HW3 - Ranking Webpages

Logan Bartels

October 18, 2020

Note

All scripts were tested by using "testLinks.txt" as the staring input file instead of "uniqueTwitterLinks.txt".

Q1

Answer

```bash
#!/bin/bash

input="uniqueTwitterLinks.txt"

while IFS= read -r line
do
  fileName=$(echo -n "$line" | (md5sum))
  curl "$line" > "$fileName.html"
  echo "$line" >> "$fileName.html"
  echo "$fileName.html" >> "fileNames.txt"
done < "$input"
ls -l
```

Listing 1: Bash script used to parse unique list of Twitter links and download their content to an output file.

```bash
#!/bin/bash

input="fileNames.txt"

while IFS= read -r line
do
  python3.8 -m justext -s English -o "processed$line" "$line"
  echo "processed$line" >> "processedFileNames.txt"
done < $input
ls -l
```

Listing 2: Bash script used to parse the list of hashed file names to open them and remove html markup.

Discussion

The script in listing 1 parses the unique list of links line-by-line. It then hashes the URI into an appropriate file name and stores it in a variable, “fileName”. Curl is then used on the link to download the content and direct the output to the value in the “fileName” variable with “.html” appended to the end of the file name. The URI is appended at the end of the hashed file to keep track of the original URIs. The appended URI in the hashed file can be easily viewed in a text editor. The hashed file name is then added to a file that lists the hashed file names, “fileNames.txt”.

The script in listing 2 reads the list of hashed file names line-by-line. For each line, JusText is ran on the file listed in the line to remove html markup. The output is directed to the same filename with “processed” added as a prefix to keep track of files put through JusText. The processed file name is then added to a file that lists the processed file names, “processedFileNames.txt”.

Q2

Answer

```
#!/bin/bash

input="processedFileNames.txt"

while IFS= read -r line
do
  count=$(grep -c 'Trump' "$line")
  if [ $count -eq 0 ]; then
    continue
  else
    grep -c 'Trump' "$line"
    grep -l 'Trump' "$line"
  fi
done <$input
```

