Jeffery Dirden
ITAI 2373
June 23, 2025

L04 Journal

Throughout this lab, I learnt a lot about how text data can be turned into something machine-readable. At first, I thought it would be simple just to transform words into numbers. But it turns out there's much more to it than that.I began with simple text preparation, which is essentially cleaning up the text so that it is ready for analysis. Lowercasing, deleting punctuation, and tokenizing were straightforward enough, but I had no idea how vital it was to additionally remove stop words and perform stemming. These few steps can significantly improve data quality.

Building a Bag of Words (BOW) model from scratch taught me how to represent text simply by counting word occurrences. It was interesting to see how a simple word count can still provide useful information, but I also saw some limits, such as how it ignores word order and context. The examples of phrases with the identical words but vastly different meanings made me realize that BOW isn't ideal. Next, I learned about TF-IDF, which was fascinating. Unlike BOW, TF-IDF assists by weighing words based on their importance in a document in relation to the entire collection. This helped me grasp why certain words are more relevant than others when attempting to classify text or identify commonalities.

The n-gram analysis represented another step ahead. Bigrams and trigrams capture more context than single words, which made me realize how important phrase-level knowledge is. It's lot closer to how humans read language, and I can see how it would help text analysis. The most interesting thing for me was learning about word embeddings. The thought that words may be represented as vectors in a space with related words close together was fantastic. I also got to see how basic math operations

on these vectors can represent word relationships, such as "king" minus "man" plus "woman" equals "queen." That was really cool!

Throughout this assignment, I grew increasingly conscious of the trade-offs between various strategies. While embeddings are powerful, they are also more complex and resource intensive. On the other hand, BOW and TF-IDF are simpler and easier to comprehend, but they do not capture as much meaning. Finally, focusing on ethical considerations was critical. I had no idea how biases in training data could be carried over into these algorithms, potentially leading to unfair or detrimental results. It reminded me that building AI ethically requires thinking beyond accuracy and performance. Overall, this lab helped me build a solid basis in text representation. I feel more prepared to tackle natural language processing challenges and grasp the advantages and disadvantages of various approaches. The hands-on exercises and visualizations helped me understand everything. I'm excited to continue studying, particularly about emerging methods such as contextual embeddings and transformers.