Honours Computer Science Thesis

rally, a one stop-shop for all reddit data

by

Kevin J. Eger

B.Sc. Hons., The University of British Columbia, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE COLLEGE OF GRADUATE STUDIES

(Interdisciplinary Studies - Optimization)

THE UNIVERSITY OF BRITISH COLUMBIA

(Okanagan)

April 2010

Abstract

Reddit is *the front page of the internet*, a slogan the company has coined and rightfully lived up to. It is a website which brings together members of all communities in a similar style to a typical forum but with much more structure and a lot more traffic. Due to the open nature of reddit, it generates a large amount of traffic, averaging over 200 million unique visitors a month. With such traffic screams the demand for data analysis through a human-interpretable medium which this thesis covers. Data analysis on reddit has been done before however this thesis focuses on bringing the data gathered in to a easily consumable format. We will explore the implementation and results of querying the reddit API, generating aggregate statistics, querying large data dumps of historic reddit data with *Google BigQuery* and the use of unsupervised machine learning to draw powerful conclusions.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Work on this thesis was widely facilitated with help from Dr. Ramon Lawrence through weekly meetings where ideas and progress were discussed extensively. It is also important to acknowledge Dr. Jeff Andrews for his support in advising on machine learning techniques which were implemented as described later.

Chapter 1

Introduction

To best understand this thesis and the work done, it is necessary to first be introduced to the relevant technologies and key terms which will be heavily referenced and built upon.

1.1 Reddit

A type of online community where users vote on content.

1.1.1 Overview

Reddit is a a news and entertainment website whose content is sustained by members of the community. Users submit text posts or direct links similar to a typical forum setting. Registered users can vote on submissions bringing order to the posts yielding an ordered online bulletin board. Furthermore, what makes Reddit unique is that content is subsectioned into different areas of interest called "subreddits". Some of the top subreddits include *movies*, *funny*, *AskReddit*, *food* and *news*. As of March 3rd, 2016 Reddit had 231,625,384 unique users a month viewing a total of 7,517,661,034 pages. The company was founded 10 years ago and has quickly become the most central place on the internet to partake in conversation or consume a wide array of content.

1.1.2 History

The company was founded by two new graduates of the *University of Virginia*, Steve Huffman and Alexis Ohanian, in June 2005 [Gua05]. After a couple years of growth, Reddit's traffic exploded and the service went viral. The creators were quick to release Reddit Gold, which offered new features and usability improvements providing the company with a primary source of income.

1.1.3 Community

Reddit thrives on its open nature and diverse content fully generated by the community [Atl14]. The demographics Reddit serves allows for a wide range of subject areas thus having the ability for smaller communities to digest their niche content. Subreddits provide a very unique opportunity by raising attention and fostering discussion that may not be seen as mainstream and covered by other news or entertainment mediums.
Reddit as a company and as a community has been known for several philanthropic projects both short and long term. A few of notable efforts are as follows:

- Users donated $185,356 to Direct Relief for Haiti after the earthquake that struck the country in January 2010
- Reddit donates 10% of it's yearly annual ad revenue to non-profits voted upon by its users [Red14]
- Members from Reddit donated over $600,000 to DonorsChoose in support of Stephen Colbert's March to Keep Fear Alive [Don10]

Bibliography

[Atl14] Ama: How a weird internet thing became a mainstream delight, 2014 [cited March 3, 2016]. → pages 2

[Don10] Welcome redditors!, 2010 [cited March 3, 2016]. → pages 2

[Gua05] A new website makes it easier to sift the mountains of news content online - and learns what you like, 2005 [cited March 3, 2016]. → pages 1

[Red14] Decimating our ads revenue, 2014 [cited March 3, 2016]. → pages 2