**CSCI 572 Information Retrieval and Web Search Engines**

**Homework 4: Indexing the Web using Solar**

**Name : Anubhav Jindal**

**USC ID : 5963113610**

**Steps followed to complete the Assignment:**

1. Installations and Environment Setup
   1. I am using a Macbook machine which has a Unix based Operating System, so I didn’t install Ubuntu.
   2. Solr installation:
      1. Downloaded the Solr Binary release as a zip file from <http://www.apache.org/dyn/closer.lua/lucene/solr/8.0.0/solr-8.0.0.zip>
      2. Unzipped the zipped file which gave me a solr-8.0.0 directory
   3. Downloaded Yahoo html files obtained from a crawl that provided to us as a Google drive link and placed them in the solr-8-0-0 directory.
   4. Solr Setup:
      1. Opened Terminal and change the current directory to **solr-8.0.0**
      2. Executed **bin/solr** start to start solr
      3. Executed **bin/solr create -c csci572** to create a core.
      4. Executed **bin/post -c myexample -filetypes html yahoo/** to index html files in the yahoo folder
      5. Accessed [**http://localhost:8983/solr/**](http://localhost:8983/solr/)and verified the working of solr.
      6. Ensured that the **<str name="df">\_text\_</str>** line is uncommented in solrconfig.xml
2. EdgeList computation:
   1. Wrote a program in Java using **JSoup Library** and HashMaps to compute the EdgeList from the provided Yahoo html pages.
   2. The EdgeList was written to a **edgeList.txt** file for further use.
3. PageRank computation:
   1. **NetworkX** was already installed on my machine as part of Anaconda framework.
   2. Wrote a python script which used NetworkX to compute the PageRanks.
   3. The **edgeList.txt** and **DiGraph** were used to compute PageRanks.
   4. The parameters used were as follows:
      1. **alpha=0.85, personalization=None, max\_iter=30, tol=1e-06, nstart=None, weight='weight',dangling=None**
   5. The results of the script were stored in **external\_pageRankFile.txt** for further use.
4. Setting up Solr for PageRank algorithm compatibility
   1. Copied external\_pageRankFile.txt to solr-8.0.0/server/solr/csci572/data/
   2. Modified managed-schema for using pageRankFile
      1. **<FieldType name="external" keyField="id" defVal="0" class="solr.ExternalFileField" valType="pfloat" />**
      2. **<field name="pageRankFile" type="external" stored="false" indexed="false" />**
   3. Modified solarconfig.xml file to reload external file
      1. **<listener event="newSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader" />**
      2. **<listener event="firstSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader" />**
5. Reloaded Solr Interface using the reload button in the Core Admin section.
6. For using PageRank algorithm instead of default algorithm use the **parameter: pageRankFile desc**
7. Created **Custom Interface** to the solr search engine using **solr-php-client**, **PHP, HTML and CSS.**
8. Ran the queries as described in the homework and computed the overlap.

**SCREENSHOTS**

* 1. Screenshot of initial page where I enter the query in input box

**A picture containing screenshot

Description automatically generated**

* 1. Page which shows the results for Lucene(Default).

**A screenshot of a social media post

Description automatically generated**

* 1. Page which shows the results for PageRank

**A screenshot of a social media post

Description automatically generated**

* 1. The actual webpage that opens when clicking on one of the result’s link

**A screenshot of a social media post

Description automatically generated**

**EXPLANATION WHY SOME PAGES HAVE HIGHER PAGERANK THAN OTHERS**

PageRank is an algorithm that is used to rank webpages in order to analyze which webpages are more important than the others. It is computed based on the number of links to a particular webpage (in-links) and the number of links out from the webpage (out-links). The more in-links a webpage has, the higher is the PageRank, the lesser out-links a page has, more is the PageRank. Also, higher the PageRank of the website which contributes to the in-link, higher will be the PageRank. Due to these reasons, some pages have higher PageRank than other pages.