Listing 3: Bash script used search for the term “Trump” in each processed html document; output directed to trumpFiles.txt at the command line.

```
4
processeda2138c569bd81e4ab695709d4d1356ef *-.html
11
processed80c370c72695a2973d74db93392a408f *-.html
1
processedb825be96da9e6d855c01ee546dbd8911 *-.html
16
processed9ad00d0bf6c823c662b40d926d82ee9e *-.html
```

```
2
processed320bc5a2d7a47d33847ea5a8edef304c *-.html
3
processedcc882fd241fd5ec23263921f25a1e6d9 *-.html
2
processed6b313aa7a7994100c900b688617af52f *-.html
6
processed1bff7fcc55aedab60b256e3dc5909c9d *-.html
6
processed1e5ef570c0376e21b50a468b5c9aa9cc *-.html
15
processed2c2ed100cabd5983974b5be6598ed90b *-.html
25
processedc759033e1fb3b93ddd923a131cd2e907 *-.html
10
processedb37d51d5fe2839c22cbec3f019d8a3b3 *-.html
4
processed33fbc4fcef5ab7acdec1be26f42cbdc8 *-.html
25
processed37ba13f0881c3f39c827c4d25b773345 *-.html
17
processeded27126f954b047b022aaa55951894c3 *-.html
1
processedadb449ccb4a453d7bb40680801425c63 *-.html
3
processed2140c4ee125188be7653cb1772739cdb *-.html
3
processed4941f6b20f5eca1af62aa26e0d420d28 *-.html
3
processed89a43168b11d73570e54196bf9e233ce *-.html
4
processeddd565711a0035fa838bfddb37b046b64 *-.html
1
processed18c130ac52932c8fa441fb65318907bc *-.html
5
processedfceda3a3cb4ac9dea3c5eeca16fc1786 *-.html
7
processed77e5cc94db3655e21d43dfe982a6cc27 *-.html
5
processede62734180cb216bfd35de2aabfbd29f6 *-.html
13
processed066fe13c6497a50db75601b48674bf8d *-.html
3
processeddca50095224291f1b96123075b0d7348 *-.html
1
processedfb54a96fbb663b7616966bee92c932aa *-.html
98
```

```
processed6252335432eaa0f5d355cb16845e45f8 *-.html
1
processed80eccb5c49fa159b99816c248149980d *-.html
1
processeddda15352641d88833c706c69525c0f14 *-.html
1
processedc1578e9e1e8baaa28cfbfd41c7b5445f *-.html
11
processed7ccf7a9080112db2036f9f8b5f4f3ffa *-.html
7
processedfab28e11d5b92f196ce5fc2a9f222cb0 *-.html
2
processedc2e2727ca0665df039768314a216929a *-.html
22
processed88248f8e054180cd0fcd0fc3c5b5ffce *-.html
9
processedc5f2908c47e3f074021fda33792be4c3 *-.html
2
processedd398df90267df34a57bb3dc8f22ea258 *-.html
14
processed27fc7e55f80adc7f8dbf1a5d66f78944 *-.html
5
processed351a0d20783529dfd3f441436aec432e *-.html
4
processed93aecb73aa3c3fc82c2135d7811a1229 *-.html
14
processedcd0ed0ce45016d1e968635161efd8583 *-.html
12
processed3b4e8406fdace6318c6c4e2c033bc034 *-.html
8
processed388dc0455085ccc1f23d9f940fa27bb1 *-.html
18
processedd1f1d38c7292eb0873e48375a4b60b2d *-.html
6
processed2cc65abc9f4e47291b817d0ca92f2df6 *-.html
2
processed97d03a07d6264496f0091c8e9b8f3c99 *-.html
10
processedb51538fc547963ec92d8a4f0ad10e48c *-.html
2
processedde66d72ddcefffd12ed9611d09cda72e *-.html
1
processed7582183c1cecb2b83dc605413eee85f4 *-.html
7
processedef1c18624e32ce3574c7164b7445e86a *-.html
6
processed48d2fe430a23289068c312ed69b6e45b *-.html
```

```
27
processeda3607aee498e2a68a50550a3f28234c8 *-.html
27
processed188746e94500ce6ba5a438f3cd5dcf4c *-.html
3
processed985ef698abb807fe3eae747ea9e0b542 *-.html
3
processed0424f214e622221649d657789ae9ff69 *-.html
30
processed7cc6d51310315987696521e2f32df765 *-.html
30
processedc70b07cddd9090d4a6605491d8534f7c *-.html
3
processede947ba48efc7bf5bbaaea2e741b44635 *-.html
2
