R.package

corpus

・https://cran.r-project.org/web/packages/corpus/index.html

■ raw data に整形しておく

■ corpus_frame() で「corpus data frame object」形式のデータとして保存

title	text

■ text_tekens() でトークン化

■ text_filter()

・オプションを指定することで各種整形ができる

■ text_ntoken()

・token の数

■ text_ntype()

type の数

■ text_nsentence()

文の数

■ text_stats()

・上三つをまとめて行う

■ term_stats()

・各用語が、コーパス・データ中のいくつのサブコーパスに含まれるか

```
term_stats(data)
```

・オプションで ngram も同様に

```
term_stats(data, ngrams = 5)
```

・特定の語を含む ngram も同様に
　・グラム数の範囲指定可能
　・何語目に含むか指定可能

```
term_stats(data, ngrams = 2:3, types = TRUE,
           subset = type1 == "dorothy" & !type2 %in% stopwords_en)
```

■ text_locate() で　KWIC 検索

・stemmer オプションでステミング可能
・複数の keyword の指定可能

■ text_sample() で同様にランダムに検索可能