RUTGERS
UNIVERSITY

redreaper

Project Number
S16-036

Team
Julian Esteban, Kanav Tahilramani, Matthew Chatten, Sujay Bandarpalle

Advisor
Dr. Shantenu Jha

April 15, 2016

Submitted in partial fulfillment of the requirements for senior design project

Electrical and Computer Engineering Department
Rutgers University, Piscataway, NJ 08854

Abstract

The goal of this project is building a webapp providing analytical analysis of online social media behavior. Specifically, due to the open nature of its data and a public API, the site Reddit was chosen as a focus of the project. Reddit is a semi-anonymous public platform; users are identified by self-chosen usernames, and can post comments or links to a wide variety of separate communities called “subreddits”. It is these two key elements – users and subreddits – which this project focuses on. The RedReap tool takes as input a username or the name of a subreddit, and outputs a wide-ranging analysis of it.

Three key categories of analysis and data are produced by this tool. The first category is general statistical data, such as number of submissions, popularity of comments, details of posting history, etc. This is accomplished through fairly simple techniques - accessing the appropriate dataset and counting up the relevant values. The second and most important category is content analysis. This includes calculating the complexity of language used, detecting personally identifiable information, and determining the positive or negative sentiment of posts. The final category is metadata analysis, focused on revealing how variables like post time, post score, and comment length relate to each other.

The aim of providing all this analysis is to give the user of the webapp an understanding both of the individual Reddit user or subreddit and of just how much data and personal details can be extracted and compiled from public social media sites. It combines the personal interest of seeing information on how you or other people speak with the educative experience of seeing what natural language processing and metadata can reveal.

Table of Contents

1. Introduction

1.1. Background Information

Reddit is an independent social media site, owned by “reddit Inc.”, which calls itself “the front page of the internet.”[1] It is essentially a network of discussion pages, organized by topic, called “subreddits” in which users can post links or discussion topics for other users to see, comment on, and vote on. The key the website is it’s “karma” system, in which users can “upvote” or “downvote” content, both submissions and comments, resulting in community-driven “front pages” of curated content. With a weighting algorithm that heavily favors new content over old, popular submissions will eventually drop off of the front of a subreddit to be replaced by new material.

Another factor that makes Reddit relatively unique among social sites, and which led to its choice as the subject of this project, is that almost all subreddits are completely public, and moreover all user histories are public. While Reddit can simply be browsed, to vote on or submit content a user account with a unique username must be created. These usernames are almost never the users’ real names, but they are still trackable and associated with only one person. Given the open nature of such data and the existence of a solid API for accessing it, Reddit is the perfect test case for analyzing online social behavior in aggregate.

1.2. State of the Art and Relevant Projects

Because Reddit is so ideal for this type of data mining and analysis, this project is not the first to attempt the task. There are a host of websites and services dedicated to Reddit data – for example, FiveThirtyEight has made a “Reddit Ngram” available.[2] (It displays frequency of word or phrase use within a corpus of Reddit comments from 2007 to 2015, much like the Google “NGram” does for Google’s collection of digitized books.) Other notable examples include KarmaDecay[3], a website which can perform reverse image searches on Reddit, and Karmawhores[4], which tracks the top users of Reddit by their karma scores. However, none of these do any analysis of individual users or subreddits, which is the goal of this project.

There are three existing websites which perform some sort of analysis on either users or subreddits. The first is roadtolarissa’s “Reddit Comment Visualizer”[5], which can graph a user’s comments by number, length, or karma according to subreddit over time. This is useful and fairly well-displayed data, but it is also simply metadata. The most you can do is see general commenting trends. This project is aimed more at specific pieces of data and analysis which go deeper into the actual content of a user’s history.

The second is Redective[6], which retrieves metadata for both users and subreddits upon request. However, the output is very limited, and not displayed in a particularly appealing or even understandable manner. It displays a small amount of basic information, counts submissions by subreddit, lists the most frequent words used, and displays the top scoring submissions. Our project aims to be more user-friendly, include more detailed information, and focus more on detailed content analysis rather than simple metadata.

The third and final similar work in this field is the more impressive SnoopSnoo[7], which again allows for analysis of both users and subreddits, though for subreddits only very basic information is presented. The user analysis includes a solid selection of basic data, such as best and worst comments, as well as a good selection of metadata like posts by day, posts by time, and posts by subreddit organized and displayed in visually appealing graphs. Unlike every other work mentioned so far, SnoopSnoo actually tries to extract some information from user's comments using natural language processing, displaying items such as what the user has explicitly said they like or the user's stated religious beliefs, whenever available. This is closer to what we hope to accomplish with our own project, but it is still limited. We intend to do deeper and wider natural language processing and content analysis, along with an even better selection of metadata and regular data analysis. We also intend to provide much more analysis for subreddits.

