Ruben Valenzuela  
Professor Patricia McManus

Course: ITAI 2373

24 May 2025

**L04 Journal – Ruben Valenzuela – ITAI 2373**

**Introduction**

In Lab 04, we explored foundational text representation techniques used in Natural Language Processing (NLP). These included Bag of Words (BOW), Term Frequency-Inverse Document Frequency (TF-IDF), N-grams, and word embeddings. The lab was structured to allow a hands-on comparison between sparse and dense approaches to text representation, using Python within Google Colab to experiment with transforming raw text into numerical vectors.

**Bag of Words, TF-IDF, and N-grams**

The BOW model served as a basic yet effective introduction to text vectorization. It highlights word frequency without considering word order or context. While this approach is easy to implement, its limitation lies in its lack of semantic understanding.  
TF-IDF improved upon BOW by weighing word importance within a corpus, reducing the influence of commonly used words that may not offer much contextual value. This method proved especially helpful when identifying unique keywords in documents.  
N-grams provided a solution to some of the shortcomings in BOW and TF-IDF by capturing partial word sequences. Bigram and trigram models offered improved context sensitivity but at the cost of increased dimensionality and complexity.

**Word Embeddings and Dense Representations**

Word embeddings, such as Word2Vec or GloVe, represented a major advancement in NLP by enabling words with similar meanings to have similar vector representations. Unlike sparse models, word embeddings are dense and low-dimensional, allowing for better generalization and semantic understanding.  
In the lab, experimenting with pre-trained embeddings showed how context and relationships between words can be modeled more naturally. Embeddings are especially useful in deep learning models where capturing nuance and meaning is essential.

**Reflections and Insights**

This lab reinforced the importance of choosing the appropriate text representation based on the problem domain. While BOW and TF-IDF remain valuable for simpler tasks and quick prototyping, more advanced applications benefit greatly from embeddings.  
I was particularly surprised by the effectiveness of embeddings in capturing semantic similarities, something that is almost impossible with traditional sparse representations. The difference in classification results and vector visualization offered clear proof of the evolution and power of modern NLP techniques.

**Conclusion**

Overall, Lab 04 was instrumental in deepening my understanding of text representation in machine learning. It served not only as a technical exercise but also as a conceptual bridge between traditional NLP models and more sophisticated neural approaches. Going forward, I am eager to explore contextual embeddings like BERT and investigate how transformers are redefining state-of-the-art NLP.