CSCI572 Assignment 4 Report:

Name: Prerana Teligi Harapanahalli Math

1. Steps:

- 1. Operating System used: MacOS (High Sierra). No need to install Linux VM
- **2. Solr installation:** Followed the instructions on http://lucene.apache.org/solr/guide/7_3/solr-tutorial.html and installed solr-7.3.0 by downloading solr-7.3.0.zip file from http://www.apache.org/dyn/closer.lua/lucene/solr/7.3.0
- **3.** Unzipped the file on Desktop. This is the entire path to the directory /Users/prerana/Desktop/solr-7.3.0
- **4. Solr setup:** cd into solr-7.3.0 directory. Followed instruction mentioned on http://www-scf.usc.edu/~csci572/2018Spring/hw4/IndexingwithTIKAV3.pdf
 - a. **bin/solr** start Solr server
 - b. **bin/solr create -c assignment4** creates new core named assignment4 inside /Users/prerana/Desktop/solr-7.3.0/server/solr/
- 5. Downloaded the NBC_News data from google drive link provided_https://drive.google.com/drive/folders/1ZPWjQZMB7xeegjaeZfzqeJKq9oKuukio and unzip the HTML files and store those in the following path /Users/prerana/Desktop/solr-7.3.0/NBC_News/
 Store the provided CSV file named UrlToHtml_NBCNews.csv in this path /Users/prerana/Desktop/solr-7.3.0/
- **6.** Use this command to index the HTML files is NBC_News folder using Solr bin/post -c assignment4 -filetypes html NBC_News/
- 7. You can view the indexing by going to this link http://localhost:8983/solr/ and selecting your core i.e assignment4 and to see how TIKA parser parsed the html pages and what fields were created by using TIKA. Select the query option, and just submit the default query (*:*). Click on overview option to view your core statistics. Before that to configure default querying using Solr, uncomment the str element with name "df", and replace the value with _text_, where we specify that the default query field has to be "_text_". Now when you click on Execute Query button on Solr Query page it returns the index results with "numFound":15725 (i.e. number of documents).
- **8. assignment4.php:** This file is created to send the form query to Solr through Apache server and retrieve the results. We also need to download solr-php-client in order to run solr service from php client (i.e. from our php file) by doing cloning this git repo using **git clone_** https://github.com/PTCInc/solr-php-client.git command. We also need to download XAMPP for

running Apache server or you can also use the default server on Mac. I used XAMPP. After starting Apache on XAMPP, place the assignment4.php file and solar-php-client folder in htdocs folder of XAMPP i.e. /Applications/XAMPP/xamppfiles/htdocs/.

- **9. PageRank:** Wrote Java code to create edgelist.txt by providing NBC_News folder containing HTML files and UrlToHtml_NBCNews.csv file as input. Using JSoup, urlFileMap and fileUrlMap functions. Imported **JSoup jar** into eclipse and ran the java code.
- 10. Wrote python code to generate external pageRankFile.txt file from edgelist.txt.
- **11.** Type **sudo pip install networkx** command in command line to install networkx python library. This is used to run the python file. I have named my python file as **networkX.py**. and used these parameters: alpha=0.85, personalization=None, max_iter=30, tol=1e-06, nstart=None, weight='weight', dangling=None. Run this to generate external_pageRankFile.txt.
- 12. After generating external_pageRankFile.txt file place it in /Users/prerana/Desktop/solr-7.3.0/server/solr/assignment4/data/ folder.
- 13. Modify managed-schema.xml by adding the following lines:
 <fieldType name="external" keyField="id" defVal="0" class="solr.ExternalFileField" valType="pfloat" />
 <field name="pageRankFile" type="external" stored="false" indexed="false" />
- 14. Define listeners for query in solrconfig.xml by modifying these lines listener event="newSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader"/> listener event="firstSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader"/>
- 15. Add the following inside **solrconfig.xml** for ANDing the query terms for better retrieval results. <**requestHandler name=''/select'' class=''solr.SearchHandler''> <lst> tag <str name=''q.op''>AND</str>**
- **16.** Now reload the index, by going to the **Solr Dashboard UI ->Core Admin** and clicking on the "**Reload**" button.
- **17.** Go to this URL http://localhost/assignment4.php to view you search application. Now you can give any query term and retrieve results from Solr.
- **18.** Run the queries given in assignment description.

2. why some pages have higher PageRank values than others?

- PageRank algorithm assigns a numeric weight to each link element of DOM in a webpage for measuring its relative importance. This is achieved by measuring the number of outgoing links for a particular website. We create a graph with the list of edges pointing to the page (i.e. edgeList.txt). We use this to find importance of any given page. PageRank determines the weight

of the page and then outputs it in given order (Highest first(descending)). Some pages have more weight than others and hence they have higher PageRank values than others.

