Natural Language Processing

Worksheet in processing Arabic data (Qalqalah)

In this worksheet we will use regular expressions to extract information from Arabic text. We will design a python program that reads the Qur'an text file (*quran-simple.txt*) to extract all words from the Qur'an that contain the phenomenon of Qalqalah. The Qalqalah is defined in (https://mukhtas.wordpress.com/2012/02/12/qalqalah/) as:

Qalqalah is a method of pronouncing certain letters that have a sukoon on them (or when one stops at these letters). They require the tone to be strong and produce an 'echo'-like sound. In general, the Qalqalahs can be divided in to 2 parts – ***Qalqalah Sughra & Qalqalah Kubra.***

The letters of Qalqalah are:د ج ب ط ق

Qalqalah Kubrah:

The 'Strongest' version of Qalqalah is when one stops at one of the letters of Qalqalah whilst, the letter has a Shadda on it. For instance: (هُنَالِكَ الْوَلَايَةُ لِلَّهِ الْحَقِّ)

Another version of Qalqalah Kubra is when one stops at the letter of Qalqalah that is at the end of a word (without a shadda). For instance: (مِّنَ اللَّهِ ذِي الْمَعَارِجِ).

Qalqah Sughrah:

In this case, the silent letter of Qalqalah appears in the middle of a word. This will create a minor echo as opposed to Qalqalah Kubra: For instance: (وَيَرْزُقْهُ مِنْ حَيْثُ لَا يَحْتَسِبُ).

Write a python program that do the following:

1- Reads the Qur'an text file () 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 Quran 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- Design a regular expression (pattern) that captures the Quranic words with Qalqalah sughra, and store the results in a new list. (How many words in the Qur'an that contains Qalqalah sughra? Use the Unicode table attached with worksheet.

4- Print the first 50 words from the Qalqalah Sughra list.

5- Print the different words (types) of that contains Qalqalah and captured by the regular expression in text file. How many words are stored in the file?