IR Assignment #3

There are several general steps of pre-processing you need to perform on the sample document collection:

1. Tokenization: tokenize the each document contents into tokens.
2. Normalization: normalize the tokens from step 1, by removing individual punctuation marks and converting tokens into lower cases.
3. Stopword removal of common English stopwords.
4. Stemming: stem the tokens back to their root form.:

Find the size of the dictionary (Vocabulary size which is number of unique words) before and after the processing.

You can use Java libraries (CoreNLP or snowball or Lucene) or NLTK for python.

Dataset :

1. TIME.ALL