Report for Indexing the Web using Solr

Started with the initial setup of Virtual Box and booted it with Ubuntu 16.04 and installed Solr and started solr. Created a new core called ‘homework’ in order to index the downloaded HTML web pages for NYTimes. After making few changes to managed\_schema as mentioned in the HW documents and started indexing. Opened Solr Admin and tested few queries to understand TIKA parser.

In order to create the extract the outgoing links from the downloaded HTML files, created a EdgeList file using Java and Jsoup package, which in turn is the input for Pagerank. Executed the Pagerank algorithm with NetworkX package in python to use a different ranking algorithm. Used Python for coding the algorithm with the following parameters ( alpha=0.85, personalization=None, max\_iter=30, tol=1e-06, nstart=None, weight='weight', dangling=None as mentioned in the Solr exercise) and created external\_PageRankFile.txt which contains the ranks for the URLs which will be used to determine the query results for Pagerank algorithm.

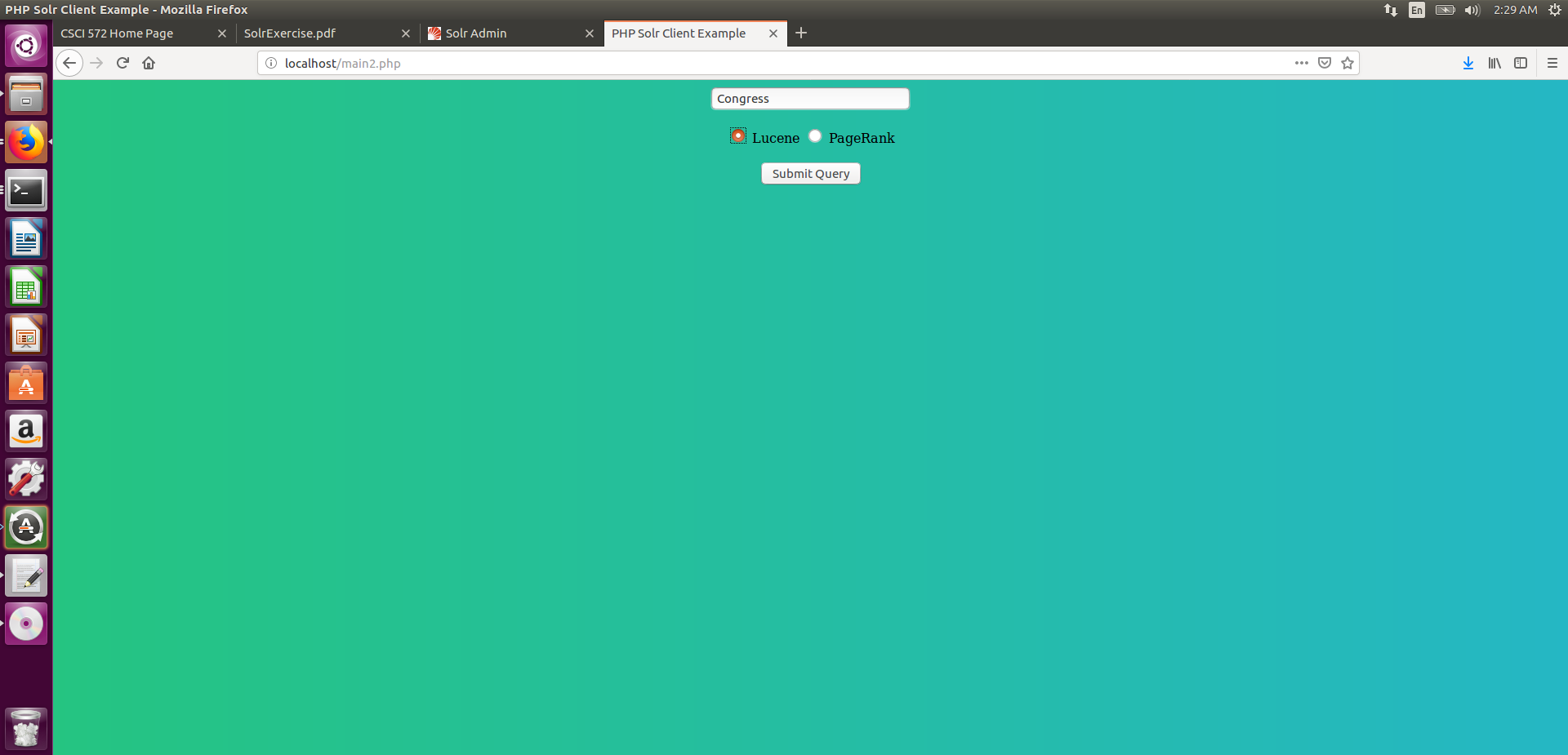
Installed Apache and PHP in order to create a web page with input for query, choice of algorithm and a submit button as front end and used Solr as backend. Integrated the PHP code with Solr using the core ‘homework’ I created. When a query is entered, we can retrieve results from either Lucene, deafault algorithm used by Solr or Pagerank, using the external\_PageRankFile.txt. The top 10 results are displayed with ID, Title, Description and URL. Also, the actual NYTimes web page of that news can be retrieved by clicking on the respective details.

