BIA-660-C. Haodong Zhao

Q2: Define a class to analyze a document. What kind of words usually have high frequency? Write your analysis.

A: From the test text, the top 5 frequent words are ‘it’, ‘to’, ‘the’, ‘a’ ,’and’. Therefore, in this text, the high frequent words’ type are pronoun, article, conjunction and preposition. All of these words are commonly used words in our life, so I thinks this kind word are not very valuable for text analysis, should be stop words in text mining.

Q3: (Bonus) Create Bigrams from a document. Are you able to find good phrases from the top N bigrams? Write down your analysis in a document.

A: From the test text, I find the top 5 frequent bigrams ‘it and’, ‘out of’, ‘a watch’, ‘that she’, ‘to her’. We can find that there is only 1 real phrases ‘out of’, I think if we have enough stop words, this analysis should be able to find good phrases.