Homework 3 - Text Representation and Retrieval

Due: 2359hrs on 1 May 2020

Instructions:

- The instruction plagiarism policy applies.
- You have to implement the assignment in Python.
- You can use libraries for reading and perform image operations. However, the core of the retrieval algorithm has to be implemented from scratch. In case of a doubt make sure you confirm with the TA's.
- For any doubts make a comment on google classroom or email TA with Subject like [MCA HW-3] Doubt.
- Submission Instructions:
    - All submissions must be inside a zip file named a3_{name}_{rollnumber}.zip containing report.pdf and a folder named src with all your scripts.
    - All the scripts must be uploaded in .py format. Make a separate script for each subquestion. Naming convention example question1_1.py for the first subquestion of question 1.
    - Write all the scores and your analysis of it in the report.pdf

---

Question 1 - Implement Word2Vec

Implement Word2Vec from scratch on the 'abc' corpus present in NLTK. The paper gives a good description of the algorithm. You are free to use any standard Deep Learning package like Keras/TF/PyTorch for constructing the neural network and training.

Visualize the generated word embeddings after every epoch. You can use tSNE for reducing the dimensions of embeddings. Refer this blog for details on visualizations.

Write a brief description of the algorithm and comment on changes observed in visualization during the training process.

Question 2 - Document Retrieval using Query expansion

In this question you will perform on a text document retrieval system on the Medline Collection dataset. Boilerplate code for performing retrieval and evaluation is provided in the Assignment payload in directory src/Problem_2. You are required to implement relevance feedback and

relevance feedback with query expansion in relevance_feedback.py file to demonstrate its effect on retrieval performance.

A set of instructions and sudo code for both algorithms is provided in file Problem_2_instructions.pdf. README.txt provides instructions on how to run the code. Please refer to these documents before your start.

Report the retrieval score for both algorithms over 3 iterations. Comment on the changes in performance, are they inline with what you expected, explain briefly?