トピックモデル

～ 1週間の献立を考える ～

Tokyo.R #76 LT

2019/03/02

自己紹介

- 名前 : もらとりあむお
- Twitter : @moratoriamuo271
- 趣味 : ・飲酒 (ビールと日本酒)
  - ・ 横浜DeNAベイスターズ
  - ・バドミントン, カラオケ, ボードゲーム
- 所属 : ４月から渋谷で働きます…

モチベーション

料理において、献立の決定は面倒くさい!!

(プログラミングで名前付けが大変なように？)

↓

レコメンドエンジンを作ってしまおう!!

- 同じものばかりオススメされても飽きるし、栄養も偏る
- レシピをクラスタリングして、クラスタ別でオススメする

レコメンドエンジンの作り方

レシピデータの収集　{rvest}と{stringr}

↓

ワードクラウドで可視化　{wordcloud}

↓

文書ターム行列の作成　{RMeCab}と{tm}

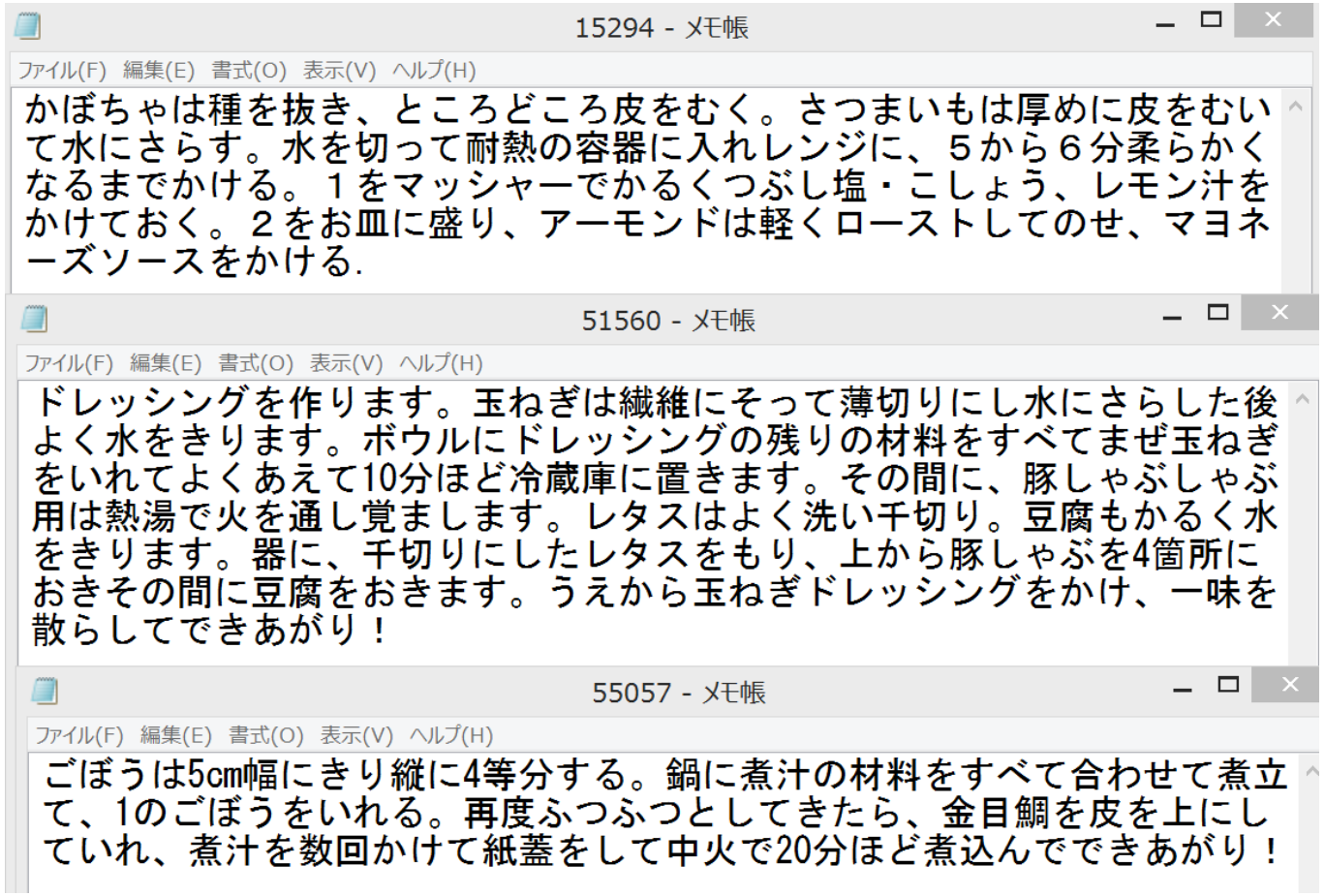
↓

LDAモデルの適用とトピック数の決定 {topicmodels}と{ldatuning}

↓

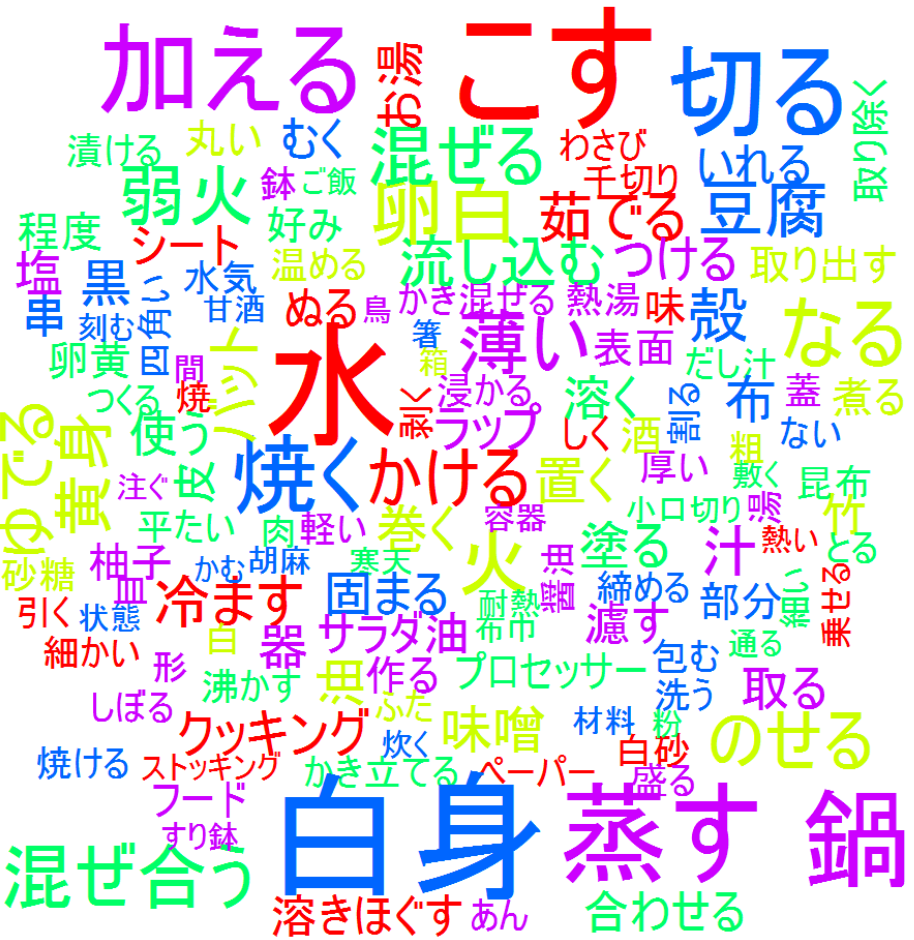
トピックによるレシピの分類とレコメンド　{tidytext}と自作関数

レシピデータの収集



クックパッドの「今日のご飯・おかず」カテゴリから500件収集

ワードクラウドで可視化



トピックモデルの説明

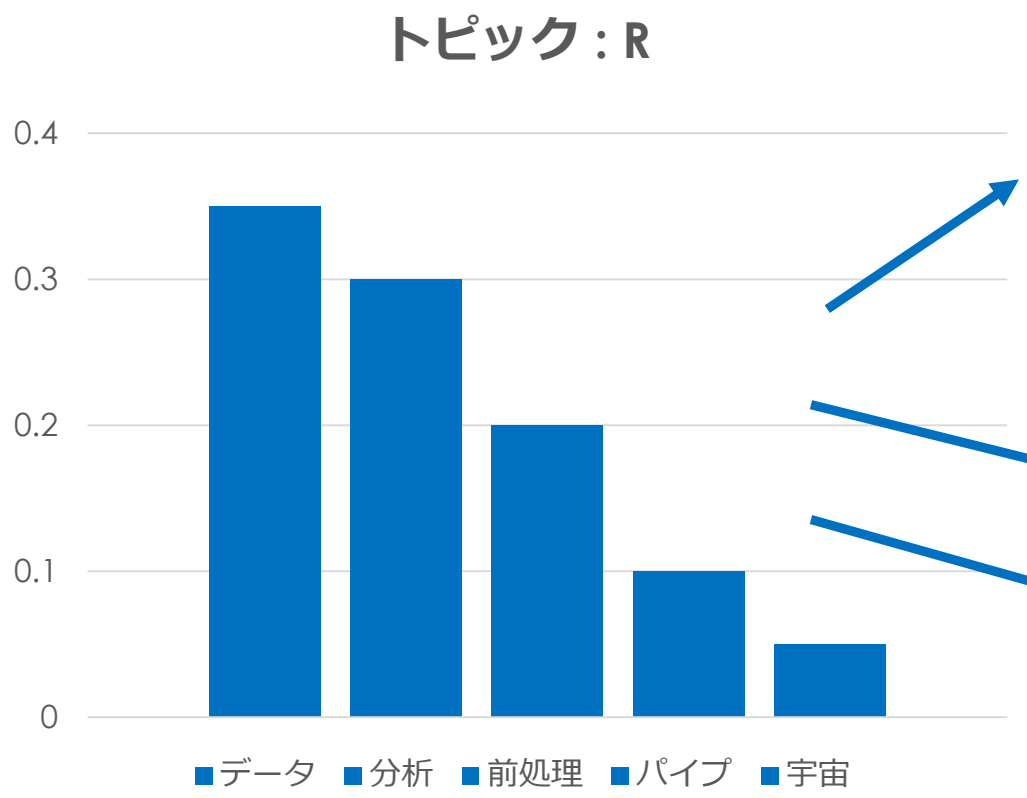
- **文書が生成される過程をモデル化した確率モデル**

