# **CSCI 572 Spring 2017 Homework 4 Report**

Name : Arpit Nigam

Email : [arpitnig@usc.edu](mailto:arpitnig@usc.edu)

USCID : 1933743387

1. Steps for completing this assignment:
2. Started with initial setup of Ubuntu environment with proper dependencies for Solr, Java, PHP, Python etc.
3. Created a new core for this assignment called “myexample” using below command:

bin/solr create –c myexample

1. After creating core modified managed-schema which exists under /myexample/conf

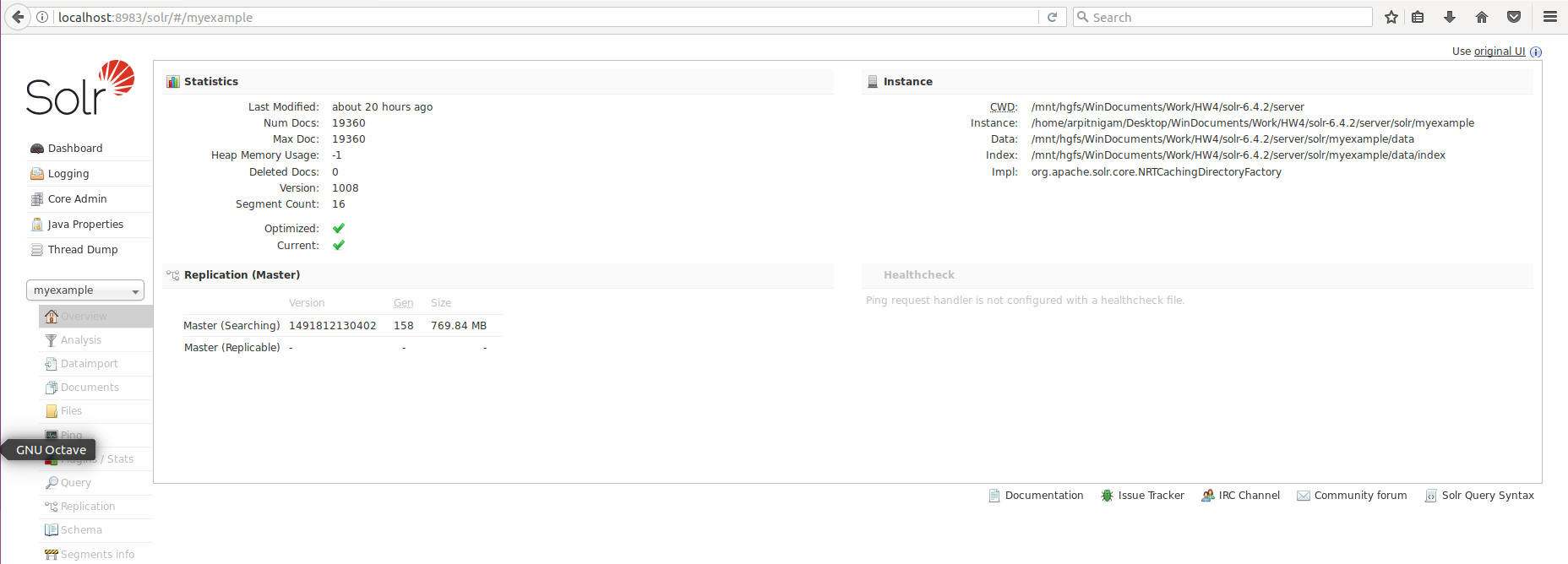




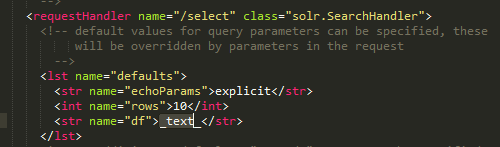
1. After this I indexed all the downloaded webpages. Out of 19362 files 19360 files got successfully indexed using below command:

bin/post -c myexample -filetypes html ../NBCNewsData/NBCNewsDownloadData/NBCNewsDownloadData/

1. After indexing my Solr UI started working on <http://localhost:8983/solr/#/myexample> and I tested it by firing some dummy queries. Below is screenshot of the Solr UI:



1. In next step, I made changes to the /myexample/conf/solrconfig.xml to that I can default query on \_text\_ field which we created in Step c.



