データ解析
第十四回「トピックモデル：Latent Dirichlet Allocation」

鈴木　大慈
理学部情報科学科
西八号館 W707 号室
s-taiji@is.titech.ac.jp

今日の講義内容

- トピックモデル
- LDA (Latent Dirichlet Allocation)
- 日本語 Wikipediade で LDA

構成

トピックモデル

- ゲッツェの 1 点でドイツが世界制覇
- フィオレンティーナは DF ゴンサロ・ロドリゲスとの契約を延長

トピックモデル

- ゲッツェの 1 点でドイツが世界制覇
- フィオレンティーナは DF ゴンサロ・ロドリゲスとの契約を延長

どちらもサッカーにまつわる話だとわかる．しかし,「サッカー」という単語は出ていない．
→ 文章に表れる単語がサッカーに関係する記事で目にするものばかり．
→ トピック：単語頻度の傾向．同じ記事に現れやすい単語は同じトピックに属するだろう．

今日は 文章クラスタリング にトピックモデルを用いるが，音楽や画像などトピックモデルは様々なデータ形式に適用可能．

Bag of Words

Bag of words: 出現した単語の頻度を並べたベクトル.
各要素が各単語の頻度に対応.
単語数分だけの次元になるため高次元になりやすい. 自然言語処理では 100 万次元はザラ.

$$\begin{pmatrix} \text{“please” の出現頻度} \\ \text{“credit” の出現頻度} \\ \vdots \\ \text{“money” の出現頻度} \end{pmatrix}$$

一つの文章を Bag of wards で表現. 文章をベクトルとして表せる.

Bag of Words (続き)

我々が用いるデータ
文章数×単語数の単語頻度行列．単語数は簡単に数百万とかになる．
基本的に単語頻度行列は**超スパース** (ほとんどの要素が 0).

	単語 1	単語 2	単語 3	···	単語 N
文章 1	4	8	0	···	2
文章 2	2	0	2	···	1
文章 3	2	4	0	···	0
⋮					

この表だけ からトピックを抽出し文章をトピックに分類 (クラスタリング)

準備: 多項分布

多項分布：単語頻度の生成確率を表すのに用いる．n 単語中，単語 i が x_i 回現れる確率:

$$(x_1, \ldots, x_k) \sim \mathrm{Mult}(\beta; n)$$

$$\Leftrightarrow \qquad P(x_1, \ldots, x_k | \pi) = \frac{n!}{x_1! \ldots x_k!} \beta_1^{x_1} \ldots \beta_k^{x_k}.$$

ただし，$\beta = (\beta_1, \ldots, \beta_M)$ は $\sum_{i=1}^{M} \beta_i = 1$, $\beta_i \geq 0$ なる確率ベクトル．

サイコロの目のでやすさ．

Dirichlet 分布

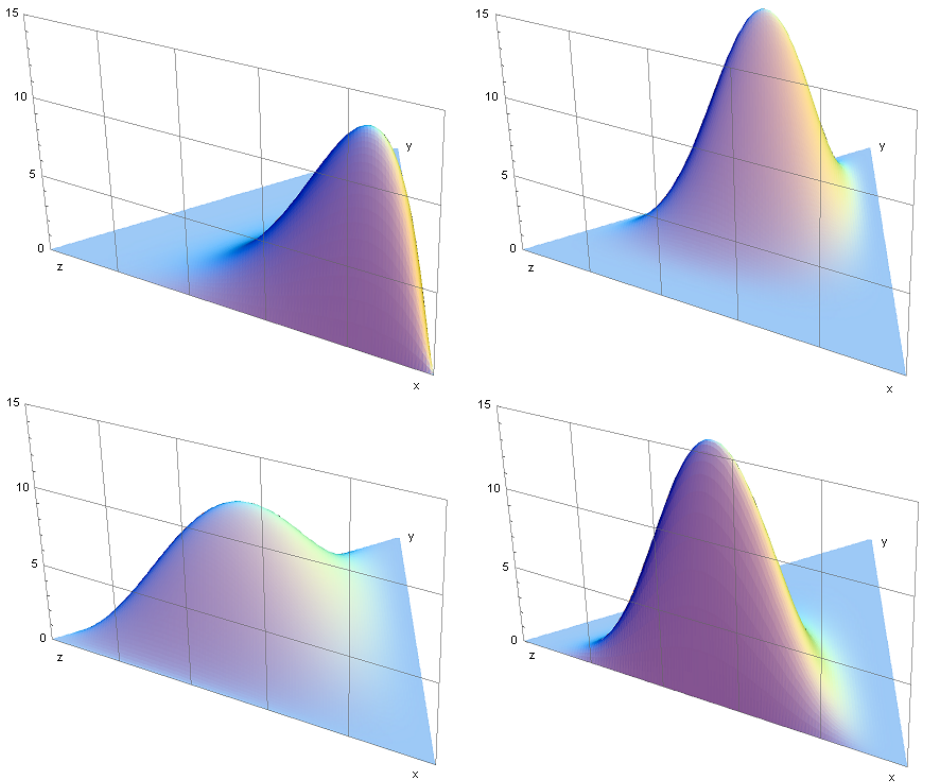
Dirichlet 分布：確率ベクトル β の生成確率を表すのに用いる (ベイズ推定).

