Predict rating of an application on a given review

Arukala Mounika

Abstract—The aim of the paper is to predict rating of an application if given a review. To achieve this goal we used regression model which takes each review as a vector along with their scores.
Also, I implemented Spam detection technique on the data set to identify Spam or inconsistent reviews
.

I. INTRODUCTION

Reviews are an important units in today's e-commerce applications, such as targeted advertising, personalized marketing and information retrieval. In recent years, the importance of contextual information has motivated generation of personalized recommendations according to the available contextual information of users. Compared to the traditional systems which mainly utilize users' rating history, review-based recommendation hopefully provide more relevant results to users. We introduce a review-based recommendation approach that obtains contextual information by mining user reviews.
The proposed approach relate to features obtained by analyzing textual reviews using methods developed in Natural Language Processing (NLP) and information retrieval discipline to compute a utility function over a given item. An item utility is a measure that shows how much it is preferred according to user's current context.
In our system, the context inference is modeled as similarity between the users reviews history and the item reviews history. As an example application, we used our method to mine contextual data from customers' reviews of movies and use it to produce review-based rating prediction. The predicted ratings can generate recommendations that are item-based and should appear at the recommended items list in the product page. Our evaluations suggest that our system can help produce better prediction rating scores in comparison to the standard prediction methods.
Reviews play a vital role when it comes to using applications. When a naive user just starts to use an application, the first thing he would do is to look at the reviews to get to know about it's use and importance. On the other hand, users reviews help them understand the further changes that are to be made to improve the application. In my paper, firstly I collected reviews of four applications and identified the inconsistent data.

II. DATA COLLECTION

I have collected reviews and their respective scores of the following four applications namely,
1.Uber
2.Tinder
3.Prizma
4.Yelp.

Later, I categorized them into positive and negative. Below I am mentioning the count of each category reviews that have been collected.
Positive Uber reviews: 127
Negative Uber reviews: 114
Positive Tinder reviews: 316
Negative Tinder reviews: 323
Positive Prizma reviews: 527
Negative Prizma reviews: 111
Positive Yelp reviews: 146
Negative Yelp reviews: 141

Sample Reviews:
1. TINDER

Positive: This app is great and all, but it ruined a relationship that my friend had.
Score: 5

Negative: Filled with American photoshoot fake profiles.
Score: 1

2. YELP

Positive: Always the first place I look for what is happening in the community or new food to try, just Awesome!!
Score: 5

Negative: Annoyed i have to download to use the website.
Score: 1

3. PRIZMA

Positive: Its magic but slow.
Score: 4

Negative: Please bring back the old tears effect. I miss the old tears effect.
Score: 1

4.UBER
Positive: Good but more more functionality
Score: 5
Negative: Worst cab booking app in the world. I request everyone not to download this app
Score: 2

Data Cleaning

When considering how to clean the text, we should think about the data problem we are trying to solve. For many problems, it makes sense to remove punctuation.

On the other hand, in this project, we are tackling a sentiment analysis problem, and it is possible that "!!!" or ":-(" could carry sentiment, and should be treated as words. In my project, for simplicity, I removed the punctuation altogether.

I replaced numbers with space, but there are other ways of dealing with them that make just as much sense. For example, we could treat them as words, or replace them all with a placeholder string such as "NUM".

To remove punctuation and numbers, I used a package for dealing with regular expressions, called re. The package comes built-in with Python; no need to install anything.

Firstly, I made sure the reviews were all in text format. Whenever I happened to see anything other than text, I just followed the steps discussed above.

In the next step, I found anything that is not a lowercase letter (a-z) and converted it into lower case so that it would help me identify similar words and eliminate them if necessary.

On the other hand, anything that is NOT a lowercase letter (a-z) or an upper case letter (A-Z), was replaced with a space.

Later, I dealt with stop words (frequently occurring words that don't carry much meaning) removal. Such words are called "stop words"; in English they include words such as "a", "and", "is", and "the". Conveniently, there are Python packages that come with stop word lists built in. I looked into the stop words and eliminated them which helped me concentrate less on very less useful words.

.

METHODOLOGY.

SEMANTIC ANALYSIS

In linguistics, semantic analysis is the process of relating syntactic structures, from the levels of phrases, clauses, sentences and paragraphs to the level of the writing as a whole, to their language-independent meanings. It also involves removing features specific to particular linguistic and cultural contexts, to the extent that such a project is possible. Semantic analysis can begin with the relationship between individual words. This requires an understanding of lexical hierarchy, including hyponymy and hypernymy, meronomy, polysemy, synonyms, antonyms, and homonyms. It also relates to concepts like connotation (semiotics) and collocation, which is the particular combination of words that can be or frequently are surrounding a single word. This can include idioms, metaphor, and simile, like, "white as a ghost."

Some knowledge bases not only list obvious affect words, but also assign arbitrary words a probable "affinity" to particular emotions. Statistical methods leverage on elements from machine learning such as latent semantic analysis, support vector machines, "bag of words" and Semantic Orientation — Pointwise Mutual Information. More sophisticated methods try to detect the holder of a sentiment (i.e., the person who maintains that affective state) and the target (i.e., the entity about which the affect is felt).