1. In next step, I created a web app using PHP and Solr-PHP-Client that given users to search via web interface and it return the search result in a pretty format that includes:
   * + 1. Title (clickable and takes you to page on web)
       2. Id: unique identifier used while indexing data
       3. Description
       4. External URL (clickable and takes you to page on web)

Screen shots attached separately

1. In next step, using java program and one external jar for **JSoup** I extracted out all the links from the downloaded webpages and generated a file edges.txt which consists various entries and each entry correspond to link from one page to another.
2. Once this edges.txt file was ready, using Python program and one of library called **networkx** created a new file which consists of page rank score corresponding to each downloaded page which was there in edges.txt. Below is the screenshot of code snippet which shows various parameter that I have used for generating the page rank score.



1. In the next step, I modified the managed-schema again so that this page rank score can be used to get search results. Below are the changes

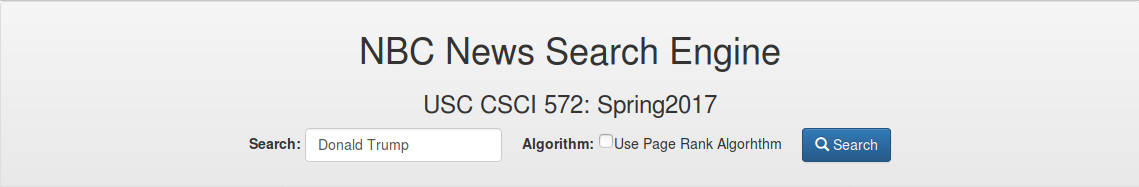


1. Similarly, I modified the solrconfig.xml and added 2 new listeners that will be using while using the Page Rank algorithm



1. After this did some tests on Solr UI by passing parameter for sort field as **pageRankFile desc**
2. Altered the web app so that page rank algorithm can also be used for searching as optional check box.

If checkbox is not selected it uses internal algorithm (Apache Lucene) and if it is selected it uses the Page Rank Algorithm to display search results.



1. Below is the table for top 10 results from internal and page rank algorithm corresponding to each query. Right after top 10 results there is one row which tell the number of overlaps between the two algorithms.

