Homework 4 – Indexing the Web using Solr

Steps:

Step 1: Basic Installation:

- Installed Ubuntu on my Windows machine using Oracle VM VirtualBox.
- Installed Apache Solr using the tutorial and made sure Solr was up and running.
- Downloaded and extracted all the crawled files from google drive for Mercury News**.**

Step 2: Indexing the HTML files:

- Started the Solr server using **bin/solr start** in stand-alone mode.
- I created a core myexample using the **command bin/solr create –c myexample**.
- I edited the **managed-schema** file in the **conf** folder so that the content of all the data file given can be mapped properly.
- The Solr Admin page showed that all the files were indexed.
- Executed certain queries in Solr Admin page to find the relevant results.
- Deduced Solr is working as expected.

Step 3: XAMPP (Apache), PHP installation:

- Installed XAMPP and PHP in Ubuntu for the development purpose of PHP and to test it on the Apache web server.
- Cloned the GitHub repository for solr-php-client and placed it in the same directory as the PHP code.
- PHP code was modified accordingly to show the input fields for user to enter the query, then to receive the query at Server end, format and send it to Solr, fetch the Solr results, display the results by parsing the JSON content.

Step 4: Generating the Edge list file:

- Used JSoup library in a Java project/class. It first created a mapping csv which has both the mapping csv files. Used fileUrlMap and urlFileMap data structures using the combined mapping file.
- Extracted the outgoing links of each web page by traversing the directory to create the edgeList.txt file.
- A directed graph which contains web page files as vertices and edge between two indicates a link between them.

Step 5: PageRank Computation using NetworkX library:

- Implemented a python utility to compute the page rank values using the NetworkX library.
- The method parses each line from edgelist.txt and computes the page rank values using the following configuration :
  alpha=0.85, personalization=None, max_iter=30, tol=1e-06, nstart=None, weight='weight', dangling=None.
- external_pageRankFile.txt was created with the following format: <document_id>=<page_rank_score>.

Step 6:

- Placed the external_PageRankFile.txt in Data folder of the created core 'myexample'
- Modified managed-schema file to add a new field type as 'external' and field name as 'pageRankFile'.
- Modified the solrconfig.xml file to include listeners, to reload the external file when a new search is done.

Step 7:

- Modified the PHP code to handle the external page rank algorithm.
- For page rank, added additional parameters array as ‘sort’ => ‘pageRankFile desc’ before sending the query request to Solr.
- Displayed the title, URL, ID and Description for each result of the query and Title and URL were made clickable.
- First 10 results are displayed. The URL for the web page file name was obtained by parsing the mappings csv file.

Explanation for higher page rank values

The pages with large number of incoming links especially a page that is linked to by many pages with high PageRank receives a high rank value itself.

These may be the most popular pages or home pages which are referred to by other pages and yet may not be the most relevant to the given query.

Ten results produced for both Page rank method and default solr method for all the queries:

1. **Donald Trump:**

Lucence	Page Rank
https://www.mercurynews.com/2018/09/28/kimberly-guilfoyle-didnt-join-donald-trump-jr-s-twitter-attacks-on-christine-blasey-fords-credibility/	https://www.mercurynews.com/2018/10/15/drugmakers-may-have-to-disclose-prices-of-medicine-in-tv-ads/
https://www.mercurynews.com/tag/celebrity-politics/page/2/	https://www.mercurynews.com/2018/10/15/san-jose-inmate-found-dead-in-cell-at-main-jail/
https://www.mercurynews.com/2016/08/31/who-is-kevin-macdonald-alt-right-californian-retweeted-by-donald-trump-jr/	https://www.mercurynews.com/2018/10/15/vote-now-best-halloween-candy-ever-heres-the-sweet-16/
https://www.mercurynews.com/2018/08/30/political-cartoons-donald-trump-blasts-google/	https://www.mercurynews.com/2018/10/15/drugmakers-may-have-to-disclose-prices-of-medicine-in-tv-ads/
https://www.mercurynews.com/2018/08/28/political-cartoons-donald-trump-and-the-john-mccain-tributes/	https://www.mercurynews.com/2018/10/15/drugmakers-may-have-to-disclose-prices-of-medicine-in-tv-ads/
https://www.mercurynews.com/2018/09/21/opinion-evangelicals-puzzling-devotion-to-donald-trump/	https://www.mercurynews.com/2018/10/15/pge-socal-utility-cut-power-over-wind-concerns/
https://www.mercurynews.com/2018/07/18/did-queen-elizabeth-wage-brooch-warfare-against-donald-trump/	https://www.mercurynews.com/2018/10/15/drivers-its-25-mph-or-less-in-school-zones-roadshow/
https://www.mercurynews.com/2018/10/05/ivanka-trump-self-proclaimed-advocate-for-women-is-fine-with-brett-kavanaughs-confirmation-report-says/	https://www.mercurynews.com/2018/10/15/pge-socal-utility-cut-power-over-wind-concerns/
https://www.mercurynews.com/2018/09/21/opinion-evangelicals-puzzling-devotion-to-donald-trump/	https://www.mercurynews.com/2018/10/15/letter-is-it-any-surprise-the-dmv-screwed-up-motor-voter/

