BFlag Progress Report

Overview

Significant progress has been made in data cleaning and preprocessing, with John Fox and John B completing the initial phase. Our team's focus included removing image-based questions, eliminating nonsensical words and symbols, fixing contractions, addressing punctuation and non-alphanumeric characters, converting all letters to lowercase, lemmatizing, and removing stopwords. We have also been moving forward on the web app development, topic clustering, and topic modeling.

Data Cleaning and Preprocessing

John F and John B collaborated to execute a comprehensive data cleaning and preprocessing strategy. This involved the meticulous removal of questions presented as images and the elimination of nonsensical words or symbols. Further enhancements included fixing contractions, handling punctuation, and removing non-alphanumeric characters. Standardization was achieved by converting all letters to lowercase. The lemmatization process, coupled with the removal of stopwords, contributed to refining the dataset for subsequent analysis.

Progress Made Thus Far

In addition to data cleaning, our team has moved into the text clustering and topic mining phase. The goal is to identify overarching topics within the current dataset. Subsequently, these topics will be assigned human-readable names and employed in training our predictive model. The text clustering portion will primarily focus on clustering the Jeopardy questions based on the resulting topics from the topic mining portion using the scikit-learn module in Python. The questions will first be vectorized with TF-IDF weighting, followed by a linear dimensionality reduction method for sparse data called Truncated Singular Value Decomposition (TruncatedSVD). Then the class sklearn.decomposition.MiniBatchKMeans will be used to cluster the questions by the mined topic.

Remaining Tasks

Moving forward, our primary objective is to continue with the text clustering and topic mining process. The identified topics will undergo human interpretation, and each will be given a name that resonates with its content. These labeled topics will play a pivotal role in training our predictive model for more accurate and insightful outcomes. In addition, they will be critical to determining the optimal amount of clusters for the text clustering portion. Once complete, we can implement our topic analysis logic into our web app.

Challenges and Issues

One notable challenge we encountered was the decision-making process between lemmatization and stemming during data preprocessing. Through trial and error, we determined that while stemming offered speed and simplicity, the accuracy achieved through lemmatization outweighed these benefits. This informed our decision to prioritize lemmatization for its more precise results, even if it involves a more intricate process. Since we are working with a large dataset of Jeopardy questions, the KMeans class from the sklearn.cluster package tends to perform slowly since it iterates through the entire dataset. As a result, the class sklearn.cluster.MiniBatchKMeans is used to split the dataset into batches of a certain size set by the batch_size parameter. In addition, since the TfidfVectorizer allocates the most RAM when default parameters are set, min_df is set to 0 and num_features is set to 1000 (subject to change over course of project). The sklearn.decomposition.PCA class takes at least a minute to perform dimensionality reduction for the entire dataset so in order to speed up the program, the sklearn.decomposition.TruncatedSVD class is used since the TfidfVectorizer returns a sparse matrix.

Conclusion

The completion of data cleaning and preprocessing lays a solid foundation for the subsequent phases of our project, and we are now transitioning into the text clustering and topic mining tasks. Challenges encountered during decision-making processes have been successfully addressed, and we remain focused on delivering a high-quality predictive model based on the refined dataset.