|  |  |
| --- | --- |
| **Internal** | **Page Rank** |
| 035b6cbb-904d-4dcb-acc1-b39b55555dd9.html | 1a5004b1-54c4-42df-8b32-0554c23a8e56.html |
| 2c45d221-46c9-4cf3-920d-c3a87485dfb8.html | 1a84f4cf-3476-49c9-9a70-42bebb1e2b8e.html |
| 61238700-526c-4e3f-9f71-69d152cc0d87.html | 1a865507-26e5-4c86-8162-4a074b2e0070.html |
| cacc07c9-689c-4d38-819f-bbc797bdc7f6.html | 1ac15740-29d6-483d-92d1-de85aea4668c.html |
| dc3ce7bc-56b0-4dd5-b2de-f421d8167c5c.html | 1af4372e-69ee-41c1-a1ed-3eb297f28b9d.html |
| 46550da7-0f82-42bc-840b-e533e1d10a74.html | 1b19831d-3e0c-4f0c-934f-a3704ec4822f.html |
| 4af864b4-e543-4ccb-b5ed-aff04ac2ba0c.html | 16e89cb5-4324-4fc8-b316-bd5c4413e343.html |
| 44659258-cfcb-4edb-ba77-6018c26bae37.html | 17903f12-efd6-40f9-be21-bdf53a58abec.html |
| 2f73e6ee-2725-4f3c-874e-bb13fc5c2ad7.html | 17c16e35-c51c-4585-a25b-6b186e4e5e24.html |
| ea483d6c-53c2-476f-948d-97d6cd0cb394.html | 197646ef-afff-474b-95a1-97ba83576e7e.html |
| **Brexit Overlap** | **0** |
| 619371a9-8b61-4931-9f10-2aa323789d72.html | 1a2e0b7f-be86-4dea-bdce-b12e3cb8d6b7.html |
| f77ce5da-0913-44a7-b8ef-a1357a0dae66.html | 19797df0-5fef-442a-86fa-44006b094c0d.html |
| 0ece91f9-a3bb-43c8-8f94-e1fa30daa331.html | 19ab6407-1718-40d5-a097-5540140efc5e.html |
| ffddfb44-e65b-4fb6-9a2f-630262352dfe.html | 5125c2cc-e0cf-4b73-a8ca-0f9727b48aba.html |
| 6931aa4a-3275-42d4-8075-73ee419a1d4e.html | 4f17ba71-016e-4728-a45a-39b4e371690b.html |
| 92f16f45-b494-4494-9c6e-6eb420bf96b0.html | 5217deef-cb4e-475b-b914-f752348b54b3.html |
| 92b21a9a-d9c3-4275-ab92-2c17c6ef3522.html | 540e1b02-7eb7-4411-9660-68553ea5151e.html |
| 801de5e3-3c70-46f5-9854-20909c366d8b.html | 1b9de6a3-b481-48b0-9362-3be9a301ba04.html |
| 47353bda-6fe2-4cf8-8431-0643863e4001.html | 498490ba-1424-4973-b1ee-1d376d597da2.html |
| 14aa1fbe-c131-43d9-a22e-bbe7f05e9a87.html | 47353bda-6fe2-4cf8-8431-0643863e4001.html |
| **NASDAQ Ovelap** | **1** |
| 2d79f0ea-4ccf-4bd7-be75-26d64a67fee8.html | 1a0182cf-93cc-4342-bcce-471e6967a733.html |
| 67dbcb60-1e63-42d0-bd4d-4a988236c493.html | 1a865507-26e5-4c86-8162-4a074b2e0070.html |
| f2d01e8b-5f47-4a93-a2db-dc09b5b6aeb6.html | 1b0d382f-e05b-467a-8819-8006e8b0dec2.html |
| 23dd7a9a-0edc-429f-8a17-76f6769e9528.html | 16ab4c07-96fb-42a4-ac10-0c33db8d0500.html |
| 0eb00e96-d3fb-469d-a91c-6b6ca4d724b2.html | 16bc3ebe-8f48-4a82-9efb-1b49b263189d.html |
| 8923f191-028a-44ad-b09c-e1c599b1cd28.html | 16ce6a88-0477-41fd-98bb-cd15b0a3b2e0.html |
| 24afcc01-1665-4e74-9b1a-87a343ad3e4d.html | 17f4c209-c491-4185-b303-042cd3bccda9.html |
| b8a4e7d0-1070-484f-bc35-315f04ebeaff.html | 18ae1128-4dfd-4b38-9fa3-ef3cbcc1c283.html |
| 5a786e10-ccac-4e7a-9e1e-58647b285d12.html | 194e4c8a-a12d-4cb1-b424-6b2181f962f9.html |
| e32397a2-a09f-4135-bd67-46e9ece82bb9.html | 1bd8d02b-378a-4ed2-ac76-3e31d6037ff2.html |
| **NBA Overlap** | **0** |
| d7f88191-453c-4cc1-8cad-25c9f2541bd0.html | 1a433560-2bdb-47aa-9d79-a660d242235f.html |
| af7031ce-74be-4b88-81f4-84056b26206b.html | 16b7b1ba-e51b-430b-83e5-edb5698c5de7.html |
| a10ab0ed-e53d-4b4f-8508-cfbd1577ad32.html | 182018a7-d7d4-4c7c-9da4-26b31811b176.html |
| a4c6131e-e962-4d9b-8808-649ca9030de7.html | 18d752f9-2229-42b6-aaa8-ca3b12c99587.html |
| ac902771-644d-4027-93ee-39fdb5ff8a4b.html | 195818ee-1cd2-4752-bdef-c17f74ee304c.html |
| a431a5d1-ec3c-4e7d-9f29-faa727fa76d8.html | 15839935-9ba5-4ddf-9826-ffb6c25017a9.html |
| e0c17f48-c22b-4399-996a-faab4ea3c087.html | 16089f9e-e528-4b24-9fee-b90664d22c3a.html |
| e63b1d46-2084-4634-8b59-76c7487ab9a0.html | 03cfade2-ebb4-4398-aa04-baf671d9bd21.html |
| e7b0bcc8-260b-438d-8950-41ec392e8905.html | 0057568c-1b23-463d-b655-ef822dfc40ae.html |
| 85f874d4-1527-49d9-9386-ac088edd6177.html | 0099071f-0358-4613-bd5b-9ef866b4be18.html |
