ECE 219

Project One: Classification Analysis on Textual Data

April 2021

Team member placeholder

Summary (TBD)

Question 8:

a) Why are GLoVE embeddings trained on the ratio of co-occurrence probabilities rather than the probabilities themselves?
Compared to the raw probabilities, the ratio of co-occurrence probabilities does a better job at distinguishing relevant words from irrelevant words, and it is also better at identifying two relevant words.

b) In the two sentences: "James is running in the park." and "James is running for the presidency.", would GLoVE embeddings return the same vector for the word running in both cases? Why or why not?
No, the GLoVE embeddings would not return the same vector because the ratio of co-occurrence probabilities is different in the two sentences based on the different words matrix with the word running.

c) What do you expect for the values of,||GLoVE["queen"] - GLoVE["king"] - GLoVE["wife"] + GLoVE["husband"]||2,||GLoVE["queen"] - GLoVE["king"]||2 and ||GLoVE["wife"] - GLoVE["husband"]||2 ? Compare these values.
Given the vector equation king − queen = husband − wife, the first value will be zero. Second will be wife - husband, and third value is king - queen.

d) Given a word, would you rather stem or lemmatize the word before mapping it to its GLoVE embedding?
Lemmatize because it is generally more informative, whereas stem may not return an actual word after word reduction.

Question 9: For the binary classification task distinguishing the "Computer Technology" class and "Recreational Activity" class

1) Describe a feature engineering process that uses GLoVE word embeddings to represent each document. You have to abide by the following rules:

   - A representation of a text segment needs to have a vector dimension that CANNOT exceed the dimension of the GLoVE embedding used per word of the segment.
   - You cannot use TF-IDF scores (or any measure that requires looking at the complete dataset) as a pre-processing routine.
   - In each document, there are specific phrases that have more important topical words than others. They are highlighted by "Keywords: . . . " or "Subject: . . . ". Use the average embedding of these words for each document representation.
   - To aggregate these words into a single vector consider normalizing the final vectors.

2) Select a classifier model, train and evaluate it with your GLoVE-based feature