Gregory Livingston

ITIA-2373

6-26-2025

Reflection on Text Representation Methods

In this lab, I found out that machines can't naturally understand language like people do, so we have to translate our words into numbers. This process is called text representation, and it's a key first step in making machine learning models work with text.

We began with the Bag of Words approach. It simply counts how many times each word appears but doesn't pay attention to the order of the words. That's a big limitation. For example, the sentences "The dog ate my homework" and "The homework ate my dog" are completely different, but BOW gives them the same result. That really showed me how word position matters.

After that, we used TF-IDF, which improves on BOW by giving more weight to unique or meaningful words while reducing the weight of common words like "the." This helps highlight important terms, especially in something like a movie review where words like "boring" or "fantastic" tell us more than filler words do.

We also looked at n-grams, which take word pairs or triples instead of just single words. This helps preserve some of the meaning that gets lost with BOW. For instance, the phrase "not bad" actually means something positive, but BOW splits the words and misses the point. Using bigrams, we can treat "not bad" as one unit.

At the end, we explored word embeddings like Word2Vec. These turn words into number vectors that reflect their meaning and how they relate to other words. For example, the model can understand that “king” and “queen” are connected. This method is way more powerful than the others because it captures the context and meaning of words.

Reflect

From this lab, I realized that there are many ways to turn text into numbers, and each has its pros and cons. BOW is easy but misses the meaning. TF-IDF helps highlight important words. N-grams help preserve word order, and word embeddings capture deeper meaning.

Now I get why cleaning up text before using it is so important. It’s kind of like cooking—if you don’t get your ingredients ready the right way, the final dish won’t come out good. The same goes for giving messy text to a machine learning model.