トピックごとに単語を生成する確率分布があり、単語の集合である文書はそれぞれトピック(トピック分布)を持ち、それらによって文書が生成されていくモデル

- **ユニグラムモデル → 混合ユニグラムモデル → LDA** と拡張

- 文書だけでなく、画像や購買履歴、ネットワークのデータにも応用可能。

ユニグラムモデル



＜文書１＞

データの**分析**をする工程の９割は**前処理**だ。汚い**データ**…

＜文書２＞

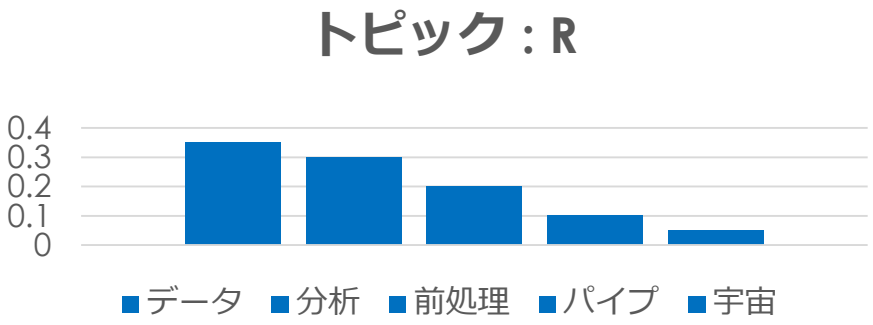
パイプ演算子は**分析**の工程を見やすくする。**パイプ**はいいぞ…

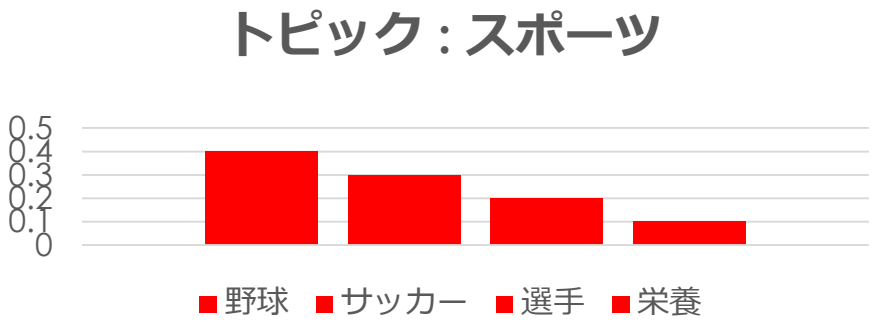
＜文書３＞

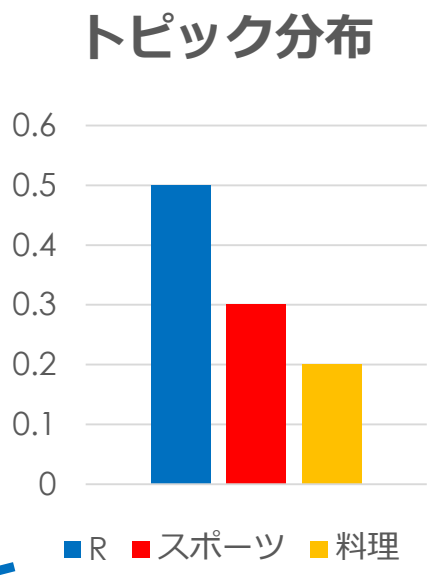