processed4fabb307842dbe523cca002a5b327153 *-.html
2
processed01d3a4163f80dc8ee8f8da9df57ae589 *-.html
3
processed5255b197d0b77df9bcdd1b5a76dbccc1 *-.html
8
processedc9583650143f6d5d584dad7360e4a0d1 *-.html
1
processed61c5cbd58f22203a212ab38d30a86653 *-.html
7
processed6d24f2f63702f4915d9519dc0b872f99 *-.html
2
processed49eb2efc9738e11c1a6c3fdc9f4f0b57 *-.html
8
processed1c5a5f3f8a7225fe989328bcbf516711 *-.html
14
processed883a4c6d8e31c3022fff25e062ffd066 *-.html
14
processedadd01bf4b2ea14eaf65294229507642a *-.html
14
processed74dfd99c9e00a89b54d09f3128d94ae9 *-.html
14
processed999b32bb8efe4cbbff081758999fbe97 *-.html
14
processed2b85efe174c95d26281ca03bfc3b3097 *-.html
4
processed734d675a13cbfbc0d852f8e008db5a79 *-.html
7
processed11cb82c8297ad54b94a49eb6b7fc9d7b *-.html
13
processedf800de7ae3eb7e05ec463a5ebabb1e7a *-.html
1
```

```
processed0abb61c1aa5ac0a3b48532e9a42efcc1 *-.html
1
processed8e3555620fc376ba1bf0b3a76ddd6ad6 *-.html
5
processeded96973fcaf868c9a899580364O2f477 *-.html
19
processed6655875af1819bd693b8fd8aba42cc43 *-.html
36
processedec66548508fc94909f6bf0f9271c67fc *-.html
6
processedbf8f1434e4bae4a5e9b96dbd56da264a *-.html
8
processed9584dce59cc1a9ff0634e4aa79d0aad1 *-.html
36
processeda450ee0438c6e0ce87d1cdedf06211e9 *-.html
2
processeda6f5e59a8aa34afa940a9acbdd9e71b4 *-.html
5
processed3b9b69cb90fb06621a874ef462e6f824 *-.html
5
processed34651d6b86a1de371719697fb79cf679 *-.html
9
processed401e807be373fa108141d2c043e354fa *-.html
21
processed4f35ee3e81f1a46ab2fea8e20fda0211 *-.html
2
processed8e8ee481de8ad70e57da460ed7898792 *-.html
10
processedf409ed5d209e96df462851cfc33ca432 *-.html
10
processed33f11252982c160086616b53858199da *-.html
8
processedde4a7604a17932b2423b4b2f48f5c7dc *-.html
26
processed605be833312b218f6d71a5833060ba71 *-.html
2
processed5da9777b7c4746d20ddaca94ae06c393 *-.html
2
processeda8e5f6d9c79a89df5803b7f25e7ec781 *-.html
12
processede0ddd1a0688b4778478154926c4ac5fd *-.html
5
processedd16364392d9c8d855becd15563e72710 *-.html
2
processed0714fc3b62d65c90c6ae8f47497dbdcf *-.html
9
processeda65835504679ba94b2e90f2802e31f8d *-.html
```

```
1
processed0b6fdb71bf716afcf63313e4f8c845b3 *-.html
9
processedb03a0bfea8cdf47cde4f6208b3cdd561 *-.html
1
processed3593bd23cf1cf6d546929e96cdf47821 *-.html
9
processedf2ff16a5055fd3c6ef14b516f8db1675 *-.html
1
processed5bdfa33d0158293f52425f9a59fc6e22 *-.html
1
processed0772518e7ca0de2779cb0455b848e7b5 *-.html
1
processed7f7a64597b2549fdf70f6bb2f4cc6e91 *-.html
1
processed93684e6b920bc7e7b368a6d7e640b51b *-.html
2
processed7263a12400702825ea7eb72e4590a419 *-.html
7
processed3e3bf924132391695188df9d7e6c2cc3 *-.html
2
processedd5d6a870ef1bbd5aa9e3ffa588eba762 *-.html
20
processed4c7f2f93d64ed0be25e8f973a91189cd *-.html
1
processedb8e48eaf071117eeff6daa1c85742a63 *-.html
10
processed668498d8e8ae38656bfb4ef11ead4b36 *-.html
5
processed0a1a81f3af871633d6097aa714b05b0a *-.html
6
processed0dfde3df2a7994b1cb55b8f46f6d4bc8 *-.html
7
processed798fb1547ed1aba3839d28fd08338a08 *-.html
1
processed7e786a4d51d2481c636a9fbb8d5ded8f *-.html
4
processedad4fb130b63271267f5257903962b963 *-.html
19
processedfdd62fe4a35ea88f39c36e073249aa6f *-.html
154
processed7f2cbb78c4a8c37cac78d0f4c70458e0 *-.html
3
