CSCI 4502 Progress Report

Dimitrios Economou
Josh Weaver
Thomas Horn

April 6, 2015

1 Abstract

Reddit is a large and rapidly growing social news and content aggregation Web site. In fact, Reddit has evolved from the front page of the internet to a self-contained community with a lot of internal discussions about various subjects[9]. On Reddit, users submit content which can be external links or text posts to various areas of interest or topics called *subreddits*. In general, we will attempt to cluster subreddits using different metrics and clustering methods. Clustering subreddits could be potentially useful for a subreddit recommender system, but out project is mainly an exploration of clustering subreddits in different ways. In particular, we will cluster subreddits based on how topically diverse their comment threads are. Weninger et al.[10] used topic modeling to show that comment threads on Reddit form topical hierarchies. It would be interesting to carry out a similar analysis to explore topical differences between subreddits or types of subreddits. For example, discussion-based subreddits may be more topically diverse than image-based subreddits.

2 Dataset

Thanks to \u \Stuck_in_the_Matrix of redditanalytics, we have access to all of the submission objects on Reddit and about five months worth of comment data (from September 2014 through January 2015). Since we will probably only use the comment data, we will focus on describing that. The data is organized into five files where each line is a JSON object representing a single comment. The fields of each JSON object are (among other things) the number of upvotes and downvotes (and the difference representing a score), the author, the subreddit, the creation time, the link or text if it is a self-post, and its parent comment id. In total, the dataset of five months of comments is 113.4 GB.

2.1 Cleaning the data

There are several fields in the comment JSON objects that we are not interested in, so we have gone through the dataset and removed these fields.

To get started, we decided to restrict the dataset to 645 popular subreddits. The list of subreddits and a subjective analysis of whether they are discussion-based (self), have a high, medium, or low density of link submissions, or are image-based (imgur) was done by Siege Media[1]. We removed all the comments not posted in one of these subreddits and added in a field to describe what type of subreddit they fall under. We can potentially use this new field as a class label.

Further, there are many comments which were posted but were deleted by the author or some other reason. We removed these comments as we need the comment bodies for our analyses.

2.2 Basic facts about the data

We will now describe some basic statistical facts about the test dataset, which is a reduction of one of the months of comment data and is what we are first running our algorithms on. Statistical facts about the bigger dataset will be given in the final report.

Stat	Comments per thread	Comments per subreddit
Count	82727	373
Mean	12	2681
Std	53	7921
Min	1	1
Q_1	1	122
Q_2	4	489
Q_3	9	1578
Max	4681	80523

3 Progress

We have written scripts to clean the data and used them. We have done basic statistical analyses of the data to get to know it better. We have attempted clustering subreddits based on where users comment using hierarchical clustering.

We have written scripts to process the comment bodies. This process involves breaking the comments up into words, obtaining their stems, removing stopwords, removing non-words, counting unique words (actually stems), removing words that appear only once in any given subreddit, etc. We made extensive use of Python's Natural Language Toolkit (NLTK) to do this. We have spent a lot of time manipulating the data and putting it into different formats, including putting it in a SQLite database. However, we have not really needed to use the database yet.

After processing the comments, we have clustered subreddits based on words most frequently found in them. We have tried k-means clustering using cosine similarity. The results have been promising. Using k-means with 15 clusters yields some intuitive clusters. For example, we have (clusters are labeled with integers and these integers are not important)

2: bikebuilders, triathlon, motorcycles, cyclocross, adventures, bmx, bicycletouring, cycling, bicycling

3: pcgaming, boardgames, gamedev, gamernews, gaming, vita, patientgamers, truegaming, pcmasterrace, iosgaming, gamingsuggestions, snes, retrogaming, gamingnews, xbox, xboxone, xbox360, zelda

4: microsoft, dotnet, technews, design_critiques, graphic_design, programminghorror, learnprogramming, netsec, css, kickstarter, oculus, photoshop, programming, coding, androiddev, learnpython, geek, javascript, linux, java, web_design, androidapps, softwaregore, technology, software, learnjavascript, webdev, tech

7: ramen, cheesemaking, fromscratch, budgetfood, food, 52weeksofcooking, recipes, nutrition, slowcooking, grilling, steak, cookingforbeginners, foodhacks, tonightsdinner, smoking, fitmeals, spicy, eatsandwiches

8: techsupport, computers, gamingpc, windows8, sysadmin, homelab, usenet, buildapc, retrobattlestations, hardware, talesfromtechsupport, 24hoursupport, htpc, applehelp, networking

9: financialindependence, freelance, woweconomy, investing, personalfinance, advertising, tax, digitalnomad, portfolios, economy, hwstartups, socialmedia, finance, smallbusiness, austrian_economics, bigseo, wallstreetbets, startups, business, marketing

11: movies, scifi, dvdcollection, bestofnetflix, southpark

12: winemaking, beertrade, wine, cider, mead, alcohol, retailporn, bourbon, cocktails, beer, beerporn,

but we also have clusters such as

10: askphilosophy, businessschool, premed, geology, expats, science, nursing, askscience, emergencymedicine, cogsci, energy, acting, medicine, math, careerguidance, learnspanish, resumes, gradadmissions, computervision, motivation, vfx, chemistry, languagelearning, architecture, academiceconomics, college, artificial, medicalschool, latin, jobs, cscareerquestions, pharmacy, theydidthemath, compsci, physicaltherapy, consulting, statistics, russian, engineering, education, findapath, linguistics, neuroscience,

which doesn't make much sense. Specifying more clusters, however, yields some more structure such as the cluster

3: learn_arabic, learnspanish, languagelearning, latin, learndutch, italianlearning, russian, linguistics.

