Title: Reflective Journal — Lesson 05 (ITAI2377)

Author: Alfredo Garza
Course: ITAI2377 - Data Science and Artificial Intelligence
Date: 11-July-2025

Challenges Encountered

One of the first issues I ran into during this lab was with mismatched column names in the dataset. The file used a column called `'descr'`, but the code was expecting `'text'`, which kept throwing `KeyError` exceptions. It was a small detail, but it caused a lot of confusion until I caught it. Another hiccup came when trying to save the cleaned data; Python wouldn't write the output file because the `data/` directory didn't exist yet. I had to dig through the error messages, rename columns to keep things consistent, and add some code to create the folder before saving with `to_csv()`.

I also struggled a bit when I tried splitting the code into multiple files. Importing helper functions from the `utils` folder caused some errors at first. It turned out to be a problem with the working directory and `sys.path`, which I fixed after some trial and error. These kinds of problems reminded me how easy it is for small changes to break the whole pipeline — and how important it is to understand the flow of the entire system.

Insights Gained

This lab really helped me appreciate the value of having a clear, repeatable preprocessing pipeline, especially in the context of generative AI. At first, cleaning the text through steps like tokenization, stopword removal, and lemmatization felt like a routine task. But once I saw how much clearer and more meaningful the embeddings became afterward, it clicked. Messy data doesn't just stay messy; it confuses models and makes everything downstream harder to interpret.

Working with TF-IDF vectors was another eye-opener. It showed me how raw text can be transformed into something models can work with, not just numbers, but features that reflect what's important in the text. Before this, I saw preprocessing as just a necessary chore. Now I see it as the foundation for getting any kind of useful output from an AI system.

Connection to Generative AI

This lab made it clear just how much generative AI depends on having clean, well-prepared input. Whether the goal is to generate text, write code, or summarize content, the model can only be as good as the data it's fed. In this case, the cleaned text after removing noise and applying some basic NLP techniques formed the foundation for everything that followed, from clustering to similarity searches.

It also made me realize that preprocessing isn’t just about fixing errors or cleaning up a messy dataset. It’s actually part of the creative process. By approaching the data with a bit more linguistic awareness, beyond just running filters and regular expressions, I made it easier for the AI to make sense of the content and respond in more useful ways. That shift in mindset really helped me see preprocessing as more than a technical task; it’s an essential part of building something smart and meaningful.