CS 2492-01: Unstructured Information Processing

HW 5 (Given Oct. 19, 2020; Due Oct. 26, 2020)

Your answers must be entered in Google Classroom by midnight of the day it is due. If the question requires a textual response, you can create a PDF and upload that. The PDF might be generated from MS-WORD, LaTeX, the image of a handwritten response, or using any other mechanism. Code must be uploaded and may require demonstration to the TA. Numbers in the parentheses indicate points allocated to the question.

You are to perform the comparison of sentiment classification using a Bag-of-Words approach and a Word Embedding approach. To obtain a Word2Vec model, you can use nltk. nltk makes it quite easy and the following allows you to obtain a Word2Vec model using the brown corpus.

from nltk.corpus import brown model = gensim.models.Word2Vec(brown.sents())

Should you require can also download the brown corpus separately from `https://www.kaggle.com/nltkdata/brown-corpus`. Once you have a trained model,

1. Find the 3 words which are closest to the word "University". Write out the closest words, the respective embedding, and the distance between the embedding of "University" and the closest 3 words. **(20 Points)**

2. Find the 3 words which are closest to the word "Nation". Write out the closest words, the respective embedding, and the distance between the embedding of "Nation" and the closest 3 words. **(20 Points)**

3. Assume that the following data is given to you as an example of a positive sentiment - "This is a great movie with lots of suspense" and another example is given to you as an example of a negative sentiment - "The movie was predictable and quite boring". Based on this and your trained Word2Vec model, is "There was no suspense at all" a positive or a negative sentiment. Show all work. **(30 Points)**