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

**Assignment #3: Indexing The Web using Solr**

**Name : Naman Avasthi**

**USC ID : 9216625985**

**Steps Followed to Complete The Assignment:**

1. Since I’m using a MacBook Pro (Unix machine) running macOS Sierra there was no need to install Linux.
2. **Installing Solr**
   * As per the instructions, I visited <http://lucene.apache.org/solr/quickstart.html> and further went to the download link on <http://www-eu.apache.org/dist/lucene/solr/6.2.1/> and downloaded solr-6.2.1.zip
   * The zip file downloaded was unzipped to extract the solr directory, this directory was then moved to Assignment Directory: /Users/NamanAvasthi/Documents/USC/Fall 2016/IR/Assignments/3/solr-6.2.1
   * This directory will be referred to as $root
3. **Setting up Solr**
   * Terminal was opened and changed the directory to $root
   * Instructions mentioned on <http://www-scf.usc.edu/~csci572/2016Fall/hw3/IndexingwithTIKAV3.pdf> was followed
     1. **bin/solr start**
        + starts the solr server
     2. **bin/solr create –c IR\_Assignment3\_Core**
        + creates the new core for this assignment named IR\_Assignment3\_Core. This is created at: $root/server/solr/IR\_Assignment3\_Core   
          this will be referred to as $core
     3. **bin/post –c IR\_Assignment3\_Core –filetypes html /Users/NamanAvasthi/Desktop/crawl\_data\_example**
        + this will index the html files in the crawl\_data\_example folder which contains all the data downloaded from the google drive link provided on the course website: <https://drive.google.com/drive/folders/0B3A22b95zplldUEtU3BzT2V4WTQ>
     4. [**http://localhost:8983/solr/**](http://localhost:8983/solr/) **was checked for the core overview and default queries were tested**
     5. **Configure Solr to default querying**
        + <str name=”df”>\_text\_</str> was added/uncommented from one of the <requestHandler> tags
4. **Creating php Code**
   * The php code was written to make the form submission page with radio buttons to select between the two ranking algorithms.
5. **PageRank Ranking Algorithm Setup**
   * **Writing Java Code For Creating edgelist.txt**
     1. Before we write the code for creating edgelist.txt we combine all the data into one file.
     2. Copy all the .csv data to one file, called mastercsvfile.csv and all the html data to one folder called AllNewsData
     3. Write the code to construct map\* data structure
     4. Use the map\* files, urlFileMap and fileUrlMap to construct the edgelist
     5. Use the edgelist to create the edgelist.txt file
   * **Writing python code to convert edgelist.txt to external\_pageRankFile.txt**
     1. Download and install the NetworkX python library from <https://pypi.python.org/pypi/networkx/>
        + Unzip this file
        + Open terminal, cd to this directory and type

$ python setup.py install

* + 1. Use the edgelist.txt to construct the external\_pageRankFile.txt
    2. Use the parameters :

alpha=0.85, personalization=None, max\_iter=30, tol=1e-06, nstart=None, weight='weight',dangling=None

* + **Copying external\_pageRankFile.txt to $core/data**
  + **Add Field for PageRank in managed-schema**
    1. Add in schema as

<fieldType name="external" keyField="id" defVal="0" class="solr.ExternalFileField" valType="pfloat" />

<field name="pageRankFile" type="external" stored="false" indexed="false" />

* + **Define listeners for query in solrconfig.xml**
    1. Add the following right before updateHandler

<listener event="newSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader" />

<listener event="firstSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader" />

* + 1. Add the following inside <requestHandler name="/select" class="solr.SearchHandler"> <lst> tag

<str name="q.op">AND</str>

1. **Reload Solr localhost and run queries for PageRank**
   * **Use parameter : pageRankFile desc**
2. **For php the radio button will add the additional required parameters**
3. **Run the queries described on :** <http://www-scf.usc.edu/~csci572/2016Fall/hw3/Homework3.pdf>
4. **Construct the overflow graph for the same**
5. **Appropriate screenshots were taken along the way**

