑮耐久消費財⑯開発⑰各種重金属類⑱割合⑲還元⑳危機㉑基本㉒区部㉓傾向㉔経済㉕計画㉖結果㉗建設㉘見込み㉙現状㉚高度成長㉛今後㉜産業廃棄物㉝仕事㉞自治体㉟自然㊱自動車㊲質㊳従来㊴心臓部㊵人類㊶推進㊷政府㊸整備㊹清掃㊺清掃工場㊻生活㊼生産㊽生存㊾宣言㊿粗大ごみ(51)増加(52)大半(53)大量消費(54)大量生産(55)段階(56)地域住民(57)都市(58)都民(59)東京(60)日本(61)熱(62)廃棄物(63)廃酸(64)廃品回収義務(65)廃油(66)排出量(67)爆発(68)発生(69)反映(70)腐食(71)変化(72)保全(73)埋立処理場(74)密着(75)目(76)問題(77)有害ガス(78)理解(79)累積

図４　８月６日読売新聞社説の概念ネットワーク

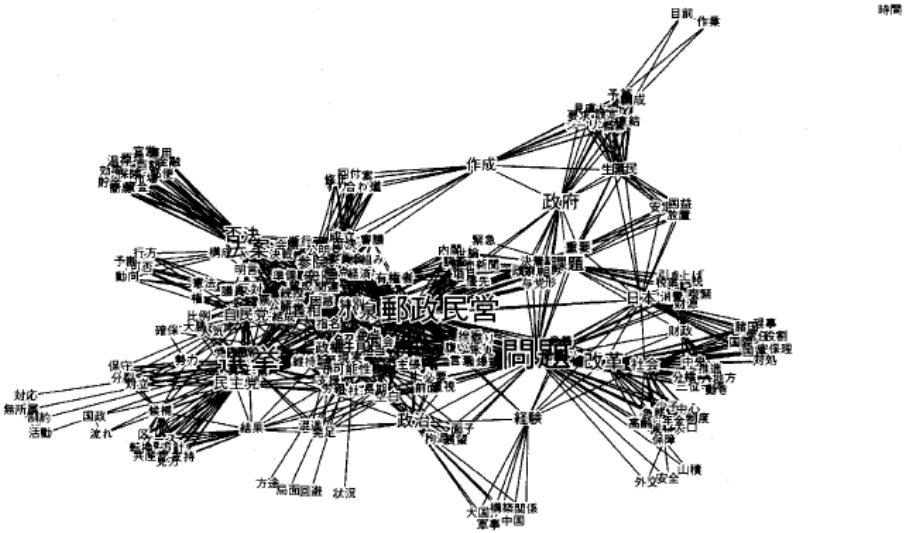


図５　８月６日朝日新聞社説の概念ネットワーク

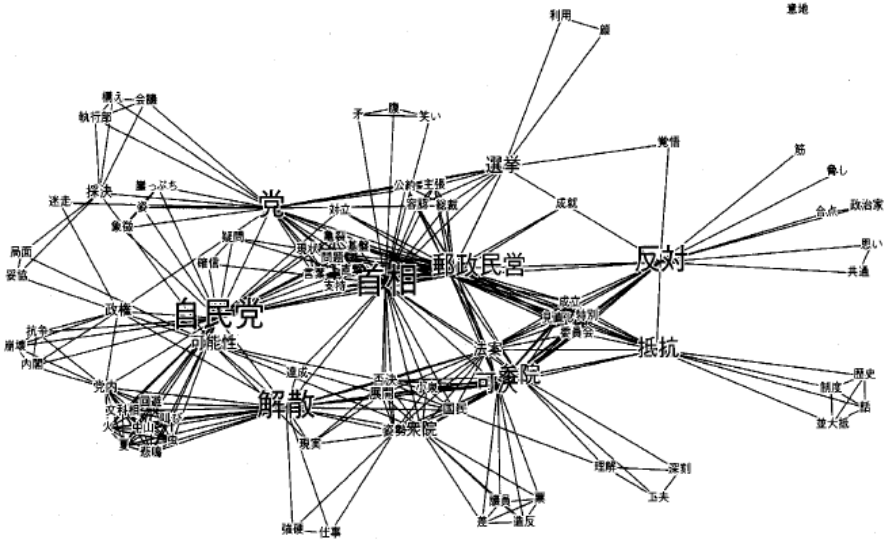
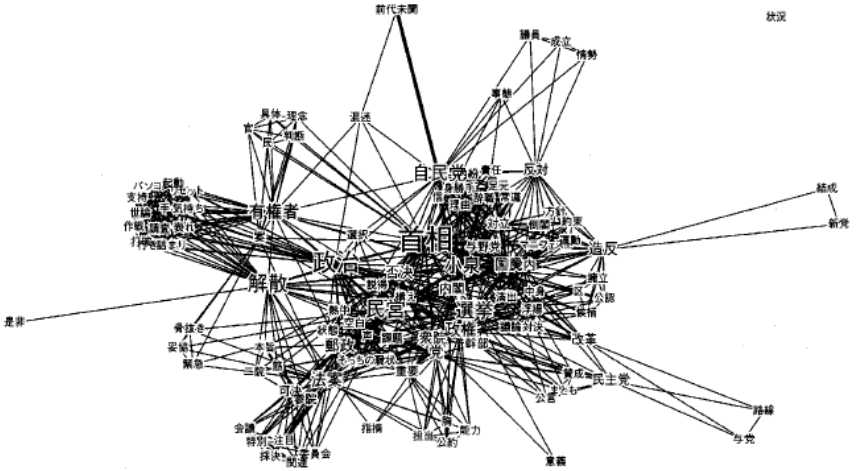


図６　８月６日毎日新聞社説の概念ネットワーク



2005年8月6日の三大紙の社説における単語の共起グラフ（鈴木 2006）

学史研究における利用

- 左古輝人, 2010,「社会の科学とテクストマイニング ―ホッブズ『リヴァイアサン』を題材に」『人文学報. 社会学』45: 73-98.
- 鈴木努, 2010,「近代文語のテキストマイニング」, 概念史・概念分析研究会（2010年12月19日）.