| **Snapchat Overlap** | **0** |
| 2c42b257-1f91-4eae-8cc5-0f41effbee29.html | 19e0f2e1-fc62-4650-8d45-2d404b90b88c.html |
| e3b635f2-487a-42c8-bfe3-97a1afa6b44d.html | 19e6a74d-9eca-48d1-ac0c-b5d159e4f078.html |
| 25b8a702-17f0-4c4f-9d9b-8c347de2b066.html | 19f19bf6-525c-4715-ab0b-dc73d2f13395.html |
| d27d0384-616a-4588-af26-a21e0b5be5d1.html | 19f3b300-30e8-43aa-817a-294b05a876c7.html |
| 70298026-cb94-4740-81aa-a5ce9c45b38e.html | 19f9927b-04de-4792-88c0-d8ed6e7805b1.html |
| e1bfb199-b5e3-4413-9b88-eea733a94ba2.html | 19fc2f74-7a2a-487a-9f98-ad37f89d2ec4.html |
| a51bc697-8d5f-41f0-8cfd-0c103cecdca5.html | 1a19c2be-c4b9-46f1-b6d1-023421b85438.html |
| 297635ec-dda2-40ca-85b9-632114a1760f.html | 1a1ea222-ca26-4408-8bee-037001f11a07.html |
| a63ae513-fe56-4540-a839-70f9c8e05251.html | 1a2bc7b1-d5d8-4a0f-9c86-421e82a8e98f.html |
| 9f255dda-8527-4068-96e6-bd4593471bde.html | 1a2e0b7f-be86-4dea-bdce-b12e3cb8d6b7.html |
| **Illegal Immigration Overlap** | **0** |
| 33eb2dbc-3efa-44d5-87ff-12ae64ffcbe0.html | 19e0f2e1-fc62-4650-8d45-2d404b90b88c.html |
| 8e7351b7-6999-427a-bf70-7381b52bf199.html | 19e6a74d-9eca-48d1-ac0c-b5d159e4f078.html |
| 76589e3e-97ef-4f4c-a643-3bfb1420c71d.html | 19f19bf6-525c-4715-ab0b-dc73d2f13395.html |
| ff8826e8-5174-40e0-b999-4efc22adaef5.html | 19f3b300-30e8-43aa-817a-294b05a876c7.html |
| 56317d55-8606-475f-9659-2a1fdd48a589.html | 19f9927b-04de-4792-88c0-d8ed6e7805b1.html |
| 2517d47e-70bb-458d-940f-425878123bde.html | 19fc2f74-7a2a-487a-9f98-ad37f89d2ec4.html |
| 03ed641c-8d2d-4a60-8f06-6ddd9ba18d66.html | 1a00ad0d-8190-4d01-8014-516114d1ab44.html |
| 61e38367-e9db-4129-b625-7c7859277a30.html | 1a017792-66cb-40b0-b252-1675f78036a4.html |
| 06d66823-dbed-43a7-a437-07a2a8a8f658.html | 1a0182cf-93cc-4342-bcce-471e6967a733.html |
| c8300a0c-3490-4e0b-ad21-7ddc5e16bd45.html | 1a0bf444-cb37-480a-9222-d215e87871ee.html |
| **Donald Trump Overlap** | **0** |
| b7e88093-ed7b-4751-9503-af2cc8ea814a.html | 19e0f2e1-fc62-4650-8d45-2d404b90b88c.html |
| 36474563-085a-4c2f-9b34-caa1bf29c106.html | 19e6a74d-9eca-48d1-ac0c-b5d159e4f078.html |
| 70fdd57d-67cb-4c4f-a5b0-b6d699637ea5.html | 19f19bf6-525c-4715-ab0b-dc73d2f13395.html |
| 28cdb474-34f0-48c4-9900-c423aeed1eef.html | 19f261d1-4ee6-4655-bd31-97f0e0599636.html |
| 3f5af4c4-69cc-4d6c-a727-bc03df2ac247.html | 19f9927b-04de-4792-88c0-d8ed6e7805b1.html |
| ad3958bb-6d89-4d2d-b767-ba8819a03985.html | 19fc2f74-7a2a-487a-9f98-ad37f89d2ec4.html |
| 4c821d32-d588-4661-8bf5-931b2cfb29b1.html | 1a00ad0d-8190-4d01-8014-516114d1ab44.html |
| f0463bcb-cbdb-4819-ab09-5702b6ea6fea.html | 1a0bf444-cb37-480a-9222-d215e87871ee.html |
| 6f4cc215-2f21-45da-8011-3fec118e9b7d.html | 1a0dab55-290b-4153-af84-4cca07213f61.html |
| 9521c97d-265b-48a2-9b2f-51bd93cef532.html | 1a16a90d-7b99-43d4-8982-8ecad9c16604.html |
| **Russia Overlap** | **0** |
| 8c987f61-5620-4ae3-814e-bce029b6baaa.html | 1a2e0b7f-be86-4dea-bdce-b12e3cb8d6b7.html |
| 5cfc90c7-8949-4cb3-a6b4-82755eb5ba5f.html | 1a6015e6-f286-4743-9757-5be804719841.html |
| b17a8d1f-3f8f-4c74-897e-03a98a48d317.html | 1af7cd17-7082-4abe-a5f3-3dc9be164c8b.html |
| 8f5f6963-5d57-4e01-ad09-f8d910ccee11.html | 1b32215d-b991-4d1c-85d3-a8664b843fc1.html |
| 00e08ef4-472e-42ca-b9b9-cb30153d933c.html | 16b7b1ba-e51b-430b-83e5-edb5698c5de7.html |
| 2a2bc981-3a07-42f4-88d3-c9672bdc5771.html | 16c9b069-3dbd-46ba-99dc-2a4dd8579911.html |
| abbe6c8a-775d-411f-93cc-eb6627d756e4.html | 179ffc15-d538-4c0e-8230-b29f6a1ffb76.html |
| 8826e497-6e0f-4384-9869-bc3e8ca0f08d.html | 17ad9b22-f712-4651-8398-5e7b6de97089.html |
| f890b005-a5d5-4b0e-885b-93d263a65aca.html | 18653e05-b68a-463d-b0f6-07f0e5f5d8ee.html |
| 29caabc4-cf76-4bd2-915f-538681dd41f4.html | 1878c8b2-bdb2-4f3e-af58-4e69ec2581b2.html |
| **NASA Overlap** | **0** |