Reason for Some Pages having Higher Page Ranks

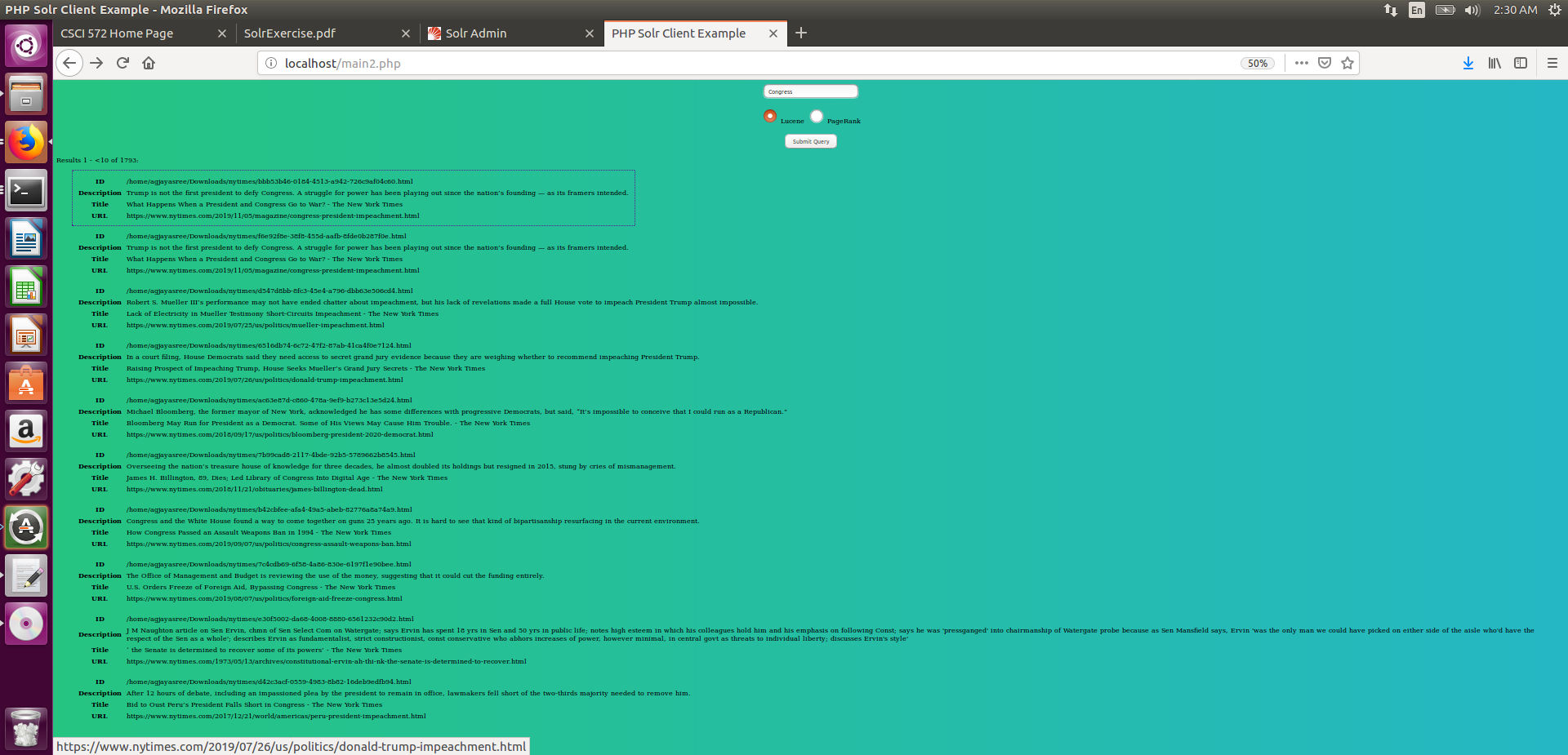
Pagerank algorithm uses the number of incoming and outgoing links to the respective URL and the rank is calculated using number of incoming links and quality of the links, either pages like Home page, pages on the navigation bar tend to have higher incoming links than a specific news link and hence will have a higher rank or even if a page has fewer incoming links but these links are from important pages, then they will also have higher ranks.

