10-401 Midterm Report

Authors:
Manik Panwar (mpanwar)
Caleb Eddy (cweddy)

Abstract

It has been shown before that tweets can be an abundant source of information for bag-of-word learning models for sentiment analysis (1), as well as other classification-oriented approaches in machine learning (2). Using a bag-of-words approach to create a movie-recommender system has also been done before (3). The idea that we had follows closely from this idea - our goal in this project is to use a Naive Bayes classifier to provide movie genre recommendations to twitter users, based upon a collection of their tweets. At this point in the project, we are in the process of processing the data to get it ready to train the classifier.

1 Description of Problem

The project is titled: Twitter Movie Genre Recommender

All of the raw data for this project can be found at the github link: (https://github.com/sidooms/MovieTweetings). There is a connected rating system between twitter and IMDB that allows twitter users to publicly rate a movie on a scale from 1 to 10, and this dataset contains their twitter id's, what movies they ranked, and the genre's these movies belong to (there can be more than one per movie). Our goal with this model is to find, for each user, the genre that they rank most highly (with ratings averaged over each genre), and from there, to train the network on a collection of each user's tweets, with their "favorite genre" as the class label. After that we can use this model to query a twitter user and from their tweets, recommend a movie genre to them. The following is some useful literature we found related to Twitter and Natural Language Processing all of which have stuff that can be applied to our project.

- Kouloumpis, E., Wilson, T., Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg!. Icwsm, 11(538-541), 164.
- Severyn, A., Moschitti, A. (2015, August). Twitter sentiment analysis with deep convolutional neural networks. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 959-962). ACM.
- Gao, M., Zhang, X. (2015). A Movie Recommender System from Tweets Data (no journal of publication - this is an undergraduate student project paper that can be found at: http://cs229.stanford.edu/proj2015/299_report.pdf)
- https://www.cs.rutgers.edu/ pazzani/Publications/IPSJ.pdf

2 Implementation (Building the Data Set)

A major part of the project is building a dataset of user tweets to movie genre ratings. For this we divided up the work into getting the the twitter userid to IMDB movie ratings and twitter userid to user tweets.

For the first part roughly speaking, we start with two types of tuples: (user_id, movie_id, rating), and (movie_id, genre_1, genre_2, ..., genre_n), where any given movie can belong to one or more genres. From there, we construct a dictionary in which user_id's are used as keys, and the value contained at each user_id key is itself a dictionary, which contains genre's as keys and lists of movie ratings as values. We use this to take the average movie rating by genre for each user, and take the maximum of that, to obtain the movie genre that that user has ranked the highest, on average. This genre is then used as the class label for that user's tweets to be paired with during training. Currently, we have 48348 users with ratings.

For the second part, we have to get user data from Twitter based on their user ID. For this we use a Python Twitter wrapper API to query Twitter and build a framework to get Public Tweets from User Timelines from their UserID's. Essentially we will query the most recent tweets and take the top 25 (variable number, will tweak and see how model behaves) tweets from their Timeline. This means setting up an application with Twitter to get credentials to access this API and also means we are rate limited with the number of queries we make so we have to split up the data gathering in parts if at any point Twitter gives us number of requests exceeded for day error. Getting familiar with this was a challenge but now we have adequate familiarity with this.

3 Implementation (Machine Learning model)

We've been working on building the data set so we don't have our initial results yet. Our first approach is to use a Naive Bayes bag of words model and train on roughly 70 percent of the dataset and use the rest as the test set. Once this is done we will incorporate more more context into our model by doing a bigram and a trigram model on the data set as well and using interpolation to combine the three probabilities to come up with our Naive Bayes probability.

Another approach to the problem which is a reach goal will be to incorporate Neural Networks into this and use Keras to build a Neural Network to do the classification and compare the Naive Bayes and Neural Network models.

4 Implementation Technical Details and Dependencies

The project is being done in Python 2.7. We are using Github to collaborate. We have listed the Github link to the project along with some of the main dependencies.

- Github for the project: https://github.com/manikpanwar/TWGR
- https://github.com/sidooms/MovieTweetings
- https://github.com/bear/python-twitter
- NLTK for Naive Bayes

5 Going Forward

We aim to have our initial results from the Naive Bayes model by the end of the week.

After that we will look at expanding the model and applying interpolation to the Naive Bayes probabilities by adding in a Bigram and a Trigram model to include more context.

Once the above is done we will be tweaking the models and seeing how the accuracy is changed, using different types of smoothening for the data and possible working on the side to see how a Neural Network would tackle the problem.

We will also look at different ways of pre processing the tweets and seeing if taking particular things out (User Handles, Links, Stop words) will affect accuracy.

The following set will be used for the stop words: http://www.ranks.nl/stopwords