Cairo University
Faculty of Computers & Artificial Intelligence
Processing of Formal and Natural Languages
(2023/2024)

Assignment 2

Delivery Notes:

- This is a group assignment of 3 members (at most)
- All students should work and fully understand everything in the code.
- Due date is on 24/3/2024.

Assignment Details:

In this assignment, you are required to Generate Document and implement TFIDF on it

- Data: you should search for function to generate a document if it gets any phrase
  Generate more than one document in different fields.
- Processing on data:
  1. Cleaning data from each symbol or character doesn't contain to the data.
  2. Normalization: make all the data to lower case
  3. Tokenization: split the data to words
  4. Lemmatization or Stemming: return each word to origin.
  5. Stop words: remove stop words from the data.
- Unique words: - Get the unique words from the data.
- TFIDF: the output of the assignment gets feature vector for each document.
  1. Get TF for each word for all documents.
  2. Get IDF for each word.
  3. Multiply TF * IDF
  4. Get Normalized TFIDF

  You should use (sklearn) equations.

Bonus : if the team apply TFIDF from Scratch code and built in (sklearn) code with the same result