3. Screenshots:

Solr folder on downloaded and unzipped on Desktop:

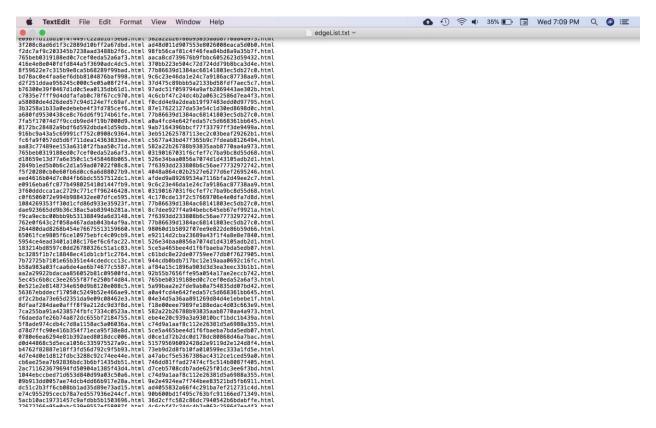


Solr indexing:

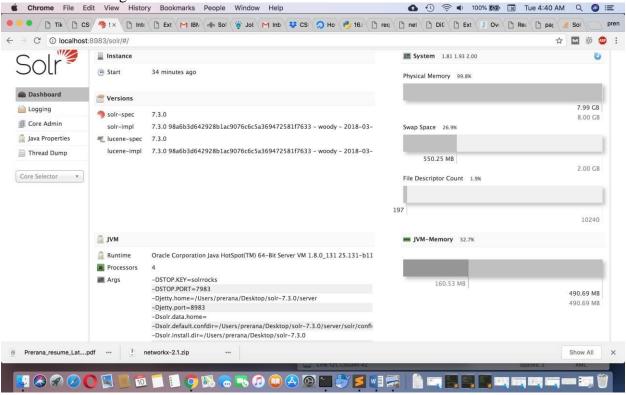
Created new core 'assignment4'
PreranasMacBook:solr=7.3.0 prerana\$ bin/post -c assignment4 -filetypes html NBC_News/
java -classpath /Users/prerana/Desktop/solr=7.3.0/dist/solr-core=7.3.0.jar -Dauto=yes -Dfiletypes=html -Dc=assignment4 -Ddata=files -Drecursive=yes org.apache.solr.util.SimplePostTo
ol NBC_News/ Created new core 'assignment4'
PreransAedockok:solr-7.3.0 prerans bin/post -c assignment4 -filetypes html NBC.N
java -classpath /Users/prerans/Desktop/solr-7.3.0/dist/solr-core-7.3.0.jar -Dautool.NBC.News/
SimplePostIool.version 5.0.0

Posting files to [base] url http://localhost:8983/solr/assignment4/update...
Entering auto mode. File endings considered are html
Entering recursive mode. File endings considered are h

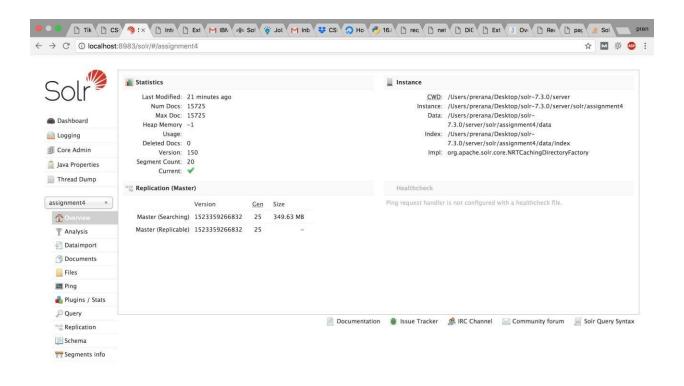
edgeList.txt:

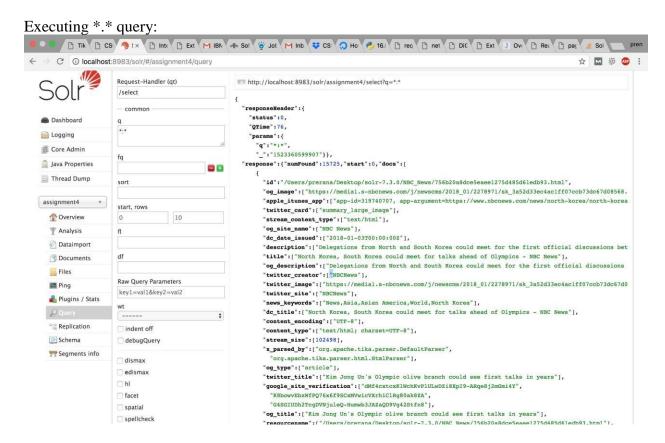


Solr Server running on Localhost:

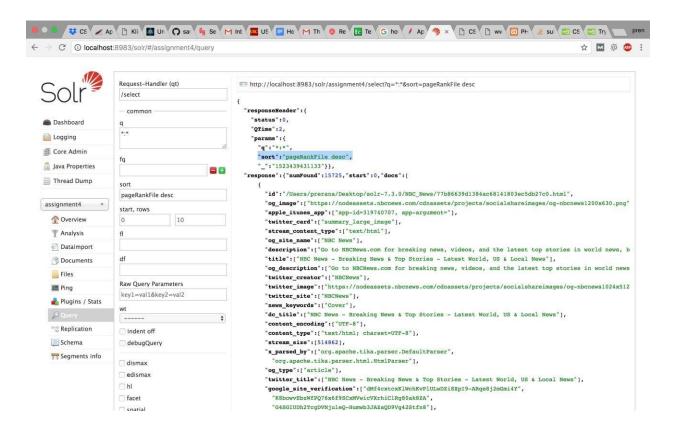


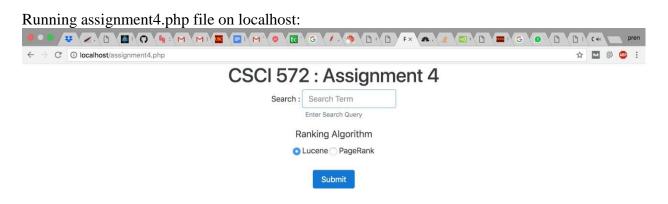
Page displayed after selecting Core. i.e. assignment4:



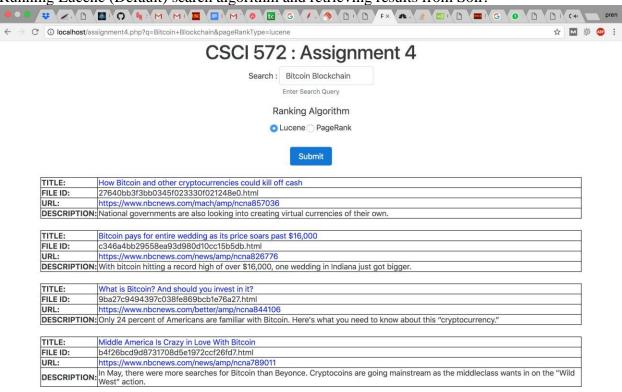


Executing query after sorting it in descending order using PageRank Algorithm:

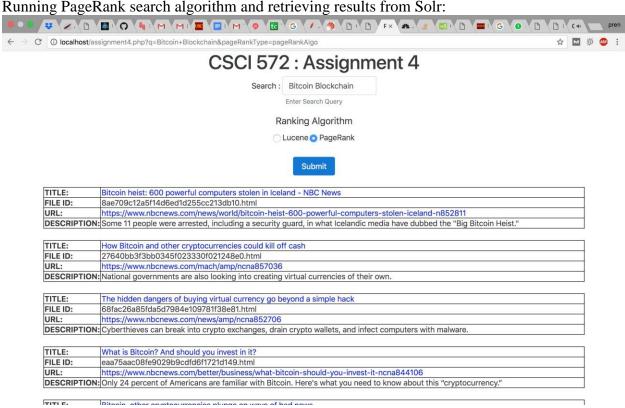




Running Lucene (Default) search algorithm and retrieving results from Solr:



Running PageRank search algorithm and retrieving results from Solr:



Page displayed after opening top link for Lucene (Default) Algo for Bitcoin Blockchain query:



4. Results:

Bitcoin Blockchain	
Lucene	PageRank
https://www.nbcnews.com/mach/amp/ncna8 57036	https://www.nbcnews.com/news/world/bitco in-heist-600-powerful-computers-stolen- iceland-n852811
https://www.nbcnews.com/news/amp/ncna82 6776	https://www.nbcnews.com/mach/amp/ncna8 57036
https://www.nbcnews.com/better/amp/ncna844106	https://www.nbcnews.com/news/amp/ncna8 52706
https://www.nbcnews.com/news/amp/ncna78 9011	https://www.nbcnews.com/better/business/ what-bitcoin-should-you-invest-it-ncna844106