**QUERY RESULTS**

|  |  |
| --- | --- |
| **NATO** | |
| **LUCENE** | **PAGERANK** |
| http://www.wsj.com/amp/articles/nato-rejects-russian-air-safety-proposal-for-planes-in-baltic-region-1474391644 | http://www.nbcnews.com/storyline/isis-terror/pope-francis-says-world-war-after-isis-linked-church-attack-n617846 |
| http://www.nbcnews.com/storyline/ukraine-crisis/ukraine-crisis-u-s-bulgaria-begin-nato-military-drills-n322946 | http://www.nbcnews.com/storyline/isis-terror/isis-sought-nuclear-material-dirty-bomb-officials-believe-n527626 |
| http://www.nbcnews.com/storyline/ukraine-crisis/nato-will-set-commands-soothe-nervous-eastern-europe-n296946 | http://www.nbcnews.com/watch/nbc-news/attorney-michael-browns-family-doesnt-accept-dojs-self-defense-theory-408780355703 |
| http://www.nbcnews.com/widget/video-embed/734439491911?autoPlayLiveVideo=true&playerType=Onsite%20Embedded | http://www.nbcnews.com/watch/nbcnews-com/brinksmanship-or-bluster-debt-limit-debunked-140175427977 |
| http://www.wsj.com/articles/nato-rejects-russian-air-safety-proposal-for-planes-in-baltic-region-1474391644 | http://www.nbcnews.com/watch/nbc-news/protests-return-to-ferguson-412623427637 |
| http://www.nbcnews.com/storyline/isis-terror/300-isis-fighters-killed-u-s-afghan-operation-nangarhar-nato-n627146 | http://www.nbcnews.com/storyline/ukraine-crisis/vladimir-putin-says-russias-military-might-has-no-match-n309651 |
| http://www.nbcnews.com/storyline/iraq-turmoil/nato-summit-obama-faces-crossroads-fight-against-isis-n192996 | http://www.nbcnews.com/watch/nightly-news/torture-photos-from-inside-syria-123447363726 |
| http://www.nbcnews.com/storyline/ukraine-crisis/russia-launches-large-scale-military-exercises-disputed-territories-n317776 | http://www.nbcnews.com/storyline/isis-terror/iraqi-forces-enter-central-fallujah-liberate-key-areas-isis-n594241 |
| http://www.wsj.com/articles/the-united-states-must-be-the-worlds-policeman-1474412665 | http://www.nbcnews.com/news/us-news/why-does-u-s-use-russian-rockets-launch-its-satellites-n588526 |
| http://www.nbcnews.com/storyline/ukraine-crisis/army-commander-ben-hodges-ukraine-crisis-needs-more-weapons-n300856 | http://www.nbcnews.com/storyline/ukraine-crisis/ukraine-lawmaker-manhandles-prime-minister-sparking-parliament-brawl-n478836 |

|  |  |
| --- | --- |
| **Dow Jones** | |
| **LUCENE** | **PAGERANK** |
| http://www.wsj.com/articles/SB120365584684285449 | http://www.nbcnews.com/business/economy/mansions-sale-oil-s-rough-ride-hits-houston-s-millionaires-n525486 |
| http://www.wsj.com/amp/articles/stand-alone-reit-grouping-seen-unlikely-to-whet-some-funds-appetites-1474659953 | http://www.nbcnews.com/business/markets/dow-drops-100-points-surprise-rise-crude-inventories-n640721 |
| http://www.wsj.com/amp/articles/dows-buffett-ceiling-stock-refuses-to-cross-berkshires-8-a-second-threshold-1472062768 | http://www.nbcnews.com/business/markets/amp/boe-stimulus-sledgehammer-lifts-world-stocks-uk-shares-hit-1-n623441 |
| http://www.nbcnews.com/storyline/brexit-referendum/amp/brexit-fallout-stocks-set-continue-decline-after-u-k-vote-n599581 | http://www.nbcnews.com/storyline/brexit-referendum/brexit-wiped-2-trillion-wiped-markets-standard-poors-n599411 |
| http://www.wsj.com/articles/dows-buffett-ceiling-stock-refuses-to-cross-berkshires-8-a-second-threshold-1472062768 | http://www.nbcnews.com/business/markets/global-stocks-dip-fed-view-only-partly-soothes-markets-n647311 |
| http://www.wsj.com/articles/stand-alone-reit-grouping-seen-unlikely-to-whet-some-funds-appetites-1474659953 | http://www.nbcnews.com/storyline/hong-kong-protests/hong-kong-unrest-weighs-stock-markets-dow-drops-n213931 |
| http://www.nbcnews.com/storyline/brexit-referendum/brexit-fallout-stocks-set-continue-decline-after-u-k-vote-n599581 | http://www.nbcnews.com/storyline/2015-year-in-review/after-wild-ride-stock-markets-end-year-almost-where-they-n488696 |
| http://www.wsj.com/articles/foundation-taps-leader-for-philadelphia-newspapers-1474354841 | http://www.nbcnews.com/card/london-mayor-no-need-panic-n598171 |
| http://www.nbcnews.com/tech/security/irs-cyberattack-total-more-twice-previously-disclosed-n526846 | http://www.nbcnews.com/business/markets/dow-drops-150-points-weak-tech-auto-stocks-n621686 |
| http://www.wsj.com/articles/bayer-deal-for-monsanto-caps-reshaping-of-seed-and-pesticide-sector-1473864494 | http://www.nbcnews.com/storyline/hong-kong-protests/protesters-seek-slumber-hong-kongs-nooks-crannies-n219301 |

