R
R.package

ngram

・文字列を n-gram に切り分ける
・https://cran.r-project.org/web/packages/ngram/index.html

■注

・一文一行になってないと、文をまたいで n-gram を生成してしまう。

■データの読み込み

・テキストファイルの読み込み

```
readLines()
```

・特定のフォルダー内の、特定の拡張子のファイルを一度に読み込む

```
multiread()
multiread( パス , extention=" 拡張子 ")
```

■大文字小文字・句読点の前処理 preprocess()

```
preprocess( データ , case="lower", remove.punct=T)
```

■ tm パッケージとの関連

・tm パッケージでは、データは、「Corpus object」という特殊なフォーマットのデータフレームになっている
・ngram のプログラムはそのデータを直接扱えない
・以下のような命令で読み込む

```
concatenate(lapply( データ , "[", 1) )
```

■ n-gram の作成 ngram(データ , n= グラム数)

・作成されたものは、独自の「ngram オブジェクト」として保存される。

■作成された n-gram オブジェクト全体の出力 print(オブジェクト , output="full")

・output="truncated" とすると、全部は出ない。

■一覧表形式で出力 get.phrasetable(オブジェクト)

・入れ子式にすると便利

get.phrasetable(ngram(テキストデータ , n= グラム数))

■ n-gram **表現を個別にすべて出力** get.ngrams(**オブジェクト）**

・入れ子式にすると便利

get.ngrams(ngram(テキストデータ , n= グラム数))

■疑似的文字列の生成 rcorpus(**単語数** , alphabet=letters[**何文字から** : **何文字まで使って**], maxwordlen= **最大単語長**)

■以下古いメモ