

## Homework 3

---

Due 05/12/2021 midnight

Due Wednesday, 05/12/2021 midnight (by email at [nlp.qc.cuny@gmail.com](mailto:nlp.qc.cuny@gmail.com)). **Please attach your write-up report as a separate attachment.** Do not include it in the zip file with all the code.

**Important:** Submissions that do not follow the submission guidelines will not receive full credit. If you are providing a link to the source files, please place the files on **Queens College Google Drive**. Do not use any other locations or services. Please include your report as a separate attachment in the email.

- Feel free to talk to other members of the class in doing the homework. You should, however, write down your solution yourself. Please try to keep the solution brief and clear.

### 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.

### **Submission**

Please include all the required code files in a tarball and email the tarball and the report to `nlp.qc.cuny@gmail.com` using subject line CSCI381/CSCI780 Homework 3:

- The tarball should include the Python code along with a README file that has instructions on how to run it in order to obtain the answers to questions in Part 2.
- The report that should be attached separately to the same email should include 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.