1.3. Problem Addressed by Project

The core problem addressed by RedReap can be looked at narrowly or broadly. In a narrow sense, Redreap's goal is to provide meaningful analysis and data on Reddit users and subreddits. In a broad sense, Redreap's goal is to give its users an understanding how social media data in aggregate can reveal more than one might expect.

1.3.1. Narrowly Defined Objective

As discussed in the section on the State of the Art, existing services using Reddit data do not match up to the wide space of possibilities in natural language processing or even the extensive metadata Reddit provides. As existing solutions to the problem of "I want to see analysis and data on a user or subreddit." go, there is significant room for improvement, and that is where Redreap comes in. Redreap will provide broader and deeper analysis for both users and subreddits, by collecting more metadata, by doing more in-depth natural language processing, and by presenting this data in an understandable manner.

1.3.2. Broadly Defined Objective

This is the era of both social media and Big Data, and yet most regular users of social media don't have a good grasp on quite what that means. Social media companies themselves, such as Facebook, are constantly sifting through and analyzing the data available to them on their users, but users themselves don't have such an opportunity. Because Reddit is an open, public site with a robust API, Redreap can give its users a taste of what is currently possible in social media analysis. Especially important is the natural language processing component, which can uncover personal information, personal habits and proclivities, and more, but whose power is generally unknown to the public. This lack of understanding can be rectified by Redreap, by showing just how much can be revealed by automatic analysis of social media users and communities.

1.4. The Adopted Approach

The basic approach is to build a website that meets the defined objectives. The problem is best solved by a website because, at its core, Redreap is built to meet the requests of users. Users of the site request analysis of either Reddit users, by their Reddit username, or of Reddit subreddits, by the subreddit's name. Because actually analyzing these inputs requires a large amount of computation time, and because there are over 36 million Reddit users and over 850 thousand subreddits[8], the results cannot all be precomputed. Thus, the output depends on the user's input and on having access to Reddit through its API, making a website approach ideal.

2. Methods / Results / Approach

The following sections detail the work done to complete the objectives of the project, the standards and guidelines followed during that process, and the results of that work.

2.1. Methods

The nature of the project problem meant that a software solution was required. Therefore, the execution of the project has centered entirely on programming goals and challenges.

2.1.1. Basic Project Framework

Any website requires some sort of "programming stack" that will handle the front-end and back-end operations necessary. For this project, the MEAN stack[9] was chosen, which includes mongoDB, express, AngularJS, and NodeJS. It was chosen for its ease of use, and because of previous experience with the framework. For coordination of the project coding, Github[10] was chosen and a repository was made with all group members included. This allowed for effective version control, and made it easy for individual group members to make their own branches of the project to implement features or test out new ideas. The project is designed such that it can run on any computer with the proper software installed – the software required is included in the repository and a few simple commands will install it – so development can be done on any device. The project can also then be deployed wherever is best at the moment, with an eye towards eventually having it run on its own server.

2.1.2. Implementation Challenges

Key implementation challenges first included setting up the full stack of software such that a website would actually load that could be interacted with. This part of the project took up much of the initial weeks, focusing on authenticating API calls to Reddit, storing and retrieving data from the database correctly, and integrating a variety of tools. Most notably, integrating NLP tools proved challenging, since they had their own dependencies. There was also considerable work done to ensure that the program would run on a variety of setups.

The second major challenge area is in performance. Analysis has two major chokepoints: API calls, and NLP processing. Reddit's API only allows for up to two API calls per second, and limits the size of each API call fairly strictly, so a user with an extensive history can require over 20 API calls in order to gather all the necessary data. Furthermore, the "tokenization" process involved in NLP analysis can take a lot of computing power and time, especially if the user in question tends to write longer comments or submissions. These problems were tackled by setting up a complex authentication system for API calls which allows for an increased number of calls per second, by separating the API calls from the main thread of the program so that other tasks can proceed while Redreap waits for more response from Reddit, and by optimizing the NLP analysis done by the program.

A third major area of challenge was general debugging and code cleanliness, troubles common to all software projects but still notable nonetheless. Memorable bugs include one that caused analysis of users to run twice, drastically increasing response times, and edge cases in NLP or API responses that caused unexpected results or errors. Working with a four person team also meant taking special care to document implemented features and keep a structure to the overall project that allowed for simple additions of new features or editing of old ones. This meant taking care to comment code, and at one point taking the time to reorganize the main analysis code in order to make it more modular.

