Out date: 02/04/2022
Due date: 27/05/2022

The given 10 txt files contain various Arabic tweets. Each file consists of two columns: tweet ID and tweet string. Most tweets contain unwanted emoji or hashtags which should be excluded first. For some tasks you can consider some files as training data and the remaining files as testing data.

Task 1 [8 points]: Select a training data to build a bigram language model that can help you in text sequence generation for sequence data. You should implement a **python class with multiple methods** to do the following jobs:

1. The class name is LanguageModel.
2. A constructor to get the text file name and open that file.
3. A method called Clean to process the text (Tokenization, Lemmatization)! Don't remove stop words.
4. A method called LMBigram to build 2-D Language Model Matrix with Laplace smoothing using NLTK. This method saves generated matrix in an instance variable.
5. A method called Run that takes a part of sentence from user and returns the expected next word using the constructed matrix from the previous step.

Task 2 [6 points]: Using NLTK and/or Camel libraries to build "Named Entity Extraction" Model that can be used to extract name of person, country, organization, events…etc. The model can be trained on training data and tested on different data.

Task 3 [8 points]: Using google BERT word embedding model to build sentiment analysis model.

Task 4 [8 Points]: prepare presentation for one of the following NLP tasks:

- Fake news detection.
- Machine Translation (Google Translator).
- Topic classification.
- Transformer and Word embedding.
- Tashkeel in Arabic language
- Question similarity

What to submit:

1. A final report that describes your work for each task including assumptions, model architecture.
2. three python files for tasks 1, 2 and 3 as Jupyter Notebooks.
3. Power Point Presentation.