COMP 7970

Special Topics: Natural Language Processing

Instructor: Shubhra ("Santu") Karmaker

Assignment #2: Implementation of Word2Vec [100 points]

⚠ **Notice:** This assignment is due **Monday, September 15, 2021, at 11:59pm**.

Please submit your solutions via Canvas (https://auburn.instructure.com/). You should submit your assignment as a **typeset PDF**. Please do not include scanned or photographed equations as they are difficult for us to grade.

1. Implementation of Word2Vec [50 pts]

Word2vec is a two-layer neural net that processes text by "vectorizing" words. Its input is a text corpus, and its output is a set of vectors: feature vectors that represent words in that corpus. Now, let's implement a custom Word2vec model following the paper "**Efficient Estimation of Word Representations in Vector Space**". You may use any programming language you would like but use of any library for training word2vec is restricted (not allowed). Please note below points for your implementation:

a. Dataset: Uploaded on the Canvas
b. Before implementing Word2vec, remember to clean your text.
c. Consider below points
    i. Embedding size=300
    ii. Use Skip-gram model to train
    iii. Epochs=10

Train and save your model in this step.

2. Find Similar Words [20 pts]

Load the saved model from the previous step and find out top 10 similar words for

a. "Coffee"
b. "Pasta"
c. "Tuna"
d. "Cookies"

3. Word Analogies [30 pts]

Load Glove 300d vector file. Available at: https://nlp.stanford.edu/projects/glove/ and solve below analogies

i. Spain is to Sapnish as Germany is to ____
ii. Japan is to Tokyo as France is to ____
iii. Woman is to Man as Queen is to ____
iv. Australia is to Hotdog as Italy is to ____

Use cosine similarity between the word vectors to solve the analogies.

Dataset

Your dataset is updated on the canvas. This dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all ~500,000 reviews. Reviews include product and user information, ratings, and a plain text review. We have sampled random 300000 from the main dataset. For the embedding, consider the text review field of the dataset.