$$\beta \sim \mathrm{Diri}(\alpha)$$

$$\Leftrightarrow \qquad p(\pi|\alpha) = \frac{\prod_{i=1}^{n} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{n} \alpha_i)} \prod_{i=1}^{n} \beta_i^{\alpha_i - 1}.$$

ここで $\alpha = (\alpha_1, \ldots, \alpha_n)$ は $\alpha_i > 0$ なる n 次元ベクトル．また，β は確率ベクトルでなくてはいけない.

Dirichlet 分布の図



Wikipedia「ディリクレ分布」の記事より.

準備: 多項分布と Dirichlet 分布の共役性

多項分布に従う確率変数 $X = (X_1, \ldots, X_k)$ の実現値 $x = (x_1, \ldots, x_k)$ $(\sum_{i=1}^{k} x_i = n)$ を観測した．多項分布のパラメータ π をベイズ推定したい．事前分布に $\mathrm{Diri}(\alpha)$ を持ってきた時，事後分布はどうなるか？

準備: 多項分布と Dirichlet 分布の共役性

多項分布に従う確率変数 $X = (X_1, \ldots, X_k)$ の実現値 $x = (x_1, \ldots, x_k)$ $(\sum_{i=1}^{k} x_i = n)$ を観測した．多項分布のパラメータ π をベイズ推定したい．事前分布に $\mathrm{Diri}(\alpha)$ を持ってきた時，事後分布はどうなるか？

答え：**Dirichlet 分布**

$$
\begin{aligned}
p(\pi|x) &= \frac{p(x|\pi)p(\pi|\alpha)}{\int p(x|\pi)p(\pi|\alpha)\mathrm{d}\pi} \\
&\simeq \underbrace{(\pi_1^{x_1} \ldots \pi_k^{x_k})}_{\text{尤度の部分}} \times \underbrace{(\pi_1^{\alpha_1 - 1} \ldots \pi_k^{\alpha_1 - k})}_{\text{事前分布の部分}} \\
&= \pi_1^{x_1 + \alpha_1 - 1} \ldots \pi_k^{x_k + \alpha_k - 1} \\
&= \mathrm{Diri}((x_1 + \alpha_1, \ldots, x_k + \alpha_k)).
\end{aligned}
$$

事前分布のパラメータ $= (\alpha_1, \ldots, \alpha_k)$
→事後分布のパラメータ $= (x_1 + \alpha_1, \ldots, x_k + \alpha_k)$

LDA: Latent Dirichlet Allocation

アイディア: 似たトピックの文章は出てくる単語も似ているだろう.

トピック数 K, 単語数 M, 文章数 N とする.

- トピックごとに単語の出現確率

$$\beta^{(k)} = (\beta_1^{(k)}, \ldots, \beta_M^{(k)}) \quad (k = 1, \ldots, K)$$

を設定. トピックの数 K は文章数 N よりずっと小さいとする.