福沢諭吉『文明論之概略』第二章の共起グラフ

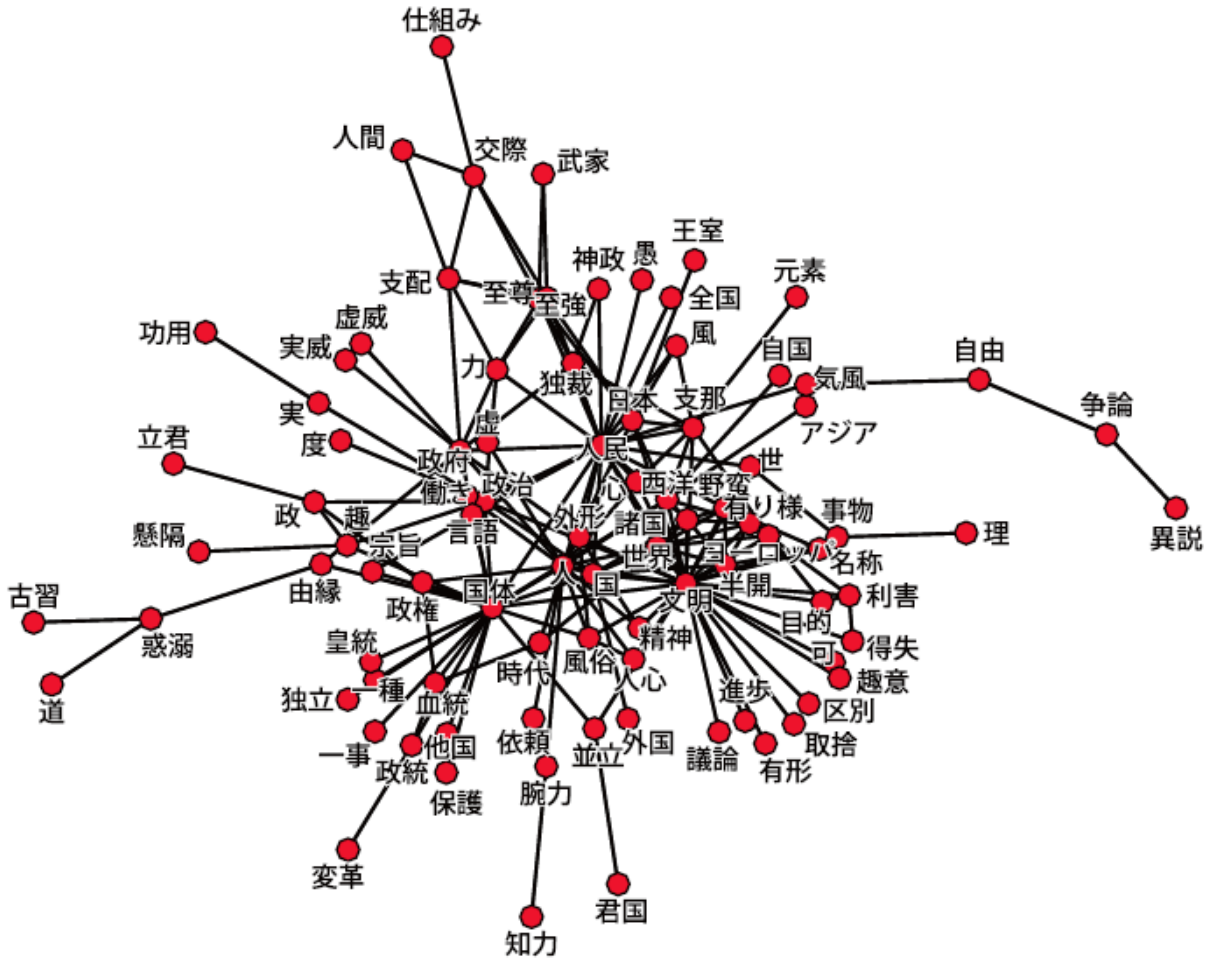


図 3: 共起頻度 3 以上の関係の最大連結成分

福沢諭吉『文明論之概略』第二章の共起グラフ

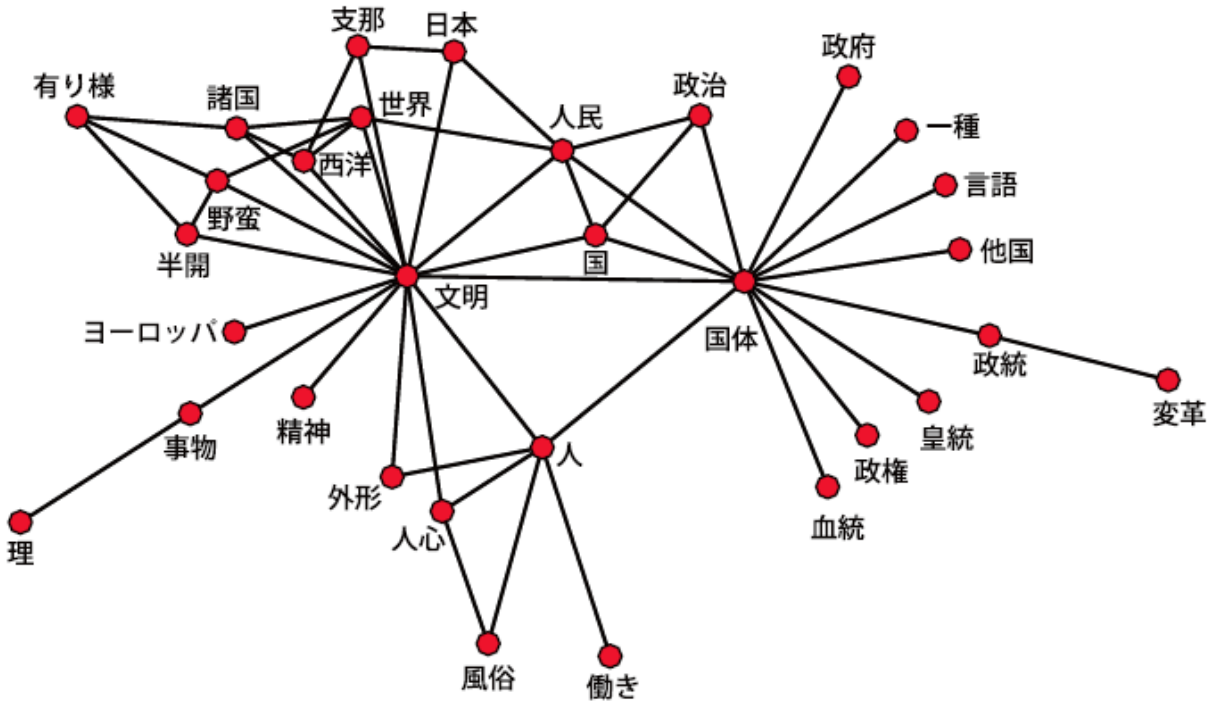


図 5: 共起頻度 5 以上の関係の最大連結成分

丸山真男『「文明論之概略」を読む』との比較

『「文明論之概略」を読む』の章立て

- 第三講　西洋文明の進歩とは何か
- 第四講　自由は多事争論の間に生ず
- 第五講　国体・政統・血統

→丸山の読解のポイント：文明論、自由論、国体論

共起グラフを見ると「文明」と「国体」は中心的な位置にあるが「自由」はそうではない

→丸山の読解の特徴である可能性

社会科学における語の共起グラフの利用

- Osgood, C. E., 1959, “The Representational Model and Relevant Research Methods,” I. d Sola Pool ed., *Trends in Content Analysis*, Urbana: University of Illinois Press, 54–78.

  の随伴分析（Contingency Analysis）が嚆矢

本日のお題

- 電子的なテキストデータの蓄積により共起グラフを用いた分析の適用範囲は飛躍的に拡大している
- 各種APIを利用化
- Rのパッケージではtwitter APIを利用するためのtwitteRパッケージで手軽にtwitterの分析が可能

twitteR

- http://cran.r-project.org/web/packages/twitteR/index.html
- 他のパッケージ同様にRにインストールします
- twitterクライアントとして閲覧や投稿が可能
- 語の検索やフォロー関係を調べることが可能
- 今回は特定のユーザーのつぶやきを取得するのに使います

橋下徹（大阪市長2011.11～
・前大阪府知事）のツイートを分析してみる

ユーザータイムラインの取得

```
library(twitteR)
UTL <- userTimeline("t_ishin", n = 3200)
#最大3200件取得できるが、結果はtwitter APIに依存
data <- twListToDF(UTL)
#Rで分析しやすいようにデータフレームに変換
write.csv(data, "data.csv")
#後日使うためにCSVで保存したり
save(data, file = "t_ishin_TL.RData")
#Rdataとして保存したり（再度読み込むときはload関数）
```

今回のデータ

- どれくらい前のツイートまでさかのぼれるかはtwitter API次第なのですが
- 今回は実際には2回に分けて取得したデータを合併して使っています

```
> dim(data)
[1] 3171   10
```

3171件取得できた

```
> data[3171,"created"]
[1] "2011-08-24 00:34:52 UTC"
```

一番古い日時（日本時間ではなく協定世界時）

日本時間にする

```
> data[1,"created"]
[1] "2012-06-26 21:57:56 UTC"
> attr(data$created,"tzone") <- "Asia/Tokyo"
> data[1,"created"]
[1] "2012-06-27 06:57:56 JST "
```

