第 5 章: 発見
(5.1. テキストデータ)

今井耕介 著

『社会科学のためのデータ分析入門 (QSS)』

2026-03-09

5.1 テキストデータ

Federalist Papers の著者推定

- **Federalist Papers**: 米国憲法の批准を促すために書かれた 85 編のエッセイ。
- ハミルトン、マディソン、ジェイの 3 人が執筆したが、一部の論文については著者が不明確であった。
- **目標**: テキスト分析を用いて、著者不明の論文の執筆者を推定する。

5.1.1 テキストデータの処理

tm パッケージによる前処理

- R の tm パッケージを使用して、文書集合（コーパス）を作成し、クリーニングを行う。

```
library(tm)
corpus.raw <- VCorpus(DirSource(directory = "federalist", pattern = "fp"))
corpus.prep <- tm_map(corpus.raw, content_transformer(tolower))
corpus.prep <- tm_map(corpus.prep, stripWhitespace)
corpus.prep <- tm_map(corpus.prep, removePunctuation)
corpus.prep <- tm_map(corpus.prep, removeNumbers)

# ステミング前の DTM（著者推定用）
dtm.pre <- DocumentTermMatrix(corpus.prep)

# ステミングありの DTM（トピック発見用）
corpus <- tm_map(corpus.prep, removeWords, stopwords("english"))
corpus <- tm_map(corpus, stemDocument)
dtm <- DocumentTermMatrix(corpus)
```

5.1.2 文書単語行列 (DTM)

DTM の作成

- 文書を「単語の出現頻度のベクトル」として表現する。

```
# 行列のサイズを確認
dim(dtm)
```

```
## [1]   85 4849
```

```
# 一部を表示
inspect(dtm[1:5, 1:5])
```

```
## <<DocumentTermMatrix (documents: 5, terms: 5)>>
## Non-/sparse entries: 0/25
## Sparsity           : 100%
## Maximal term length: 7
## Weighting          : term frequency (tf)
## Sample             :
##           Terms
## Docs       abandon abat abb abet abhorr
##   fp01.txt       0    0   0    0      0
##   fp02.txt       0    0   0    0      0
##   fp03.txt       0    0   0    0      0
##   fp04.txt       0    0   0    0      0
##   fp05.txt       0    0   0    0      0
```

5.1.3 トピックの発見

ワードクラウドによる可視化

```
library(wordcloud)
dtm.mat <- as.matrix(dtm)
# 第 12 論文の頻出単語
wordcloud(colnames(dtm.mat), dtm.mat[12, ], max.words = 20)
```

tf-idf による重要語の抽出

- **tf-idf**: 特定の文書で頻出するが、他の文書ではあまり使われない単語を重視する指標。

```
dtm.tfidf <- weightTfIdf(dtm)
dtm.tfidf.mat <- as.matrix(dtm.tfidf)
# 第 12 論文の重要語トップ 10
head(sort(dtm.tfidf.mat[12, ], decreasing = TRUE), n = 10)
```

```
##     revenu contraband     patrol      excis      coast      trade          per
## 0.01905877 0.01886965 0.01886965 0.01876560 0.01592559 0.01473504 0.01420342
##        tax       cent     gallon
## 0.01295466 0.01257977 0.01257977
```

5.1.4 著者推定の予測

著者の特徴語

- ハミルトンとマディソンで使い方が異なる「機能語」（upon, there 等）に注目する。

```
# ステミング前の行列を使用
dtm.pre.mat <- as.matrix(dtm.pre)
# 出現頻度（1000 語あたり）に変換
tfm <- dtm.pre.mat / rowSums(dtm.pre.mat) * 1000
words <- c("although", "always", "commonly", "consequently",
           "considerable", "enough", "there", "upon", "while", "whilst")
tfm <- tfm[, words]

# 既知の著者のラベル付け (Hamilton=1, Madison=-1)
hamilton <- c(1, 6:9, 11:13, 15:17, 21:36, 59:61, 65:85)
madison <- c(10, 14, 37:48, 58)
```

回帰による著者予測

```
author <- rep(NA, nrow(tfm))
author[hamilton] <- 1; author[madison] <- -1
author.data <- data.frame(author = author[c(hamilton, madison)],
                          tfm[c(hamilton, madison), ])

# 4 つの単語を用いた線形回帰モデル
hm.fit <- lm(author ~ upon + there + consequently + whilst,
             data = author.data)
# 著者不明の論文 (disputed) の予測
disputed <- c(49, 50:57, 62, 63)
pred <- predict(hm.fit, newdata = as.data.frame(tfm[disputed, ]))
pred # 全て負の値であればマディソンの可能性が高い
```

```
##    fp49.txt    fp50.txt    fp51.txt    fp52.txt    fp53.txt    fp54.txt
## -0.99831799 -0.06759254 -1.53243206 -0.26288400 -0.54584900 -
0.56566555
##    fp55.txt    fp56.txt    fp57.txt    fp62.txt    fp63.txt
##  0.04376632 -0.57115610 -1.22289415 -1.00675456 -0.21939646
```

5.1.5 交差妥当性 (Cross-Validation)

LOOCV の実行

- 1 つのデータを除いてモデルを学習し、除いた 1 つを予測することを繰り返して、モデルの汎化性能を評価する。

```
n <- nrow(author.data); hm.classify <- rep(NA, n)
for (i in 1:n) {
    sub.fit <- lm(author ~ upon + there + consequently + whilst,
                  data = author.data[-i, ])
    hm.classify[i] <- predict(sub.fit, newdata = author.data[i, ])
}
# 正答率
mean(hm.classify[author.data$author == 1] > 0) # Hamilton
```

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

```
mean(hm.classify[author.data$author == -1] < 0) # Madison
```

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

5.1.6 まとめ

まとめ

- **テキストマイニング**: 構造化されていない文章を数値化し、分析可能にする。
- **前処理**: 小文字化、ストップワード削除、ステミングなどが重要。
- **DTM / tf-idf**: 単語の重要度を重み付けする。
- **著者推定**: 単語の使用パターンの違いを利用して、未知の文書の執筆者を統計的に予測できる。
- **交差妥当性**: モデルの信頼性を客観的に評価する。