Homework #1

(CSE 582; modified 1/31/2023)

Task description: learn the code of Word2Vec and try to use it to train word embeddings on a text corpus

Duriation: 1/30/2023 ~ 2/15/2023

Word2Vec official code:
https://storage.googleapis.com/google-code-archive-source/v2/code.google.com/word2vec/source-archive.zip

What you should do:
1, download and study the Word2Vec code. Particularly, find the code that implements the CBOW algorithm, and write your comments for some important lines of code. Some questions you may keep in your mind while diving into the code

- How does CBOW compose context embeddings?
- How does it compute word probability given context?
- How does it implement negative sampling?
- Any other parameters apart from the word embeddings and context embeddings?
- What input format does Word2Vec require?

2, Train word embeddings using Word2Vec's command on the following data:
http://www.statmt.org/lm-benchmark/1-billion-word-language-modeling-benchmark-r13output.tar.gz

3, Once you get word embeddings, try to compute the top-10 most similar words for query words "cat", "she", "like"

4, upload the code with your comments, and the top-10 similar words to your github; submit the github URL by the deadline 11:59pm on 2/15/2023 (EST) to Canvas. Pls refer to TA, Shravya Chillamcherla (sjc6752@psu.edu), for more details on how to submit.