年月日や年月を文字列でベクトル化

```
> date <- format(data[,"created"],"%Y-%m-%d")
> year_month <- format(data[,"created"],"%Y-%m")
```

データの内容

公式RTは含まれない。非公式は含まれる

```
> data[1,]
text
1 ということで、朝日新聞さん、僕の質疑は機械的ですが、代理人の河合弁護士が、個別議案についての提案説明、質疑をしっかりとやってくれます。政府が相手なら総会屋になっても良いのですが、やはり関電も民間会社ですから、そこまではね。
  favorited  replyToSN               created  truncated  replyToSID
1     FALSE       <NA> 2012-06-27 06:57:56      FALSE        <NA>
                  id  replyToUID
1 217738491948302336        <NA>
                                                             statusSource
1 <a href="http://twipple.jp/" rel="nofollow">ついっぷる/twipple</a>
  screenName
1    t_ishin
```

簡単な分析

```
> table(data[,"replyToSN"])

    hirokook     inosenaoki Kageyama_hideo     namatahara   seiji_ohsaka
           1              1              3              1              1
watanabe_miki     ytsuji2001
           2              1
```

@つきの「返信」はあまりしていない。

対話より発信にtwitterを使っている？

簡単な分析

```
> table(data[,"replyToSN"])
```

岡田裕子（美術家） 猪瀬直樹 陰山英男 田原総一朗 逢坂誠二（衆議院議員）

hirokook	inosenaoki	Kageyama_hideo	namatahara	seiji_ohsaka
1	1	3	1	1

watanabe_miki	ytsuji2001
2	1

わたなべ美樹 辻よしたか（大阪市会議員）

@つきの「返信」はあまりしていない。

対話より発信にtwitterを使っている？

@つきツイートの内容例

> data[which(data[,"replyToSN"] == "Kageyama_hideo"),c("text","created")] #陰山英男

text

2850 @Kageyama_hideo 渡辺さんも言われているように現在の委員体制では教委行政を仕切れない。追認機関かご意見番にしか過ぎません。常勤、高報酬が必要。さらに公選にするのか。僕は反対で、首長に権限と責任を負わせ、そして教委に首長を抑制する権限を与える方向ではないかと思います。

2851 @Kageyama_hideo 今回の条例では、政治が現場の人事権・予算権を握ることを目的としていません。ピラミッド組織とは、今の教育委員会体制です。校長と住民に権限・財源を渡す。知事と教委は大きな全体方針を決める。今の体制の抜本的変更。その上で今の委員会をどうするのか？

2852 @Kageyama_hideo 陰山先生、昨日はありがとうございました。非常に建設的な議論になったかと思います。これからは対案です。スケジュール観も重要です。これ次第ですから峠はまだまだですよ。文科省を頂点とする教育行政システムを自立型学校組織に変えることができるか。

