Miguel Montano
Capstone Project 2 Proposal

Topic Modeling Yelp Reviews To Infer Sentiment Causes

Problem: Utilizing Yelp reviews with polarizing opinions (1,2 stars for negative and 4,5 stars for positive), create a model to extract from each group sets of topics that are being discussed when users write reviews with strong sentiments.

Client Interest: A customer's perspective of an individual business' strengths and weaknesses are of extreme importance in a variety of industries, and with the availability of a ranking system built by consumers themselves (yelp star ratings), further insights can be drawn by analyzing the reviews that harbor polarizing opinions. This would give businesses the ability to extract concrete objectives from commentary, in terms of positive attributes to reinforce and negative attributes to mitigate; as opposed to a simple aggregation of star ratings, which lacks explicit or constructive criticism.

Data Source and Acquisition: Data has been acquired from an online posting on Kaggle containing a subset of Yelp's businesses, reviews, and user data. It was originally put together for the Yelp Dataset Challenge, and I will be primarily utilizing the reviews themselves grouped by business type and star rating.

Approach: Python packages SpaCy, NLTK, and Gensim will be utilized in preprocessing the text data, assisting in the removal of stopwords, tokenization, lemmatization, and vectorization. I will be using the Latent Dirichlet Allocation (LDA) topic modelling approach from the Gensim package, as well as Mallet's implementation of the LDA algorithm via Gensim. Finally, I will calculate model perplexity and topic coherence, and visualize the extracted topics with the pyLDAvis package to view the volume and percentage contribution of each topic to get an idea of how important each one is.