NLP Assignment 1

My implementation details and assumptions

Q1.) For finding the no. of sentences and words in the file, I simply used the sent_tokenize() and word_tokenize functions respectively.

Q2.) For counting the no. of words starting with consonants and words starting with vowels , I first stored all the words in a list and then checked the first character of each the words for consonant or vowel also the character can be in lower case as well as in upper case.

Q3.) To list all the email ids I used the regular expression \S+@\S+ which means that there need to be one or more non-whitespace characters before @ and after @ and then used the function re.findall() to find the email ids in the file.

Q4.). Here first I extracted all sentences from the file using sent_tokenize and then for each sentence used word_tokenize and stored all the words of the sentence in a list and compared the first word of the sentence i.e words[0] in the list containing words of that sentence with the word to searched .

Here I have assumed that the word entered by the user to be searched is case sensitive.

Q5.) Here first I extracted all sentences from the file using sent_tokenize and then for each sentence used word_tokenize and stored all the words of the sentence in a list and compared the last word of the sentence i.e words[-2] in the list containing words of that sentence with the word to be searched . Here words[-1] contains the punctuation mark at the end of the sentence.

Here my assumption is that the word entered by the user to be searched is case sensitive . Also I have assumed that the sentences are ending with a word and then immediately followed by a punctuation mark which is treated by sent_tokenize() as sentence terminator.

Q.6.) Here I first extracted all the words in the file into a list and then started counting the occurrences of the word input by the user.

Here again I have assumed the word entered by the user to case sensitive.

INPUT FORMAT :

When giving file name as input to the program it has to be specified as:

alt.atheism\\<file_name> or comp.graphics\\< file_name>

depending upon the parent folder of the file . I have hardcoded the folder path before alt.atheism or comp.graphics according to the folder location on my system.