- 各文章ごとに**トピックの割合** $\pi^{(d)} = (\pi_1^{(d)}, \ldots, \pi_K^{(d)})$ を設定.
その文章の単語頻度はトピック割合で混合された混合多項分布から発生:

$$(x_1^{(d)}, \ldots, x_M^{(d)}) \sim \sum_{k=1}^{K} \pi_k^{(d)} \underbrace{\mathrm{Mult}(\beta_k; n^{(d)})}_{k \text{ 番目のトピックの分布}}$$

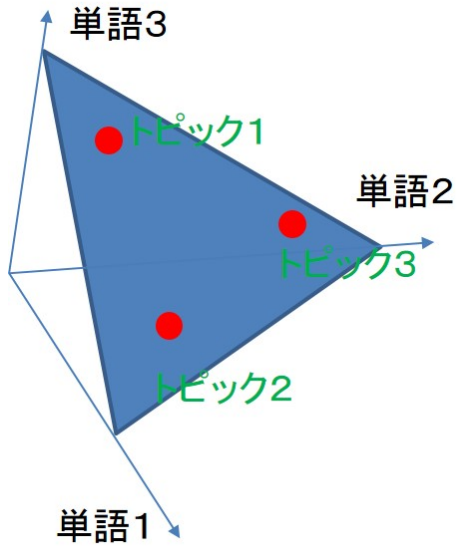
$$= \mathrm{Mult}\Big(\sum_{k=1}^{K} \pi_k^{(d)} \beta_k; n^{(d)} \Big).$$

ただし, $n^{(d)}$ は文章 d に現れた総単語数.

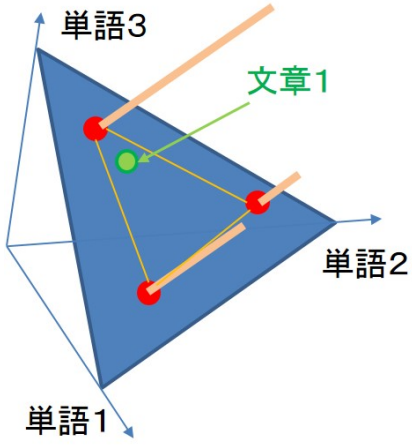
この $\{\pi^{(d)}\}$ と $\{\beta^{(k)}\}$ をデータから推定.

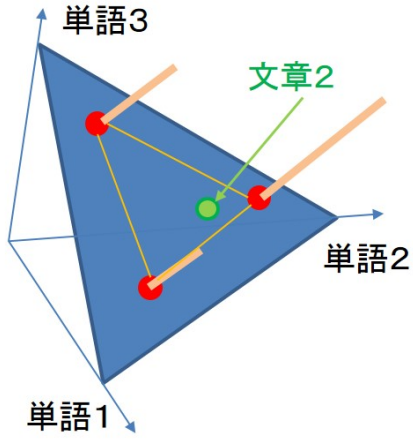
※トピックの割合が似ている文章は内容も似る.

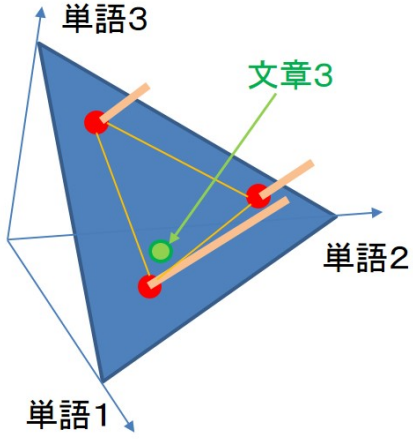
LDA のモデルを図示



各トピックは単語の出現頻度で特徴付けられる．







各文章における単語の出現頻度はトピックの混合で決まる．

尤度関数

観測値: 単語頻度行列 $X = (x_1^{(d)}, \ldots, x_M^{(d)})_{d=1}^N$ ($N \times M$: 文章数×単語数)

$x_i^{(d)}$ は単語 i が文章 d に現れた回数.

尤度:

$$p(X|\{\beta_k\}_{k=1}^K, \{\pi^{(d)}\}_{d=1}^N) = \prod_{d=1}^N p\left(x^{(d)}|\sum_{k=1}^K \beta^{(k)}\pi_k^{(d)}\right).$$

なお，$p\left(x^{(d)}|\sum_{k=1}^K \beta^{(k)}\pi_k^{(d)}\right)$ は各文章において単語頻度 $x^{(d)}$ が観測される確率. つまり，多項分布 $\mathrm{Mult}\left(\sum_{k=1}^K \pi_k^{(d)}\beta_k; n^{(d)}\right)$ の確率質量関数.