**QUERY RESULTS**

|  |  |
| --- | --- |
| **Venezuela** | |
| **LUCENE** | **PAGERANK** |
| https://www.yahoo.com/news/anti-maduro-protests-continue-venezuela-slideshow-wp-132714046/photo-p-venezuela-opposition-leader-governor-photo-132714450.html | https://www.yahoo.com/news/science/ |
| https://www.yahoo.com/news/venezuela-braces-more-protests-election-slideshow-wp-143135783/photo-p-venezuela-president-nicolas-maduro-photo-143135961.html | https://www.yahoo.com/news/operation-cobra-untold-story-cia-officer-trained-network-agents-found-soviet-missiles-cuba-100005794.html |
| https://www.yahoo.com/news/power-water-scarce-looting-venezuela-185817628.html | https://www.yahoo.com/news/photos-malnourished-venezuelans-hope-urgently-needed-aid-arrives-soon-184505075.html |
| https://www.yahoo.com/news/anti-maduro-protests-continue-venezuela-slideshow-wp-132714046.html | https://www.yahoo.com/news/beto-o-rourke-jumping-into-2020-presidential-race-235309920.html |
| https://www.yahoo.com/news/engineer-satellite-suggests-fire-caused-venezuela-outage-212757391.html | https://www.yahoo.com/news/trump-a-selfproclaimed-champion-of-americans-detained-abroad-quiet-on-former-marine-held-in-russia-130000990.html |
| https://www.yahoo.com/news/anti-maduro-protests-continue-venezuela-slideshow-wp-132714046/photo-p-demonstrator-holds-sign-doubles-photo-132714790.html | https://www.yahoo.com/news/the-top-10-undercovered-stories-of-2016-100009854.html |
| https://www.yahoo.com/news/anti-maduro-protests-continue-venezuela-slideshow-wp-132714046/photo-p-people-place-crosses-representing-photo-132714223.html | https://www.yahoo.com/news/maduro-urges-peaceful-opposition-vote-venezuela-012524217.html |
| https://www.yahoo.com/news/anti-maduro-protests-continue-venezuela-slideshow-wp-132714046/photo-p-opposition-supporters-carry-materials-photo-132714045.html | https://www.yahoo.com/news/u-top-court-backs-trump-travel-ban-targeting-142429044.html |
| https://www.yahoo.com/news/anti-maduro-protests-continue-venezuela-slideshow-wp-132714046/photo-p-demonstrator-boy-wearing-helmet-photo-132714648.html | https://www.yahoo.com/news/cias-communications-suffered-catastrophic-compromise-started-iran-090018710.html |
| https://www.yahoo.com/news/anti-maduro-protests-continue-venezuela-slideshow-wp-132714046/photo-p-demonstrator-sits-discarded-stove-photo-132714988.html | https://www.yahoo.com/news/close-house-race-republican-karen-handel-gets-help-paul-ryan-220451019.html |

|  |  |
| --- | --- |
| **Senate** | |
| **LUCENE** | **PAGERANK** |
| https://www.yahoo.com/news/senate-skinny-repeal-obamacare-falls-apart-senate-floor-mccain-defects-072042708.html | https://www.yahoo.com/news/ |
| https://www.yahoo.com/news/senate-judiciary-committee-vote-kavanaugh-slideshow-wp-170920887/photo-p-senate-judiciary-committee-chairman-photo-170920546.html | https://www.yahoo.com/news/originals/ |
| https://www.yahoo.com/news/senate-judiciary-committee-vote-kavanaugh-slideshow-wp-170920887/photo-p-senate-judiciary-committee-member-photo-170920673.html | https://www.yahoo.com/news |
| https://www.yahoo.com/news/senate-judiciary-committee-vote-kavanaugh-slideshow-wp-170920887/photo-p-senate-judiciary-committee-members-photo-170920113.html | https://www.yahoo.com/news/russians-targeted-reagan-presidential-campaign-heres-cia-stop-100048764.html |
| https://www.yahoo.com/news/jeff-session-testifies-senate-hearing-slideshow-wp-191620787/photo-p-senate-intelligence-committee-chairman-photo-191620949.html | https://www.yahoo.com/news/politics/ |
| https://www.yahoo.com/news/protesters-across-country-oppose-gops-slideshow-wp-133309283/photo-p-senate-democrats-gather-senate-photo-150009242.html | https://www.yahoo.com/news/tagged/skullduggery |
| https://www.yahoo.com/news/kavanaugh-ford-testify-senate-judiciary-slideshow-wp-145913238/photo-p-senate-judiciary-committee-member-photo-213813950.html | https://www.yahoo.com/news/topics |
| https://www.yahoo.com/news/paul-ryan-premature-say-whether-house-pass-senate-healthcare-bill-174352395.html | https://www.yahoo.com/news/habitat-for-sale-an-oil-and-gas-group-calls-the-tune-at-the-interior-department-215244528.html |
| https://www.yahoo.com/news/cbo-likely-say-millions-lose-coverage-senate-health-care-bill-215135603.html | <https://www.yahoo.com/news/maverick-republican-william-weld-looking-run-malignant-narcissist-trump-1427>09863.html |
| https://www.yahoo.com/news/trump-senate-republicans-look-like-fools-dont-dump-60-vote-rule-135937103.html | https://www.yahoo.com/news/politics |