Flow for Query : Congress

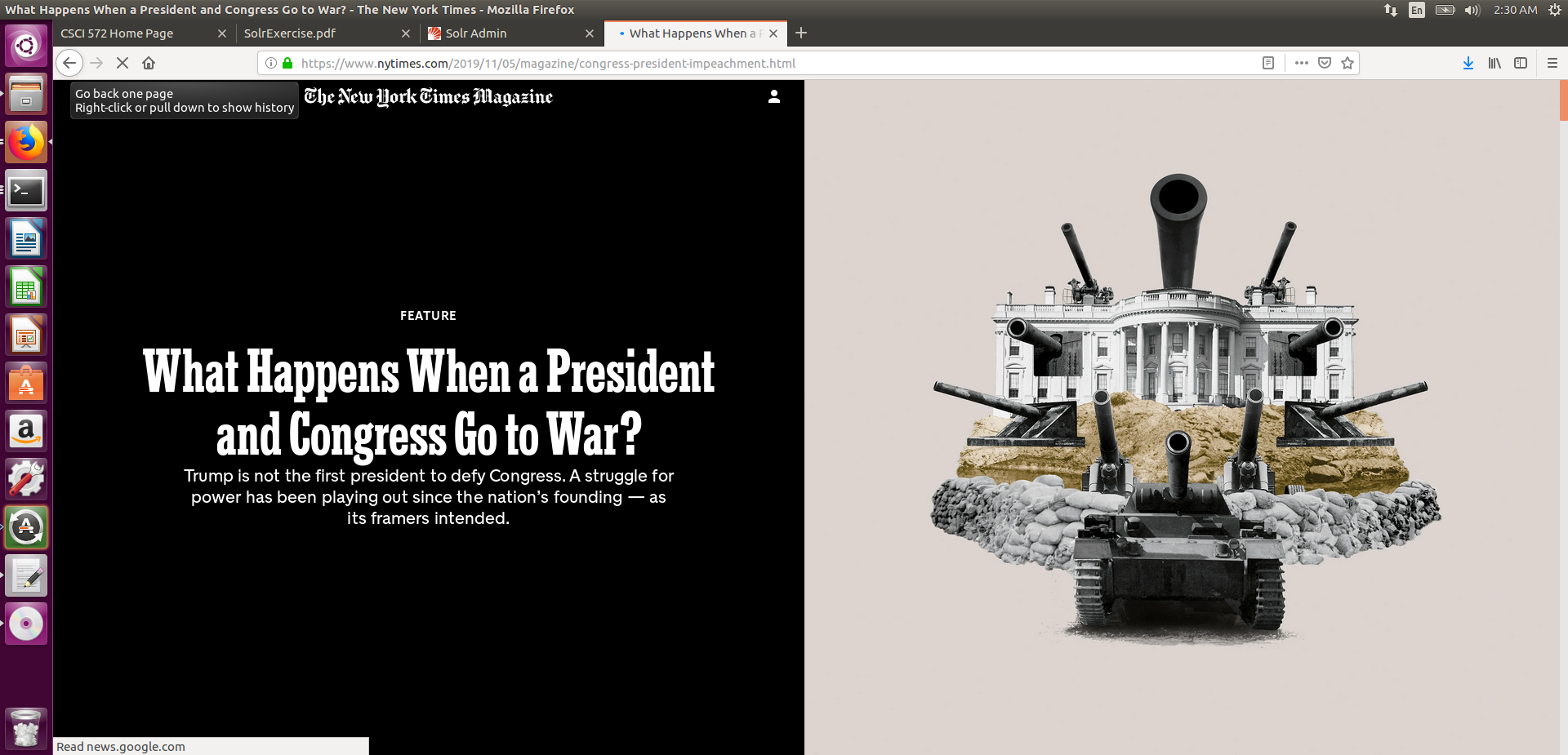
Home Page & Entering Query