推定

尤度:

$$p(X|\{\beta_k\}_{k=1}^K, \{\pi^{(d)}\}_{d=1}^N) = \prod_{d=1}^N p\left(x^{(d)}|\sum_{k=1}^K \beta_k \pi_k^{(d)}\right).$$

LDA のやること：$\{\beta_k\}_{k=1}^K$ と $\{\pi^{(d)}\}_{d=1}^N$ を**ベイズ推定**．これらに Dirichlet 事前分布を置く．π も β も確率ベクトルであることに注意.
（β_k は先の図でいう赤い点の位置，$\pi^{(d)}$ は黄色い棒の長さに対応）

LDA から得られるもの：事後分布 $p(\{\beta_k\}_{k=1}^K, \{\pi^{(d)}\}_{d=1}^N|X)$.
例えば事後分布の期待値を各パラメータの推定量として採用.

- $\{\hat{\beta}_k\}_{k=1}^K$ から各トピックがどんな単語を生成しやすいかがわかる.
  → 単語の属性
- $\{\hat{\pi}^{(d)}\}_{d=1}^N$ から各文章がどんなトピックの内容かわかる.
  → 文章の属性

こうして，単語や文章の性質が **LDA** から自動的にわかってしまう！

LDA の注意点

1. 多項分布を混合しても多項分布なので，ドキュメントが 1 つの時は LDA のモデルは単なる多項分布.
2. 沢山ドキュメントがあったとき，それぞれの単語出現頻度を推定していてはパラメータの自由度が多過ぎる.
   → 少数のトピックに分けて実質的なパラメータを減らしている.
3. ベイズ事後分布を使えば新しく文章が来てもそのトピック割合を推定できる.

ベイズ事後確率の推定方法

難しいので本講義では割愛.
いろいろな方法がある.

- Gibbs サンプリング (事後確率からパラメータをサンプリング)
- 変分ベイズ法 (事後確率を近似計算)
- Collapsed 変分ベイズ法
- Collapsed Gibbs サンプリング

いずれにせよ解析的に事後確率は求まらない．計算機を使って近似なりサンプリングなりをする.

R で LDA

library(topicmodels) や library(lda) で利用可.

X は単語頻度行列，K はトピック数として，

```
> LDA(X, K)
```

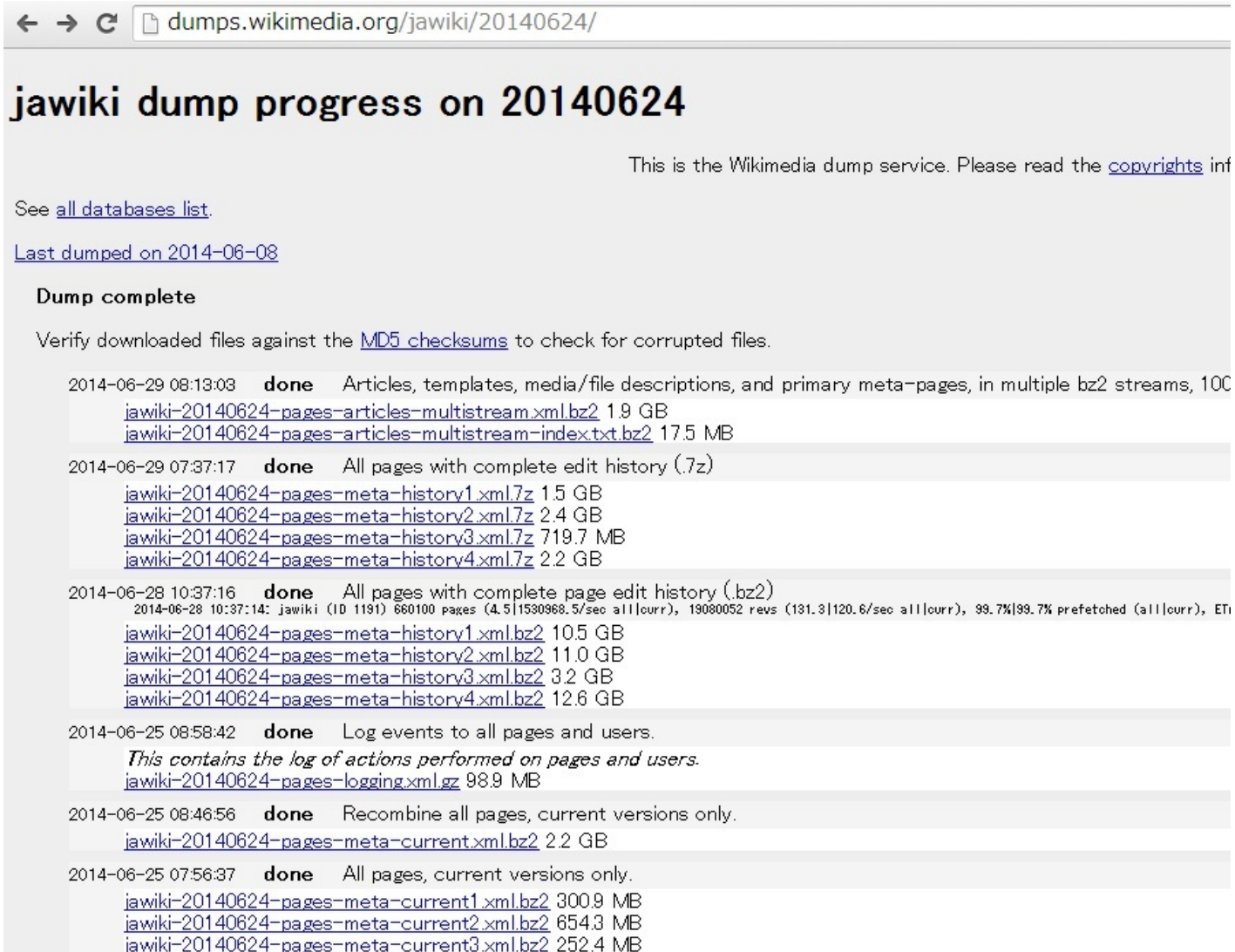
でよい.