|  |  |
| --- | --- |
| **Rio Olympics** | |
| **LUCENE** | **PAGERANK** |
| http://www.nbcnews.com/storyline/2016-rio-summer-olympics/more-1-000-u-s-spies-brazil-protecting-rio-olympics-n623186 | http://www.nbcnews.com/feature/short-take/video/migrants-surge-across-the-mexican-border-587780163988 |
| http://www.nbcnews.com/storyline/2016-rio-summer-olympics/feyisa-lilesa-ethiopian-runner-makes-defiant-protest-gesture-rio-olympics-n635761 | http://www.nbcnews.com/dateline/video/somebodys-daughter-part-12-254048323538 |
| http://www.nbcnews.com/storyline/2016-rio-summer-olympics/dustin-johnson-skip-rio-olympics-over-zika-fears-n606381 | http://www.nbcnews.com/feature/flashback/video/five-years-ago-the-space-shuttle-program-ended-723717187838 |
| http://www.nbcnews.com/storyline/2016-rio-summer-olympics/blackgirlmagic-african-american-women-take-spotlight-rio-olympics-n635101 | http://www.nbcnews.com/storyline/zika-virus-outbreak/bracing-zika-tale-two-cities-n608546 |
| http://www.nbcnews.com/storyline/2016-rio-summer-olympics/rio-olympics-shaunae-miller-s-dive-denied-allyson-felix-gold-n631636 | http://www.nbcnews.com/storyline/zika-virus-outbreak/mosquitoes-can-infect-their-eggs-zika-n639646 |
| http://www.nbcnews.com/storyline/2016-rio-summer-olympics/google-searches-reveal-your-burning-questions-about-olympics-n628341 | http://www.nbcnews.com/storyline/zika-virus-outbreak/where-s-zika-most-likely-u-s-answer-may-surprise-n540281 |
| http://www.nbcnews.com/storyline/2016-rio-summer-olympics/escaping-olympics-rio-s-magical-urban-jungle-n629371 | http://www.nbcnews.com/dateline/video/obsession-part-5-758012483957 |
| http://www.nbcnews.com/storyline/2016-rio-summer-olympics/zika-virus-won-t-be-big-problem-olympics-rio-officials-n587316 | http://www.nbcnews.com/news/us-news/hacker-guccifer-2-0-releases-more-dnc-docs-including-tim-n647921 |
| http://www.nbcnews.com/storyline/2016-rio-summer-olympics/british-athlete-robbed-olympics-officials-say-n633446 | http://www.nbcnews.com/tech/video-games/sims-removes-gender-barriers-video-game-n584516 |
| http://www.nbcnews.com/storyline/2016-rio-summer-olympics/serena-williams-zika-team-usa-top-rio-social-buzz-n620876 | http://www.nbcnews.com/feature/short-take/video/how-is-security-for-el-chapo-different-this-time-601950275760 |