2.1.3. Time Constraints and Their Impact

There is a very wide variety of features we could implement in our project. Because the end goal is simply informative analysis of Reddit users and subreddits, the actual implementation of that analysis can take many different shapes. In the beginning stages of the project the team produced multiple pages of various ideas that could be implemented, knowing that we would have to prioritize and choose only those that were feasible and most valuable. And indeed, there are many ideas that will be left unimplemented.

The other key step which has been prevented by time constraints is testing a full deployment of the project online. Releasing a public Redreap website would mean serious bug testing, as there would almost undoubtedly be new issues either discovered by the influx of users or caused by them. It’s difficult to judge how much interest there would be ahead of time, but there’s a chance that server capacity would be overwhelmed since analysis is computationally expensive.

2.1.4. Required Knowledge Base/Tools for Project

Because this is a software project where all team members need to contribute code, the key knowledge required is basic programming practices. How to perform basic actions like creating functions, how to install various sets of software, how to operate in a coding environment with databases, websites, and API calls, and more are all required. Furthermore, this project also meant learning new programming tools, like NLP or various Javascript frameworks. The exact amount of learning varied from team member to team member, but all had to do a fair bit of study in order to execute the project.

2.1.5. Work Done by Team Members

The project can essentially be divided into two main stages: the group stage, and the individual stage. In the group stage, the entire team worked together to produce proposals, planning documents like the list of possible types of analysis, etc. This also extended to the early stages of the project where setting up a working programming stack and website framework to implement features on was the goal. During this stage the team generally operated by setting up specific meeting times where everyone would join a Skype call and discuss the issue of the day, either troubleshooting problems, discussing ideas, or working together on some sort of programming solution.

The second stage of the project was the individual stage, where each team member focused in implementing specific features of the project. Julian Esteban focused on implementing NLP analysis features and overall performance. Kanav Tahilramani focused on overall performance, fixing bugs, and NLP design. Matthew Chatten focused on data collection and metadata analysis. Sujay Bandarpalle focused on the front-end display of the results.

2.2. Use of Standards

a. Reddit API

   a. Documented at https://www.reddit.com/dev/api

   b. Required for retrieving data necessary for analysis

   c. OAuth2[12] – SSL protocol for API authentication

b. Stanford NLP

   a. Documented at http://nlp.stanford.edu/software/

   b. Selection of open source software for NLP analysis

   c. Project use centered on CoreNLP[13]