|  |  |
| --- | --- |
| **Democrats** | |
| **LUCENE** | **PAGERANK** |
| https://www.yahoo.com/news/midterms/house | https://www.yahoo.com/news/ |
| https://www.yahoo.com/news/midterms/house/seat | https://www.yahoo.com/news/originals/ |
| https://www.yahoo.com/news/democrats-rolling-2018-midterms-message-next-week-191517316.html | https://www.yahoo.com/news |
| https://www.yahoo.com/news/will-democrats-filibuster-gorsuch-republicans-dont-think-so-111241581.html | https://www.yahoo.com/news/russians-targeted-reagan-presidential-campaign-heres-cia-stop-100048764.html |
| https://www.yahoo.com/news/white-house-tries-tag-democrats-blame-possible-shutdown-203618659.html | https://www.yahoo.com/news/ann-coulter-lunatic-trump-challenged-2020-right-165049018.html |
| https://www.yahoo.com/news/democrats-hope-better-deal-message-will-resonate-190849661.html | https://www.yahoo.com/news/politics/ |
| https://www.yahoo.com/news/democrats-2018-many-candidates-many-candidates-162905320.html | https://www.yahoo.com/news/habitat-for-sale-an-oil-and-gas-group-calls-the-tune-at-the-interior-department-215244528.html |
| https://www.yahoo.com/news/democrats-vow-fight-lifetime-keep-brett-kavanaugh-off-supreme-court-051334537.html | https://www.yahoo.com/news/maverick-republican-william-weld-looking-run-malignant-narcissist-trump-142709863.html |
| https://www.yahoo.com/news/mcconnell-plan-fails-will-democrats-republicans-work-together-fix-health-care-220450895.html | https://www.yahoo.com/news/politics |
| https://www.yahoo.com/news/democrats-wrest-power-away-republicans-state-level-183309710.html | https://www.yahoo.com/news/democrat-predicts-house-will-get-trumps-tax-returns-complications-142301949.html |

|  |  |
| --- | --- |
| **Republicans** | |
| **LUCENE** | **PAGERANK** |
| https://www.yahoo.com/news/midterms/house | https://www.yahoo.com/news/ |
| https://www.yahoo.com/news/midterms/house/seat | https://www.yahoo.com/news/tagged/photos/ |
| https://www.yahoo.com/news/mcconnell-plan-fails-will-democrats-republicans-work-together-fix-health-care-220450895.html | https://www.yahoo.com/news/originals/ |
| https://www.yahoo.com/news/grim-warnings-white-house-republicans-ahead-election-171354295--election.html | https://www.yahoo.com/news |
| https://www.yahoo.com/news/democrats-wrest-power-away-republicans-state-level-183309710.html | https://www.yahoo.com/news/russians-targeted-reagan-presidential-campaign-heres-cia-stop-100048764.html |
| https://www.yahoo.com/news/republicans-rush-obamacare-repeal-and-replace-through-house-committees-233538837.html | https://www.yahoo.com/news/politics/ |
| https://www.yahoo.com/news/gop-obamacare-replacement-narrowly-passes-out-of-committee-after-3-republicans-defect-145519718.html | https://www.yahoo.com/news/tagged/when%20presidents%20lead/ |
| https://www.yahoo.com/news/will-democrats-filibuster-gorsuch-republicans-dont-think-so-111241581.html | https://www.yahoo.com/news/tagged/skullduggery |
| https://www.yahoo.com/news/house-republicans-reveal-obamacare-replacement-bill-013510100.html | https://www.yahoo.com/news/world/ |
| https://www.yahoo.com/news/down-ticket-5-republicans-finally-000000015.html | https://www.yahoo.com/news/science/ |

