Yelpilyzer - Sentiment Analysis Based On Yelp Reviews

Power of Natural Language Process

Author: Yisheng Cai, Advisor: Ralph Morelli

Introduction

Natural Language Processing (NLP) is one of the newest fields of computer science and it provides quantitative insights into big data.

In this project, the program utilizes NLP to train itself over user sentiments behind Yelp reviews on restaurants. A trained program is able to understand reviews and predict a numerical ratings for future reviews.

Approach

The key of vectorization is to help the machine understand the meaning of words, quantitatively, for which a tf-idf conversion is introduced, as shown in Figure-2.

Term Frequency-inverse Document Frequency (tf-idf), is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The corpus, in this case, is the set of all Yelp reviews.

The program tokenizes reviews and removes meaningless English words from the input vector to improve accuracy. Some most frequent words are displayed below in Figure-3:

Support Vector Machine

The machine learning model of this project is a Support Vector Classifier/Machine. The algorithm partitions data into five classes, 1-star (most negative) to 5-stars (most positive) respectively. The advantages of using SVC are:

- Effective on high dimensional spaces and when number of dimensions is greater than number of samples
- Versatile because different Kernel functions can be specified for the decision function

Multiple kernels for this partitioning algorithm are shown in Figure-4. After a few iterations, linear kernel out-performs other kernels on accuracy.

Results

The result table below shows promising accuracy on training and predicting sentiment over 1,000,000 reviews.