created

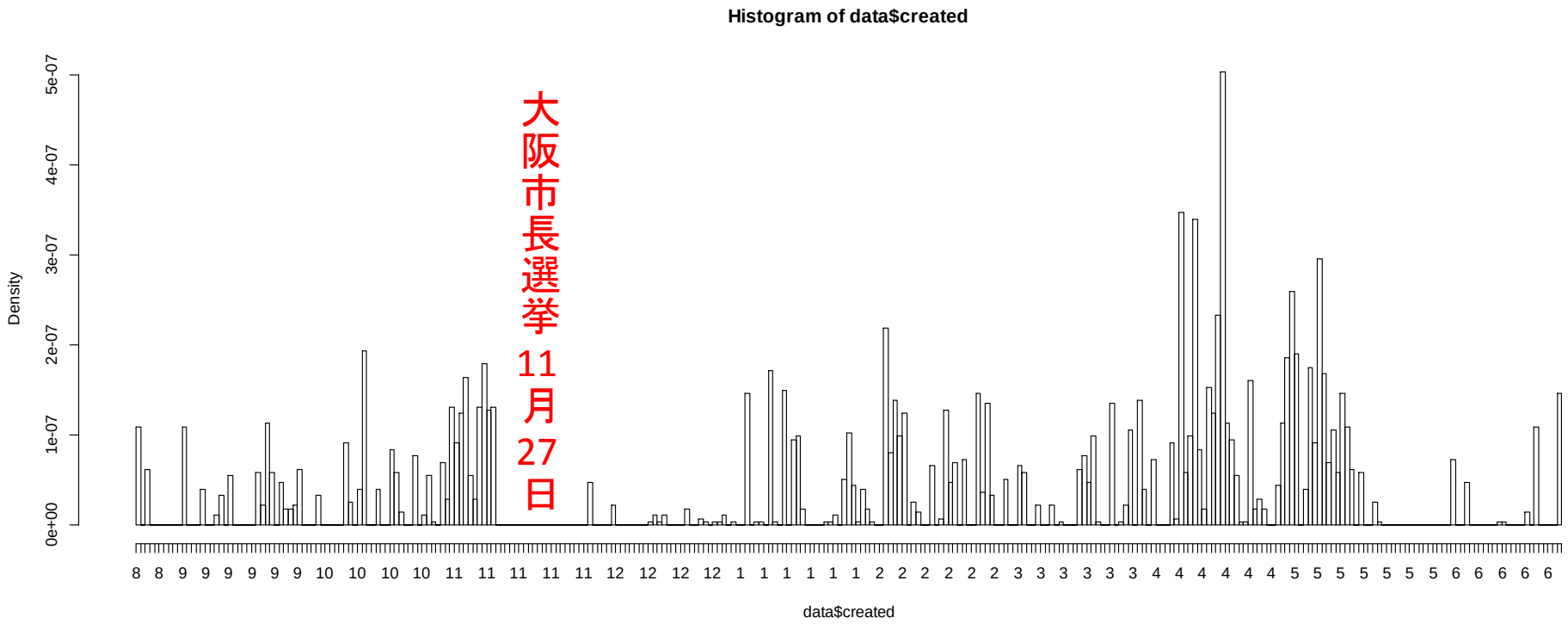
2850 2011-10-12 13:26:49

2851 2011-10-12 13:23:35

2852 2011-10-12 13:20:00

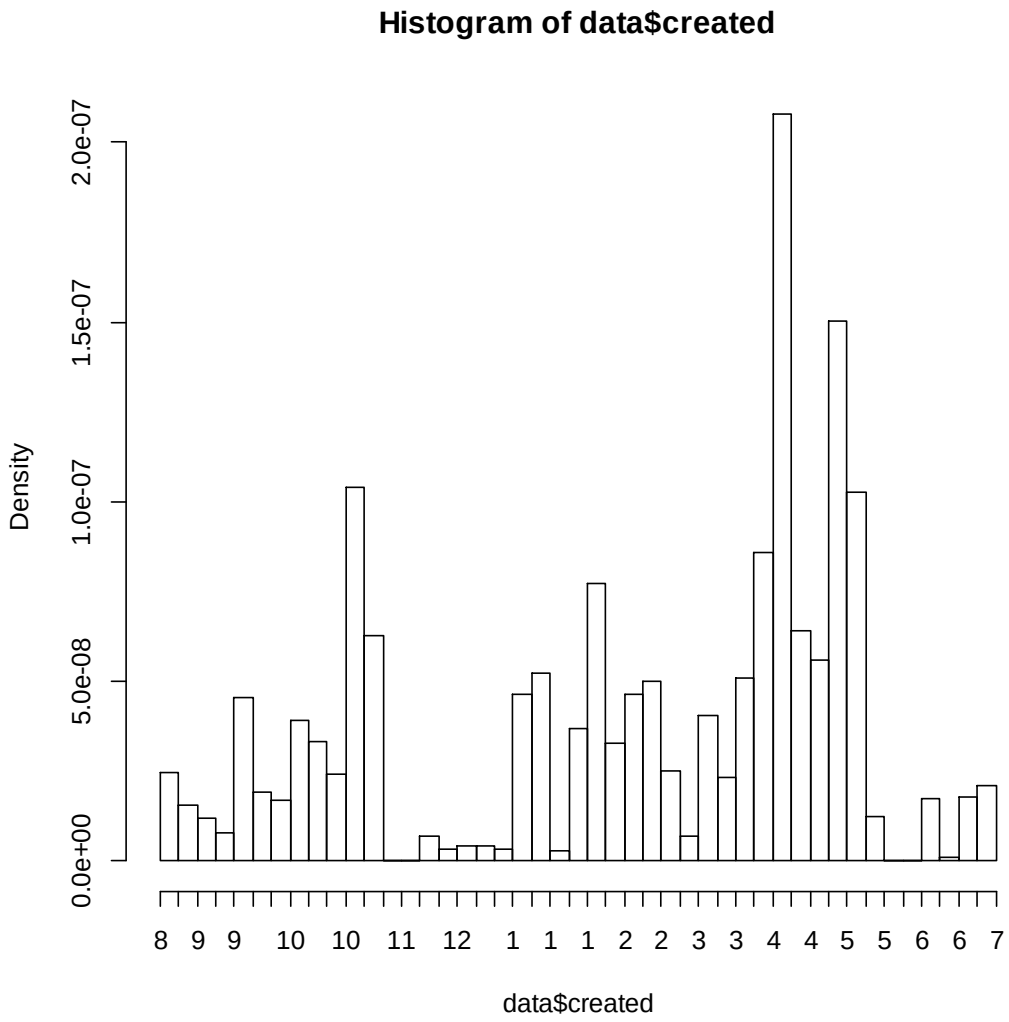
ツイート数の変動（日ごと）

```
> hist(data$created, breaks = "days")
```



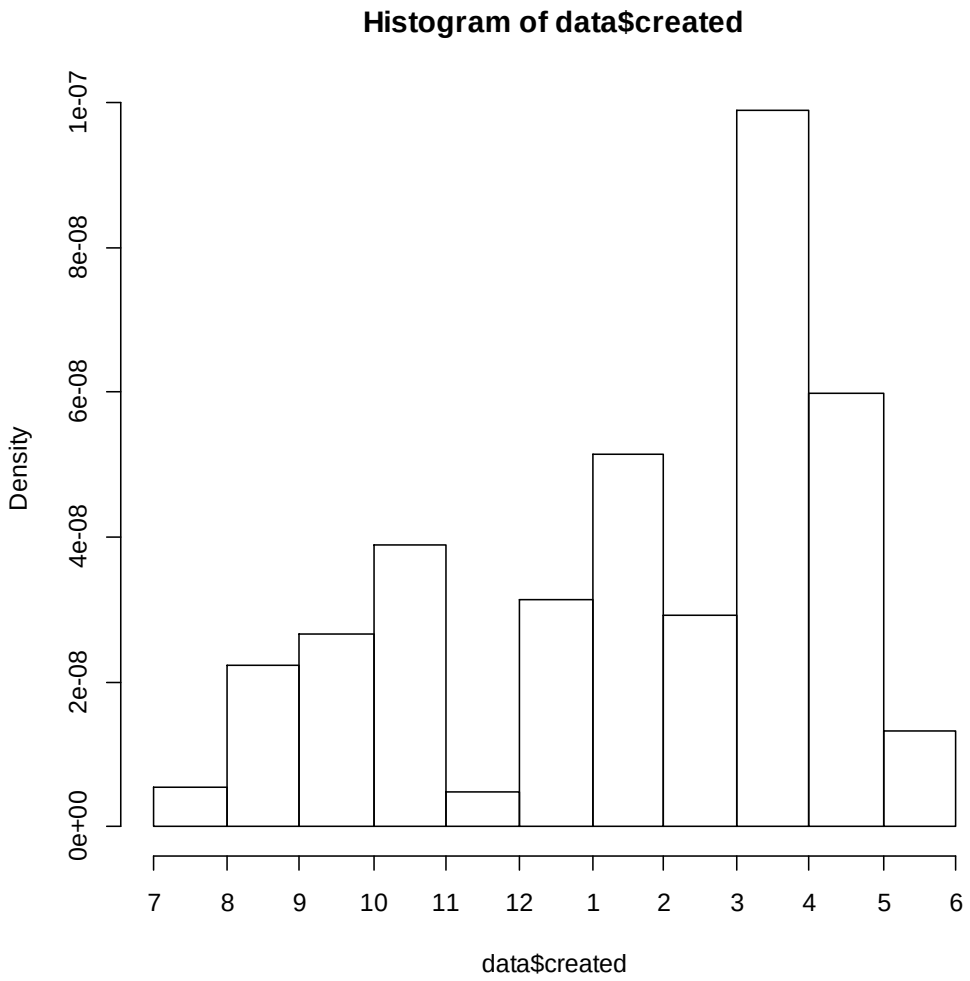
ツイート数の変動（週ごと）

> hist(data$created, breaks = “weeks")



ツイート数の変動（月ごと）

> hist(data$created, breaks = “months")



MeCabのインストール

- MeCabは日本語の形態素解析ソフト
- 辞書に従って、文を単語に分割し品詞などの分析をしてくれる
- http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

からインストーラをダウンロードしてインストールします（今回もWindowsで進めます）

RMeCabのインストール

- RMeCabはRからMeCabを使うためのパッケージ
- 開発者の石田基広さんのサイト（http://rmecab.jp/wiki/index.php?RMeCab）からzipファイルをダウンロードして[パッケージ]→[ローカルにあるzipファイルからのパッケージのインストール]

RMeCabで形態素解析

```
> library(RMeCab)
> #今回は文字コードの問題？でいったんCSVに書き出して再読み込みした
> write.csv(data, "data.csv")
> data <- read.csv("data.csv", header = TRUE)
> res <- RMeCabDF(data, "text", 1)
>#RMeCab(分析対象DF, 列名, 形態素原形を返すとき1)
>
> #今回は名詞のみ取り出します
> #ツイートごとに使われた名詞をリスト化
> nlist <- list()
> for (i in 1:length(res)) {
+  nlist[[i]] <- res[[i]][names(res[[i]]) == "名詞"]
+ }
```

名詞の生起頻度をチェック

```
> words <- sort(table(unlist(nlist)), decreasing = TRUE)
> words[1:300]
```

の	こと	大阪	政治	これ
2040	1659	1501	988	878
教育	僕	それ	会	的
850	828	777	729	603
よう	行政	市長	税	委員
597	592	582	580	547
地方	今	さん	者	市
513	506	503	501	491
選挙	何	@	氏	人
486	470	459	452	444

形式名詞、代名詞、記号、接尾語などがまじっている

リプライ機能を使わない@付きのツイートは意外に多い？

ストップワードを指定

>write.csv(words, "words.csv“)

とりあえずCSVに書き出して…

Excelで取り除きたい語を編集…今回は頻度1の語も取り除く

stopwords.csvの名前で保存

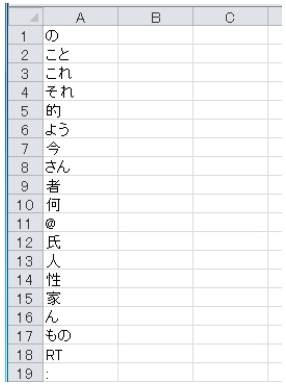
>stopwords <- read.csv("stopwords.csv", header = FALSE)

>keywords <- setdiff(names(words),as.character(stopwords[,1]))

ストップワードとの差集合をとり、キーワードにしました

> length(keywords)

[1] 3439



	A
1	の
2	こと
3	これ
4	それ
5	的
6	よう
7	今
8	さん
9	者
10	何
11	@
12	氏
13	人
14	性
15	家
16	ん
17	もの
18	RT
19	:

←stopwords.csvはこんな感じ

ツイートごとに生起したキーワードをリスト化

```
> keywords_list <- list()
> for (i in 1:length(res)) {
>  keywords_list[[i]] <- intersect(nlist[[i]], keywords)
> }
> keywords_list[[1]]
> sort(table(unlist(keywords_list)), decreasing = TRUE)[1:300]
```

高頻度の語をチェック…

上位300語

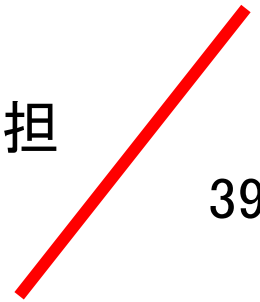
大阪	僕	政治	会	行政
793	660	648	537	432
教育	市長	必要	選挙	委員
431	380	367	347	342
日本	問題	市	制度	維新
332	313	311	296	282
税	府	議論	批判	地方
260	250	247	243	233

・・・・・・・

・・・・・・・

・・・・・・・

自称	成長	存在	分担	経験
40	40	40	40	39



ここで切ろう（基準はフィーリング）

生起行列を作成

行項目をツイート（3171ツイート）、列項目をキーワード（299語）とする生起行列を作成

```
> occurrence <- matrix(0, nrow = length(keywords_list),ncol =
+ length(topwords))
> colnames(occurrence) <- topwords
```

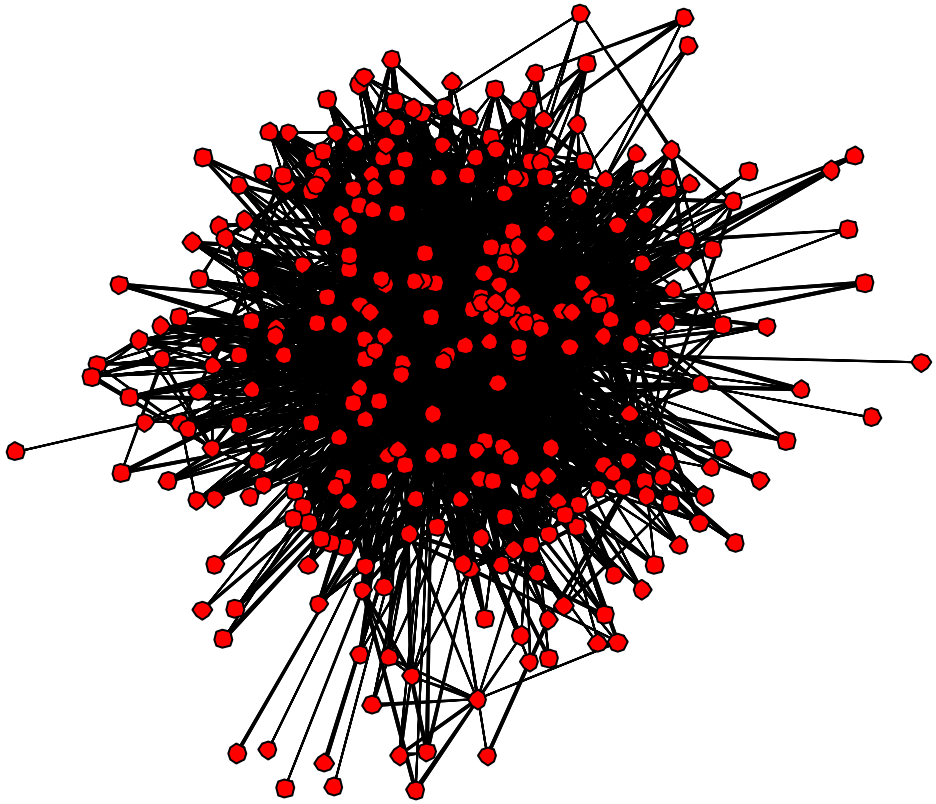
隣接行列（共起数行列）を作成

キーワード間の共起回数を要素とする行列を作成

```
> adj <- t(occurrence) %*% occurrence
```

これを隣接行列としてsnaパッケージで描画してみると…

```
> library(sna)
> gplot(adj, gmode = "graph", displayisolates = FALSE, thresh = 10)
```



よく分からない…

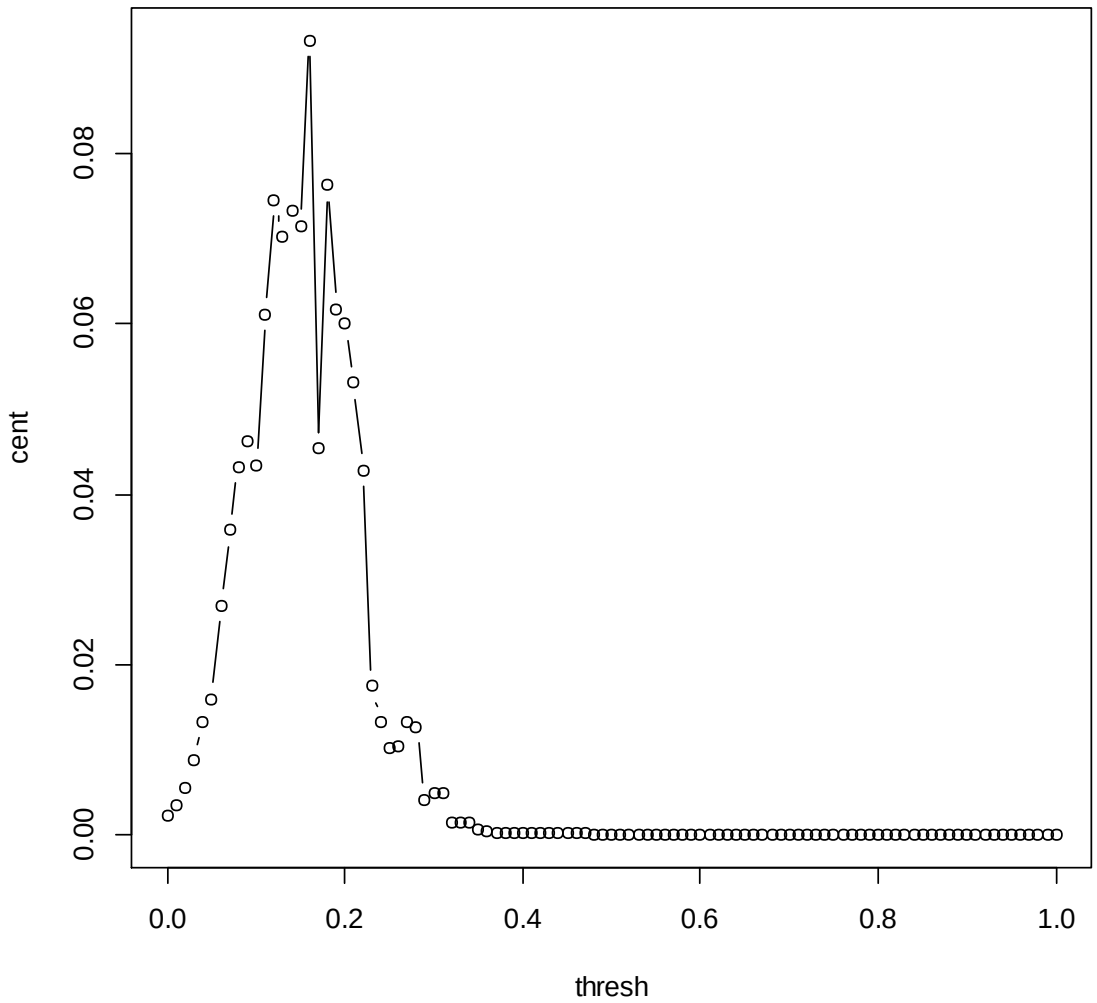
リンクの基準をどうするか

- 共起回数に閾値をもうける（先の図では10）
- 共起回数の期待値をもとに有意性検定…など
- 今回は、生起行列からキーワード間の相関行列を求め、媒介中心性の集中度が大きくなる閾値を設定する
- 理由：相関行列で頻度の影響を低減、集中度を基準にすることで中心的（媒介的）な語を見つけやすくするため