processed20a10076cb9a7c72bd0a477a26fea0ea *-.html
4
processed6cbeb8272f605328351644ab1fde3519 *-.html
17
```

```
processed21b980231398b5afc3efdd9a17f1daf5 *-.html
11
processed9f1eaf0cb9164f3c52a60b3269cab6a3 *-.html
5
processedb0f4df9acaab3848756e392e712ed4a1 *-.html
2
processed6be9f004e252ba9ab1c6bdfd6b7bdf13 *-.html
3
processed5c46da0f39b8122c30cb8777698513b2 *-.html
1
processed0634841493ac7eb2e3e61cfa7ea4e534 *-.html
11
processeda7df78bf4a62b774c10cf82996f9cc8a *-.html
12
processeda68ebbaa8158d511f1a6cb2429fd06e6 *-.html
2
processedf02accb7fe56597bfbb1444e44b3eb95 *-.html
11
processed29def6cc29230ac560eacda0b921dcc7 *-.html
5
processed69ae6aafb63a06c11c21e4079669dde8 *-.html
5
processedc6d71f3f4580184d3e8f01f9978a5746 *-.html
5
processed563ffd3cb57af290bef84d41e67b1f47 *-.html
4
processedf1f3651e7b20f75e9777f8107ad95d7e *-.html
4
processed522b4e15a87b3129b66da1780caf25d2 *-.html
8
processed8f6d0483521d0b68156ac09fa8c55e4d *-.html
8
processededede3858df6a513b80bef53daff37b24f *-.html
1
processed4ba6529df092e33587b6dae9cb6b36f5 *-.html
1
processed907c96a92c9af620cbf93bdaa178bea5 *-.html
5
processed782e7bd1c76d6c934abc84f06086c257 *-.html
6
processed41405da02da77ee74770581e79bec2bc *-.html
58
processed4c9110e53509e6dedb3c7201c4a2c95d *-.html
19
processed0f77a437437082e69b15aab3425e3563 *-.html
18
processeddaf4fd4ea95929781721ad7c0a685419 *-.html
```

```
7
processed62c05f9c957f3c578abe300041fa2b22 *-.html
7
processed7ca86cbfa4b70547394a1a51542ae006 *-.html
18
processed35c26a39e832dbc6040b1227e79819be *-.html
13
processed160f6cc5588d9be0e51074c4eddb8672 *-.html
1
processede945f074cef69efee0e875a0534defa9 *-.html
1
processed918898f319f7cb57a1d913769a77ecff *-.html
78
processedd27fcee32d99bceb4c1cb9a2000e2d5f *-.html
1
processed35421f1d666623f9492c7908b7c9390a *-.html
12
processed53db1fcbb638564a61b8f7d7fd8f5dcd *-.html
4
processede659bc80119f4f0315448a145779f9b4 *-.html
10
processed49bb767f32d2ed9abe5b96ebfc2fba56 *-.html
5
processed9c60a6762bc006e73ab2eb353df2c5ba *-.html
4
processed2a73b186fcc4310d7bd3cd8fb9ca96f0 *-.html
6
processed017edec92055e26f2ec1efcf4b1314a5 *-.html
19
processed8fab5d19043a4551f168931286e76134 *-.html
8
processed97a3c2a73625ed22d2c787690709059a *-.html
7
processed47c9cf197f1541831ba76f4af3d5606b *-.html
2
processed650e79cbd1a8e85fd1a59bfca4e19cac *-.html
2
processed938cfe2bc7ce5acd1d3f601adf1bf4f9 *-.html
1
processed8e46fe83c5783297ae3f7b3832e5efff *-.html
8
processed31c31ebfe8c1c7c8903101069f029efe *-.html
1
processed8f11f1db8650ae8e729a9155a464d6fc *-.html
4
processed4572f21ddba36437ca512db58c3a0fc1 *-.html
9
```

```
processedb915cca3f96e5d84ee9a6202a8a59c8e *-.html
13
processed428111085e13129c2983dfef9d3886f9 *-.html
7
processede4792b80c1adcb385af177b8173a94fa *-.html
26
processede56063df5f5c72bbb389e946925be9c1 *-.html
2
processed33c2fe1206e6dfa896aab30e19c6b220 *-.html
29
processed3181831e7679f961a44ec29b6e4c672c *-.html
13
processed98be348d214b19eb9fc77fb6701ce5d6 *-.html
5
processedbd7c8e65453bd61f8a92fac938cb73ba *-.html
14
processedab96e36150f99555e6d859791a99d23c *-.html
6
processedfac3a042cc76842b443302d44714ffbd *-.html
13
processed3326d54bab2b997198cb8dbc94d1702c *-.html
31
processed6eda695dd1cc38ffc141581ba88636bf *-.html
```