2. LA Lakers:

Lucence	Page Rank
https://www.mercurynews.com/2018/05/06/nba-playoffs-2018-golden-state-warriors-kevin-durant-game-4-highlights-video-new-orleans-pelicans-stephen-curry-score-schedule-time/	https://www.mercurynews.com/2018/10/15/sponsored-one-of-a-kind-fresh-renovated-orinda-home-with-mt-diablo-and-hillside-views/
https://www.mercurynews.com/2017/12/06/6-acre-brentwood-brush-fire-burning-uphill-near-skirball-center/	https://www.mercurynews.com/2018/10/15/heres-how-the-warriors-can-make-opening-night-a-success/
https://www.mercurynews.com/2018/05/05/nba-playoffs-2018-golden-state-warriors-new-orleans-pelicans-game-schedule-channel-stream-draymond-green-gsw-stephen-curry-rajon-rondo/	https://www.mercurynews.com/2018/10/15/prep-football-rankings-week-10-bay-area-news-group-top-25/
https://www.mercurynews.com/2018/05/07/nba-playoffs-2018-golden-state-warriors-new-orleans-pelicans-andre-iguodala-draymond-green-hamptons-five-5-death-lineup-schedule-time-gsw/	https://www.mercurynews.com/2018/10/09/and-kerri-walsh-jennings-new-partner-is-beach-star-gets-defensive/
https://www.mercurynews.com/tag/warriors-live/	https://www.mercurynews.com/2018/10/15/heres-how-the-warriors-can-make-opening-night-a-success/
https://www.mercurynews.com/2018/05/08/golden-state-warriors-new-orleans-pelicans-houston-rockets-stream-time-channel-nba-playoffs-2018-western-conference-finals-kevin-durant/	https://www.mercurynews.com/2018/10/14/oakland-raiders-jon-gruden-derek-carr-news-seattle-seahawks-nfl-london-fire-contract-salary-years-roster-las-vegas-stadium-schedule/
https://www.mercurynews.com/2018/01/22/jason-kidd-fired-as-milwaukee-bucks-head-coach/	https://www.mercurynews.com/2018/10/15/here-are-the-odds-on-the-name-of-meghan-markle-and-prince-harrys-baby/
https://www.mercurynews.com/2018/10/12/live-preseason-updates-warriors-vs-lakers-friday-at-730-p-m/	https://www.mercurynews.com/2018/10/15/heralded-stage-phenomenon-barber-shop-chronicles-comes-to-sf-bay-area/
https://www.mercurynews.com/2018/10/13/golden-state-warriors-vs-los-angeles-lakers-highlights-video-lebron-james-news-contract-salary-rumors-stephen-curry-twitter-steve-kerr/	https://www.mercurynews.com/2018/10/15/paul-allen-microsoft-co-founder-seahawks-owner-dead-at-65/
https://www.mercurynews.com/2018/10/15/heres-how-the-warriors-can-make-opening-night-a-success/	https://www.mercurynews.com/2018/10/15/warriors-joe-lacob-on-patrick-mccaws-absence-i-dont-really-understand-it/

