**Assignment # 1**

**Consider the following raw text and perform the following tasks.**

Muhammad Ali Jinnah was a barrister, politician and the founder of Pakistan. Mr. Jinnah served as the leader of the All-India Muslim League from 1913 until the inception of Pakistan on 14 August 1947, and then as the Dominion of Pakistan's first governor-general until his death. In Pakistan, he is recognized as Quaid-i-Azam!, which means "Great Leader" and Baba-i-Qaum which means "Father of the Nation". Jinnah was trained as a barrister at Lincoln's Inn in London, England. By 1940, Jinnah had come to believe that the Muslims of the subcontinent should have their own state to avoid the possible marginalised status they may gain in an independent Hindu–Muslim state. In that year, the Muslim League, led by Jinnah, passed the Lahore Resolution, demanding a separate nation for British Indian Muslims. Naïve people do not know what he has done for the country

**Submission should be in the form ONE python Jupyter notebook.**

**Task # 1:** Copy the above text into a notepad file named QUAID.TXT and save the fil.

**Task # 2:** Read the file and tokenize the text using the built-in python function.

**Task # 3:** Read the file and tokenize the text in the file using NLTK.

**Task # 4:** Read the file and tokenize the text in the file using spaCy.

**Task # 5:** Read the file and tokenize the text in the file using Gensim.

Hint for Task 2 - 5:

<https://towardsdatascience.com/5-simple-ways-to-tokenize-text-in-python-92c6804edfc4>

**Task # 6:** Read the file and tokenize the text in the file using [Penn] Treebank tokenizer.

Hint 1: <https://rdrr.io/cran/tokenizers/man/ptb-tokenizer.html>

Hint 2: <https://search.r-project.org/CRAN/refmans/tokenizers/html/ptb-tokenizer.html>

**Task # 7:** Read the file and tokenize the text in such a way that each word is written in a separate line.

**Task # 8:** Convert the text into lower case.

**Task # 9:** Use Porter stemmer to perform stemming on each word

**Task # 10:** Use Snowball stemmer to perform stemming on each word

**Task # 11:** Remove diacritics from the text.

**Task # 12:** Write a program for spelling correction task using a look-up based approach.

**Task # 13:** Write a program for calculating the accuracy and error rate of your look-up based approach.