Listing 4: Output of files containing the term "Trump"; each file is preceded by the number of times "Trump" appears in the file.

```
processed62523
processedc7590
processed80c37
processed9ad00
processed2c2ed
processedb37d5
processeded271
processed066fe
processed7ccf7
processed27fc7
```

Listing 5: List of files chosen to compute TF-IDF values for.

List of URIs referenced in table 1.

1. `https://finance.yahoo.com/news/win-presidency-001318552.html?soc_src=social-sh&soc_trk=tw&tsrc=twtr`
2. `https://apnews.com/article/election-2020-virus-outbreak-seniors-florida-michael-pence-b8bbfd3a87dc290a84b9ed9915a4cf63`
3. `https://abcnews.go.com/Politics/candace-owens-blexit-group-pays-`

attendees-travel-trumps/story?id=73531036

4. https://apnews.com/5e833a62e9492f6a66624b7920cc846a
5. https://apnews.com/article/election-2020-joe-biden-donald-trump-pennsylvania-lawsuits-15e9dfeede4ddee5086611f0dd7b63a0
6. https://apnews.com/article/virus-outbreak-donald-trump-a6c145029afb7a28
7. https://apnews.com/ed5453fa2078982dba31919b8c1e274f
8. https://dailycaller.com/2020/10/09/commission-presidential-debates-board-members-anti-donald-trump/
9. https://hillreporter.com/moodys-analytics-a-joe-biden-presidency-would-create-7-million-more-jobs-than-trumps-81808
10. https://news.yahoo.com/judge-battleground-state-tosses-trump-220917951.html?soc_src=hl-viewer&soc_trk=tw

Table 1: TF-IDF values for "Trump" using Bing as the corpus

TF-IDF	TF	IDF	URI Number
0.2	0.026	7	1
0.1	0.020	7	8
0.1	0.017	7	10
0.1	0.016	7	5
0.1	0.015	7	7
0.08	0.012	7	2
0.08	0.012	7	3
0.08	0.012	7	4
0.08	0.012	7	9
0.07	0.01	7	6

