L05-AWS Machine Learning University Module 2 Lab Exploration

Martin Demel

Department of Science, Technology, Engineering & Math, Houston Community College

ITAI 2376 Deep Learning in Artificial Intelligence

Patricia McManus

February 26th, 2025.

Introduction

Completing the four labs from AWS MLU Module 2 gave me an in-depth opportunity to practice natural language processing (NLP) and deep learning techniques. My work spanned fundamental text preprocessing, feature extraction through bag-of-words (BoW) and TF–IDF methods, employing GloVe embeddings for semantic understanding, and ultimately constructing a recurrent neural network (RNN). In this journal, I detail the key takeaways from each lab, the challenges I encountered, and the ways I envision applying these techniques in my customer service domain.

Body

Lab 1 introduced me to the foundational aspects of NLP, particularly, how raw text can be analyzed and processed in Python. It was fascinating to generate a word cloud from the example text. As I watched the key words float into place, I was struck by how visual methods can make even large bodies of text more understandable. At first, it seemed almost magical that lines of code could automatically create this snapshot of text importance.

Simultaneously, I was reintroduced to techniques like *stemming* and *lemmatization*. Although I knew they were supposed to reduce words to a root form, I never fully appreciated how the nuances like “running” vs. “ran” vs. “run” could matter so much to a machine. Seeing how spaCy handled part-of-speech tagging and named entity recognition reminded me that text data is a goldmine of structured information. This laid a foundation for why **robust preprocessing is so important**.

By the time I reached **Lab 2**, I already felt more confident about text cleaning, so diving into bag-of-words (BoW) methods felt like a natural next step. One of the most eye-opening parts of this lab was realizing that there are multiple ways to numerically represent text. For instance, I initially thought that counting words might be enough. Then I saw how TF–IDF (term frequency–inverse document frequency) can highlight less frequent but contextually rich words. It was a moment of realization that not all words are created equal. For someone in customer service, the ability to spotlight unusual or distinctive terms, could be invaluable.

As I experimented with binary vectors versus word counts, I confronted the subtle interplay between model complexity and interpretability. Word counts felt straightforward. They simply tell you how often a word appears. TF–IDF, however, seemed more sophisticated, reminding me that a truly valuable word might be rare across the entire corpus but extremely relevant to a specific message.

Lab 3 thought me of word embeddings via GloVe. Lab 3 highlighted a deeper semantic understanding of words. Testing car against truck and bike was a particularly memorable exercise. Initially, I wondered if the model might produce unpredictable similarities, however, the measured closeness of car to truck compared to bike demonstrated that embeddings actually capture real-world relationships.

Lab 4 required assembling the lessons from previous labs to build and train an RNN for text classification. I learned to handle variable-length sequences via padding, and I fine-tuned hyperparameters (e.g., epochs and learning rate) to avoid overfitting. Integrating GloVe embeddings into the RNN pipeline also demonstrated how using pre-trained word vectors speeds up convergence and boosts accuracy.

Conclusion

The four labs collectively taught me how text data is transformed from plain strings to meaningful vectors ready for deep learning. They gave me a hands-on appreciation for the building blocks, from cleaning text (Lab 1), through vectorization (Lab 2), semantic embeddings (Lab 3), to full-fledged sequential modeling (Lab 4). Each stage pushed me to connect theory with implementation, while also prompting ideas about how to streamline customer interactions at scale—perhaps through intelligent message routing or real-time sentiment detection. That was something that was always on top of my head.

My biggest takeaway is how often success in NLP depends on careful preprocessing. Even the most sophisticated model can stumble if the text is full of noise. These labs didn’t just teach me about words and vectors; they highlighted how advanced machine learning can genuinely transform the ways we interact with our customers, bridging the gap between technology and genuine human support.

Resources:

GeeksforGeeks. (2025, February 11). *Introduction to recurrent neural networks*. GeeksforGeeks. https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/

GeeksforGeeks. (2022, September 16). *Tokenization Using Spacy library*. GeeksforGeeks. https://www.geeksforgeeks.org/tokenization-using-spacy-library/

GeeksforGeeks. (2024, January 3). *Pretrained Word embedding using Glove in NLP models*. GeeksforGeeks. https://www.geeksforgeeks.org/pre-trained-word-embedding-using-glove-in-nlp-models/