ITAI 2373 – Natural Language Processing

Student: Natalia Solorzano W207818526

Professor: Narges DeBary

Spring 2025 Course Number: 16579

**A06: Bag of Words (BoW), N-gram, Word Embeddings (Word2Vec), and TF-IDF**

**Introduction**

In Natural Language Processing (NLP), transforming textual information into a numerical format is essential for analysis and machine learning. This assignment focuses on three primary text vectorization techniques: Bag of Words (BoW) with N-grams, Word2Vec embeddings, and Term Frequency-Inverse Document Frequency (TF-IDF). These methods each offer different advantages when it comes to capturing word patterns, frequency, and semantic meaning.

**Bag of Words (BoW) and N-grams**

The Bag of Words approach simplifies text by converting it into a matrix of token counts, disregarding grammar and word order. While effective for capturing raw word frequency, this model lacks contextual awareness. To address this limitation, the N-gram method extends BoW by analyzing sequences of n words, enabling better understanding of adjacent word relationships.

Implementation and Observations: A set of sample sentences was used to generate a BoW matrix, followed by a bigram representation. While the basic BoW technique performed well in identifying frequently used words, it failed to capture contextual meaning. In contrast, N-gram modeling provided richer insights by including local word pairings.

**Word Embeddings (Word2Vec)**

Unlike BoW and N-grams, Word2Vec creates vector representations based on the semantic similarity between words. It uses shallow neural networks to learn word associations from large corpora, allowing it to embed words in a continuous vector space where similar terms have similar coordinates.

Implementation and Observations: Word2Vec was applied to the same dataset to examine its semantic capabilities. The model was able to cluster related words closely, reflecting deeper linguistic relationships. However, its effectiveness depends heavily on the size and quality of the training corpus.

**TF-IDF (Term Frequency-Inverse Document Frequency)**

TF-IDF enhances the BoW model by scaling word frequency relative to how commonly a word appears across the entire dataset. This approach helps reduce the impact of common but non-informative terms, focusing instead on words that carry meaningful weight in specific documents.

Implementation and Observations: The TF-IDF model successfully prioritized unique and informative words while minimizing the influence of frequently repeated terms. This made it particularly useful for applications like keyword extraction and text classification, where distinguishing important content is critical.

**Key Learnings**

-BoW is straightforward but lacks nuance in representing meaning.

-N-grams add value by preserving local context through word sequences.

-Word2Vec captures conceptual similarities, but it requires more data and computational resources.

-TF-IDF is effective in emphasizing significant words and suppressing noise in text data.

**Conclusion**

Effective text preprocessing is foundational to NLP success. Each technique explored—BoW, N-grams, Word2Vec, and TF-IDF—offers a unique way to structure textual data for downstream tasks. Understanding their strengths and trade-offs allows us to tailor models to specific objectives, from document classification to semantic analysis. This assignment reinforces the importance of selecting the right representation strategy to enhance AI-driven outcomes.

**References**

* Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (3rd ed.).
* Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space.* arXiv preprint arXiv:1301.3781.
* Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.