Natural Language Processing

Worksheet: Counting Words

In this worksheet we will use Lists, Sets and Frequency distribution functions to count words and generate frequency distribution for texts of Arabic and English. We will design a python program that reads the Qur'an text files in Arabic (*quran-simple.txt*) and in English (*en.pickthall.txt*). Then, the program will do the following:

Part 1: Processing Arabic text.

1- Reads the Qur'an text file (*quran-simple.txt*) and store the Quran text in a variable. (How many characters in the Qur'an text?)

2- Use the nltk.word_tokenize(text) to tokenize the Qur'an text and store the tokenized text in a list. (how many tokens (words) in the Qur'an according to the given Qur'an text file?

3- Define a new variable of type set and store the Qur'an list of words in it? (how many different words in the Qur'an?)

4- Design a function that computes the lexical diversity of the Qur'an's vocabulary? (What is the lexical diversity obtained?)

5- Design a function that prints on a file the frequency distribution of the Qur'an words? (what is the frequency of the top 10 words?)

Part 2: processing English text

6- Repeat the tasks (1-5) above to process the English translation of the Qur'an (*en.pickthall.txt*) use the same functions that were used to process Arabic text to handle the English text.

7- Compare between the results obtained by task 1 and task 2.

Part 3: Processing non-vowelized Arabic text

8- Design a function that received a vowelized Arabic word and returns its equivalent non-vowelized Arabic word form. The function iterates on the words characters, if the character is a short vowel or diacritic then it is removed from the word.

Short vowels and diacritics in Arabic and there Unicode characters are:

ً : \u064B (تنوين الفتح)	ٌ : \u064C (تنوين الضم)	ٍ : \u064D (تنوين الكسر)
َ : \u064E (الفتحة)	ُ : \u064F (الضمة)	ِ : \u0650 (الكسرة)
ّ: \u0651 (الشَّدّة)	ْ : \u0652 (السُّكون)	

9- use the function that removes short vowels and diacritics from Arabic words generate a list of the non-vowelized forms for all words in the Qur'an list. Then, repeat entries (1-5) from task 1. Compare the results obtained with results obtained from task 1 before.