CS 495 - Introduction to Web Science

Fall 2014

Assignment 3

by

Eric Littley
UIN: 00821698

October 2, 2014

Instructor

Dr. Michael Nelson

Department of Computer Science
Old Dominion University

Honor Pledge

I pledge to support the Honor System of Old Dominion University. I will refrain from any form of academic dishonesty or deception, such as cheating or plagiarism. I am aware that as a member of the academic community it is my responsibility to turn in all suspected violations of the Honor Code. I will report to a hearing if summoned.

Signed: Eric Littley

Contents

1 Introduction

The task for assignment three is to download the web pages associated with the 1000 URIs found in assignment two. This data will then be used to demonstrate basic algorithms that can be used by search engines to return desirable results. The concepts and algorithms explored are Term Frequency (TF), Inverse Document Frequency (IDF), and TFIDF.

2 Design

This assignment was completed with two bash scripts: downloadURI.sh and extractData.sh. The first bash script, was responsible for downloading and parsing the 1000 URIs gathered in assignment two. The script loops through a file and downloads each URI using wget. These file names were stored in a directory called rawsites. Each file was named by the md5 hash of the URI that the file contained. A hash table was created storing the hash along with the URI associated with that hash. The script then looped through all the raw html files and used lynx to get rid of the html markup, these processed files were saved in a directory named processedsites.

The second script calculates the TF, the IDF, and the TFIDF, for the seach word "computer" using the processed URIs. The script loops through the processed files and searches each file for the term, and keeps track of the number of times the term was found. It also counts the words in each file in which the term was found. The word count and the term count are used to determine the TF of each file containing the term. The eqautions for IDF, TF and TFIDF are shown below:

TDC represents "total docs in corpus" and DT represents "docs with term":

$$IDF = log_2(\frac{TDC}{DT})$$

The denominator,"docs with term", of the IDF value was calculated using a google search of "computer" this returned 840,000,000 results. The website, http://www.worldwidewebsize.com/ was used to get an estimate of the size of "total docs in corpus" which was estimated to be about 40,000,000,000.

The TF was determined by using the following formula: TC is term count and WC is total word count of the document:

$$TF = \frac{TC}{WC}$$

The TFIDF is determined by simply multiplying the TF and IDF together. The script stored the top ten URIs based on TFIDF values. Then these URIs were copied into a page ranker website: http://www.bulkpagerank.com/. The website ran all ten URIs twice with no difference. The results were saved to a .csv file. The results of both the script and page rank website can be seen in the next section.

3 Results

3.1 TFIDF

TFIDF	TF	IDF	URI
.2965	.0532	5.5737	http://computerscienceforanyone.cbfeed...
.1081	.0194	5.5737	http://www.ebay.com/sch/i.html?_nkw=Co...
.1036	.0186	5.5737	http://www.ebay.com/sch/i.html?_nkw=Co...
.1003	.0180	5.5737	http://www.calchamber.com/Headlines/Pa...
.0975	.0175	5.5737	http://cs.utdallas.edu/graduate/grad-d...
.0863	.0155	5.5737	http://codeorg.tumblr.com/post/8926728...
.0824	.0148	5.5737	http://Code.org/
.0730	.0131	5.5737	http://technews.ninja/news/ask-slashdo...
.0646	.0116	5.5737	http://www.edutopia.org/blog/computer-...
.0596	.0107	5.5737	http://dominoaward.topplers.org/

3.2 Page Rank

Rank	URL
1.0	http://Code.org/
0.5	http://cs.utdallas.edu/graduate/grad-d...
0.0	http://www.edutopia.org/blog/computer-...
0.0	http://www.ebay.com/sch/i.html?_nkw=co...
0.0	http://www.ebay.com/sch/i.html?_nkw=co...
0.0	http://www.calchamber.com/headlines/pa...
0.0	http://technews.ninja/news/ask-slashdo...
0.0	http://dominoaward.topplers.org/
0.0	http://computerscienceforanyone.cbfeed...
0.0	http://codeorg.tumblr.com/post/8926728...

*The page rank has been normalized.

3.3 Comparison

The order in which the URIs appear in the page rank differes drastically from how they appear in the TFIDF table. However, since most of the URI's were very specific pages it was unlikely that many would be ranked. The TFIDF value calculated would most likely be a much better sreturning desired results if we constrained the search to the 1000 URIs downloaded.