|  |  |
| --- | --- |
| **Patriot Movement** | |
| **LUCENE** | **PAGERANK** |
| https://www.yahoo.com/news/jeffersons-tree-liberty-blood-schoolchildren-171918999.html | https://www.yahoo.com/news |
| https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094/photo-p-wing-supporter-patriot-prayer-photo-212529706.html | https://www.yahoo.com/news/tagged/through-her-eyes |
| https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094/photo-p-wing-supporter-patriot-prayer-photo-212529025.html | https://www.yahoo.com/news/crackdown-opioids-victims-people-need-live-100058361.html |
| https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094/photo-p-counter-protester-argues-police-photo-212529993.html | https://www.yahoo.com/news/admissions-scam-adds-insult-injury-minority-applicants-202405114.html |
| https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094/photo-p-left-wing-counter-protesters-photo-212529339.html | https://www.yahoo.com/news/death-penalty-outlawed-203116351.html |
| https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094/photo-p-far-protester-receives-treatment-photo-212529076.html | https://www.yahoo.com/news/ilhan-omar-and-andre-carson-respond-to-horror-of-new-zealand-massacre-of-muslims-171412117.html |
| https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094/photo-p-counter-demonstrators-held-back-photo-212529207.html | https://www.yahoo.com/news/latest-students-delhi-protest-cleaner-143006709.html |
| https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094/photo-p-police-push-protesters-back-photo-212529889.html | https://www.yahoo.com/news/photos-in-gaza-women-walk-thin-line-between-hope-and-despair-132405730.html |
| https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094.html | https://www.yahoo.com/news/2016-pictures-news-slideshow-wp-100040397.html |
| https://www.yahoo.com/news/wing-patriot-prayer-rally-counter-slideshow-wp-212529094/photo-p-counter-protester-raises-arms-photo-212529025.html | https://www.yahoo.com/news/trump-mocks-beto-o-rourkes-crazy-hand-gestures-183443014.html |

|  |  |
| --- | --- |
| **Oscar 2019** | |
| **LUCENE** | **PAGERANK** |
| https://www.yahoo.com/news/spike-lee-pushes-back-trump-spat-continues-oscar-speech-162915413.html | https://www.yahoo.com/news/ |
| https://www.yahoo.com/news/skullduggery-power-vice-000835154.html | https://www.yahoo.com/news/tagged/photos/ |
| https://www.yahoo.com/news/topics/star-wars | https://www.yahoo.com/news/originals/ |
| https://www.yahoo.com/news/trump-bashes-spike-lee-racist-hit-oscars-123716258.html | https://www.yahoo.com/news |
| https://www.yahoo.com/news/topics/royal-family | https://www.yahoo.com/news/russians-targeted-reagan-presidential-campaign-heres-cia-stop-100048764.html |
| https://www.yahoo.com/news/kamala-harris-trump-racist-185542673.html | https://www.yahoo.com/news/ann-coulter-lunatic-trump-challenged-2020-right-165049018.html |
| https://www.yahoo.com/news/the-names-behind-the-college-cheating-charges-141742295.html | https://www.yahoo.com/news/politics/ |
| https://www.yahoo.com/news/topics/technology | https://www.yahoo.com/news/tagged/when%20presidents%20lead/ |
| https://www.yahoo.com/news/gore-us-getting-close-political-shift-climate-change-223933667--politics.html | https://www.yahoo.com/news/us-grounds-boeing-737-max-amid-growing-safety-200009938.html |
| https://www.yahoo.com/news/heres-know-orlando-victims-000000001.html | https://www.yahoo.com/news/tagged/skullduggery |

