R

quanteda を使った例

```
install.packages("quanteda", dependencies = T)
library(quanteda)
```

サンプル：学習者コーパス NICEST 1.0 の学習者データ 78 ファイル

- ・フォルダー内に、テキストデータのためのテキストファイルが複数入っている
- ・作業ディレクトリーをそのフォルダーに設定してある状態で

readtext パッケージの を使ってテキストを読み込む

```
install.packages("readtext", dependencies = T)
library(readtext)
```

コーパスデータ化する

- ・文書変数の追加修正 docvars()
- ・文書属性の追加修正 metadoc() こちらは分析対象とならない備考欄的信息
- ・reference

<https://github.com/koheiw/workshop-IJTA/blob/master/documents/corpus.md>

https://quanteda.io/articles/pkgdown/examples/quickstart_ja.html

コーパスデータから条件で絞り込んで、サブ・セットを作成

- ・絞り込みの例
 - ・見出し > 数字
 - ・見出し == " 文字列 "

コーパスデータの概要を見る

```
> summary(nicestJ1)
```

Corpus consisting of 78 documents:

	Text	Types	Tokens	Sentences
JAN0001_P1B.txt	116	214	12	
JAN0001_P2B.txt	138	268	17	
JAN0001_P3B.txt	97	169	11	
JAN0001_P4B.txt	68	99	8	
JAN0001_P5B.txt	120	262	16	
JAN0001_P6B.txt	114	224	13	
JAN0001_P7B.txt	121	268	18	
JAN0001_P8B.txt	71	108	8	
JAN0002_P1A.txt	98	170	15	

文書変数（属性情報）を追加する

中身を見る

- ・このままだと、すべて出力される
- ・特定のデータを見る場合は、何番目のデータを見るかを [番号] で指定する。

KWIC 検索する

トークン化する

(実際は、これをしなくても文書行列を直接作成すればよい)

- ・結果は、各文書を要素とする リスト形式で保存される
- ・句読点を除くオプション
- ・数字を除くオプション
- ・stopwords を除くオプション
- ・token 化したものをすべて小文字にする `tokens_tolower()`
- ・token 化したものを stem 化する `tokens_wordstem()`

文書行列の作成 document-feature matrix (DFM)

- ・ tm より新しいので、こっちのほうがいい (Welbers et al 2017)
- ・オプション
 - ・ステム化
 - ・句読点の削除
- ・対象は、トークン化した オブジェクト だが、
 - ・トークン化しなくても自動でトークン化してくれる
- ・機能語などの頻出単語 (stopwords) を除くオプション `remove = stopwords("english")`
- ・文書をグループ化することもできる `groups = " 見出し "`

文書行列の閲覧

大文字 V

文書行列の概要を見る

単語頻度一覧表

- ・デフォルトは、上位 10 語
- ・オプションで数字をつけるとそこまで

ワードクラウドの作成

いくつかの単語をまとめてグループにし、そのグループに該当するものの頻度を調べる。

- ・例えば、接続語句の リスト のグループを複数作っておいて、それぞれのグループに属する接続語句の頻度をかぞえる。
- ・グループを「辞書」と呼ぶ。コマンド名は `dictionary()`
- ・ `View(connectives.dfm)` で見る。

(以下、データフレーム名が違うので、注意。そのままコピペでは動きません。)

文書行列を並べ替えて語彙頻度一覧作成

```
dfm_sort( 文書行列 ) [ 何行目, 頻度順に単語を何位まで ]
```

コーパスの結合 + するだけ

```
nigest20.corpus <- nigestJAN.corpus + nigestNTV.corpus
```

文書に特徴的な単語を選び出す keyness

文書の類似度 textstat_simil(文書行列)

- ・ 類似度を測る方法には、オプションが何種類もある。

文書間の距離を測る

語彙的分散を観察する

- ・ kwic() との組み合わせで使う
- ・ 分布の図を絶対値にするオプション scale = "absolute"

ランダムサンプリング

- ・ Reference

https://quanteda.io/articles/pkgdown/quickstart_ja.html

http://i.amcat.nl/lda/1_textanalysis.html

<https://quanteda.io/articles/pkgdown/examples/plotting.html>