|  |  |
| --- | --- |
| **Pokemon Go** | |
| **LUCENE** | **PAGERANK** |
| http://www.nbcnews.com/tech/video-games/amp/how-pokemon-go-s-new-buddy-system-works-n646841 | http://www.nbcnews.com/tech/tech-news/feds-recall-1-million-overheating-samsung-galaxy-note-7s-n649066 |
| http://www.nbcnews.com/tech/video-games/how-pokemon-go-s-new-buddy-system-works-n646841 | http://www.nbcnews.com/health/health-news/new-study-shows-flu-vaccines-might-be-aiming-wrong-target-n559256 |
| http://www.nbcnews.com/tech/video-games/pokemon-go-players-fuming-after-update-resets-their-account-level-n620816 | http://www.nbcnews.com/health/health-news/silver-obamacare-lining-69-monthly-premiums-n134506 |
| http://www.nbcnews.com/tech/video-games/pokemon-go-eludes-cloning-attempts-big-game-studios-executives-n626116 | http://www.nbcnews.com/tech/tech-news/pokemon-no-go-why-driving-lizards-don-t-always-go-n616411 |
| http://www.nbcnews.com/tech/video-games/amp/pokemon-go-eludes-cloning-attempts-big-game-studios-executives-n626116 | http://www.nbcnews.com/health/kids-health/heres-reason-get-puppy-kids-pets-have-less-anxiety-n469591 |
| http://www.nbcnews.com/tech/video-games | http://www.nbcnews.com/tech/tech-news/apple-s-new-san-francisco-office-signals-change-valley-n530856 |
| http://www.nbcnews.com/news/world/amp/pokemon-go-hunters-south-korea-thwarted-war-north-korea-n608426 | http://www.nbcnews.com/feature/college-game-plan/usc-students-bring-late-night-munchies-directly-libraries-dorms-across-n618146 |
| http://www.nbcnews.com/tech/video-games/pokemon-go-creators-working-remove-problematic-pokestops-n619856 | http://www.nbcnews.com/storyline |
| http://www.nbcnews.com/tech/video-games/amp/pokemon-go-players-fuming-after-update-resets-their-account-level-n620816 | http://www.nbcnews.com/tech/mobile/no-romance-these-apps-want-find-you-best-friend-n525696 |
| http://www.nbcnews.com/tech/mobile/amp/iran-bans-pokemon-go-mobile-game-even-launched-there-n625156 | http://www.nbcnews.com/tech/video-games/sims-removes-gender-barriers-video-game-n584516 |

|  |  |
| --- | --- |
| **California Wild Fires** | |
| **LUCENE** | **PAGERANK** |
| http://www.wsj.com/amp/articles/how-the-exxon-case-unraveled-1472598472 | http://www.nbcnews.com/business/business-news/low-gas-prices-may-ease-economic-pinch-will-they-last-n589926 |
| http://www.nbcnews.com/business/business-news/low-gas-prices-may-ease-economic-pinch-will-they-last-n589926 | http://www.nbcnews.com/pop-culture/movies/star-wars-force-awakens-world-premiere-wows-fans-celebs-n480111 |
| http://www.nbcnews.com/pop-culture/movies/star-wars-force-awakens-world-premiere-wows-fans-celebs-n480111 | http://www.nbcnews.com/business/business-news/bringing-green-seaweed-emerges-new-cash-crop-maine-n473096 |
| http://www.nbcnews.com/business/business-news/bringing-green-seaweed-emerges-new-cash-crop-maine-n473096 | http://www.wsj.com/news/science |
| http://www.wsj.com/articles/how-the-exxon-case-unraveled-1472598472 | http://www.wsj.com/asia |
| http://www.wsj.com/articles/in-flood-ravaged-louisiana-small-businesses-struggle-to-recover-1474050071 | http://www.wsj.com/amp/articles/how-the-exxon-case-unraveled-1472598472 |
| http://www.wsj.com/news/science | http://www.wsj.com/europe |
| http://www.wsj.com/ | http://www.wsj.com/ |
| http://www.wsj.com/europe | http://www.wsj.com/articles/how-the-exxon-case-unraveled-1472598472 |
| http://www.wsj.com/asia | http://www.wsj.com/articles/in-flood-ravaged-louisiana-small-businesses-struggle-to-recover-1474050071 |