単語頻度行列はスパース行列フォーマットを用いる．詳しくはそれぞれのパッケージのドキュメントを参照.

構成

データの入手元

2014 年現在の日本語 Wikipedia の記事データ.
`http://dumps.wikimedia.org/jawiki/20140624/` から
jawiki-20140624-pages-articles1.xml.bz2 をダウンロード.



データの作り方

Python を用いて記事を抽出．記事を一つの文章として単語頻度行列を作成．
Python モジュールの Gensim を用いてコーパス (Bag-of-words, 単語頻度行列) を作成．

環境：Python3.4.1 + Numpy1.8.1 + Scipy0.14.0, Windows 7, 64bit

Gensim の WikiCorpus を継承した JaWikiCorpus というクラスを作成し，コーパスを生成 (スクリプトファイルを参照)．
その際，python コードの中から MeCab を呼び出して日本語文章を分かち書きし，名詞だけを取り出し単語辞書とした．

```
> python jawikicorpus_make.py
   jawiki-20140624-pages-articles1.xml.bz2 jawiki1
```

分かち書き

MeCab による分かち書き.
入力：文章を書きましょう.
出力：

```
文章  名詞,一般,*,*,*,*,文章,ブンショウ,ブンショー
を  助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
書き  動詞,自立,*,*,五段・カ行イ音便,連用形,書く,カキ,カキ
ましょ  助動詞,*,*,*,特殊・マス,未然ウ接続,ます,マショ,マショ
う  助動詞,*,*,*,不変化型,基本形,う,ウ,ウ
.  記号,句点,*,*,*,*,.  ,.  ,.
```

生成されたデータ：単語辞書 (単語のリスト) と記事タイトル

単語辞書最初の 10 個

```
62158 aa 202
31510 aaa 132
10543 aab 65
15293 aac 48
25269 aaron 42
19714 ab 212
32430 aba 93
19037 abba 21
45622 abbey 24
19673 abc 706
```

記事タイトル最初の 10 個

```
0 アンパサンド
1 言語
2 日本語
3 地理学
4 EU（曖昧さ回避）
5 国の一覧
6 パリ
7 ヨーロッパ
8 生物
9 コケ植物
```

生成されたデータ：単語頻度行列

単語頻度行列は Matrix Market フォーマットで吐き出した.

```
%%MatrixMarket matrix coordinate real general
59749 62999 4970557
1 867 2
1 1577 1
1 6045 1
1 9144 1
1 9393 1
1 10498 2
1 11234 3
1 11705 1
```

単語頻度行列のサイズは 59,749×62,999 (記事数 × 単語数) で 4,970,557 個の非ゼロ要素.

LDA にかけてみる

トピック数 20, Gibbs サンプリングを用いた.

```
K = 20
wiki.lda <- LDA(ssx,K,method='Gibbs',control
                    = list(burnin=2000,iter = 5000))
```

list(burnin=2000,iter = 5000) は 5000 回サンプリングして最初の 2000 個のサンプルは捨てたという意味. 計算時間は約４０分ほど.

