Capstone Project 2 - Yelp Review Topic Model - Milestone Report

Problem Statement:

Utilizing Yelp reviews with polarizing opinions, using star ratings to infer sentiment, create a latent dirichlet allocation topic model to extract sets of topics that are being discussed when users write reviews with strong sentiments.

Background:

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.

Dataset:

Data has been acquired from an online posting on Kaggle containing a subset of Yelp's business, review, 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. A dataset provided by Yelp housing the taxonomy of the categorical tags was also utilized as a resource in the grouping process.

Approach:

Initialization of the Review, Businesses, and Categories datasets was done utilizing pandas, with the Reviews being loaded into chunks, housed in a pandas JsonReader object consisting of 60 dataframes with 100,000 reviews each. Unique parent tags were then derived from the Categories dataset, which consisted of 1539 individual tags and 118 parent tags. For the purpose of narrowing down the amount of data and variety of industries the baseline topic model might encompass, only parent tags pertaining to the food industry were selected. The selected parent tags were then used to create a mask and gather the unique businesses id's from the Businesses dataset that had at least 200 reviews and belonged to the food industry. Using this list of business ids, a new dataset was created by looping through the review chunks and slicing for the rows pertaining to the desired business ids , then concatenating the sliced rows into a new pandas dataframe. This dataset consisting of reviews for food industry businesses was then split into two dataframes, one of reviews with star ratings above 3 (considered 'positive') and one of reviews with star ratings below 3 (considered 'negative').

Once the desired reviews were gathered, pre-processing began by using Gensim's simple_preprocess tool to tokenize the text by splitting the reviews (currently one long string) into lists of strings of individual words, make them lowercase, and remove special characters. Then if the

individual tokens were not in Gensim's STOPWORDS library, and were over three characters in length, they would be lemmatized using NLTK's WordNetLemmatizer and placed back into a new list. So now instead of reviews consisting of one long string, they were lists of lemmatized tokens. These tokens were then used to create a mapping between words and their given integer id with Gensim's Dictionary module, and finally, the resulting dictionary object was used to create a bag of words corpus from the tokenized text. The final corpus and the dictionary mapping from id-to-word were then utilized to train a Latent Dirichlet Allocation Model using gensim's LdaModel.

Initial Results:

Lda Models for positive and negative reviews, in this case trained on a corpus consisting only of unigrams, both showed promising initial results. With the ten topics derived from each seeming to harbor associations feasibly inline with human logic, i.e. 'breakfast', 'coffee', 'brunch' all appear together, and 'service', 'wait', 'rude' also appearing together. An apparent downside to selecting such a broad range of reviews was the lack of specificity when it came to the topics themselves, as many were simply food categories, breakfast/pizzeria/burger items were all grouped together for both the negative and positive topic models, instead of drawing different sentiment specific topics.

Moving Forward:

In order to tackle the lack of sentiment-specific actionable results garnered from the initial topic models, a more specific subset of reviews will be used to train and query the lda models; this seeks to both provide utility pertinent to the domain of the interested client, and assist in reducing training/testing time while providing clearer topics. A scikit-learn pipeline will be constructed, using customized transformers built upon the existing text processing steps. A review selector will be built to slice the desired reviews by sentiment, from a list of tags or an individual businesses name or id. The selected review text will be pre-processed, with stopwords and words less than 3 characters removed, and the remaining tokens lemmatized, with the option to use the Phrases module from Gensim to create bigrams or trigrams. The pre-processed tokens will then be transformed into a vector, either bag-of-words or term frequency-inverse document frequency, which will finally be utilized to train a Latent Dirichlet Allocation model. The model will be evaluated with a combination of performance metrics and topic visualization.