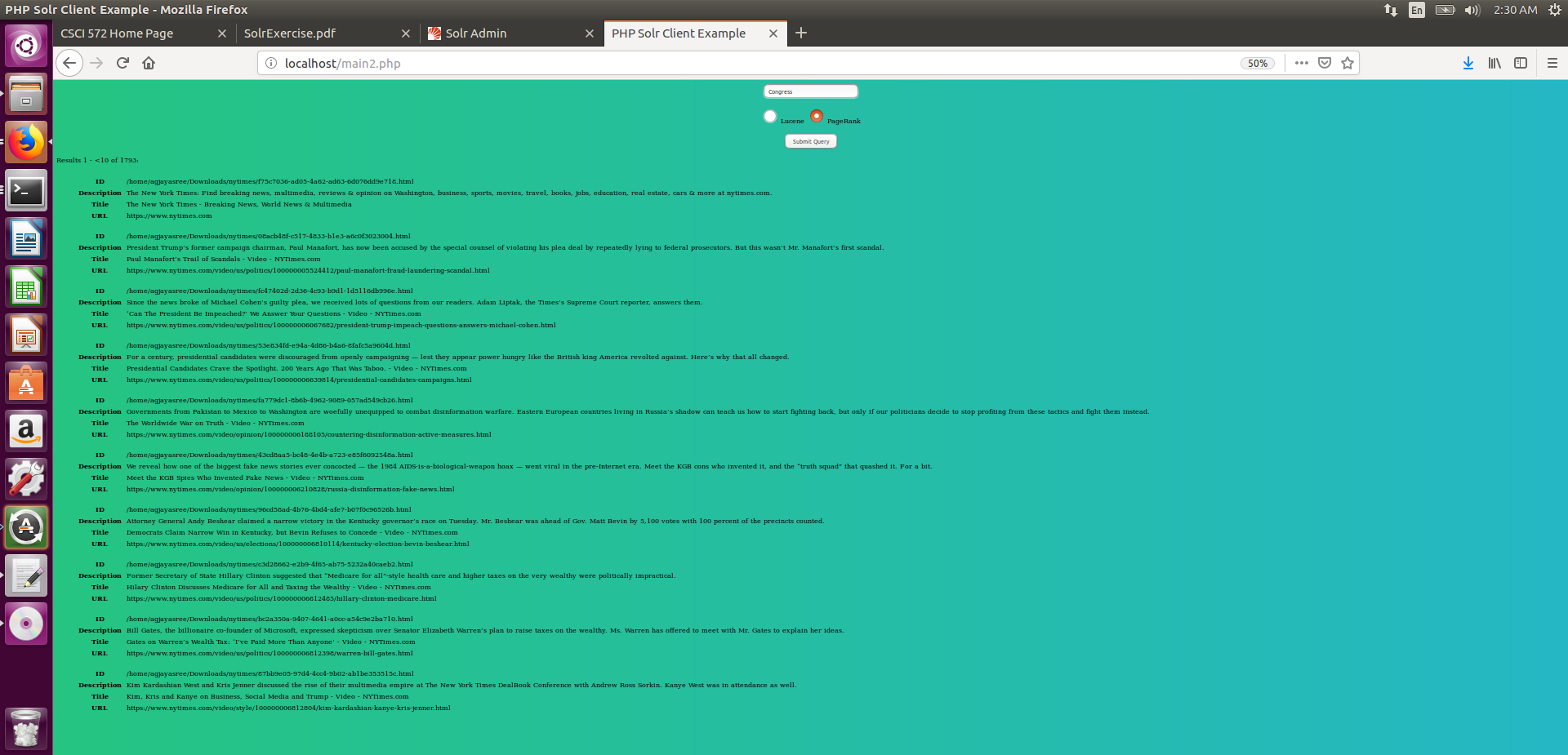
Submitting Query (Lucene)



Opening URL



Submitting Query again (Pagerank)