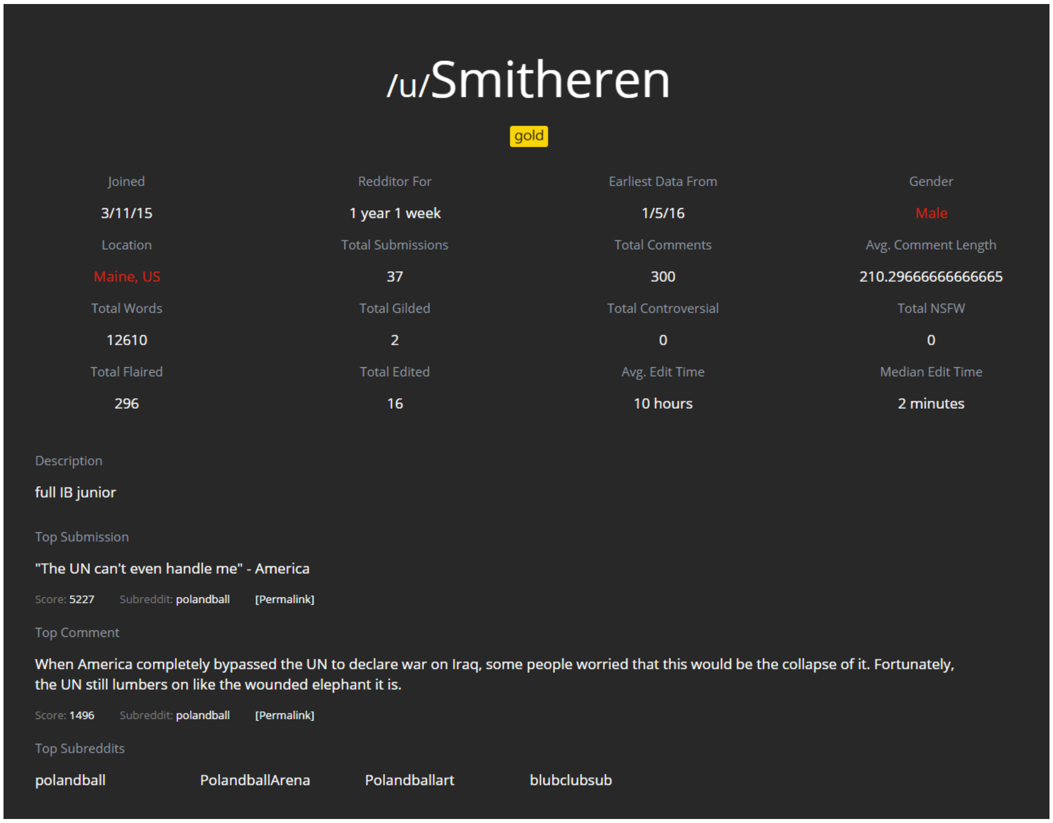
c. MEAN Stack

   a. Documented at mean.io

   b. Programming stack used for making responsive websites

2.3. Experiment / Product Results

Sample outputs:



/u/Smitheren

gold

Joined	Redditor For	Earliest Data From	Gender
3/11/15	1 year 1 week	1/5/16	Male
Location	Total Submissions	Total Comments	Avg. Comment Length
Maine, US	37	300	210.29666666666665
Total Words	Total Gilded	Total Controversial	Total NSFW
12610	2	0	0
Total Flaired	Total Edited	Avg. Edit Time	Median Edit Time
296	16	10 hours	2 minutes

Description

full IB junior

Top Submission

"The UN can't even handle me" - America

Score: 5227 Subreddit: polandball [Permalink]

Top Comment

When America completely bypassed the UN to declare war on Iraq, some people worried that this would be the collapse of it. Fortunately, the UN still lumbers on like the wounded elephant it is.

Score: 1496 Subreddit: polandball [Permalink]

Top Subreddits

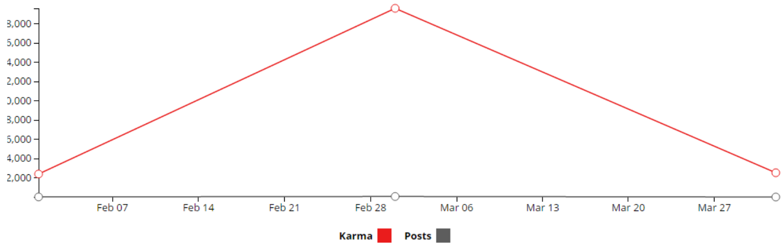
polandball PolandballArena Polandballart blubclubsub

Activity

Your account has **15712 comment karma** and **43836 link karma**.

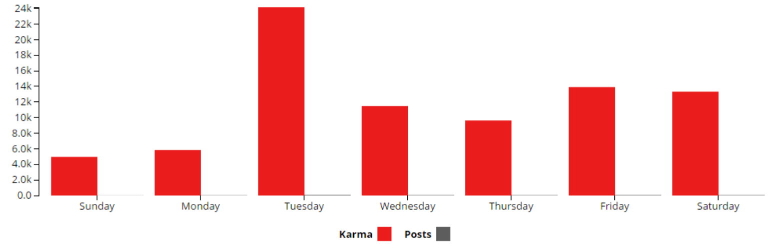
Timeline

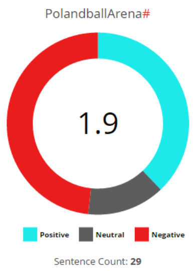
(submissions and comments combined)

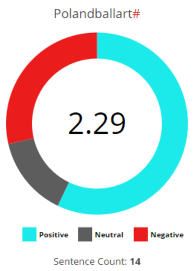


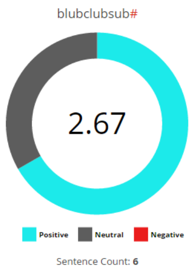
Weekly

(submissions and comments combined)









Adjectives

About **21.1%** of adjectives used are **very positive**.

perfect	great	better	better	perfect	great
perfect	worthy	beautiful	great	perfect	worthy
beautiful	happy	exciting	perfect	beautiful	happy

About **52.9%** of adjectives used are **positive**.

