NLP MVP: Tweet Reccomender

Goal

The goal of this project is to use NLP to create a tweet recommendation system that would recommend 2-3 similar tweets to a given tweet. This would allow users to explore relevant tweets, expanding reach beyond the tweet and its author.

Data & Process

The data is collected using sns scrape. I collected data for two distinct topics: nightmares and caviar delivery app customer complaints. The focus right now is on building the pipeline and then the best topic will be selected.

Each dataset is 1000+ rows and each consists of a tweet and its corresponding information (number of likes, user id,etc…). For the purpose of this project only the tweet field is used to focus on NLP.

Work Completed

Pre-Processing:

In pre-processing the data, I used regex to remove the following fields:

- Hashtags
- Emojis

- Mentions
- Non alphanumeric characters
- Hyperlinks

Topic Modeling:

To further prepare the data for topic modeling, I used Spacy to remove stop words, and lemmatized tokens.

To begin building a topic model, I tested out:

- Different vectorizers: CountVectorizer vs TfidfVectorizer. I tuned the min_df parameter of the vectorizer to filter out words that didn't occur in more than 1% of the dataset; this helped get rid of nonsensical words like 'ahhh' or 'agghh'. I also tuned the max_df to filter out words that occur too much like "dream" or "sleep" and don't add any value to the topic modeling.
- Combined with different dimensionality reduction techniques: PCA, LSA, NMF and LDA.
- Different numbers for topic components (ranging from 2 to 10).

Early Findings:

So far, it seems like the TfidfVectorizer. combined with NMF is yielding the best results, with 5 as the number of components yielding separate-enough topics.

Here are some of the initial sample topics determined from the 5 top words:

- **Missing an Exam:** school, high, class, start, teacher, exam
- **No Mask in a Crowd:** mask, wear, forget, people, store, public
- **Vivid Dreams:** scary, friend, real, house, shit, weird
- **Being Late for Work:** work, job, stress, late, customer, wake
- **Junko Enoshima:** play, junko, enoshima, dangle, rope, game

Next Steps

I will continue to test out different numbers of topics modeled to make the topics even more interpretable and distinct. I will use the result to build the recommendation system and re-tune accordingly based on the obtained recommendations.