List of URLs

|  |
| --- |
| Query 1 - Cannes |
| Lucene |
| https://www.nytimes.com/2019/05/25/movies/cannes-film-festival-winners-parasite.html https://www.nytimes.com/2019/05/25/movies/cannes-film-festival-winners-parasite.html https://www.nytimes.com/2018/04/12/movies/cannes-lineup-netflix-boycott.html https://www.nytimes.com/2019/05/24/movies/cannes-almodovar-kechiche.html https://www.nytimes.com/2019/05/16/movies/cannes-film-festival-jim-jarmusch.html https://www.nytimes.com/2019/05/21/movies/cannes-film-festival-antonio-banderas-pedro-almodovar.html https://www.nytimes.com/2019/05/27/movies/cannes-oscars.html https://www.nytimes.com/2019/05/20/arts/cannes-film-festival-black-director.html https://www.nytimes.com/2019/05/22/movies/quentin-tarantino-margot-robbie.html https://www.nytimes.com/2019/05/22/movies/dicaprio-pitt.html |
| Pagerank |
| https://www.nytimes.com/2019/11/07/movies/nyc-this-weekend-film-series.html https://www.nytimes.com/interactive/2019/10/15/t-magazine/nick-cave-artist.html https://www.nytimes.com/interactive/2019/10/15/t-magazine/nicolas-ghesquiere-louis-vuitton.html https://www.nytimes.com/interactive/2019/10/15/t-magazine/rachel-weisz-acting-movies.html https://www.nytimes.com/2019/10/30/movies/bong-joon-ho-parasite.html https://www.nytimes.com/2019/11/01/movies/taika-waititi.html https://www.nytimes.com/interactive/2019/10/15/t-magazine/shigeru-ban.html https://www.nytimes.com/interactive/2019/10/15/t-magazine/the-greats.html https://www.nytimes.com/interactive/2017/obituaries/notable-deaths/jack-doroshow https://www.nytimes.com/interactive/2016/obituaries/notable-deaths/abbas-kiarostami |

Number of Overlaps - 0

|  |
| --- |
| Query 2 - Congress |
| Lucene |
| https://www.nytimes.com/2019/11/05/magazine/congress-president-impeachment.html https://www.nytimes.com/2019/11/05/magazine/congress-president-impeachment.html https://www.nytimes.com/2019/07/25/us/politics/mueller-impeachment.html https://www.nytimes.com/2019/07/26/us/politics/donald-trump-impeachment.html https://www.nytimes.com/2018/09/17/us/politics/bloomberg-president-2020-democrat.html https://www.nytimes.com/2018/11/21/obituaries/james-billington-dead.html https://www.nytimes.com/2019/09/07/us/politics/congress-assault-weapons-ban.html https://www.nytimes.com/2019/08/07/us/politics/foreign-aid-freeze-congress.html https://www.nytimes.com/1973/05/13/archives/constitutional-ervin-ah-thi-nk-the-senate-is-determined-to-recover.html https://www.nytimes.com/2017/12/21/world/americas/peru-president-impeachment.html |
| Pagerank |
| https://www.nytimes.com https://www.nytimes.com/video/us/politics/100000005524412/paul-manafort-fraud-laundering-scandal.html https://www.nytimes.com/video/us/politics/100000006067682/president-trump-impeach-questions-answers-michael-cohen.html https://www.nytimes.com/video/us/politics/100000006639814/presidential-candidates-campaigns.html https://www.nytimes.com/video/opinion/100000006188105/countering-disinformation-active-measures.html https://www.nytimes.com/video/opinion/100000006210828/russia-disinformation-fake-news.html https://www.nytimes.com/video/us/elections/100000006810114/kentucky-election-bevin-beshear.html https://www.nytimes.com/video/us/politics/100000006812485/hillary-clinton-medicare.html https://www.nytimes.com/video/us/politics/100000006812398/warren-bill-gates.html https://www.nytimes.com/video/style/100000006812804/kim-kardashian-kanye-kris-jenner.html |