3. Star Wars:

Lucence	Page Rank
https://www.mercurynews.com/tag/star-wars/	https://www.mercurynews.com/2018/10/08/san-jose-earthquakes-hire-big-time-coach/
https://www.mercurynews.com/2017/12/14/fans-line-up-for-star-wars-the-last-jedi-at-the-tech-museum/	https://www.mercurynews.com/2018/10/15/49ers-packers-pregame-te-dwelley-added-green-bay-missing-wrs/
https://www.mercurynews.com/2018/10/08/disneyland-star-wars-galaxys-edge-preview/	https://www.mercurynews.com/2018/10/15/paul-allen-microsoft-co-founder-seahawks-owner-dead-at-65/
https://www.mercurynews.com/2018/10/08/disneyland-star-wars-galaxys-edge-preview/	https://www.mercurynews.com/2018/10/15/vote-now-best-halloween-candy-ever-heres-the-sweet-16/
http://www.ocregister.com/disneyland-will-sell-alcohol-for-the-first-time-when-star-wars-land-opens-in-2019	https://www.mercurynews.com/2018/10/15/sponsored-one-of-a-kind-fresh-renovated-orinda-home-with-mt-diablo-and-hillside-views/
http://www.ocregister.com/disneyland-is-quietly-removing-seating-and-planters-for-star-wars-land-access	https://www.mercurynews.com/2018/10/15/paul-allen-microsoft-co-founder-seahawks-owner-dead-at-65/
https://www.mercurynews.com/2018/08/01/star-wars-filmmaker-j-j-abrams-sends-his-first-ever-tweet-heres-what-it-said/	https://www.mercurynews.com/2018/10/15/heres-how-the-warriors-can-make-opening-night-a-success/
https://www.mercurynews.com/2018/08/22/star-wars-actress-defies-racists-with-return-to-instagram/	https://www.mercurynews.com/2018/10/14/these-actors-favorite-co-stars-their-horses/
https://www.mercurynews.com/2018/10/15/17-ways-to-save-money-on-a-trip-to-disneyland/	https://www.mercurynews.com/2018/10/15/at-40-san-francisco-girls-chorus-sounds-as-young-as-ever/
http://www.ocregister.com/disneyland-is-quietly-removing-seating-and-planters-for-star-wars-land-access	https://www.mercurynews.com/2018/10/15/vote-now-best-halloween-candy-ever-heres-the-sweet-16/

4. Lebron James:

Lucence	Page Rank
https://www.mercurynews.com/tag/lebron-james/page/2/	https://www.mercurynews.com/2018/10/15/drug makers-may-have-to-disclose-prices-of-medicine-in-tv-ads/
https://www.mercurynews.com/tag/lebron-james/	https://www.mercurynews.com/location/california/bay-area/east-bay/alameda-county/
https://www.mercurynews.com/2018/08/04/president-trump-rips-lebron-james-says-i-like-mike/	https://www.mercurynews.com/2018/10/15/paul-allen-microsoft-co-founder-seahawks-owner-dead-at-65/
https://www.mercurynews.com/2018/08/06/lebron-james-son-bronny-headed-to-crossroads-reports-say/	https://www.mercurynews.com/2018/10/15/sponsored-one-of-a-kind-fresh-renovated-orinda-home-with-mt-diablo-and-hillside-views/
https://www.mercurynews.com/tag/cleveland-cavaliers/	https://www.mercurynews.com/2018/10/15/heres-how-the-warriors-can-make-opening-night-a-success/
https://www.mercurynews.com/2018/05/30/nba-finals-2018-golden-sta-warriors-lebron-james-cleveland-cavaliers-game-1-time-schedule-roster-stephen-curry/	https://www.mercurynews.com/2018/10/15/prep-football-rankings-week-10-bay-area-news-group-top-25/
https://www.mercurynews.com/2018/08/04/president-trump-rips-lebron-james-says-i-like-mike/	https://www.mercurynews.com/2018/10/15/heres-how-the-warriors-can-make-opening-night-a-success/
https://www.mercurynews.com/2018/10/10/report-lebron-james-wont-face-warriors-in-san-jose-friday/	https://www.mercurynews.com/2018/10/14/oakland-raiders-jon-gruden-derek-carr-news-seattle-seahawks-nfl-london-fire-contract-salary-years-roster-las-vegas-stadium-schedule/
https://www.mercurynews.com/2018/10/01/did-javale-mcgee-upstage-lebron-james-in-their-laker-debuts/	https://www.mercurynews.com/2018/10/15/drugmakers-may-have-to-disclose-prices-of-medicine-in-tv-ads/
https://www.mercurynews.com/2018/06/12/lebron-james-son-california-high-school/	https://www.mercurynews.com/2018/10/15/raiders-report-card-jon-grudens-1-5-start-is-very-ugly/

