Cairo University
Faculty of Computers & Artificial Intelligence
Processing of Formal and Natural Languages
(2021/2022)

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 Apr 21st until 11:55 pm
- No late submission is allowed.
- Submission will be on blackboard
- No submission through e-mails.
- The submitted files should be named Ass2_firstStudentID_SecondStudentID_ SecondStudentID.ipynb
- **Do not send your code** to anyone**,** so that no other student would take your files andsubmit it under their names.
- **In case of Cheating, you will get a zero grade whether you give the code to someone or take the code from someone or from the Internet**
- A discussion will be held for some teams. As for the rest, **the Assignment will be graded without discussion using your submission on blackboard**. So, for a fair evaluation, make sure that your notebook **has a clear and visible output** and that your code **is clean and understandable**.

Assignment Details:

In this assignment, you are required to implement a **Google-like search engine** using a **word2vec pretrained model** using the following instructions:

- **Loading Data:** Download **20 different documents** in **5 different domains** using **Wikipedia API** in python. Each document has to be at least **one page (500 Words)**
- **Download Model:** Download a **pretrained word2vec model**. You are not restricted to a specific model. Feel free to download any model pretrained on any data as long as it's a word2vec model.
- **Training:** Use your pretrained model to create a word embedding for each word in each

document then **create a final embedding representation for each document** using whatever method you want (Ex: Average). Then **save the final representation** for each document in a file on your hard disk.

- **Testing:** Enable the user to **enter any sentence** in your search engine, generate its embedding then calculate the similarity between the sentence and all your documents using whatever **similarity measure** you want and the **document embeddings** you created. Then display your documents search results **sorted descending based on their similarity to the input sentence** (Exactly like google). Repeat this step 3 times with 3 different inputs that maps to **3 different domains**.

Note that each instruction is part of the grading criteria. So make sure to follow all the instructions so you don't miss any grades.