Number of Overlaps - 0

|  |
| --- |
| Query 3 - Democrats |
| Lucene |
| https://www.nytimes.com/2018/09/04/us/politics/ayanna-pressley-massachusetts.html https://www.nytimes.com/2018/11/07/us/politics/election-history-firsts-blackburn-pressley.html https://www.nytimes.com/2019/09/29/us/politics/pelosi-trump-impeachment.html https://www.nytimes.com/2019/10/16/us/politics/democrats-impeachment.html https://www.nytimes.com/2018/01/23/us/politics/impeach-trump-democrats-tom-steyer.html https://www.nytimes.com/2019/09/24/us/politics/trump-democrats.html https://www.nytimes.com/2019/10/28/us/politics/trump-impeachment-subpoena-house.html https://www.nytimes.com/2019/09/11/opinion/black-voters-democrats-2020-election.html https://www.nytimes.com/2019/11/05/opinion/trump-democrats-2020.html https://www.nytimes.com/2019/11/01/us/politics/house-republican-box-prank.html |
| Pagerank |
| https://www.nytimes.com https://www.nytimes.com/video https://www.nytimes.com/video/opinion/100000006064751/coal-miner-to-trump-coal-mining-isnt-coming-back.html https://www.nytimes.com/video/us/politics/100000006067682/president-trump-impeach-questions-answers-michael-cohen.html https://www.nytimes.com/video/us/politics/100000006639814/presidential-candidates-campaigns.html https://www.nytimes.com/video/arts/music/100000006686500/lil-tecca-ransom.html https://www.nytimes.com/video/us/elections/100000006810114/kentucky-election-bevin-beshear.html https://www.nytimes.com/video/us/politics/100000006812485/hillary-clinton-medicare.html https://www.nytimes.com/2019/11/09/opinion/sunday/bloomberg-versus-Trump-2020.html https://www.nytimes.com/2019/11/09/us/politics/bernies-sanders-michael-bloomberg-billions.html |

Number of Overlaps - 0

|  |
| --- |
| Query 4 – Patriot Movement |
| Lucene |
| https://www.nytimes.com/2019/08/14/opinion/joe-walsh-trump-primary.html https://www.nytimes.com/1976/08/27/archives/lewis-michaux-92-dies-ran-bookstore-in-harlem.html https://www.nytimes.com/2019/05/15/us/politics/us-iraq-embassy-evacuation.html https://www.nytimes.com/2019/01/07/obituaries/moshe-arens-dead.html https://www.nytimes.com/2019/10/27/obituaries/john-conyers-jr-dead.html https://www.nytimes.com/2018/12/29/obituaries/shehu-shagari-dead.html https://www.nytimes.com/2019/10/22/sports/football/jets-patriots.html https://www.nytimes.com/2019/09/23/opinion/democrats-republicans.html https://www.nytimes.com/2019/10/30/opinion/alexander-vindman-trump-ukraine.html https://www.nytimes.com/2019/09/20/us/data-privacy-fbi.html |
| Pagerank |
| https://www.nytimes.com https://www.nytimes.com/video https://www.nytimes.com/video/opinion https://www.nytimes.com/video/opinion/100000006525694/stonewall-the-making-of-a-monument.html https://www.nytimes.com/2019/11/01/arts/television/new-to-stream-netflix.html https://www.nytimes.com/video/arts/music/100000006686500/lil-tecca-ransom.html https://www.nytimes.com/video/opinion/100000006788354/nike-running-mary-cain.html https://www.nytimes.com/video/opinion/100000006210828/russia-disinformation-fake-news.html https://www.nytimes.com/video/world/middleeast/100000006777985/syria-isis-alhol-turkey-kurds.html https://www.nytimes.com/video/world/middleeast/100000006789926/us-kurds.html |

Number of Overlaps - 0

|  |
| --- |
| Query 5 - Republicans |
| Lucene |
| https://www.nytimes.com/2019/10/24/us/politics/senate-republicans-trump.html https://www.nytimes.com/2019/11/04/us/politics/trump-republicans-impeachment.html https://www.nytimes.com/2019/11/04/us/politics/trump-republicans-impeachment.html https://www.nytimes.com/2019/10/31/us/politics/house-impeachment-partisan-vote.html https://www.nytimes.com/2019/10/04/us/politics/republicans-trump-defense.html https://www.nytimes.com/2019/11/09/us/politics/republican-strategy-impeachment-trump.html https://www.nytimes.com/2019/09/30/opinion/republicans-trump-impeachment.html https://www.nytimes.com/2019/11/06/us/trump-senate-republicans-courts.html https://www.nytimes.com/2019/11/06/us/trump-senate-republicans-courts.html https://www.nytimes.com/2019/10/08/us/politics/republicans-trump-syria.html |
| Pagerank |
| https://www.nytimes.com https://www.nytimes.com/video/us/politics/100000006067682/president-trump-impeach-questions-answers-michael-cohen.html https://www.nytimes.com/video/us/elections/100000006810114/kentucky-election-bevin-beshear.html https://www.nytimes.com/video/us/politics/100000006812485/hillary-clinton-medicare.html https://www.nytimes.com/2019/11/09/us/politics/virginia-elections-democrats-republicans.html https://www.nytimes.com/2019/11/09/us/politics/mulvaney-trump-impeachment-subpoenas.html https://www.nytimes.com/2019/11/09/us/politics/republican-strategy-impeachment-trump.html https://www.nytimes.com/video/opinion/100000006780505/fintess-guillermo-morales-runner.html https://www.nytimes.com/video/us/politics/100000006067682/president-trump-impeach-questions-answers-michael-cohen.html https://www.nytimes.com/video/us/politics/100000006785328/biden-trump-economy.html |

