**Journal: Reflection on Representation in NLP**

Over this lab 04, I’ve gained a deeper understanding of how machines interpret human language through various text representation techniques. The transition from basic models like Bag of Words (BOW) to more sophisticated word embeddings was particularly eye-opening. It helped me see how representation affects both the performance and fairness of Natural Language Processing (NLP) systems.

We started with Bag of Words (BOW), which was easy to grasp. It converts text into fixed-length vectors by counting word occurrences. While simple and efficient for basic tasks like spam detection or topic classification, I realized it has limitations. BOW ignores grammar and word order, which makes it lose important contextual information. For example, “dog bites man” and “man bites dog” would be seen as the same. This insight made me appreciate the importance of moving beyond word counts in more nuanced applications.

Another important topic discussed was TF-IDF (Term Frequency–Inverse Document Frequency) improved upon BOW by adding a weighting mechanism that highlights rare but informative words. I saw how this is especially helpful in document retrieval and search engines. However, like BOW, TF-IDF still treats words independently and lacks a true sense of semantic relationships between terms. This led me to explore more advanced models.

The most impactful part of this Lab 04 was learning about word embeddings like Word2Vec. These methods map words into continuous vector spaces where semantic meaning is preserved. What fascinated me was how arithmetic operations on vectors could capture analogies, like how “king - man + woman ≈ queen.” This example showed that word embeddings encode deep relational patterns that simple frequency-based models cannot. I also noticed how this technique allows models to generalize and transfer knowledge across tasks more effectively.

Alongside the technical side, I also reflected on ethical concerns. Embeddings, while powerful, can capture and reinforce biases present in training data. For example, analogies like “doctor - man + woman ≈ nurse” show how gender stereotypes can creep into models. This made me think about how important it is to analyze and audit training data and consider fairness when deploying NLP systems in real life, such as in resume screening or content moderation.

Another key insight came from understanding the distributional hypothesis, which states that words that occur in similar contexts tend to have similar meanings. This hypothesis is the foundation of why word embeddings work. It also helped me understand why context matters so much when designing NLP pipelines.

If I had to summarize what I found most surprising, it would be how mathematical and abstract representations—like vectors—can so accurately reflect meaning and relationships in language. Seeing this both through theory and hands-on experimentation in Colab made it much more concrete for me.

Going forward, I want to learn more about contextualized embeddings like those used in BERT and GPT. Today’s models are static, meaning a word has the same vector regardless of context. But in real language, the meaning of a word like “bank” changes based on context. I’m excited to dig into these models and understand how transformers build on what we explored today.