CSCI381/CSCI780

Homework 2

Due 05/11/2020 3pm (by email). **Please attach your write-up report as a separate attachment.** Do not include it in the zip file with all the code.

1 WORD2VEC

In this assignment, you will use a training corpus news.crawl to train Word2Vec.
You should follow the instructions found at:
https://kavita-ganesan.com/gensim-word2vec-tutorial-starter-code/#.XqO4C2ZKj3h
Please use the following parameters:

1. size=150
2. window=5
3. min_count=2
4. iter=10

Questions:

1. Report similarity scores for the following pairs: (dirty, clean), (big, dirty),(big, large) , (big,small)
2. Report 5 most similar items and the scores to 'polite', 'orange'
3. Now change the parameters of your model, as follows: window=2, size=50. Answer the 2 questions above for this new model.

1.1 SUBMISSION

Please include all the required files in a tarball and email those to nlp.qc.cuny@gmail.com using subject line `CSCI381/CSCI780 Homework 1`:

1. The Python code along with a README file that has instructions on how to run it in order to obtain the answers to questions.

2. The writeup that includes the answers to the questions.

Your grade will be based on the *correctness* of your answers, the *clarity* and completeness of your responses, and the *quality* of the code that you submitted.
Please refer to the course webpage on late submission policy.