|  |  |
| --- | --- |
| **Channel** | |
| **LUCENE** | **PAGERANK** |
| https://www.yahoo.com/news/topics/fox-news-channel | https://www.yahoo.com/news/topics |
| https://www.yahoo.com/news/topics/cnn | https://www.yahoo.com/news/crackdown-opioids-victims-people-need-live-100058361.html |
| https://www.yahoo.com/news/service-charlottesville-victim-heather-heyer-call-righteous-action-172114070.html | https://www.yahoo.com/news/operation-cobra-untold-story-cia-officer-trained-network-agents-found-soviet-missiles-cuba-100005794.html |
| https://www.yahoo.com/news/topics/megyn-kelly | https://www.yahoo.com/news/skullduggery-washington-think-tank-says-no-saudi-funds-124450269.html |
| https://www.yahoo.com/news/topics/marijuana-legalization | https://www.yahoo.com/news/topics/president-trump |
| https://www.yahoo.com/news/topics/space-exploration | https://www.yahoo.com/news/rory-mcilroy-blames-pga-tour-slow-play-epidemic-004543461.html |
| https://www.yahoo.com/news/topics/health | https://www.yahoo.com/news/the-names-behind-the-college-cheating-charges-141742295.html |
| https://www.yahoo.com/news/topics/sex-crimes | https://www.yahoo.com/news/death-toll-rises-nine-collapsed-lagos-school-building-074812489.html |
| https://www.yahoo.com/news/topics/sanctuary-cities | https://www.yahoo.com/news/topics/jared-kushner |
| https://www.yahoo.com/news/topics/adam-rippon | https://www.yahoo.com/news/fire-trump-says-misspoke-russian-interference-u-s-election-192113980.html |

|  |  |
| --- | --- |
| **Wall** | |
| **LUCENE** | **PAGERANK** |
| https://www.yahoo.com/news/wall-wall-trump-defends-border-plan-kelly-suggests-changed-145144558.html | https://www.yahoo.com/news/ |
| https://www.yahoo.com/news/imagining-trumps-big-beautiful-wall-might-183350584.html | https://www.yahoo.com/news/tagged/photos/ |
| https://www.yahoo.com/news/gop-leaders-back-at-least-12-billion-for-border-wall-153042941.html | https://www.yahoo.com/news/originals/ |
| https://www.yahoo.com/news/trumps-big-beautiful-wall-collides-congress-102804089.html | https://www.yahoo.com/news |
| https://www.yahoo.com/news/border-wall-funding-threatens-government-shutdown-budget-battle-heats-180534239.html | https://www.yahoo.com/news/russians-targeted-reagan-presidential-campaign-heres-cia-stop-100048764.html |
| https://www.yahoo.com/news/on-the-border-waiting-for-the-big-beautiful-wall-225959295.html | https://www.yahoo.com/news/ann-coulter-lunatic-trump-challenged-2020-right-165049018.html |
| https://www.yahoo.com/news/sanders-calls-obamas-400k-wall-street-speaking-fee-unfortunate-150857262.html | https://www.yahoo.com/news/politics/ |
| https://www.yahoo.com/news/steve-king-has-a-model-of-the-border-wall-he-wants-to-build-to-protect-our-superior-civilization-183248446.html | https://www.yahoo.com/news/tagged/when%20presidents%20lead/ |
| https://www.yahoo.com/news/rebuilding-great-wall-china-slideshow-wp-163045771/photo-p-people-rest-working-reconstruction-photo-153045115.html | https://www.yahoo.com/news/tagged/skullduggery |
| https://www.yahoo.com/news/rebuilding-great-wall-china-slideshow-wp-163045771/photo-p-people-wait-bricks-other-photo-153045472.html | https://www.yahoo.com/news/world/ |

**OVERLAPS**

|  |  |
| --- | --- |
| **QUERY** | **NO. OF OVERLAPS** |
| Venezuela | 0 |
| Senate | 0 |
| Democrats | 0 |
| Republicans | 0 |
| Patriot Movement | 0 |
| Oscar 2019 | 0 |
| Channel | 0 |
| Wall | 0 |

There were no overlaps in any query. So, no bar graph has been drawn as per the instructions given.