Abstract

Analysis and classification of yelp reviews

This project focuses on detecting the review habits of different type of users and classification of reviews' star levels based on. The dataset used is from yelp dataset challenge, which contains 2.2M reviews made by 552K users for 77K businesses. In the first phase of the project, stratified samples were used in F-test for multiple aspects of reviews and visualization of basic exploratory statistic of the data. The goal is to have a basic idea of the data structure and provide insight on what information can be used for further classification. The next phase of the project is to build a classification model that pinpoint the star rating of a review based on the text. This has been done by building word frequency matrix from the reviews, abstract other language features, dimension reduction and finally classification by SVM. Through showing the power of the model in classifying reviews, this research shows certain features can be used in detecting sentiment in natural language.