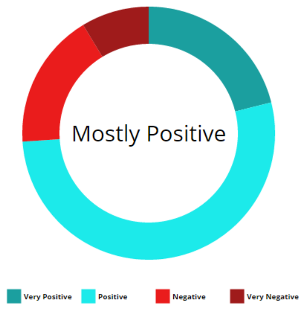
finest	new	good	full	finest	new
new	full	able	favorite	new	full
alive	rich	alive	good	alive	rich

About **17.4%** of adjectives used are **negative**.

bottom	poor	least	cold	bottom	poor
dead	violent	violent	unfinished	dead	violent
sad	bland	Nazi	bottom	sad	bland

About **8.68%** of adjectives used are **very negative**.

bad	worst	bad	worst	bad	worst
stupid	bad	worst	bad	stupid	bad
impossible	stupid	worst	worst	impossible	stupid



Language Complexity

Top 5

Polandballart#
Language Complexity Score: **8.68**
Average Sentence Complexity: **2.67**
Average Word Complexity: **6.01**

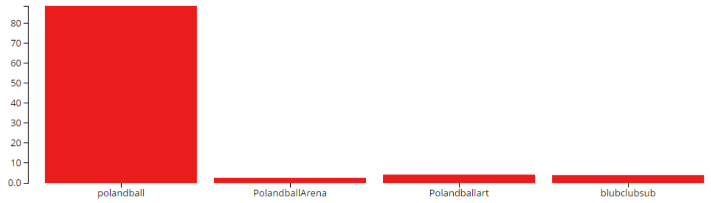
PolandballArena#
Language Complexity Score: **6.72**
Average Sentence Complexity: **2.38**
Average Word Complexity: **4.34**

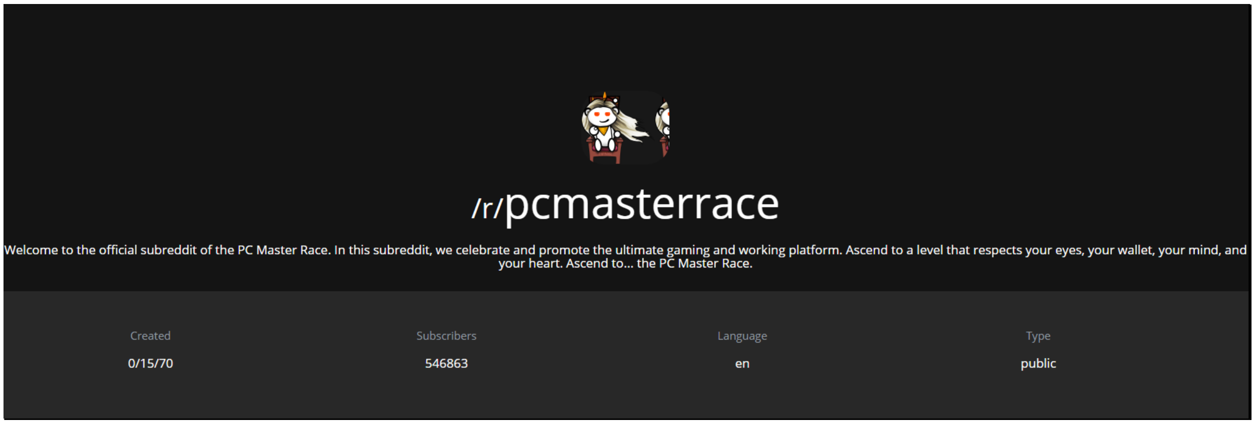
polandball#
Language Complexity Score: **6.7**
Average Sentence Complexity: **2.18**
Average Word Complexity: **4.52**

blubclubsub#
Language Complexity Score: **5.46**
Average Sentence Complexity: **1**
Average Word Complexity: **4.46**

Popularity

Top Subreddit Average Scores





/r/pcmasterrace

Welcome to the official subreddit of the PC Master Race. In this subreddit, we celebrate and promote the ultimate gaming and working platform. Ascend to a level that respects your eyes, your wallet, your mind, and your heart. Ascend to... the PC Master Race.

Created	Subscribers	Language	Type
0/15/70	546863	en	public

3. Cost and Sustainability Analysis

Even plain software has cost and sustainability concerns. Though our project did not require any upfront investment, long term sustainability is still a problem to be considered, and not just in the economic sense.

3.1. Economic Cost/Impact

Development of the project costs only the team members' time at this stage. This is not trivial, considering software developers can generally command a high salary, and many hours of work have been put into the project. At regular hourly wages, the development time cost of the project could conceivably be in the thousands of dollars. Luckily, we did not need to pay ourselves a salary, so the projects actual expenses were zero.

