Lab 4: Data preprocessing

Deadline: 24/03/2022

The current Quran dataset as-is can make doing analytics and machine learning difficult because it may have words or characters that cannot be used with the current NLP techniques. For example, tashkeel and stop words and other non-alphabetical characters.

Your task is to preprocess the original dataset and produce additional datasets according to the following requirements:

1. Ayat without extra symbols e.g. (۞) only
2. Ayat with words without tashkeel
3. Ayat with words without tashkeel and remove stop words
4. Ayat with words having tashkeel but without stop words in the dataset

At the end, we will have 5 datasets, the original and the above datasets.

Submission:

Send me a link to your Github repository after committing the IPython notebook.

Pointers:

ML Feature Extraction:
https://spark.apache.org/docs/latest/ml-features.html

For a list of stop words, use the following repository. Use this to remove stop words from the original dataset.
https://github.com/mohataher/arabic-stop-words