R

keyness

・複数の文書からなるコーパスがあったとして、
・その中の特定の文書が、ほかの残りと比べて、際立って違う言葉遣いをしていることを探る。
・二種類の文書に対して行うことで、二つの文書の相違を観察できる。
　・target と reference group

・符号付き２ｘ２の関連性スコア (association score)

quanteda::

・target で、注目したい文書を指定する。
　・数字を入れれば、文書行列中の要素番号
　・ほかには、文書行列に、各文書の属性情報をつけて置き、その属性でグループを指定する。
　　・例えば、JAN と NTV という属性をつけて、JAN をターゲットに指定して、残りの NTV を reference として比較する。
・measure = オプションで統計値を選べる。（signed というのは、プラスマイナスの符号を使うから）
　・ chi2 は、χ二乗
　・ exact は、Fisher's exact test
　・ lr は、likelihood ratio（G2）

keyness の分析例：学習者コーパス NICEST のサンプルデータを使って

■必要なパッケージ

■データの構成

・NICEST sample files というディレクトリー内に二つのサブディレクトリー
　・NTV sample 10
　・JPN sample 10
・それぞれサンプルエッセイの本文部分のみ、10 個ずつ入れてある。

■データの読み込み

・作業ディレクトリーを「NICEST sample files」に設定
　・そこをもとに、二つのサブディレクトリー内のファイを読み込む
・まず母語話者データ

■コーパスデータ化する

・コーパスの概要を確認

■言語情報を属性として追加する

```
docvars( コーパスデータ , 文書属性 ) <- 値
summary( コーパスデータ ) で内容を確認
```

・母語話者データは L1、学習者データは L2

■同様に、L2 の学習者データのテキストファイルを読み込む。

■コーパスデータ化する

```
nicestJPN.corpus <- corpus(nicestJPN.tmp)
str(nicestJPN.corpus)

## List of 4
##  $ documents:'data.frame':   10 obs. of  1 variable:
##   ..$ texts: chr [1:10] "Some people say that specialized knowledge is not important for human, however, who make todays life such a con"| truncated "You may think that young people are active and free, on the other hand olders are less active and they have muc"| truncated "Compared with past, young people nowadays do not give enough time to helping their communities.\nI guess there "| truncated "You may have experiences like this, feel nice at products in some advertisement but you buy and see it, you dis"| truncated ...
##  $ metadata :List of 2
##   ..$ source : chr "C:/Users/sugiura/Documents/* on x86-64 by sugiura"
##   ..$ created: chr "Sun Dec 01 12:04:54 2019"
##  $ settings :List of 12
##   ..$ stopwords          : NULL
##   ..$ collocations       : NULL
##   ..$ dictionary         : NULL
##   ..$ valuetype          : chr "glob"
##   ..$ stem               : logi FALSE
##   ..$ delimiter_word     : chr " "
##   ..$ delimiter_sentence : chr ".!?"
##   ..$ delimiter_paragraph: chr "\n\n"
##   ..$ clean_tolower      : logi TRUE
##   ..$ clean_remove_digits: logi TRUE
##   ..$ clean_remove_punct : logi TRUE
##   ..$ units              : chr "documents"
##   ..- attr(*, "class")= chr [1:2] "settings" "list"
##  $ tokens   : NULL

summary(nicestJPN.corpus)

## Corpus consisting of 10 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
##  JAN0002_P2A.txt   117    216       19
##
## Source: C:/Users/sugiura/Documents/* on x86-64 by sugiura
## Created: Sun Dec 01 12:04:54 2019
## Notes:
}}
```

■言語情報を属性として追加する

■二種類のコーパスデータの統合　統合コーパス AB <- 統合前コーパス A 統合前コーパス B

■文書行列を作成。句読点の削除。（小文字化は自動）

■言語属性をもとに、グループ分け

```
docvars( コーパスデータ , 文書属性 ) == 文書属性該当情報
```

・keyness の考え方として、lang 属性が L2 のものを「target」、その他のもの、つまり L1 を「reference」と位置付ける

■特徴語 keyness の算出

```
textstat_keyness( 文書行列 , ターゲット情報 )
```

feature（特徴語）	χ二乗値	p 値	ターゲットでの頻度	レファレンスでの頻度

・結果は、符号付（プラスマイナス）χ二乗値の高い順に出力される
　・で、ターゲットの方での「上位」20 語の観察
　・で、レファレンスの方での「上位」20 語の観察

■プロットする