

Yelp reviews using natural language processing

Writeup

Abstract

Through Natural Language Processing, this project will build a system that can detect the sentiment of cafe reviews and help Yelp find out which cafes have advantages and disadvantages, so they can make improvements. The project aims to provide similar cafes to your favorites through the development of a recommendation system.

Data

Yelp cafes reviews is a dataset provided by kaggle. It consist of Nearly 7,000 observation and 20 features. Important features were selected as (review of the cafe, cafe name, cafe rating).

Algorithms

- Removed the low-level information from our text in order to give more focus to the important information by stop-words removal.
- Grouping together the different inflected forms of a word so they can be analyzed as a single item by lemmatization.
- Remove Punctuations From the text to to get a vector representation of words.
- Digits removal in order to get a clean text only
- Used word tokenizer to tokenize the text.
- Used TF-IDF algorithm to transform text into a representation of weight of words.
- Used Latent semantic analysis (LSA) for topic modeling -most important words for each topic-
- Get the Subjectivity and Polarity scores to do a Sentiment Analysis.
- Get the cosine similarity score for every word in text in order to build the recommendation system for the most relative word to which cafe name.

Tools

- Jupyter Notebook will be used to create and document live code and visualization
- Pandas for data manipulation
- Sklearn for model building
- Matplotlib and Seaborn for plotting