トピックごとの主要単語

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
[1,]	"丁目"	"de"	"オブ"	"windows"	"モハ"
[2,]	"人口"	"la"	"シリーズ"	"gt"	"mm"
[3,]	"交通"	"サン"	"ゲーム"	"pc"	"クハ"
[4,]	"教育"	"cc"	"ドラマ cd"	"lt"	"km"
[5,]	"地理"	"file"	"vol"	"例えば"	"番台"
[6,]	"道路"	"フォン"	"アニメ"	"os"	"キハ"
[7,]	"中学校"	"年頃"	"名探偵コナン"	"ms"	"両編成"
[8,]	"北海道道"	"ルイ"	"機動戦士ガンダム"	"ii"	"編成"
[9,]	"高等学校"	"フランス"	"one"	"mhz"	"系電車"
[10,]	"小学校"	"ドイツ"	"ナレーション"	"mb"	"サハ"
[11,]	"年生"	"le"	"劇場版"	"vs"	"国鉄"
[12,]	"鉄道"	"マリア"	"テレビ朝日版"	"minus"	"cm"
[13,]	"行政"	"ジャン"	"ドラえもん"	"for"	"クモハ"
[14,]	"自由民主党"	"image"	"テレビアニメ"	"mac"	"形電車"
[15,]	"番地"	"パリ"	"それいけ"	"system"	"dd"

トピックごとの主要単語

	Topic 6	Topic 7	Topic 8	Topic 9
[1,]	"俳優"	"フジテレビ"	"nbsp"	"大正"
[2,]	"ジョン"	"tbs"	"km"	"となる"
[3,]	"女優"	"nhk"	"年創業"	"が発足"
[4,]	"作曲家"	"テレビ朝日"	"大正"	"暴力"
[5,]	"元プロ野球選手"	"日本テレビ"	"長野県道"	"現在"
[6,]	"プロ野球選手"	"東映"	"キハ"	"条第"
[7,]	"政治家"	"主演"	"日廃止"	"の管轄となる"
[8,]	"歌手"	"東宝"	"近畿日本鉄道"	"恋愛"
[9,]	"小説家"	"監督"	"特急"	"一部"
[10,]	"声優"	"テレビ東京"	"急行"	"犯罪"
[11,]	"ジョージ"	"シリーズ"	"往復"	"幕府領"
[12,]	"広島県道"	"映画"	"駅名"	"日町制"
[13,]	"タレント"	"松竹"	"営業キロ"	"不明"
[14,]	"サッカー選手"	"出演"	"全線"	"に改称"
[15,]	"ロバート"	"テレビドラマ"	"上り"	"郡より離脱"

トピックごとの主要単語

	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14
[1,]	"bs"	"月号"	"系統"	"回戦"	"en"
[2,]	"hd"	"講談社"	"下り"	"優勝"	"日本"
[3,]	"月発売"	"のち文庫"	"上り"	"万円"	"一方"
[4,]	"日に"	"op"	"番線"	"試合"	"という"
[5,]	"月に"	"新潮社"	"jr 東日本"	"東京"	"アメリカ"
[6,]	"com"	"作品"	"駅周辺"	"阪神"	"例えば"
[7,]	"sports"	"岩波書店"	"隣の駅"	"月場所"	"right"
[8,]	"欠番"	"原作"	"のりば"	"人目"	"英語"
[9,]	"tv"	"集英社"	"駅構造"	"巨人"	"であり"
[10,]	"sup"	"文藝春秋"	"日本の鉄道駅一覧"	"史上"	"年には"
[11,]	"沿革"	"小学館"	"利用状況"	"得点"	"または"
[12,]	"月現在"	"角川書店"	"行先"	"kg"	"そして"
[13,]	"当時"	"小説"	"路線"	"通算"	"このため"
[14,]	"億円"	"中央公論社"	"番のりば"	"年は"	"png"
[15,]	"スカパー"	"日号"	"快速"	"西武"	"建築"