5. 2018 World Cup:

Lucence	Page Rank
https://www.mercurynews.com/2018/04/26/high-school-baseball-monta-vista-beats-lynbrook-3-1/	https://www.mercurynews.com/2018/10/08/san-jose-earthquakes-hire-big-time-coach/
https://www.mercurynews.com/tag/world-cup/page/2/	https://www.mercurynews.com/2018/10/15/drug makers-may-have-to-disclose-prices-of-medicine-in-tv-ads/
https://www.mercurynews.com/tag/soccer/page/3/	https://www.mercurynews.com/2018/10/15/san-jose-inmate-found-dead-in-cell-at-main-jail/
https://www.mercurynews.com/tag/world-cup/	https://www.mercurynews.com/2018/10/15/power-ratings-remember-when-ucla-was-supposed-to-go-0-12/
https://www.mercurynews.com/2018/09/30/us-looking-for-another-historic-comeback-at-the-ryder-cup/	https://www.mercurynews.com/2018/10/13/opinion-measure-v-will-build-housing-for-san-jose-families/
https://www.mercurynews.com/2018/07/13/world-cup-final-what-time-do-france-and-croatia-play-for-soccer-title/	https://www.mercurynews.com/location/california/bay-area/east-bay/alameda-county/
https://www.mercurynews.com/2018/05/11/startup-world-cup-competition-decides-best-of-28-startups-from-around-the-globe/	https://www.mercurynews.com/2018/10/15/what-we-know-about-the-giants-ongoing-search-for-bobby-evans-replacement/
https://www.mercurynews.com/2018/07/16/world-cup-rewind-why-this-brit-and-many-like-him-are-bloody-angry-and-still-hurt/	https://www.mercurynews.com/2018/10/15/49ers-packers-pregame-te-dwelley-added-green-bay-missing-wrs/
https://www.mercurynews.com/2018/09/30/us-looking-for-another-historic-comeback-at-the-ryder-cup/	https://www.mercurynews.com/2018/10/15/sears-will-close-stores-pleasanton-santa-rosa-amazon-google-ebay/
https://www.mercurynews.com/2018/07/15/france-wins-2nd-world-cup-title-beats-croatia-4-2/	https://www.mercurynews.com/2018/10/13/oakland-as-95-7-the-game-splitting-ways-on-a-very-bitter-note/

6. North Korea

Lucence	Page Rank
https://www.mercurynews.com/2018/10/02/seoul-north-korea-estimated-to-have-20-60-nuclear-weapons/	https://www.mercurynews.com/location/california/bay-area/south-bay/santa-clara-county/san-jose/
http://wirehub-digitalfirstmedia-com.go-vip.co/two-koreas-agree-to-end-war-this-year-pursue-denuclearization	https://www.mercurynews.com/2018/10/15/sponsored-one-of-a-kind-fresh-renovated-orinda-home-with-mt-diablo-and-hillside-views/
https://www.mercurynews.com/2018/09/26/rubin-four-reasons-why-declaring-peace-on-korean-peninsula-is-a-bad-idea/	https://www.mercurynews.com/2018/10/15/paul-allen-microsoft-co-founder-seahawks-owner-dead-at-65/
https://www.mercurynews.com/2017/08/29/tivo-apple-among-gainers-as-investors-shake-off-north-korea-fears/	https://www.mercurynews.com/2018/10/15/drug makers-may-have-to-disclose-prices-of-medicine-in-tv-ads/
https://www.mercurynews.com/2018/05/14/trudy-rubin-hold-the-trump-nobel-korea-talks-will-be-a-long-haul/	https://www.mercurynews.com/2018/10/15/pge-socal-utility-cut-power-over-wind-concerns/
https://www.mercurynews.com/2018/04/19/hanson-will-syrian-airstrikes-take-us-down-slippery-slope-of-more-involvement-in-the-middle-east/	https://www.mercurynews.com/2018/10/14/woman-dies-in-solo-i-280-rollover-collision-in-redwood-city/
https://www.mercurynews.com/2018/08/01/pence-welcomes-return-of-presumed-korean-war-dead/	https://www.mercurynews.com/2018/10/15/drug makers-may-have-to-disclose-prices-of-medicine-in-tv-ads/
https://www.mercurynews.com/2018/06/18/trudy-rubin-the-nauseating-flattery-trump-lavished-on-this-mass-murderer/	https://www.mercurynews.com/2018/10/15/pge-socal-utility-cut-power-over-wind-concerns/
https://www.mercurynews.com/2018/06/24/the-southern-poverty-law-center-has-lost-all-credibility/	https://www.mercurynews.com/2018/10/15/us-eyes-west-coast-bases-for-coal-exports/
https://www.mercurynews.com/2018/09/26/rubin-four-reasons-why-declaring-peace-on-korean-peninsula-is-a-bad-idea/	https://www.mercurynews.com/2018/10/15/paul-allen-microsoft-co-founder-seahawks-owner-dead-at-65/