|  |  |
| --- | --- |
| **Donald Trump** | |
| **LUCENE** | **PAGERANK** |
| http://www.nbcnews.com/politics/2016-election/stump-donald-trumps-presidential-campaign-n453726 | http://www.nbcnews.com/widget/video-embed/769905219602 |
| http://www.wsj.com/amp/articles/donald-trump-expands-short-list-of-potential-supreme-court-picks-1474657161 | http://www.nbcnews.com/video/viewers-favorite-videos-of-2015-593784387736 |
| http://www.wsj.com/amp/articles/ted-cruz-endorses-donald-trump-1474660956 | http://www.nbcnews.com/feature/short-take/video/migrants-surge-across-the-mexican-border-587780163988 |
| http://www.nbcnews.com/storyline/data-points/poll-donald-trump-still-leads-gop-field-n490116 | http://www.nbcnews.com/news/veterans/vets-don-t-blame-obama-track-palin-s-behavior-n500716 |
| http://www.nbcnews.com/storyline/data-points/amp/new-nbc-news-survey-monkey-poll-donald-trump-still-lead-n406766 | http://www.nbcnews.com/video/president-to-crowd-thanks-obama-764158019654 |
| http://www.nbcnews.com/politics/2016-election/donald-trump-meet-mexican-president-enrique-pe-nieto-n640441 | http://www.nbcnews.com/news/latino/latino-first-ben-ray-luj-n-leads-push-more-house-n602061 |
| http://www.nbcnews.com/storyline/data-points/new-nbc-news-survey-monkey-poll-donald-trump-still-lead-n406766 | http://www.nbcnews.com/nightly-news/video/soaring-prescription-drug-prices-forcing-many-americans-to-make-tough-choices-713260099957 |
| http://www.nbcnews.com/meet-the-press/meet-press-transcript-august-16-2015-n412636 | http://www.nbcnews.com/nightly-news/video/the-incredible-moment-a-mother-hears-her-deceased-sons-heartbeat-614533699833 |
| http://www.nbcnews.com/pop-culture/tv/donald-trump-no-laughing-matter-some-late-night-comics-n622151 | http://www.nbcnews.com/storyline/hispanic-heritage-month-2016/these-latinos-will-be-honored-national-medals-art-humanities-n649061 |
| http://www.nbcnews.com/politics/2016-election/donald-trump-criticized-after-he-appears-mock-reporter-serge-kovaleski-n470016 | http://www.nbcnews.com/politics/2016-election/primaries/DC |

|  |  |
| --- | --- |
| **Harry Potter** | |
| **LUCENE** | **PAGERANK** |
| http://www.wsj.com/articles/scholastics-revenue-blessed-by-harry-potter-and-the-cursed-child-1474547511 | http://www.nbcnews.com/storyline/brexit-referendum/boris-johnson-6-times-u-k-s-top-diplomat-offended-n609126 |
| http://www.nbcnews.com/pop-culture/books/new-harry-potter-book-script-upcoming-play-gets-summer-release-n515866 | http://www.nbcnews.com/pop-culture/books/new-harry-potter-book-script-upcoming-play-gets-summer-release-n515866 |
| http://www.nbcnews.com/pop-culture/pop-culture-news/j-k-rowling-begs-harry-potter-fans-not-post-cursed-n586971 | http://www.nbcnews.com/pop-culture/books/comic-book-artist-darwyn-cooke-dies-53-n574251 |
| http://www.nbcnews.com/pop-culture/books/chair-used-harry-potter-author-j-k-rowling-auction-n550931 | http://www.nbcnews.com/video/-utterly-remarkable-harry-potter-fans-rave-about-cursed-child-701202499625 |
| http://www.wsj.com/articles/finding-the-spell-of-the-stage-on-the-page-with-harry-potter-and-the-cursed-child-1470864415 | http://www.nbcnews.com/news/europe/heir-b-b-orphans-get-boost-princess-diana-s-althorp-n580471 |
| http://www.nbcnews.com/watch/nbcnews-com/cloaking-devices-not-just-for-harry-potter-341672003525 | http://www.nbcnews.com/pop-culture/books/umberto-eco-italian-author-name-rose-dies-84-n522251 |
| http://www.nbcnews.com/pop-culture/books | http://www.nbcnews.com/feature/nerdwatch |
| http://www.wsj.com/video/nbc-snags-harry-potter-movies-in-massive-deal/CC30FB3E-DCD9-4DF9-8761-A54D916FD7B3.html | http://www.nbcnews.com/pop-culture/books/pat-conroy-prince-tides-author-dies-after-battle-pancreatic-cancer-n532261 |
| http://www.wsj.com/video/harry-potter-makes-a-magical-comeback/97B03814-913E-4E9C-86E3-7FD84938C17B.html | http://www.nbcnews.com/politics/hillary-clinton/green-party-s-jill-stein-seeks-woo-young-voters-clinton-n653211 |
| http://www.nbcnews.com/video/-utterly-remarkable-harry-potter-fans-rave-about-cursed-child-701202499625 | http://www.nbcnews.com/business/consumer/credit-card-contracts-literally-too-hard-read-most-n646696 |

