NLP Assignment - 1

Akshala Bhatnagar
2018012

1. The text file string has initially been split using /n/n. For each element in this list, [\n, -, <, >, *, #, +, =, ^, |] have been replaced with a single space. Additional spaces from the start and end of the elements are removed. Then each of the elements is broken into sentences using sen_tokenize, which breaks a string into sentences on the basis of characters like [. , ! ?]. From these sentences multiple spaces are replaced with a single space. A RegexpTokenizer is used with ‘\w+’ to split sentences to words. ‘/w+’ matches any word character equal to [a-zA-Z0-9_]. If a word is numeric it is not added in the list. Finally the length of sentence and word lists are printed. The above methods are used to classify a sentence and a word.
2. The same method is used to find words and sentences as described in the first question. After words are found a regex '^[aeiouAEIOU]' is used to find words starting with vowels and the rest start with consonants. The length of vowels and consonants list is printed.
3. The same method is used to find sentences as described in the first question. The regex '\S+@\S+' is used to find all the words

that contain @, which means that it is an email. Finally the list of emails is printed.

4. The same method is used to find sentences as described in the first question. The target word is converted to lowercase. The regex '^' + word is used to find sentences starting with the given word. The sentences are also converted to lowercase. If the target word is a number then for that condition in the lowercase sentence numbers are converted to words using num2vec and then the regex matching is done. Finally the count of the sentences starting with the target word along with the sentence itself are printed.
5. The same method is used to find sentences as described in the first question. The target word is converted to lowercase. The regex, word + '[\.!\?*]' + '$' is used to find sentences ending with the given word. Characters [. ! ?] are considered as end of sentence tokens. The sentences are also converted to lowercase. If the target word is a number then for that condition in the lowercase sentence numbers are converted to words using num2vec and then the regex matching is done. Finally the count of the sentences ending with the target word along with the sentence itself are printed.
6. The same method is used to find sentences as described in the first question. The target word is converted to lowercase. The sentences are also converted to lowercase. Python find function is used to find the word in the string. If the target word is a number

then for that condition in the lowercase sentence numbers are converted to words using num2vec and then the regex matching is done. Finally the count of the sentences ending with the target word along with the sentence itself are printed.

7. The same method is used to find sentences as described in the first question. The regex '[\?+]' + '$' is used to find questions. Sentences ending with one or more question marks are considered as a sentence. If by chance \n is present then the string is split and the last element of the list is taken as the question. Finally the questions are printed.
8. The same method is used to find sentences as described in the first question. The sentence containing the word Date is found. This sentence is split on the basis of spaces. The index of the word GMT is found in this. The word before GMT would have the time. Time is in the format HH:MM:SS, so the time word is split using ':' . The first index has the minutes and the second index has the seconds.
9. The same method is used to find words and sentences as described in the first question. The number of uppercase characters are found in the words. If the number is more than three it is considered as an abbreviation.