7. Hurricane Florence

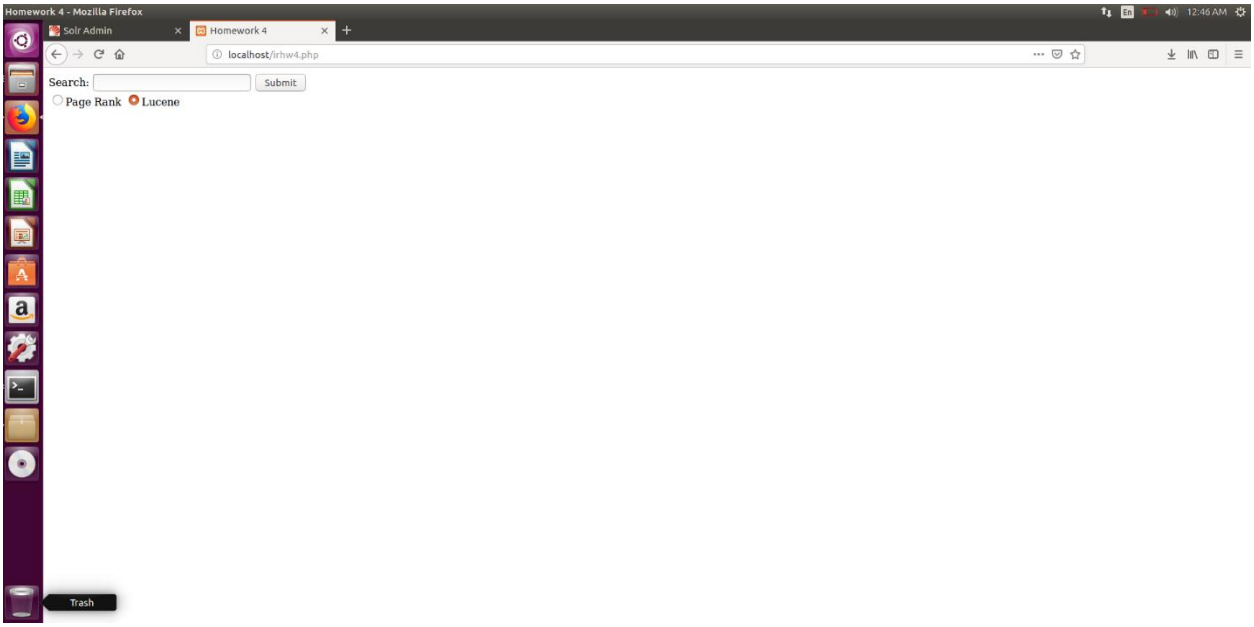
Lucence	Page Rank
https://www.mercurynews.com/2018/09/14/how-hurricane-florence-is-affecting-bay-area-travelers-and-what-they-need-to-know/	https://www.mercurynews.com/2018/10/15/vote-now-best-halloween-candy-ever-heres-the-sweet-16/
https://www.mercurynews.com/2018/09/27/beer-trending-on-social-media-after-brett-kavanaughs-references-during-testimony/	https://www.mercurynews.com/2018/10/14/these-actors-favorite-co-stars-their-horses/
https://www.mercurynews.com/2018/09/27/beer-trending-on-social-media-after-brett-kavanaughs-references-during-testimony/	https://www.mercurynews.com/2018/10/15/pge-socal-utility-cut-power-over-wind-concerns/
https://www.mercurynews.com/tag/travel/	https://www.mercurynews.com/2018/10/15/at-40-san-francisco-girls-chorus-sounds-as-young-as-ever/
https://www.mercurynews.com/tag/weather/page/3/	https://www.mercurynews.com/tag/celebrities/
https://www.mercurynews.com/2018/10/11/florida-surveys-damage-as-michael-inundates-the-carolinas/	https://www.mercurynews.com/2018/10/15/meghan-duchess-of-sussex-and-prince-harry-are-expecting-says-kensington-palace/
https://www.mercurynews.com/2018/10/10/michael-still-offshore-begins-battering-florida/	https://www.mercurynews.com/2018/10/12/bradley-cooper-reportedly-miserable-with-model-girlfriend-after-bonding-with-lady-gaga/
https://www.mercurynews.com/2018/10/12/the-latest-official-michael-responsible-for-georgia-death/	https://www.mercurynews.com/2018/10/15/drug makers-may-have-to-disclose-prices-of-medicine-in-tv-ads/
https://www.mercurynews.com/tag/pets/page/2/	https://www.mercurynews.com/2018/10/15/washington-archdiocese-releases-id-of-28-accused-priests/
https://www.mercurynews.com/2018/09/14/live-radar-images-of-hurricane-florences-path/	https://www.mercurynews.com/tag/pm-report/