集中度最大のときの相関係数を探す

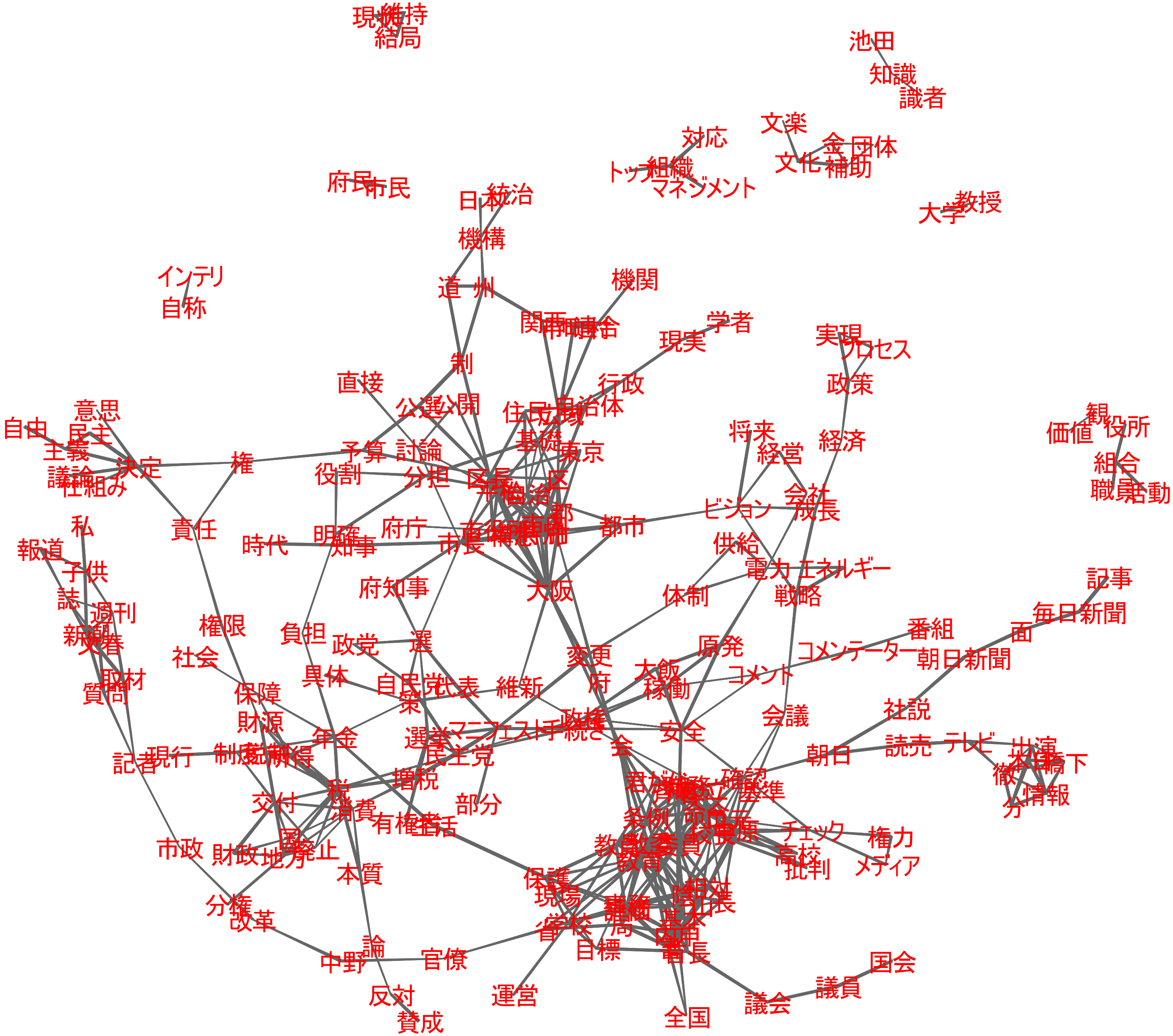
```
> thresh = seq(0,1,0.01)
> cent <- 1:length(thresh)
> cormatrix <- cor(occurrence) #相関行列作成

> for (i in 1:length(thresh)) {
+  adjmatrix <- matrix(0, 299,299)
+  adjmatrix[cormatrix >= thresh[i]] <- 1
+  cent[i] <- centralization(adjmatrix, betweenness)
+ }
> plot(thresh,cent, type = "b")
> thresh[which.max(cent)]
[1] 0.16
```

結果



相関係数0.16以上を閾値に

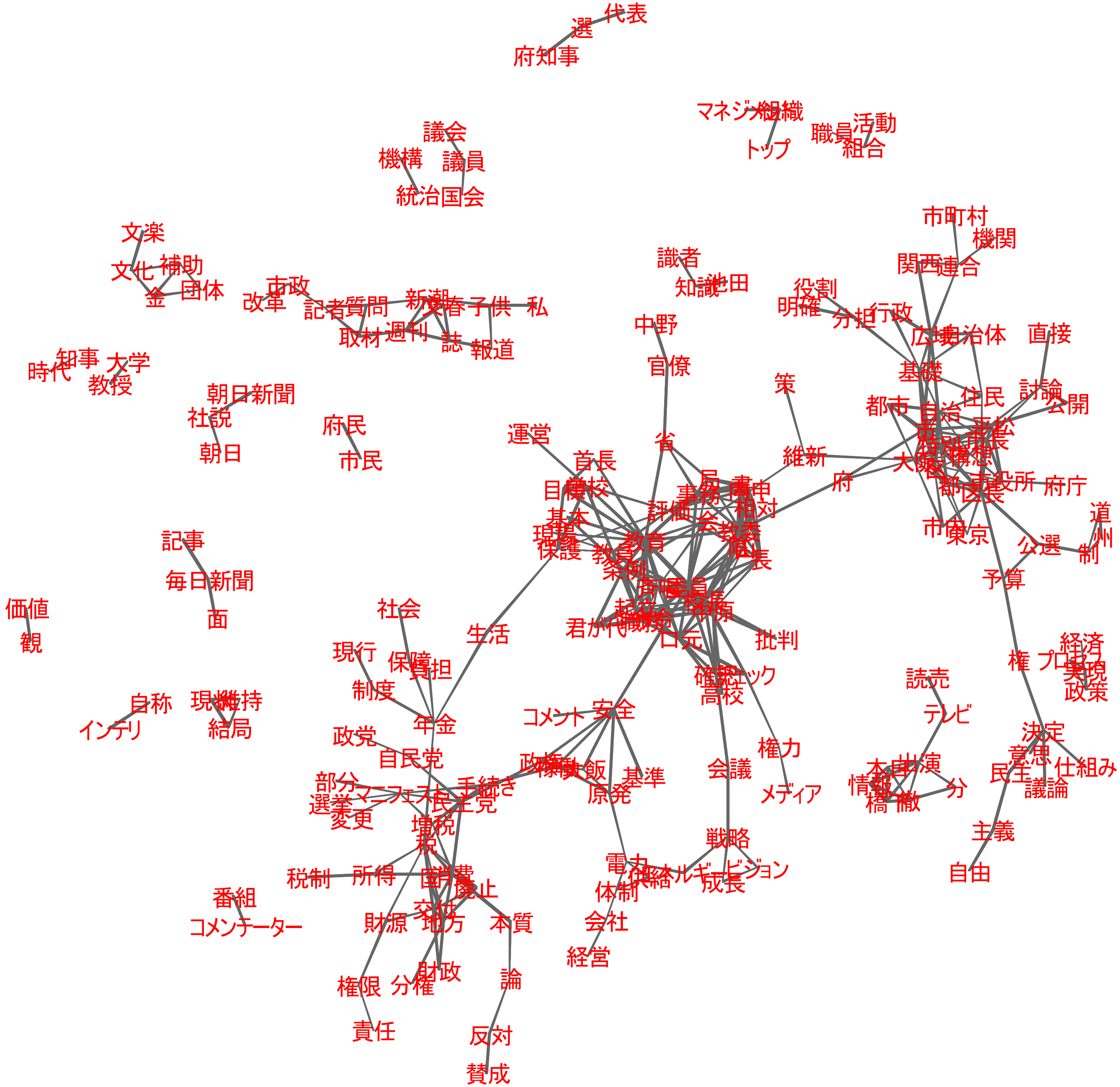
```
> adjmatrix <- matrix(0, 299,299)
> adjmatrix[cormatrix >= 0.16] <- 1
> rownames(adjmatrix) <- colnames(adjmatrix) <- topwords
>
> gplot(adjmatrix, gmode = "graph", displayisolates = FALSE,
+ displaylabels = TRUE, vertex.cex = 0, label.pos= 5,
+ label.col = "red", edge.col = "grey40")
```



