

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

© Kevin J. Eger, 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

Abstract	ii
Table of Contents	iii
List of Tables	iv
List of Figures	v
Acknowledgements	vi
Chapter 1: Introduction	1
1.1 Key Terms	1
1.1.1 Reddit	1
Bibliography	2

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 understand the relevant technologies and key terms which will be heavily referenced and built upon.

1.1 Key Terms

1.1.1 Reddit

This is an example citation [Mou04]. [BGLW08]

Bibliography

- [BGLW08] Heinz H. Bauschke, Rafal Goebel, Yves Lucet, and Xianfu Wang. The proximal average: Basic theory. *SIAM J. Optim.*, 19(2):768–785, 2008. → pages 1
- [Mou04] A. Moudafi. An algorithmic approach to prox-regular variational inequalities. *Appl. Math. Comput.*, 155(3):845–852, 2004. → pages 1