8. Paul Allen

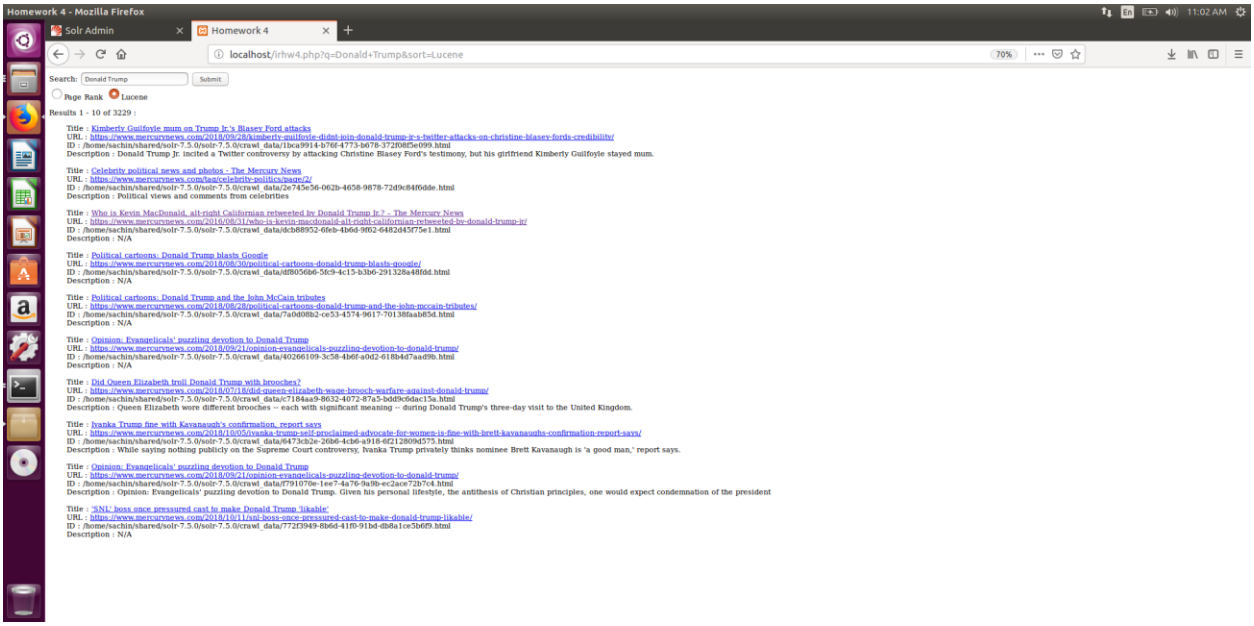
Lucence	Page Rank
https://www.mercurynews.com/2018/10/15/paul-allen-microsoft-co-founder-seahawks-owner-dead-at-65/	https://www.mercurynews.com/2018/10/15/drug makers-may-have-to-disclose-prices-of-medicine-in-tv-ads/
https://www.mercurynews.com/2018/10/15/paul-allen-microsoft-co-founder-seahawks-owner-dead-at-65/	https://www.mercurynews.com/2018/10/15/san-jose-inmate-found-dead-in-cell-at-main-jail/
https://www.mercurynews.com/2018/01/02/youtuber-logan-paul-apologizes-for-showing-body-in-japans-suicide-forest/	https://www.mercurynews.com/2018/10/15/paul-allen-microsoft-co-founder-seahawks-owner-dead-at-65/
https://www.mercurynews.com/2018/09/24/vintage-typewriters-are-booming-at-manhattan-shop/	https://www.mercurynews.com/location/california/bay-area/south-bay/santa-clara-county/san-jose/
https://www.mercurynews.com/2018/08/06/tensions-flare-in-brazil-as-it-considers-loosening-abortion-laws/	https://www.mercurynews.com/2018/10/15/sponsored-one-of-a-kind-fresh-renovated-orinda-home-with-mt-diablo-and-hillside-views/
https://www.mercurynews.com/2018/07/12/looking-up-pc-sales-rise-for-first-time-in-six-years/	https://www.mercurynews.com/2018/10/15/paul-allen-microsoft-co-founder-seahawks-owner-dead-at-65/
https://www.mercurynews.com/2018/07/20/deadly-gaza-strike-follows-death-of-israeli-soldier/	https://www.mercurynews.com/2018/10/15/drug makers-may-have-to-disclose-prices-of-medicine-in-tv-ads/
https://www.mercurynews.com/2018/09/11/renesas-to-acquire-u-s-chipmaker-idt-for-6-7-billion-3/	https://www.mercurynews.com/2018/10/15/letter-is-it-any-surprise-the-dmv-screwed-up-motor-voter/
https://www.mercurynews.com/2017/09/05/car-navigation-tech-brings-new-twists-and-turns-to-driving/	https://www.mercurynews.com/location/california/bay-area/south-bay/santa-cruz-county/
https://www.mercurynews.com/2017/10/02/roku-cuts-price-on-top-streaming-player-to-counter-apple-tv/	https://www.mercurynews.com/2018/10/14/takeaways-jones-breathes-new-life-in-goalie-controversy-as-sharks-lose-to-devils/

Screenshots that describes the whole flow:

