ASSIGNMENT-01

README

How to Run the Code?

1. Extract the files of 20_newsgroups.
2. Must have NLTK.
3. Give file path of folder 20_newsgroup dataset in a variable directory_name.
4. Then enter the folder name and file name you want to search.

Assumptions:

1. Removed stopwords for counting vowels and consonant. For example for vowel , 'a' is itself a vowel and 'of' is stopword.
2. All numbers are converted into words, treating 100 and hundred as same entity.

Preprocessing Steps:

1. Word Tokenization:
   - Cut character sequence into word tokens.

2. Sentence Tokenization:
   - On the basis of delimiter break paragraph into sentences.
   - Sentences are not extracted from the header assuming that it contains attributes of file like name, date, path etc. File with header is used in 3 questions which includes extracting list of emails, time and abbrevations.

3. Normalization:
   - Converting all text to the same case (upper or lower) is normalization.
   - Here, without header text is converted into lower case.

4. StopWords:
   - Omit out common words/stop words such as the, a, to, of, etc. for computing total number of *vowels* and *consonants*.

5. Num2Words:
   - Converting number to words.

6. Removal of Puntuations:
   - for counting total number of *words* in a corpus.

7. Removal of Header:
   - for all parts other than extraction time, emails and abbrevations.

Methodology Explained:

1. Loaded documents from dataset i.e 20_newsgroup.
2. Enter the folder name and file name.
3. Then after entering file name and folder name, it outputs:

- Total number of words,
- Total number of sentences,
- Total number of words starting with vowels,
- and Total number of words startig with consonants.

4. Total number of words are computed using function ***countWords*** in which we sent tokenize words, and it prints the length of the word.
5. Total number of sentences are computed using function ***countSentences*** in which we sent tokenize sentences and it prints the length of the sentence.
6. Total number of words starting with vowels using function ***isVowel*** and ***countStartVowel*** while total number of words starting with consonants are found using function ***isConsonant*** and ***countStartConsonant***.
7. For listing total number of emails present inside the file ***countEmails*** function is used.
8. Using ***startingWithWord*** function will print the sentences and number of sentences starting with a given word in an input file
9. Similarly, we've function for printing the sentences ending with a given word by using function ***endingWithWord*** and printing the sentences containing particular word by using ***specificWordInSentenceFile***
10. ***specificWordFile*** will print count of specific word present in a file.
11. To find if any questions present in a file ***containsQuestion*** function is used.
12. ***extractTime*** function is used to list the minutes and seconds mentioned in the date present in the file.
13. ***findAbbreviations*** function is used to list all abbrevations present in a file.
14. Header is removed using ***removeHeader*** function.
15. number2words function is used to convert numbers into words.

Note: *If file or folder not found it will throw an message "Exception Occured". Make sure to enter the correct file name and folder name.*
Functions and their inputs:

We have two types of text from a file where one is with metadata and other is without metadata, in a variable name ***dataWithMetada*** or ***data***.

Function Name	**Inputs**
removeHeader	dataWithMetadata
number2words	User entered word
countWords	Tokenize words
countSentences	Tokenize sentences
isVowel	character
countStartVowel	Tokenize words
isConsonant	character
countStartConsonant	Tokenize words
countEmails	dataWithMetada file
startingWithWord	sentences, user entered word
endingWithWord	sentences, user entered word
specificWordFile	data,user entered word
specificWordInSentenceFile	Sentences,user entered word

containsQuestion	data
extractTime	dataWithMetadata
findAbbreviations	dataWithMetadata

References:

- *https://www.nltk.org/book/ch01.html*
- *https://github.com/rain1024/slp2-pdf/blob/master/chapter-wise-pdf/%5B02%5D%20Regular%20Expressions%20and%20Automata.pdf*