CSE 4392 Special Topic: Natural Language Processing

Homework 5 - Spring 2024

Due Date: Mar 4, 2024, 11:59 p.m. Central Time

In this lecture, we learned how to embed words into a feature space of specific dimensions to densely represent a word. The core idea is that words with similar semantics or grammar will be mapped to vectors that are close in distance. We will further reinforce word embedding through three exercises and apply it in practical applications.

Problem 1 - 30%

In this problem, you will work on deriving the gradient for the negative sampling objective function used in Skip-Gram Word2Vec. The negative sampling approach is a technique to efficiently train word embeddings by sampling negative examples. Derive the gradient of the negative sampling objective function with respect to the parameters ($\mathbf{u}_t$ and $\mathbf{v}_t$) used in Skip-Gram Word2Vec.

$$y = -log(\sigma(\mathbf{u}_t \cdot \mathbf{v}_c)) - \sum_{i=1}^{K} \mathbb{E}_{j \sim P(w)} log(\sigma(-\mathbf{u}_t \cdot \mathbf{v}_j))$$

Problem 2 - 40%

In this problem, you are required to utilize **pre-trained** word embeddings (Word2Vec, GloVe, and FastText) to perform a word similarity task using the WordSim-353 dataset. Evaluate the performance of each word embedding technique in capturing semantic similarities between words.

- Word2Vec: `https://code.google.com/archive/p/word2vec/`
- GloVe: `https://nlp.stanford.edu/projects/glove/`
- FastText: `https://fasttext.cc/`
- WordSim-353: `http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/`
- **Attach your Codes**

Problem 3 - 30%

In this problem, you are required to re-implement feature extraction for the E-commerce dataset, but this time utilize advanced word embedding techniques instead of traditional Bag of Words with Logistic Linear Regression. Measure the impact on accuracy.

- Choose a modern word embedding technique for feature extraction. Options include Word2Vec, GloVe, or FastText. Select one that aligns with the dataset's characteristics and size.
- Utilize the **pretrained** chosen word embedding to convert words in the dataset into dense vectors. Ensure that the vectors capture semantic relationships between words.
- Train the model on a subset of the dataset and evaluate its performance using appropriate metrics (e.g., accuracy, precision, recall). Compare the results with the previous Bag of Words approach.
- Provide insights into how the choice of word embedding technique affected the model's accuracy compared to the Bag of Words approach. Discuss any observed improvements or challenges.
- **Attach your codes**