維持
結局
池田
知識
識者
文楽
対応
団体
文化
補助
組織
マネジメント
府民
市民
日本
統治
機構
大学
教授
インテリ
自称
道
州
機関
関西
学者
現実
実現
プロセス
政策
制
直接
行政
自治体
住民
意思
自由
民主
主義
決定
議論
権
予算
討論
役割
分担
基礎
東京
区長
区
都
将来
経営
経済
価値
観
役所
組合
職員
活動
私
報道
責任
時代
府庁
知事
市長
都市
ビジョン
会社
成長
供給
電力
エネルギー
府知事
大阪
体制
戦略
記事
毎日新聞
子供
週刊
新潮
文春
権限
負担
政党
選
番組
面
社会
取材
質問
原発
コメンテーター
朝日新聞
具体
自民党
代表
維新
変更
府
大飯
稼働
コメント
保障
財源
年金
会議
社説
選挙
マニフェスト
安全
民主党
現行
記者
制度
所得
朝日
読売
テレビ
出演
橋下
徹
情報
分
確認
基準
増税
税
交付
消費
有権者
生活
部分
条例
チェック
権力
高校
批判
メディア
市政
財政
地方
廃止
本質
教育
保護
現場
分権
改革
学校
目標
論
中野
官僚
反対
賛成
運営
全国
議会
議員
国会

もう少し切りたい…

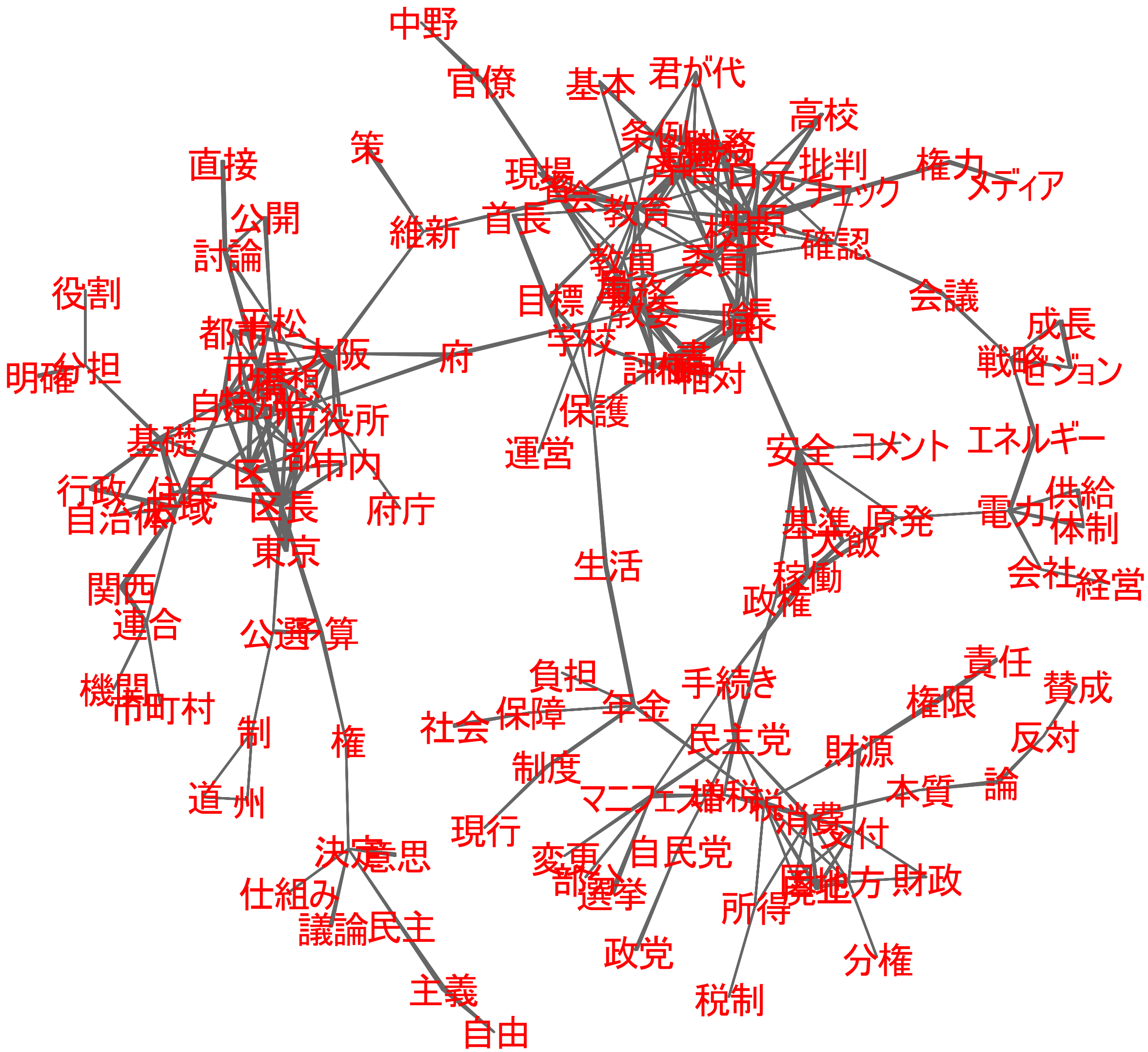
```
> adjmatrix2 <- matrix(0, 299,299)
> adjmatrix2[cormatrix >= 0.18] <- 1
> rownames(adjmatrix2) <- colnames(adjmatrix) <- topwords
>
> gplot(adjmatrix2, gmode = "graph", displayisolates = FALSE,
+ displaylabels = TRUE, vertex.cex = 0, label.pos= 5,
+ label.col = "red", edge.col = "grey40")
```

