

R

R.package

quanteda

quanteda を使った連語 (multi-word expressions) 処理

連語 (n-gram 表現) の一覧表

■ 連語の頻度と強さの一覧

- ・ 連語の長さ (グラム数) の指定オプション: `size = 数字`
- ・ 頻度の最低回数の指定オプション: `min_count = 回数`

■ 連語を token 単位にしたリストの作成

- ・ 単語の token 化と同じプログラム () のオプションの指定でできる。

```
tokens(コーパスデータ)
tokens(コーパスデータ, ngrams = グラム数の指定)
```

- ・ 実際には、数字や句読点を削除するオプションを使う。

```
remove_numbers=T, remove_punct=T
```

- ・ グラム数に範囲を指定することもできる `ngrams = 2:4`

```
nicestJAN2to4gram <- tokens(nicestJAN.corpus, ngrams= 2:4, remove_punct=T)
```

■ 途中の単語を飛ばした連語リスト (skip-gram) の作成

- ・ オプションで `skip = 数字` で飛ばす単語数を指定する。
 - ・ 複数の範囲の指定可
 - ・ `skip = 0:1` (スキップなし、つまり、普通の ngram および、一つ飛ばしの skip-gram)
 - ・ `skip = 1:2` (単語一つ、および二つ飛ばした skip-gram)
- ・ 間に「空欄」のある n-gram 表現のリストの作成が可能
- ・ ただし、どこに「空欄」があるかという情報が結果に残されない点が欠点。

連語の検索

- ・ 具体的な連語をベクトルにまとめておく

```
multiword <- c("in addition", "on the other hand", "as a result")
```

- ・ 「フレーズ」という単位で扱われるように指定する

```
phrase(multiword)
```

- 处理例：kwic 检索
-

- reference <https://quanteda.io/articles/pkgdown/examples/phrase.html>