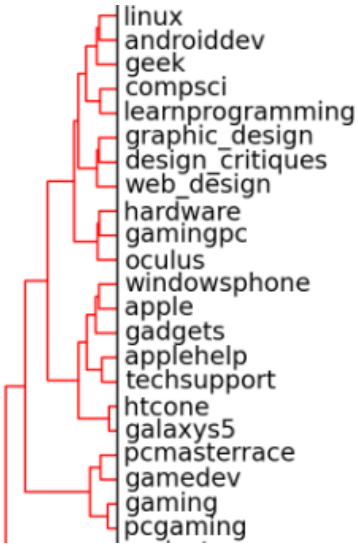
It also refines the clusters further into clusters such as

6: vegan, vegetarian
15: winemaking, wine, cider, mead.

It is exciting to see such structure revealed from a simple clustering algorithm.

Figure 3 shows a (relevant) small portion (the actual plot is too big to show for the report) of a dendrogram obtained by hierarchical clustering using Ward's method.

Figure 1: Hierarchical clustering of subreddits using a td-idf matrix (from test data), cosine similarity, and Ward's method.



4 Problems

A 113.4 GB dataset is a much larger dataset than any other dataset we have worked with. This indroduces many time- and memory-related issues. It takes a large time to clean the data and process the comments. Although performing more subtle language processing on the comment bodies yields better results, this substantially increases the computational complexity of our analyses. Even doing a relatively straightforward processing of the comment bodies takes a long time. We have also needed to learn about natural language processing in general, as well as topic modeling. For topic modeling, it is not practical to set up a database to do the work we want to do. So we have been relying on loading our data into memory for computing. We are therefore constrained by how much memory we can access.

Many clustering algorithms (such as k-means) requires the number of clusters to be specified beforehand. However, we do not have this information, so we have resorted to trial-and-error. There is an R package called NbClust that uses 30 different indices to determine an estimate for the best number of clusters and clustering scheme for a dataset. It may be worthwhile to try this on out dataset.

Different implementations of clustering and topic modeling requires putting the data into different formats. This introduces a lot of overhead when experimenting with different tools.

5 What's Next

Since the proposed work is likely a bit too extensive for the scope of this project, we will initially be doing basic analyses of submission data and comment data, crawling more comment data, searching the social network analysis literature, and refining the problems we want to tackle in depth. Once our problems are more refined, we will explore them in more detail and evaluate them. The following is a rough timeline of what we would like to accomplish.

1. Draw plots of the clusters for visualization.
2. Try different subreddit similarity metrics.
3. Try different clustering techniques.
4. Try using NbClust in the R programming language to find the best number of clusters. This may subsume the previous two items.
5. Try dimension-reduction techniques such as principal component analysis or some form of multidimensional scaling. This may be important to do for a larger dataset, more subreddits, and it may reveal some additional structure in our data.
6. Perform topic modeling on comment threads using MALLET.
7. Using topic modeling on comment threads (LDA in particular), find how difficult it is to find the topics of a comment thread in each subreddit on average (using the perplexity or log-likelihood of topic models of comment threads). Try using this as a subreddit similarity metric for clustering.
8. Perform the above on larger datasets (both in number of comments and number of subreddits). We may need to use different techniques for this to handle a large amount of data. It may become crucial to use some sort of database, such as the SQLite database we have constructed.
9. If time permits, see if it is possible to combine clustering techniques for a subreddit recommender system.

References

[1] 750 popular subreddits, categorized by industry and submission type. `http://www.siegemedia.com/popular-subreddits-by-industry`. Accessed: 2015-2-1.

[2] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[3] Vicenç Gómez, Andreas Kaltenbrunner, and Vicente López. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 645–654, New York, NY, USA, 2008. ACM.

[4] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[5] Salman Jamali and Huzefa Rangwala. Digging digg: Comment mining, popularity prediction, and social network analysis. In *Proceedings of the 2009 International Conference on Web Information Systems and Mining*, WISM '09, pages 32–38, Washington, DC, USA, 2009. IEEE Computer Society.

[6] Jong Gun Lee, Sue Moon, and Kavé Salamatian. Modeling and predicting the popularity of online contents with cox proportional hazard regression model. *Neurocomput.*, 76(1):134–145, January 2012.

[7] Clemens Meinhart. Studying User Submissions and Content on Reddit. Master's thesis, Graz University of Technology, Austria, 2014.

[8] Matthew Rowe, Sofia Angeletou, and Harith Alani. Predicting discussions on the social semantic web. In *Proceedings of the 8th Extended Semantic Web Conference on The Semanic Web: Research and Applications - Volume Part II*, ESWC'11, pages 405–420, Berlin, Heidelberg, 2011. Springer-Verlag.

[9] Philipp Singer, Fabian Flöck, Clemens Meinhart, Elias Zeitfogel, and Markus Strohmaier. Evolution of reddit: From the front page of the internet to a self-referential community? In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 517–522, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

[10] Tim Weninger, Xihao Avi Zhu, and Jiawei Han. An exploration of discussion threads in social news sites: A case study of the reddit community. In

Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '13, pages 579–583, New York, NY, USA, 2013. ACM.