TOKENIZATION

Tokenized the document collection (Use any programming language)

1. Tokenization on the Reuters data set

https://archive.ics.uci.edu/ml/machine-learning-databases/reuters21578-mld/

(It contains 22 separate files, use all files)

2. Try to find the frequency of each tokens.

3. symbols, single characters may be discarded.

STOP WORD REMOVAL

Perform stop word removal on the **Reuters data set**! For stop word removal, please use the list of stop words given on

http://www.textfixer.com/resources/common-english-words.txt.

Note: 1. You can write code on any language.

2. This process can be applied before tokenization of after creating tokens.

STEMMING

Perform stemming on the **Reuters data set**, please use the implementation of the following stemmers (Any programming Language)

Make sure that you end up list of terms that does not contain any duplicate terms.

- https://tartarus.org/martin/PorterStemmer/
- https://snowballstem.org/demo.html
- https://www.scientificpsychic.com/paice/paice.html