1. Below is the table and graph for Query Overlap by two algorithms.

|  |  |
| --- | --- |
| **Query Overlap** | |
| Brexit | 0 |
| NASDAQ | 1 |
| NBA | 0 |
| Snapchat | 0 |
| Illegal Immigration | 0 |
| Donald Trump | 0 |
| Russia | 0 |
| NASA | 0 |

Why Page Rank Score for some pages is high for some webpages?

Page Rank Score depends on the number of times a page is getting referred from some other webpages. Webpages which exists in the parent/higher hierarchy of the website are obviously to have more references compared to webpages which exists more to the lower hierarchy of the website.

Other Files:

1. external\_pageRankFile: file containing Page Ranks for different webpages.
2. solr.php: application that provides interface to user to fire queries, process the query and finally displays the result.
3. ExtactLinks.java: java program used to generate edges.txt files which contains all the directed edge between two files which basically represents link between them.
4. pageRank.py: python program which processes the edges.txt from previous program and generates a new file external\_pageRankFile containing the page rank score.
5. Screenshots from Web App:

* Screenshots for 8 Lucene Queries and 8 PageRank Queries
* Flow of Lucene and PageRank
* Search Result in which hover the title displays the URL of the hyperlink at bottom of page
* Actual webpage opened after clicking on the title