データ分析、それは**宇宙**。神。

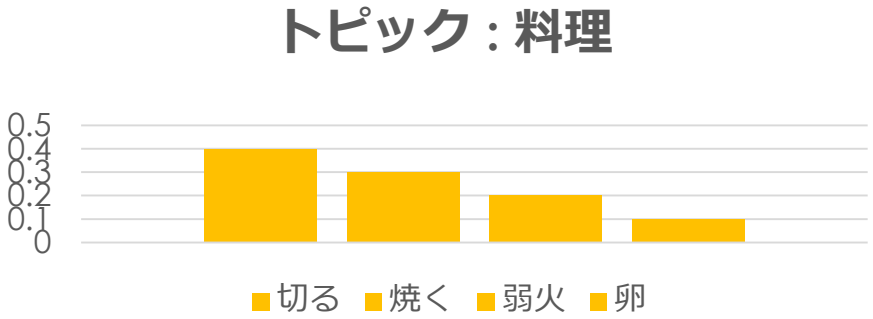
…

混合ユニグラムモデル









＜文書１＞

データの**分析**をする工程の９割は**前処理**だ。汚い**データ**…

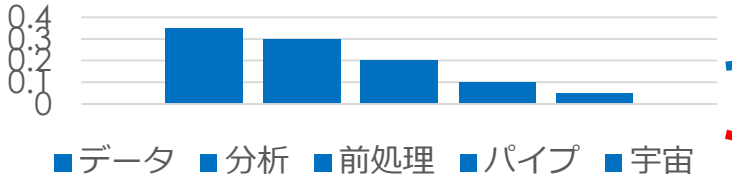
＜文書２＞

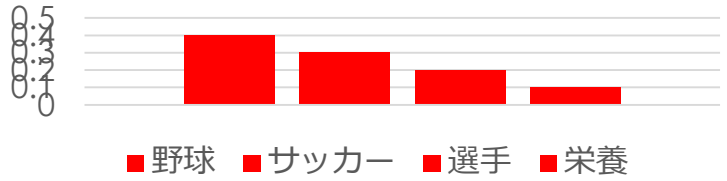
プロ**野球**の開幕に向けて若手**選手**がキャンプの…

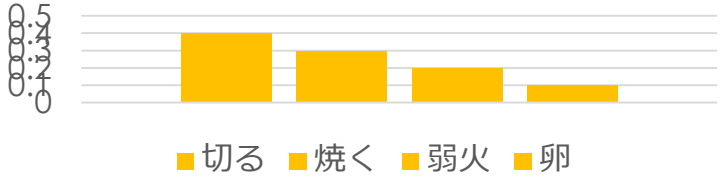
＜文書３＞

ジャガイモを細かく**切っ**て、カリッとなるよう**焼き**ます…

LDA(Latent Dirichlet Allocation)





