German University in Cairo
Media Engineering and Technology
Dr. Mohamed El Mahdy
Eng. Maged Shalaby

Natural Language Processing and Information Retrieval, Winter 2017
Programming Assignment1 (Individual)

Deadline: 06.10.2017 23:59

Your task in this assignment is to implement a bigram inverted index as discussed in Lecture 2.

There should be two steps:

a) Building the bigram inverted index.

b) Querying the index to return which documents contain a given bigram.

The following is an example sentence: "This man plays American Football", where the bigrams generated should be: (This, man), (man, plays), (plays, American), (American, Football).

Note that the order of the words matters when generating the bigrams (bigrams are consecutive pairs of words in a file).

You should use only the first 5,000 files from the dataset "single-wiki". You should have the folder "single-docs" in the same directory of your Jupyter notebook, so that you call `os.walk('single-docs')`.

You are free to use functions from the NLTK library, or implement it without NLTK.

After building the index, you should print the documents that contain the bigrams:

a) "American Football"

b) "Northern Hemisphere"

You should submit your code as a Jupyter notebook file on the assignment on the MET website, by Friday 06.10.2017 at 23:59. Do not submit the dataset; just submit the notebook file.

You should name your file as: Your ID - Your Lab Group - Your Name

Best of luck!