|  |  |
| --- | --- |
| **Brazil** | |
| **LUCENE** | **PAGERANK** |
| http://www.wsj.com/amp/articles/itau-in-talks-to-buy-citigroups-brazil-consumer-banking-operations-1474541965 | http://www.nbcnews.com/feature/short-take/video/migrants-surge-across-the-mexican-border-587780163988 |
| http://www.wsj.com/articles/itau-in-talks-to-buy-citigroups-brazil-consumer-banking-operations-1474541965 | http://www.nbcnews.com/health/womens-health/male-female-drinking-habits-becoming-more-similar-study-n468191 |
| http://www.nbcnews.com/health/health-news/scientists-test-vaccine-against-deadly-dengue-virus-brazil-n497706 | http://www.nbcnews.com/feature/flashback/video/five-years-ago-the-space-shuttle-program-ended-723717187838 |
| http://www.nbcnews.com/storyline/zika-virus-outbreak/amp/zika-doubled-birth-defect-rate-brazil-study-shows-n644196 | http://www.nbcnews.com/storyline/zika-virus-outbreak/bracing-zika-tale-two-cities-n608546 |
| http://www.wsj.com/articles/brazil-plans-simpler-rules-on-infrastructure-contracts-1473799812 | http://www.nbcnews.com/business/travel/first-ever-commercial-flight-cuba-set-take-today-n640586 |
| http://www.wsj.com/amp/articles/brazils-former-finance-minister-arrested-as-part-of-corruption-probe-1474549000 | http://www.nbcnews.com/storyline/zika-virus-outbreak/mosquitoes-can-infect-their-eggs-zika-n639646 |
| http://www.wsj.com/articles/investors-see-gold-in-distressed-properties-in-brazil-1470158015 | http://www.nbcnews.com/health/kids-health/heres-reason-get-puppy-kids-pets-have-less-anxiety-n469591 |
| http://www.nbcnews.com/storyline/zika-virus-outbreak/zika-has-been-brazil-longer-anyone-thought-study-n545111 | http://www.nbcnews.com/storyline/zika-virus-outbreak/where-s-zika-most-likely-u-s-answer-may-surprise-n540281 |
| http://www.nbcnews.com/storyline/2016-rio-summer-olympics/judge-orders-ryan-lochte-jimmy-feigen-stay-brazil-during-robbery-n632686 | http://www.nbcnews.com/science/weird-science/natural-preference-study-finds-most-kangaroos-are-left-handed-n378091 |
| http://www.wsj.com/articles/brazil-passes-the-olympic-torch-and-breathes-a-sigh-of-relief-1471832043 | http://www.nbcnews.com/feature/short-take/video/how-is-security-for-el-chapo-different-this-time-601950275760 |

**OVERLAP CHART**

**Explanation regarding why some pages have higher pagerank values**

1. Pagerank algorithm works by assigning a numeric weight to each element of a hyper linked structure of documents, such as the World Wide Web, with the purpose of measuring its relative importance. This is achieved by measuring the number of outgoing links for a particular website (Using data provided in external\_edgelist.txt).

**THE SCREENSHOTS ARE SUBMITTED AS A PDF IN THE COMPRESSED FILE**