Trevon Woods
Lab 04 - Journal
NLP - ITAI - 2373
06/25/2025

Throughout this lab I had the opportunity to gain a lot of valuable insights in regards to text representations. Firstly, my understanding of how to use BOW, TF-IDF and cosine similarity, and semantic relationships through embeddings has increased tremendously. I already understood the underlying premise behind BOW and TF-IDF but haven't seen them in action. Through this journal I will convey the key insights along with challenges, connections, questions, and comparisons, if any, that I've obtained from this lab's content. I will also touch on how I may use these techniques in future projects.

This lab started out setting up our environment which was all standard and normal. Then we got into our first big topic: Foundations and Sparse Representations. This section covered why text to number conversion is necessary, a little coding with text preprocessing and tokenization, implemented Bag of Words from scratch, and explored the limitations of sparse representations. This section was fairly straight forward. The only insights I obtained from this part of the lab was from the implementation of BOW from scratch. It was interesting to see that sentences with the same context and sentiment have totally different representations. One example from this part of the lab was “This movie is not bad.” and “This movie is bad”. They both have the same connotation but are represented by totally different vectors. This would mean that our model would not be able to accurately tell that these two sentences are the same.

Next, I gained a bunch of cool insights from the section covering TF-IDF, Cosine Similarity, and N-grams. First I got a better understanding of TF-IDF and how it works which is to count the frequency of each word in text and across all documents to determine what words are or aren't important. Then I got a better understanding on what Cosine Similarity does, I always heard the term in AI lectures and videos but the concept never really stuck. It was interesting to know that it calculates the angle between two vectors. It was also cool to see all the heatmaps and graphical representations of the data throughout the lab, it really helps to drive concepts home. We also compared BOW with TF-IDF which in my opinion didn't show much of a difference. I'm not entirely sure what I might be missing but they only seem different by fractions of a point in very limited scenarios.

Finally word embeddings, this was the section I was the most excited to dive into. The first really cool insight was the distribution hypothesis. This hypothesis states that words that appear in similar contexts tend to have similar meanings. This was intriguing because this is the case in most cases in the human language. The last interesting insight was about the comparison of sparse representations (BOW/TF-IDF) and dense representations (embeddings). From this lab I understand that BOW and TF-IDF are less efficient and used if you have a limited dataset. Furthermore, embeddings are more efficient but require larger corpus/datasets .