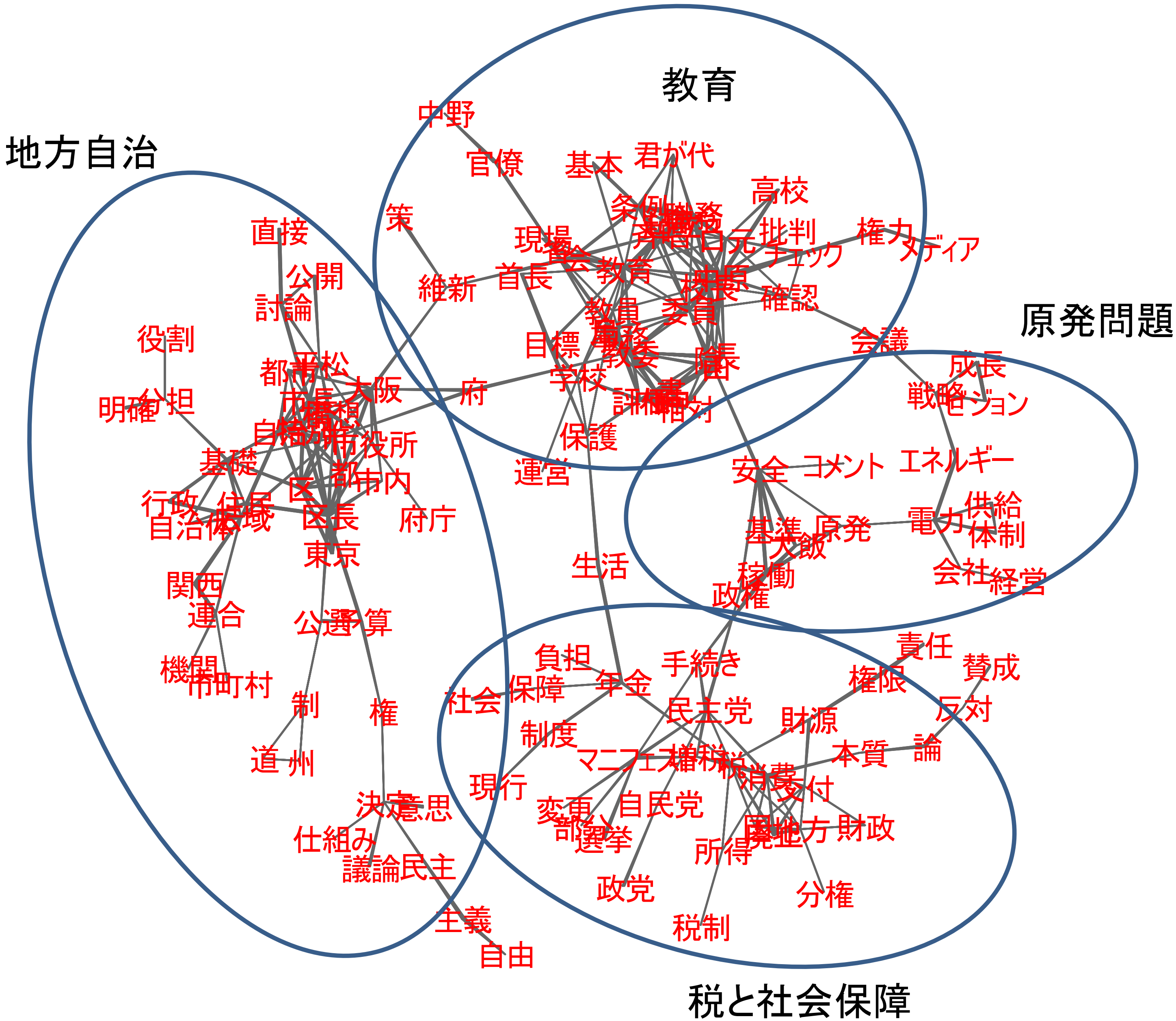


代表
選
府知事
組織
トップ
職員
活動
組合
議会
機構
議員
統治
国会
市町村
機関
関西
連合
文楽
文化
補助
金
団体
市政
改革
記者
質問
新潮
子供
私
取材
週刊
誌
報道
識者
知識
池田
役割
明確
分担
行政
広域
自治体
直接
中野
官僚
時代
知事
大学
教授
朝日新聞
社説
朝日
府民
市民
策
基礎
都市
自治
住民
討論
公開
運営
首長
省
維新
府
大阪
府庁
市役所
区長
東京
道
州
制
公選
予算
記事
毎日新聞
面
価値
観
社会
生活
現行
保障
負担
制度
年金
目標
基本
保護
評価
相対
教育
条例
君が代
口元
批判
経済
実現
政策
権
読売
テレビ
自称
インテリ
維持
結局
コメント
安全
確認
高校
権力
会議
メディア
政党
自民党
決定
意思
仕組み
民主
議論
出演
情報
分
部分
選挙
変更
手続き
基準
原発
民主党
増税
税
戦略
主義
自由
電力
体制
成長
税制
所得
番組
コメンテーター
会社
財源
地方
本質
経営
財政
権限
分権
論
責任
反対
賛成

最大連結成分だけ表示

```
> gplot(component.largest(adjmatrix2, result = "graph"),
+ gmode = "graph", displayisolates = FALSE,
+ displaylabels = TRUE, vertex.cex = 0, label.pos= 5,
+ label.col = "red", edge.col = "grey40")
```


中野
官僚
基本
君が代
高校
批判
権力
メディア
チェック
現場
首長
教育
確認
直接
策
公開
維新
討論
役割
教員
委員
会議
目標
成長
戦略
大阪
府
学校
分担
明確
相対
市役所
保護
運営
基礎
エネルギー
安全
コメント
市内
行政
区
区長
府庁
供給
自治体
電力
原発
体制
東京
生活
稼働
会社
経営
関西
政権
連合
予算
責任
手続き
賛成
負担
機関
市町村
権限
年金
社会
保障
民主党
財源
反対
制
権
制度
本質
論
道
州
現行
自民党
決定
意思
変更
仕組み
選挙
財政
所得
議論
民主
分権
政党
主義
税制
自由



地方自治
教育
原発問題
税と社会保障
中野
官僚
基本
君が代
高校
批判
権力
メディア
策
現場
維新
首長
教育
確認
会議
目標
学校
保護
運営
府
直接
公開
討論
役割
明確
分担
大阪
市役所
基礎
行政
自治体
区長
東京
府庁
関西
連合
機関
市町村
制
道州
権
公選
予算
決定
意思
仕組み
議論
民主
主義
自由
生活
成長
戦略
エネルギー
コメント
安全
原発
電力
供給
体制
会社
経営
基準
大飯
稼働
政権
責任
賛成
権限
反対
財源
本質
論
負担
手続き
年金
社会
保障
民主党
制度
現行
自民党
変更
選挙
政党
所得
税制
分権
財政
交付

感想

- センセーショナルな発言が注目されるが、ツイートのトピックは政策が中心
- あえて代名詞「僕」をキーワードにいれてみたが、特に意味はなかった…

今回やっていないこと

- 「ツイート」「ツイート」「ツィート」など表記ブレの統一
- 「朝日新聞」「朝日」など同義語の統一
- 「道」「州」「制」のように分割されしまう語や「維新の会」のような固有名詞の辞書登録

MeCabの辞書登録

- CSVファイルでユーザー辞書を作成

例）道州制,-1,-1,10,名詞,一般,*,*,*,*,道州制,ドウシュウセイ,ドーシューセー

- これをuserdic.csvという名前でC:¥に置いたとする（名前と場所は任意）
- コマンドプロンプト（管理者として実行）で"C:¥Program Files¥MeCab¥bin"に移動
- C:¥Program Files¥Mecab¥bin>mecab-dict-index.exe -d “C:¥Program Files¥MeCab¥dic¥ipadic” -u user.dic -f shift-jis -t shift-jis C:¥userdic.csv　を実行
- -u user.dic は作成する辞書の名前を指定（任意）
- 最後のC:¥userdic.csvは参照するCSVファイルの指定
- “C:¥Program Files¥MeCab¥dic¥ipadic”にあるdicrcというファイルをメモ帳などで開き

  userdic = C:¥Program Files¥MeCab¥bin¥user.dicと書き足す

感想

- 辞書作成やストップワードの指定など地味な作業が必要
- MeCabのユーザー辞書登録作業の繰り返しにはバッチファイルを作って処理するのがよい
- 時系列的な変化の分析もやってみたい