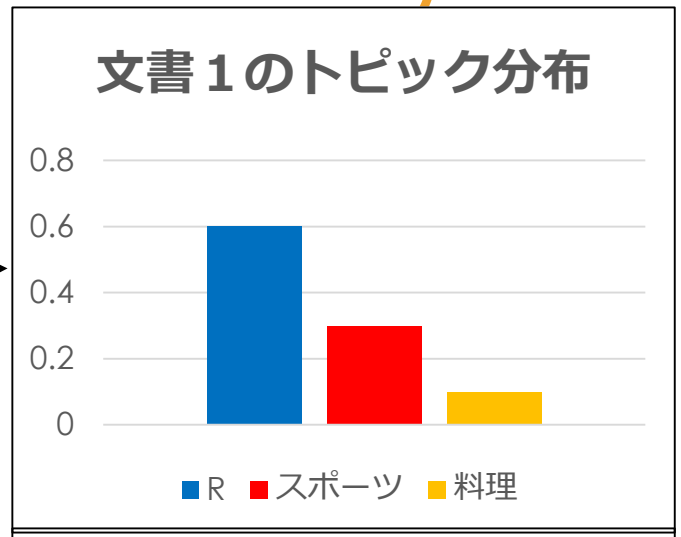


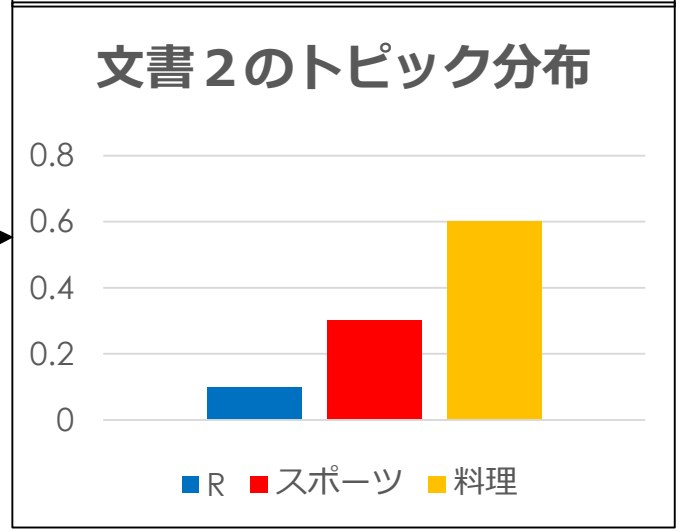
＜文書１＞

過去の**データ**を用いて、**野球**におけるバントの効果を**分析**…

＜文書２＞

スポーツ**選手**がアスリートとしての身体を維持するために**卵**料理…





トピック数の決定方法

- **パープレキシティ(perplexity)で評価**

  負の対数尤度から計算される値。testデータを使用。低い方が良いモデル。

- **{ldatuning}パッケージを使用**

  ４つの論文で提案された指標でモデルを評価。

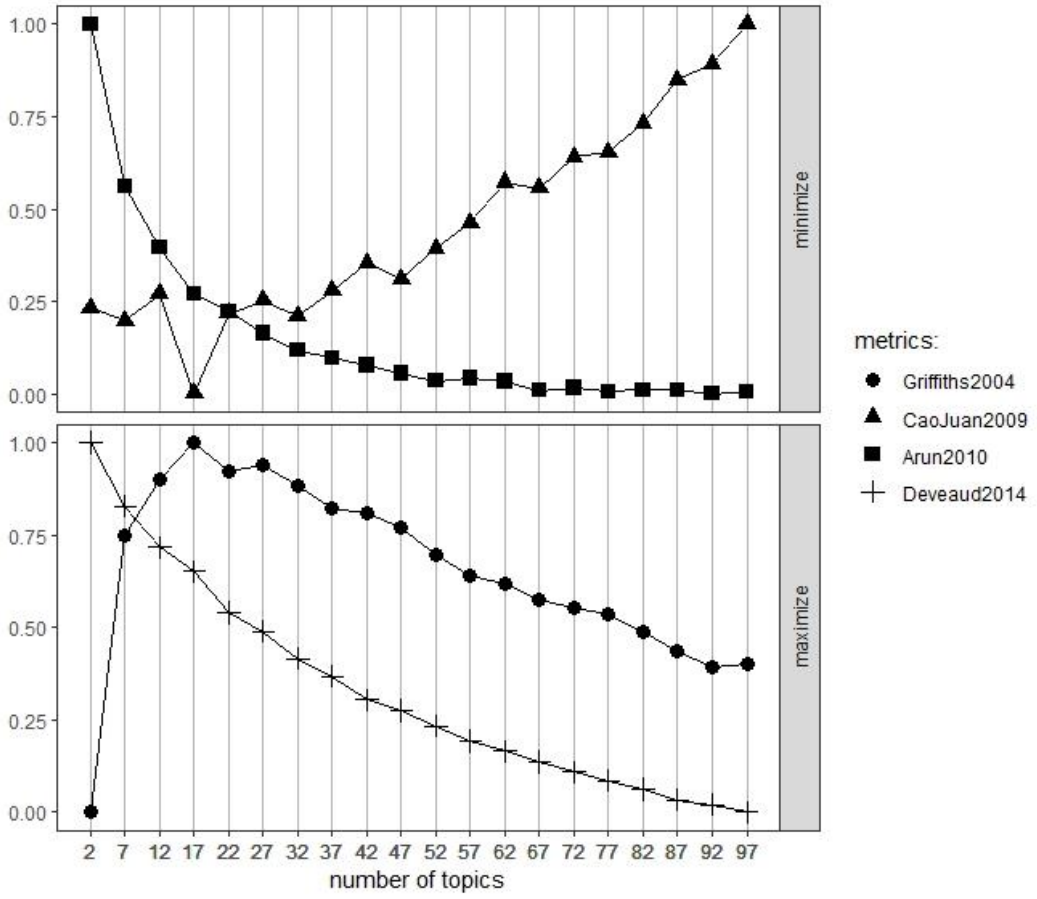
- **変分下限でモデル評価**

  変分ベイズ法を用いて推定をしている場合。

- **ディリクレ過程でトピック数もモデル化**

  階層ディリクレ過程を用いるとトピック数の推定が可能。

ldatuningとperplexity




```
result_tp12 <- LDA(DTM_nouns_verbs_train,12,method = "Gibbs")
result_tp17 <- LDA(DTM_nouns_verbs_train,17,method = "Gibbs")
result_tp22 <- LDA(DTM_nouns_verbs_train,22,method = "Gibbs")
result_tp27 <- LDA(DTM_nouns_verbs_train,27,method = "Gibbs")
```

```
perplexity(result_tp12, DTM_nouns_verbs_test)
```

```
## [1] 488.4454
```

```
perplexity(result_tp17, DTM_nouns_verbs_test)
```

```
## [1] 474.552
```

```
perplexity(result_tp22, DTM_nouns_verbs_test)
```

```
## [1] 471.9647
```

```
perplexity(result_tp27, DTM_nouns_verbs_test)
```

```
## [1] 473.9953
```

分類と解釈をしてみる

```
##     Topic 3 Topic 8 Topic 15 Topic 18
## 1    チーズ      肉        塩  キャベツ
## 2      牛乳      鶏  きゅうり      豚肉
## 3    ソース     むね     ボール    ごま油
## 4    バター      塩       胡椒      好み
## 5 コンソメ      衣     玉ねぎ        火
```