Number of Overlaps - 1

|  |
| --- |
| Query 6 - Senate |
| Lucene |
| https://www.nytimes.com/2019/10/01/us/politics/senate-trump-impeachment.html https://www.nytimes.com/2019/10/26/us/trump-senate-presidency.html https://www.nytimes.com/2019/08/03/us/politics/senate-votes-mcconnell.html https://www.nytimes.com/2019/08/22/us/politics/john-hickenlooper-senate-2020.html https://www.nytimes.com/2019/11/08/us/impeachment-senate-democrats.html https://www.nytimes.com/2019/08/31/us/politics/senate-race-2020.html https://www.nytimes.com/2019/10/24/us/politics/senate-republicans-trump.html https://www.nytimes.com/2019/10/18/us/politics/mcconnell-impeachment-senate.html https://www.nytimes.com/2019/09/05/opinion/the-argument-senate-democrats-2020.html https://www.nytimes.com/2019/09/05/opinion/the-argument-senate-democrats-2020.html |
| Pagerank |
| https://www.nytimes.com/video/opinion/100000006064751/coal-miner-to-trump-coal-mining-isnt-coming-back.html https://timesjourneys.nytimes.com/tour-type/ https://www.nytimes.com/times-journeys/tour-type/small-group/ https://www.nytimes.com/times-journeys/trip-type/politics-and-perspective/ https://www.nytimes.com/video/us/politics/100000006639814/presidential-candidates-campaigns.html https://www.nytimes.com/video/opinion/100000006188105/countering-disinformation-active-measures.html https://www.nytimes.com/video/us/politics/100000006812485/hillary-clinton-medicare.html https://www.nytimes.com/2019/11/09/us/politics/pete-buttigieg-democrats.html https://www.nytimes.com/2019/11/09/us/politics/virginia-elections-democrats-republicans.html https://www.nytimes.com/2019/11/09/us/politics/republican-strategy-impeachment-trump.html |

Number of Overlaps - 0

|  |
| --- |
| Query 7 – Olympics 2000 |
| Lucene |
| https://www.nytimes.com/2016/12/08/sports/great-moments-in-cheerleading-could-the-olympics-be-next.html https://www.nytimes.com/2019/11/01/sports/olympics/enriqueta-basilio-71-dies-first-woman-to-light-olympic-flame.html https://www.nytimes.com/video/sports/basketball/100000006374267/zion-williamsons-shoe-splits-open-during-game.html https://www.nytimes.com/video/sports/basketball/100000006374267/zion-williamsons-shoe-splits-open-during-game.html https://www.nytimes.com/video/sports/100000006009986/referees-battle-against-abuse-from-the-sidelines.html https://www.nytimes.com/video/sports/100000006009986/referees-battle-against-abuse-from-the-sidelines.html https://www.nytimes.com/video/sports/100000006086118/nike-courts-controversy-with-kaepernick-deal.html https://www.nytimes.com/video/sports/100000006086118/nike-courts-controversy-with-kaepernick-deal.html https://www.nytimes.com/video/sports/100000006008687/france-fans-celebrate-world-cup.html https://www.nytimes.com/video/sports/100000006008687/france-fans-celebrate-world-cup.html |
| Pagerank |
| https://www.nytimes.com https://www.nytimes.com/video/opinion https://www.nytimes.com/video/movies/100000006803253/terminator-dark-fate-scene.html https://www.nytimes.com/video/opinion/100000006048561/calving-season.html https://www.nytimes.com/video/opinion/100000006055559/sanctuary-city-hotline.html https://www.nytimes.com/video/opinion/100000006064751/coal-miner-to-trump-coal-mining-isnt-coming-back.html https://www.nytimes.com/video/opinion/100000006033259/between-sound-and-silence.html https://www.nytimes.com/video/world/middleeast/100000001253159/2011-arab-spring.html https://www.nytimes.com/video/opinion/100000006616228/mumbais-midnight-gardeners.html https://www.nytimes.com/video/opinion/100000006594650/stay-close.html |