トピックごとの主要単語

	Topic 15	Topic 16	Topic 17	Topic 18
[1,]	"and"	"km"	"ch"	"紀元前"
[2,]	"in"	"text"	"土曜"	"在位"
[3,]	"file"	"style"	"金曜"	"年頃"
[4,]	"to"	"東京都"	"日曜"	"天正"
[5,]	"university"	"億円"	"月曜"	"年代"
[6,]	"new"	"北海道"	"月から"	"には"
[7,]	"on"	"center"	"日から"	"慶長"
[8,]	"by"	"align"	"木曜"	"代藩主"
[9,]	"with"	"bar"	"月まで"	"在位紀元前"
[10,]	"press"	"県道"	"kw"	"万石"
[11,]	"for"	"時間"	"日本テレビ系列"	"ユリウス暦"
[12,]	"at"	"部リーグ"	"出力"	"天文"
[13,]	"en"	"大阪府"	"備考"	"寛永"
[14,]	"white"	"新潟県"	"火曜"	"世紀"
[15,]	"black"	"bull"	"fm"	"任官"

トピックごとの主要単語

	Topic 19	Topic 20
[1,]	"th"	"作曲"
[2,]	"love"	"作詞"
[3,]	"live"	"日発売"
[4,]	"アルバム"	"編曲"
[5,]	"in"	"脚本"
[6,]	"dvd"	"音楽"
[7,]	"you"	"監督"
[8,]	"cd"	"身長"
[9,]	"best"	"スタッフ"
[10,]	"to"	"に登場"
[11,]	"music"	"主題歌"
[12,]	"シングル"	"演出"
[13,]	"my"	"cm"
[14,]	"on"	"体重"
[15,]	"go"	"製作"

トピックの解釈

- Topic 3: アニメ関係
- Topic 4: コンピュータ関係
- Topic 5: 鉄道関係
- Topic 6: 有名人関係
- Topic 18: 歴史関係

トピックごとの記事タイトル（抜粋）

"Topic 3 :"
"架空の国一覧 | 岡村明美 | 佐久間レイ | 三木眞一郎 | 石田彰 | うえだゆうじ | 山口勝平 | 根谷美智子 | 広瀬正志 | 小西克幸 | 八奈見乗児 | 山口由里子 | 進藤尚美 | くまいもとこ | 関俊彦 | 千葉一伸 | 草尾毅 | 坂本千夏 | 飛田展男 | 三宅健太"

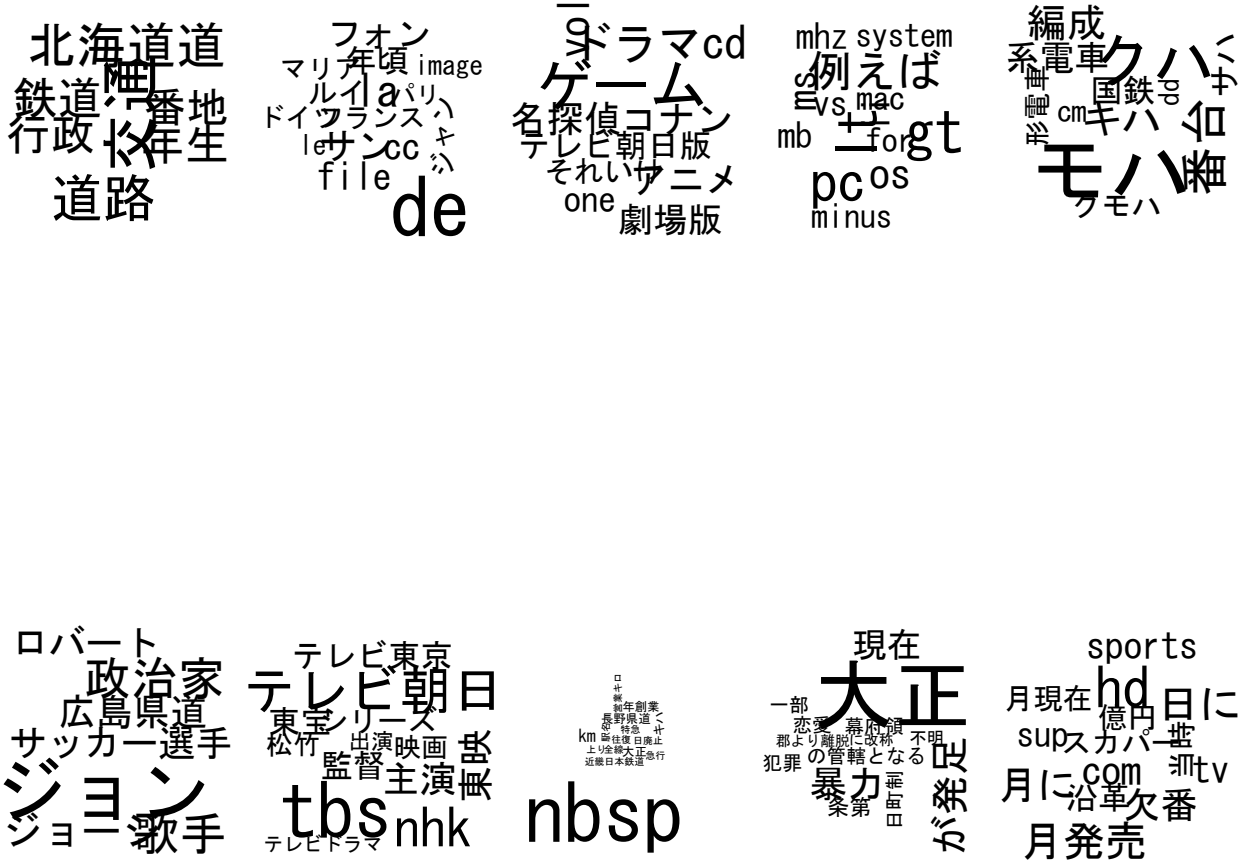
"Topic 4 :"
"Xeon | PC-9821シリーズ | 順序数 | ThinkCentre | Safari | Microsoftオン化傾向 | X68000 | Unicode一覧 E0000-E0FFF | MC68000 | .NET Framew

