

Word2Vec Implementation with Large Corpus

Due: 11/09/2024 SATURDAY 11:59 PM

Objective:

In this assignment, you will implement Word2Vec using the Skip-gram architecture with negative sampling. You will train the model on a large corpus of 1.4 million words and visualize the results.

Dataset:

- Dataset includes 14 million words. You must use at least 10K sentences from this corpus. You don't have to use all the training data but more is better.

Instructions:

- Refer to the skeleton.txt. You must implement the methods in the Word2VecImplementation class.

Guidelines:

- Use the skeleton code provided in the skeleton.txt.
- You may change the parameters as you like: max_sentences, epochs, batch_size etc.
- Submit your code in a Jupyter notebook format (.ipynb).
- Submit the Jupyter notebook **with all cell outputs visible**. Ensure that the notebook is fully run from start to finish without any errors.