- Topic 3

  グラタン, シチュー, スープ

- Topic 8

  唐揚げなど鶏肉料理

- Topic 15

  サラダ

- Topic 18

  野菜炒め

レコメンドエンジンの完成

```
#一週間分の料理(7食分)をレコメンド
frecommend7url <- function(doctop){
  seventopics <- sample(x = unique(doctop$topic), size = 7, replace = FALSE)
  rec7 <- character(7)
  for(i in 1:7){
    doctopfil <- doctop %>% filter(topic==seventopics[i])
    rec7[i] <- sample(x = doctopfil$document, size = 1)
  }
  rec7 <- rec7 %>% str_remove(".txt") %>% str_c("https://cookpad.com/recipe/", .)
  return(rec7)
}

frecommend7url(doctop)
```

```
## [1] "https://cookpad.com/recipe/3169287"
## [2] "https://cookpad.com/recipe/1370545"
## [3] "https://cookpad.com/recipe/2733198"
## [4] "https://cookpad.com/recipe/2866455"
## [5] "https://cookpad.com/recipe/2721802"
## [6] "https://cookpad.com/recipe/2430458"
## [7] "https://cookpad.com/recipe/1120158"
```

参考文献

- 岩田『トピックモデル』
- 佐藤『トピックモデルによる統計的潜在意味解析』
- 松浦『StanとRでベイズ統計モデリング』
- 小林『Rによるやさしいテキストマイニング[機械学習編]』
- 『Select number of topics for LDA model』
  https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html
- 『蒙古タンメン中本コーパスに対してのLDAの適用とトピック数の探索』
  http://kamonohashiperry.com/archives/1619
- 『[R] トピックモデル(LDA)を用いた大量文書の教師なし分類』
  https://qiita.com/YM_DSKR/items/017a5dddeb56fcdf1054

ENJOY!!