However, the long-term sustainability of Redreap would not be free. While it has currently only been run on developer computers, for a full deployment a server would need to be rented and a web domain purchased. The actual cost of this would depend on how much traffic the website received; websites can be hosted on rented servers for as little as $8 dollars a month[14], but due to the computationally intense nature of Redreap a large userbase might require a beefy server costing as much as $150 a month or more in the extreme case.

There is also the need to consider continued maintenance on the developer side. Bugs will have to be quashed, and the website will have to be monitored, and there will probably be a need to engage with the userbase or respond to queries. All of this will eat up valuable time.

3.2. Environmental Impact

As a solely software project, Redreap has a fairly limited environmental impact. The extent of it would probably be increased power usage by the server hosting the website, or perhaps by the computers of the users themselves. It’s not clear that deployment of the site would even cause an increase in power consumption at all – perhaps the power would be used for other tasks instead by the users, or perhaps the server farm has excess capacity currently being wasted.

3.3. Social Impact

This is one of the main goals of the project. As discussed earlier, by showing detailed analysis of Reddit users and subreddits built from publicly accessible data, Redreap should demonstrate to its users just how much can be learned about a person or a community through software analysis. Coupled with the knowledge that this sort of analysis is being done privately, opaquely, by every major social media company there is, Redreap should help spark insight into the new world of Big Data and how it affects them personally.

One notable example in our development came about from the “sentiment analysis” portion of Redreap’s NLP output. For testing purposes, one team member input their own Reddit username, and found that Redreap turned up with a fairly negative comment in their history. Finding that they weren’t comfortable with this in their history, they deleted the comment. This sort of response is likely to not be unique.

4. Conclusions / Summary

TBD – after presentation?

5. REFERENCES

[1] "Reddit." *Wikipedia*. Wikimedia Foundation, n.d. Web. 15 Apr. 2016. <https://en.wikipedia.org/wiki/Reddit>.

[2] "How The Internet* Talks." *FiveThirtyEight*. N.p., 18 Nov. 2015. Web. 15 Apr. 2016. <http://projects.fivethirtyeight.com/reddit-ngram>.

[3] "Karma Decay." *Karma Decay*. N.p., n.d. Web. 15 Apr. 2016. <http://karmadecay.com/>.

[4] "Top Comment Karma." *Top Reddit Karma Users*. N.p., n.d. Web. 15 Apr. 2016. <http://www.karmawhores.net/>.

[5] "Download Reddit Comment History." *Graph Reddit Comments*. N.p., n.d. Web. 16 Apr. 2016. <http://www.roadtolarissa.com/javascript/reddit-comment-visualizer/>.

[6] "Redective." - *The Reddit Search Detective*. N.p., n.d. Web. 16 Apr. 2016. <http://www.redective.com/>.

[7] "SnoopSnoo - Reddit User and Subreddit Analytics." *SnoopSnoo*. N.p., n.d. Web. 16 Apr. 2016. <http://snoopsnoo.com/>.

[8] "60 Amazing Reddit Statistics." *DMR*. N.p., 26 Feb. 2014. Web. 18 Apr. 2016. <http://expandedramblings.com/index.php/reddit-stats/>.

[9] "MEAN.IO - MongoDB, Express, Angularjs Node.js Powered Fullstack Web Framework - MEAN.IO - MongoDB, Express, Angularjs Node.js Powered Fullstack Web Framework." *MEAN.IO*. N.p., n.d. Web. 19 Apr. 2016. <http://mean.io/#!/>.

[10] "Vanak/red-reap." *GitHub*. N.p., n.d. Web. 19 Apr. 2016. <https://github.com/vanak/red-reap>.

[11] "Software." - *The Stanford Natural Language Processing Group*. N.p., n.d. Web. 19 Apr. 2016. <http://nlp.stanford.edu/software/>.

[12] "Reddit/reddit." *GitHub*. N.p., n.d. Web. 19 Apr. 2016. <https://github.com/reddit/reddit/wiki/OAuth2>.

[13] "Stanford CoreNLP." – *a Suite of Core NLP Tools*. N.p., n.d. Web. 19 Apr. 2016. <https://stanfordnlp.github.io/CoreNLP/>.

[14] "Best Web Hosting Features, Vps Hosting, Dedicated Hosting, by DreamHost – DreamHost." *Best Web Hosting Features, Vps Hosting, Dedicated Hosting, by DreamHost – DreamHost*. N.p., n.d. Web. 19 Apr. 2016. <https://www.dreamhost.com/hosting/>.