Nick Videtti

nvidetti@syr.edu

Assignment 8

Text Mining

Text Mining, Naive Bayes, Support Vector Machines

9/14/2022

| Reviewing Some Reviews | |
| --- | --- |
| **Introduction** | The information age has had an extremely large impact on the way each person lives their life, to the point where devices and knowledge of how to use those devices has become a necessity, rather than a niche interest or skill. In recent years, there are now countless music albums, podcasts, videos, TV shows, movies, books, news articles, and other forms of entertainment. The world now finds itself in what could be referred to as the “content age”.  In this “content age”, the world finds itself in a place where content creators need a way to stand out from the other countless means of entertainment. This is often addressed through ratings and reviews, which can help highly rated content stand out from the rest and can help others see what people think of the content by reading the reviews.  Ratings are usually shown as the overall average rating and are usually useful just looking at that single data point. On the other hand, while people can read a few reviews, it is likely impossible to truly read and interpret every single review. For this reason, This paper will do some analysis on the sentiment of a set of reviews for a restaurant in Syracuse. While this is not necessarily content ratings, this is still relevant since the ratings are made online, are often in abundance, and are used to help potential diners see how others enjoyed their experience.  Unfortunately, there are some reviews that are malicious. Sometimes competitors and other enemies of the content creators / business owners / etc. will write bad reviews to give bad press and make the content/company look worse than it actually is. Using that as additional motivation, this paper will also look into recognizing when a review of the restaurant is not truthful. |
| **Analysis** | Reviews for a restaurant in Syracuse were supplied in a CSV file, as well as an assigned truth value (t = truth, f = false) and an assigned sentiment (n = negative, p = positive). The data cleaning for these data involved needing to extend the number of columns of the data, and then turning the reviews into a document term matrix. The reason for extending the number of columns was that commas within the reviews seemed to cause R to see that as multiple reviews, each separated by the comma. Once this was fixed and the reviews were turned into a document term matrix, that document term matrix was combined with the appropriate truth values and sentiments to make everything a single data frame. Lastly, all variables were changed to factor variables.  Once the data were cleaned and ready to use, they were split into two separate data sources, one containing the truth values with the document term matrix, and one containing the sentiments with the document term matrix. We will refer to these as truth and sentiment, respectively.  For both truth and sentiment, each was once again split into testing and training data sets. These were created by taking a random sample of 2/3 of the row indices and assigning those corresponding rows to the training set. The testing sets were simply the remaining rows that were not used for training. For the testing sets, the actual truth values / sentiments were stored in a separate vector and removed from the testing data. This way those separate vectors can be used to test the accuracy of the data.  Naive Bayes with laplace was used to create models for both the truth and sentiment data, and confusion matrices with their corresponding accuracy were examined. Support Vector machines with costs 0.1, 1, and 10 were modeled using the linear, polynomial, radial, and sigmoid kernels. For all of these models, the accuracy is calculated and listed in descending order to find the best of these models. The accuracy of the Naive Bayes and Support Vector Machine models will be examined for both truth and sentiment. |
| **results** | For the sentiment and truth value predictions, the accuracies were as follows.  Naive Bayes - sentiment    Support Vector Machines - sentiment    Naive Bayes - truth    Support Vector Machines - truth    For both sentiment and truth, Naive Bayes seems to be much less accurate than the most accurate Support Vector Machines. These “most accurate” Support Vector Machines are those using the linear kernel, as those using polynomial, radial, or sigmoid kernels are actually less accurate than Naive Bayes. The best model for sentiment was a Support Vector Machine using the linear kernel and a cost of 0.1, which was approximately 71% accurate. The best model for truth was a Support Vector Machine using the linear kernel and a cost of either 1 or 10, which both were approximately 68% accurate. |
| **conclusions** | There were many methods attempted to find both whether the reviews were truthful and whether the reviews had positive or negative sentiment. In terms of finding truthfulness, the best method was roughly 68% accurate at determining whether a review was true or false. For sentiment, the best method was roughly 71% accurate at determining whether a review has positive or negative sentiment.  While these results are relatively good, there is still much improvement to be made, as these numbers would ideally be much closer to 100%, probably somewhere in the 90’s. This indicates that perhaps prospective consumers should not be looking at reviews unless they look at a large number of them, and that the average rating scores are a much better indicator of their expectations than a review. While reviews are more detailed, they are more prone to fallacies and are not as easy to interpret the way a rating score is. |