"Topic 18 :"
"中国帝王一覧 | 元号一覧（日本） | 天文（元号） | 従一位 | 後白河天皇 | 延暦 | 享保 | 紀元前1千年紀 | 文化（元号） | 伺候席 | 征夷大将軍 | 夏商周年表 | 宝暦 | 備前国 | 守護代 | 醍醐天皇 | 伊勢国 | 摂津国 | 相模国 | 紀元前4世紀"

トピックの解釈とよく合っている.

ワードクラウド

`wordcloud(vocs,freq)`
でワードクラウドをプロットできる (library(wordcloud) が必要)．vocs は文字列のベクトル，freq は各文字列要素の頻度 (0 以上なら良い，足して 1 である必要もない)．freq が大きいほど大きな文字になる．



構成

レポート (gam)

マサバデータで gam

- 講義情報ページ (`For_report.zip` に含まれている) にある mackgam.R を実行せよ.
- mackgam.R を修正し lon,lat の交互作用を取り込んだモデルで推定せよ.
- さらに b.depth,c.dist の交互作用を取り込んだモデルを推定せよ.
- もとのモデルと上の 2 つのモデルで UBRE (Unbiased Risk Estimator) を比較せよ．UBRE は `mack.gamadd$gcv.ubre` で得られる．(UBRE は GCV のように予測誤差の推定量である)

レポート (gam)

都道府県をクラスタリング

- 'ken-kankyo-kakou.csv' (`For_report.zip` に含まれている) を読み込み，主成分分析にかけて，第四主成分スコアまで取り出せ.
- 上で取り出した主成分スコア (4 次元) を用いて階層クラスタリング (hclust) および混合正規分布によるクラスタリング (Mclust) を適用した結果を表示せよ．また，その結果について考察を述べよ．(hclust の結果は `plot(hc)` で，Mclust の結果は `heatmap(result$z,col=redgreen(256))` のようにして表示すればよい)

レポート (LDA): 余力がある人のみ

(optional) LDA を jawiki2 データに適用し，結果を解析せよ．
(`jawiki2_short_bow.mm` が単語頻度行列，`jawiki2_short_titles_tmp.txt` が記事タイトル，`jawiki2_short_wordids_tmp.txt` が単語リスト)

レポートの提出方法

- 私宛にメールにて提出.
- 件名に 必ず「データ解析第 n 回レポート」と明記し，R のソースコードと結果をまとめたレポートを送付のこと.
- 氏名と学籍番号も忘れず明記すること.
- レポートは本文に載せても良いが，pdf などの電子ファイルにレポートを出力して添付ファイルとして送付することが望ましい (これを期に tex の使い方を覚えることを推奨します).
- 今回の提出期限は 8/8(金) まで.

※相談はしても良いですが，コピペは厳禁です.

講義情報ページ

`http://www.is.titech.ac.jp/~s-taiji/lecture/dataanalysis/dataanalysis.html`