BAG OF WORDS APPROACH

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. The bag-of-words model has also been used for computer vision.

The bag-of-words model is commonly used in methods of document classification where the (frequency of) occurrence of each word is used as a feature for training a classifier. In my project, I implemented this approach and converted the text reviews into vectors which helps to proceed forward with regression.

REGRESSION MODEL

In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent

variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables.

In this project, I used the logistic regression which is a regression model where the dependent variable (DV) is categorical. This article covers the case of a binary dependent variable—that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Cases where the dependent variable has more than two outcome categories may be analysed in multinomial logistic regression, or, if the multiple categories are ordered, in ordinal logistic regression. In the terminology of economics, logistic regression is an example of a qualitative response/discrete choice model.

TRANING DATA AND TESTING DATA

In many areas of information science, finding predictive relationships from data is a very important task. Initial discovery of relationships is usually done with a training set while a test set and validation set are used for evaluating whether the discovered relationships hold. More formally, a training set is a set of data used to discover potentially predictive relationships. A test set is a set of data used to assess the strength and utility of a predictive relationship. Test and training sets are used in intelligent systems, machine learning, genetic programming and statistics.

I divided the collected data into two parts called training data and testing data. 75 percent of the data was taken as training data and the rest 25 percent as testing data. I build a regression model using this training data. In the next step, I sent the testing data as an input to the model built and got scores for all the reviews. In the end, I compared the actual testing data scores with the scores obtained from the regression model to check the accuracy.

ACCURACY CALCULATION

Regressions differing in accuracy of prediction. The standard error of the estimate is a measure of the accuracy of predictions. Recall that the regression line is the line that minimizes the sum of squared deviations of prediction (also called the sum of squares error).

As mentioned above, I compared the actual testing data scores with the scores obtained from the regression model to calculate the accuracy. The obtained accuracy of each application is mentioned in the results section.

SPAM DETECTION

SENTIMENT ANALYSIS

Sentiment analysis (sometimes known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

INCONSISTENT REVIEWS

There are reviews where the content is written in a positive way but the score given is negative.
It could be vice-versa as well, in other words, the review written is negative but the score given was positive.

Implementation of Inconsistent Reviews:

Step 1: Calculated the sentiment scores for all the reviews

Step 2: Compared it with the actual scores

Step 3: If the sentiment score calculated is negative and the actual score is positive, I called it Spam

Step 4: If the sentiment score calculated is positive and the actual score is negative, I called it Spam/

RESULTS

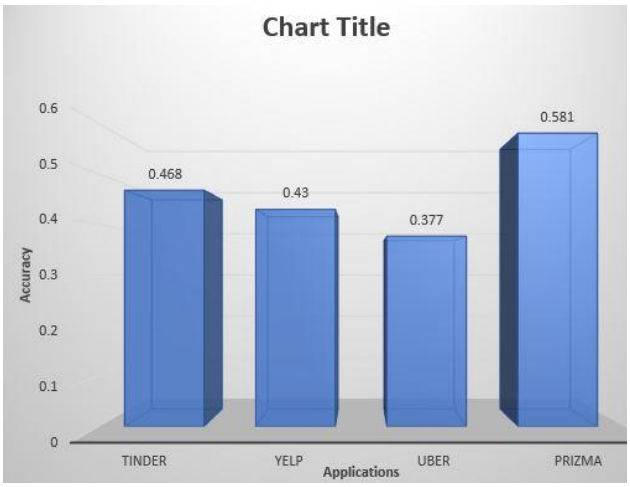
As discussed above, for each of the four applications taken, I calculated the accuracy by using the testing data, training data and regression model. The obtained accuracy for each is listed below:

1. Prizma: 0.581

2. Uber: 0.377

3. Tinder: 0.468

4. Yelp: 0.43



CONCLUSION

Ratings and reviews are some of the most important triggers in app discoverability and installs, and are still core metrics app marketers use to gauge success. App store ratings are crucial to driving rankings, discovery, downloads, updates, and in-app purchases. Ratings are an integral part of the ranking algorithm for app searches in both the App Store and the Play Store.

To conclude, once the data was collected we performed sentiment analysis, implemented bag of words approach to convert the text reviews into numeric. Later, the data was divided into training data and testing data. Using the training data regression model was built. To test the model, I sent testing data as an input to the model and checked the accuracy at the end. On the other hand, I have done Spam detection to identify spam reviews.

ACKNOWLEDGEMENT

Firstly, we would like to show our gratitude to Dr. Tung Nguyen, Professor at Utah State University for sharing his pearls of wisdom with us during this project. We are also immensely thankful to the Computer Science Department for all the extraordinary support given when it comes to working space which helped us finish the project on time.

References

[1] Wikipedia

[2] https://www.yelp.com/dataset_challenge

[3] https://www.yelp.com/topic/new-york-yelps-star-rating-system