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

Natural Language Processing and Information Retrieval, Winter 2017
Programming Assignment3 (Individual)
Deadline: 14.12.2017 23:59

Your task in this assignment is to build a trigram language model that will generate random pieces of text given a start bigram.

1 Calculating Probabilities

In this part, you will be calculating the probabilities needed to generate the random sentences in the second part. The steps are as follows:

a) (Optional) Create an (ordered) list of words from the concatenation of the first 30,000 files in `simple-wiki` (use higher/lower number of files depending on your RAM). Filter stopwords and punctuation.

b) Generate the counts of bigrams and trigrams in the data.

c) Calculate the probabilities of trigrams from the counts of trigrams and bigrams.

2 Generating Random Sentences

In this part, you will be using the previously calculated probabilities to generate random sentences. The steps are as follows:

a) Initialize a new sentence with a start bigram.

b) Given the last two words (the last bigram) in the current sentence, search for a trigram that starts with this bigram and ends with a new word, that has a probability value > `p` (try different values of `p`, 0.1 and 0.05 are good starts).

c) Concatenate the last word of the found trigram with the current sentence.

d) Repeat b) and c) until a sentence of a given length is formed.

e) Print the resultant sentences after trying different start bigrams, `p` values and sentence lengths.

You can shuffle the list before running the algorithm as not to get the same result every time.

3 Bonus (1 Marks)

Add a sentence end token to all the sentences in the data, and change the stopping condition to be reaching the end token, rather than the length of the generated sentence.

4 Submission

You should submit your code as a Jupyter notebook file on the assignment on the MET website, by 14.12.2017 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!