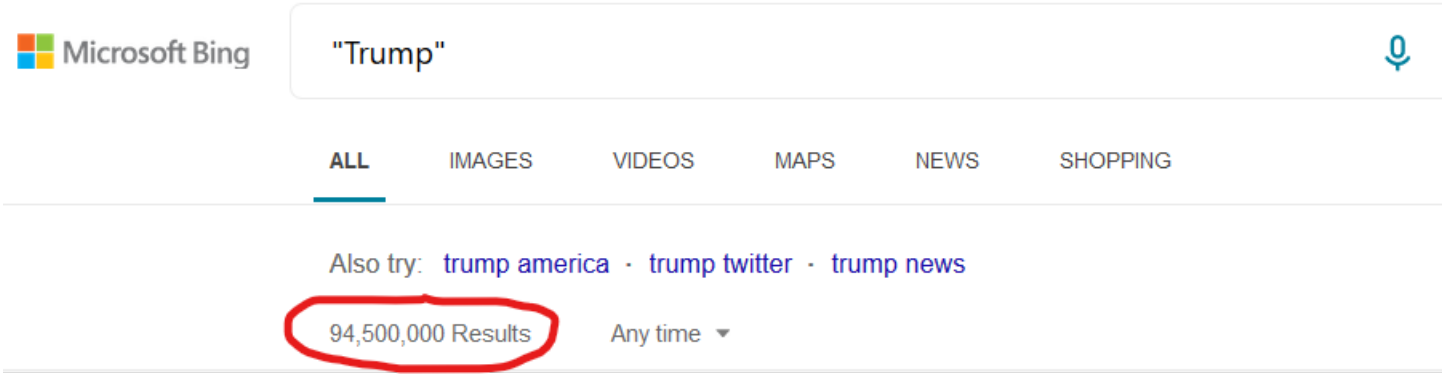

Figure 1: Results from the query "Trump" using Bing

Discussion

The potential candidates for computing TF-IDF values were determined by using the script listed in listing 3. The script in listing 3 reads the list of processed files created by listing 2 line-by-line. For each line, the script counts how many times "Trump" appears in the file listed. The value is stored in the "count" variable. A check is performed to see if count=0. If it does, then the iteration is skipped. If it does not, then "grep -c" is performed again, as well as "grep-l". This format shows how many times the term "Trump" appears in the file, and the file name. The output was directed to "trumpFiles.txt" (shown in listing 4) when the shell script was run at the command line.

The list of files used to compute TF-IDF values are shown in listing 5. The filenames are only listed to the point of being able to complete a successful search in File Explorer on Windows, or being able to successfully use tab-completion in bash.

Term frequency was calculated by using the "grep -c" command, and dividing the returned number by the value returned by the "wc -w" command used on the processed html document. The term frequency value in the table for each URI is rounded to two significant figures, with the exception of URI number six, because the numerator in its TF calculation only had one significant figure. The IDF was determined by dividing ten billion by ninety-four-million-five-hundred-thousand (shown in Figure 1). Since ten billion only has one significant figure, the final IDF answer in the table was rounded to seven. The TF-IDF values for each URI were calculated by multiplying the TF and IDF values in the table for each URI. The TF-IDF values for each URI in the table are rounded to one significant figure, each. All calculations were done manually with a TI-84 calculator.

Q3

Answer

Discussion

Page ranks were determined by `https://dnschecker.org/pagerank.php` Given that my ten URIs point to news websites, it's not surprising that their page ranks are in the 6-8 range. Five out of the ten URIs changed places. Notably, URI 10 moved up 2 rows, URI 1 moved down 1 row, URI 8 moved down 1 row, and URIs 6 and 9 switched places.

Table 2: Page Rank Values

Page Rank	URI Number
8	10
7	1
7	8
7	5
7	7
7	2
7	3
7	4
7	6
6	9

References

- JusText GitHub, https://github.com/miso-belica/jusText
- module 'cgi' has no attribute 'escape', https://github.com/trustedsec/social-engineer-toolkit/issues/721
- How to Use the grep Command on Linux, https://www.howtogeek.com/496056/how-to-use-the-grep-command-on-linux/
- How to Create/Write a Simple/Sample Linux Shell/Bash Script, https://www.instructables.com/How-to-Write-a-Linux-Shell-Script/
- HowTo: Bash For While Loop Through File Contents Script, https://www.cyberciti.biz/faq/linux-unix-appleosx-bsd-bash-loop-through-file-contents/
- Linux/UNIX: Bash Read a File Line By Line, https://www.cyberciti.biz/faq/unix-howto-read-line-by-line-from-file/
- Bash Beginner Series #8: Loops in Bash, https://linuxhandbook.com/bash-loops/
- BASH command output to the variable, https://linuxhint.com/bash_command_output_variable/
- Bash conditional statement, https://linuxhint.com/bash_conditional_statement/
- Shebang, https://bash.cyberciti.biz/guide/Shebang
- Writing output to files, https://bash.cyberciti.biz/guide/Writing_output_to_files
- Significant Figure Rules for Logarithms, https://laney.edu/cheli-fossum/wp-content/uploads/sites/210/2018/01/Significant-Figure-Rules-for-logs.pdf

- Significant Figures, https://courses.lumenlearning.com/introchem/chapter/significant-figures/
- Significant figures: Rounding and decimal places, https://en.wikipedia.org/wiki/Significant_figures#Rounding_and_decimal_places
- Pagerank Checker Tool, https://dnschecker.org/pagerank.php
- Week-06 Slides, https://docs.google.com/presentation/d/1rKj7Qdpz0GH0lHl0WYk41Lhi0C612j_0hjm4j5X8/edit#slide=id.g32fc6d3dd1_0_4