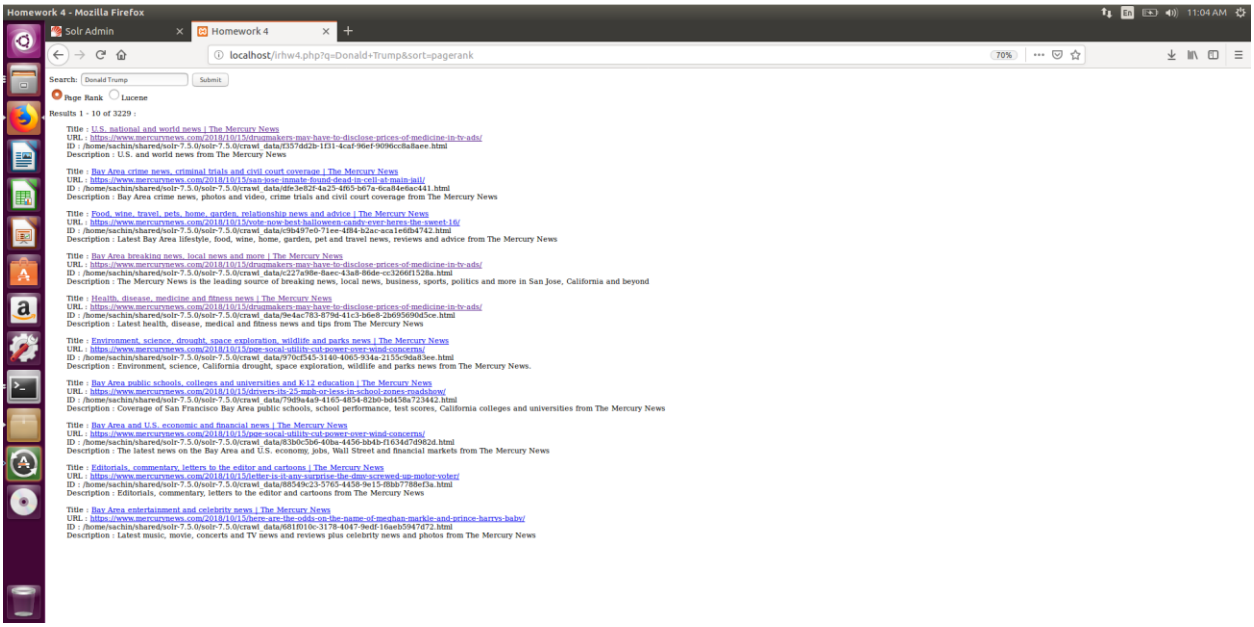
1. Initial Page



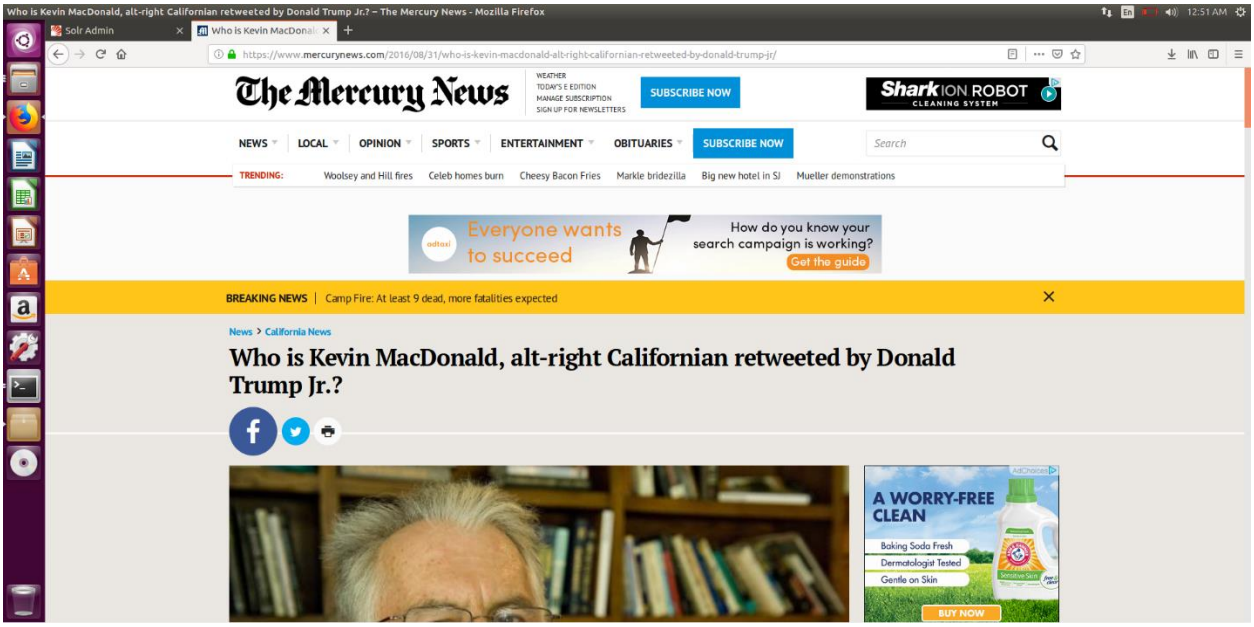
2. Lucene results:



3. Page Rank results:



4. Actual web page:



Results Overlap:

Query Number	Query	No. of Overlaps
1	Donald Trump	0
2	LA Lakers	1
3	Star Wars	0
4	Lebron James	0
5	2018 World Cup	0
6	North Korea	0
7	Hurricane Florence	0
8	Paul Allen	2