CS 495: Introduction to Web Science

Fall 2013

Assignment 3

Onapha Rattanachottiteepakorn

Submission Date: October 5, 2013

**Question 1**: Download the 1000 URIs from assignment #2. "curl", "wget", or "lynx" are all good candidate programs to use. We want just the raw HTML, not the images, stylesheets, etc.

This question required the links from last assignment, so I used twitter\_links.txt as the data file. Two scripts were written for this question: downloadURI and lynxURI. The first script download the URIs from the text file and store the preprocessed files within the directory ‘raw’. Line 10 of the script is not related to the current question, but the text file created by the line will be used later on in question 2.

The second script file removes the HTML markup from the raw files. These files are stored in the directory ‘processed’.

**Question 2**: Choose a query term (e.g., "shadow") that is not a stop words (see week 4 slides) and not HTML markup from step 1 (e.g., "http") that matches at least 10 documents (hint: use "grep" on the processed files). If the term is present in more than 10 documents, choose any 10 from your list. (If you do not end up with a list of 10 URIs, you've done something wrong).

The term chosen for this question was “Newell”, the president of Valve Corporation. The script file “searchterm.sh” was written as a solution. The script runs a grep search throughout the processed directory from question 1 and then sorted with total number of term found on the file via command uniq –c and sed –r. The script then proceeds to copy first ten results into a new text file called “onlytenline.txt”. The TF-IDF values of each webpages are of the following:

|  |  |  |  |
| --- | --- | --- | --- |
| **TF-IDF** | **TF** | **IDF** | **URI** |
| 0.0053092207 | 0.0006659 | 7.973 | http://www.tomshardware.com/news/steam-box..... |
| 0.0044146501 | 0.0005534 | 7.973 | http://www.youtube.com/watch?v=1r3v19KbRvs&feature=youtu.be&a/ |
| 0.004265555 | 0.0005350 | 7.973 | http://www.bbc.co.uk/news/technology-24275896 |
| 0.0014773969 | 0.0001853 | 7.973 | http://www.youtube.com/watch?v=C4pcqSODdQA&feature=youtu.be&a |
| 0.0014758023 | 0.0001851 | 7.973 | http://www.meristation.com/pc/noticias/nvidia-trabaja-con-valve-en-el-diseno-de-steamos/58/1904687 |
| 0.0053092207 | 0.0006659 | 7.973 | http://www.tomshardware.com/news/steam-box-steamos-nvidia-valve,24405.html#xtor=RSS-999 |
| 0.0007040159 | 0.0000883 | 7.973 | http://www.gamespot.com/news/microsoft-watching-what-valve-does-with-great-interest-6414994 |
| 0.001012571 | 0.0001270 | 7.973 | http://www.muylinux.com/2013/09/24/nvidia-liberara-documentacion-gpus/ |
| 0.0020378988 | 0.0002556 | 7.973 | http://www.zdnet.fr/actualites/steamos-valve-presente-une-distribution-gnu-linux-dediee-au-jeu-39794249.htm#xtor=RSS-1 |
| 0.0142987782 | 0.0017934 | 7.973 | http://www.kudosmercadotecnia.com/steamos-valve-volvera-a-revolucionar-el-mercado-de-los-videojuegos/ |

Table 1: The TF-IDF values

**Question 3**: Now rank the same 10 URIs from question #2, but this time by their PageRank. Use any of the free PR estimaters on the web

I utilized <http://www.prchecker.info/check_page_rank.php> as this question’s PageRank estimater. The result (Table 2) was unexpected. The highest PR went to youtube.com, but the TF-IDF of the youtube webpages were the lesser of the ranks. The highest TF-IDF ranking went to a website called ‘kudosmercadotecnia.com’, a Mexican marketing agency. The PageRank for it, however, was not even available. The webpage itself, upon visited, was broken. These results show inconsistency of their ranks. I can only conclude that term “Newell” was an unstable choice for a search term, as the results returned were unreliable.

|  |  |
| --- | --- |
| **PageRank** | **URI** |
| 0.6 | http://www.tomshardware.com |
| 0.9 | http://www.youtube.com |
| 0.9 | http://www.bbc.co.uk |
| 0.9 | http://www.youtube.com |
| 0.7 | http://www.meristation.com |
| 0.6 | http://www.tomshardware.com |
| 0.8 | http://www.gamespot.com |
| 0.5 | http://www.muylinux.com |
| 0.7 | http://www.zdnet.fr |
| N/A | http://www.kudosmercadotecnia.com |

Table 2: PageRank results from prchecker.info PR tool.