Machine Learning Final Project Proposal

Michael Ashmead (mashmea1@jhu.edu)
Maxwell Yeo (myeo1@jhu.edu)

November 4th, 2016

1 Introduction

Reddit is a social platform that allows registered users to upload text, images, and links. Users, based on their preferences, can follow various Subreddits that are each for a different topic. Currently, in order to find a Subreddit that may interest you, you have to manually search for Subreddits. We propose a method for recommending Subreddits to a user based on both user activity (users posting in various Subreddits) and descriptions of Subreddits to predict which Subreddits a user will be likely to follow.

2 Libraries

We will use the Python programming language. The only library we plan to use is NLTK. NLTK is for natural language processing and will help us remove extraneous words in the descriptions of Subreddits that won't be beneficial for determining similar Subreddits.

3 Data Source

We will be using two data sets.

1. The first data set is a collection of anonymized users identified by user ID followed by the subreddits that they are active in. A reddit user is considered active in a subreddit if they posted 10 times in the 1,000 most recent posts in that subreddit. The file has 850,000 user's posting preferences back in 2013.
   We found this data online at this URL: `https://figshare.com/articles/reddit_user_posting_behavior/874101`

2. The second data set is a collection of 1,156,310 subreddits. Each row contains a subreddit followed by features that are relevant to the subreddit like the number of subscribers, whether comments are allowed, whether the subreddit is restrained to users who are over 18, description, and much

more.
We found this data online at this URL: `http://files.pushshift.io/reddit/subreddits/`

4 Methods

In order to achieve our goals, we plan to use a Collaborate Filtering Algorithm and a Content Filtering Algorithm. In both cases we can use λ-means clustering to recommend Subreddits since this is loosely based on the paper where we got our dataset from [1]. Depending on the results of this implementation, we may attempt Naive Bayes to see if we can achieve better results since it is the most common algorithm used for recommender systems [2].

5 Milestones

We have two milestones, and a reach goal. Our two milestone are:

1. Our first goal/milestone is to use the first data source to cluster subreddits. Utilizing the users' posting preferences, we can group subreddits together based on how many users post in two particular subreddits.

2. Our second goal/milestone is to use the second data source to again cluster subreddits. This time, instead of utilizing users' posting preferences, we will parse the descriptions for each subreddit and base clustering off of word frequency in descriptions.

We have two reach goals, they are:

1. Our first reach goal is to improve how we utilize the descriptions for each subreddit. We would like to use n-grams by combining words into phrases to improve the accuracy how we calculate word frequency on our second data set. We would also like to filter out stop words, and most frequent words, because words like, "the" would appear in almost every description. There are 1,025,109 words in the dictionary, eliminating stop words and frequent words, we hope to bring this number down.

2. Our second reach goal is to combine our two models to create an even better recommendation system. We will utilize learning theory to test out different weights between the two models.

6 Outline of Final Writeup

The final writeup will include the introduction, related work, datasets and libraries used, any preprocessing of the data sets that we do, a detailed description of our technical approach, our results and analysis, our conclusion and possible future work, and a bibliography.

In the technical approach we will graphically at minimum have a table of our test results for each of the 4 goals we hope to achieve and each algorithm we use for each goal.

7 Bibliography

[1] https://arxiv.org/pdf/1312.3387v1.pdf
[2] https://arxiv.org/ftp/arxiv/papers/1511/1511.05263.pdf