Number of Overlaps - 0

|  |
| --- |
| Query 8 - Stock |
| Lucene |
| https://www.nytimes.com/2019/05/02/business/tesla-stock-fundraising.html https://www.nytimes.com/2018/12/16/technology/tech-workers-company-stock-shareholder-activism.html https://www.nytimes.com/2019/05/10/technology/uber-stock-price-ipo.html https://www.nytimes.com/2019/11/04/business/under-armour-stock-investigation.html https://www.nytimes.com/2019/09/20/technology/airbnb-employees-ipo-payouts.html https://www.nytimes.com/2019/10/24/style/rebag-clair-handbag-stock-market.html https://www.nytimes.com/2019/05/30/technology/uber-stock-earnings.html https://www.nytimes.com/2018/11/12/obituaries/david-pearson-nascar-dead.html https://www.nytimes.com/2019/04/03/business/tesla-sales-elon-musk.html https://www.nytimes.com/2019/10/28/business/stock-market-record.html |
| Pagerank |
| https://www.nytimes.com/video/opinion/100000006616246/the-king-of-fish-and-chips.html https://www.nytimes.com/times-journeys/travel/egypt/ https://www.nytimes.com/2019/11/09/opinion/sunday/stripping-sex-work.html https://www.nytimes.com/video/movies/100000006675663/brittany-runs-a-marathon-scene.html https://www.nytimes.com/2019/11/07/opinion/sunday/california-fires-trump.html https://www.nytimes.com/video/movies/100000006675663/brittany-runs-a-marathon-scene.html https://www.nytimes.com/2019/11/08/t-magazine/hawaii-restaurants.html https://www.nytimes.com/real-estate https://www.nytimes.com/2019/11/09/business/saudi-arabia-aramco.html https://www.nytimes.com/2019/11/08/sports/football/nfl-picks-week-10.html |

Number of Overlaps - 0

|  |
| --- |
| Query 9 - Virus |
| Lucene |
| https://www.nytimes.com/2019/11/09/science/seals-distemper.html https://www.nytimes.com/2019/10/25/health/predict-usaid-viruses.html https://www.nytimes.com/2019/10/28/us/politics/kay-hagan-dead.html https://www.nytimes.com/2019/02/05/us/politics/trump-hiv-aids-plan.html https://www.nytimes.com/2019/01/23/neediest-cases/daughter-cmv.html https://www.nytimes.com/2019/09/27/well/live/when-is-the-best-time-to-get-your-flu-shot.html https://www.nytimes.com/2019/11/06/health/crispr-cancer-leukemia.html https://www.nytimes.com/2019/08/13/style/sti-stigma-sexual-transmitted-infections.html https://www.nytimes.com/2019/10/14/health/ebola-gorillas.html https://www.nytimes.com/2019/02/04/us/politics/trump-hiv-aids-plan.html |
| Pagerank |
| https://www.nytimes.com/video/opinion/100000006188102/what-is-pizzagate.html https://www.nytimes.com/video/opinion/100000006188105/countering-disinformation-active-measures.html https://www.nytimes.com/video/opinion/100000006210828/russia-disinformation-fake-news.html https://www.nytimes.com/video/opinion/100000006188102/what-is-pizzagate.html https://www.nytimes.com/video/opinion/100000006210828/russia-disinformation-fake-news.html https://www.nytimes.com/video/opinion/100000006188105/countering-disinformation-active-measures.html https://www.nytimes.com/2019/11/09/science/seals-distemper.html https://www.nytimes.com/2018/11/12/opinion/russia-meddling-disinformation-fake-news-elections.html https://www.nytimes.com/2019/11/08/health/hiv-prevention-truvada-patents.html https://www.nytimes.com/2019/11/09/science/climate-change-music-sound.html |

Number of Overlaps - 1