https://www.nbcnews.com/mach/science/how-bitcoin-other-cryptocurrencies-could-kill-cash-ncna857036	https://www.nbcnews.com/news/amp/ncna8 56821
https://www.nbcnews.com/news/amp/ncna 852706	https://www.nbcnews.com/tech/tech- news/bitcoin-other-cryptocurrencies-plunge- wave-bad-news-n856821
https://www.nbcnews.com/tech/innovation/bitcoin-pays-entire-wedding-its-price-soars-past-16-000-n826776	https://www.nbcnews.com/news/amp/ncna8 26776
https://www.nbcnews.com/business/econo my/middle-america-crazy-love-bitcoin- n789011	https://www.nbcnews.com/news/amp/ncna8 13351
https://www.nbcnews.com/better/business/what-bitcoin-should-you-invest-it-ncna844106	https://www.nbcnews.com/better/amp/ncna 844106
https://www.nbcnews.com/business/business-news/hidden-dangers-buying-virtual-currency-go-beyond-simple-hack-n852706	https://www.nbcnews.com/mach/science/how-bitcoin-other-cryptocurrencies-could-kill-cash-ncna857036

Stephen Hawking	
Lucene	PageRank
https://www.nbcnews.com/mach/science/ brief-history-stephen-hawking-cambridge- cultural-icon-ncsl856456	https://www.nbcnews.com/mach
https://www.nbcnews.com/news/world/british-physicist-stephen-hawking-amongworld-s-greatest-minds-science-n702301	https://www.nbcnews.com/think
https://www.nbcnews.com/think/opinion/s tephen-hawking-s-life-was-made-possible- access-great-health-ncna856971	https://www.nbcnews.com/mach/video/this-futuristic-jet-pack-that-lets-users-fly-up-to-10-000-feet-in-the-sky-1189436995946
https://www.nbcnews.com/science/space/i -m-atheist-stephen-hawking-god-space- travel-n210076	https://www.nbcnews.com/mach/video/from- space-dream-to-space-junk-the-vanguard-1- satellite-completes-60-years-in-space-today- 1188094531644

https://www.nbcnews.com/science/space/im-atheist-stephen-hawking-god-space-travel-n210076	https://www.nbcnews.com/mach/video/why-youtube-s-algorithm-boosted-conspiracy-theory-videos-in-the-wake-of-the-parkland-shooting-1187660355671
https://www.nbcnews.com/news/world/amp/british-physicist-stephen-hawking-among-world-s-greatest-minds-science-n702301	https://www.nbcnews.com/mach/video/pi-in-the-sky-how-a-nasa-scientist-uses-pi-to-look-for-life-beyond-earth-1185847363999
https://www.nbcnews.com/mach/amp/ncn a818926	https://www.nbcnews.com/mach/science/alie n-satellites-might-offer-new-way-find-e-t- ncna857841
https://www.nbcnews.com/science/science -news/six-years-after-zero-g-flight-stephen- hawking-still-space-f6C9703084	https://www.nbcnews.com/mach/space
https://www.nbcnews.com/mach/science/what-stephen-hawking-taught-us-about-black-holes-ncna857471	https://www.nbcnews.com/mach/science
https://www.nbcnews.com/mach/science/s tephen-hawking-says-humans-must-leave- earth-within-600-years- ncna818926?cid=sm_npd_nn_fb_mc	https://www.nbcnews.com/mach/video/-a-wrinkle-in-time-is-probably-right-about-multiple-dimensions-1182239299525

Falcon Heavy	
Lucene	PageRank
https://www.nbcnews.com/mach/amp/ncna845256	https://www.nbcnews.com/mach/video/this-futuristic-jet-pack-that-lets-users-fly-up-to-10-000-feet-in-the-sky-1189436995946
https://www.nbcnews.com/mach/amp/ncna845096	https://www.nbcnews.com/mach/video/from- space-dream-to-space-junk-the-vanguard-1- satellite-completes-60-years-in-space-today- 1188094531644
https://www.nbcnews.com/mach/science/watch-spacex-s-falcon-heavy-	https://www.nbcnews.com/mach/video/why- youtube-s-algorithm-boosted-conspiracy-theory-

rocket-make-its-maiden-flight- ncna845096	videos-in-the-wake-of-the-parkland-shooting- 1187660355671
https://www.nbcnews.com/mach/scie nce/spacex-s-falcon-heavy-launch- gives-big-boost-elon-musk- ncna845256	https://www.nbcnews.com/mach/video/pi-in-the-sky-how-a-nasa-scientist-uses-pi-to-look-for-life-beyond-earth-1185847363999
https://www.nbcnews.com/mach/science/spacex-s-falcon-feat-may-be-just-prelude-main-event-ncna846446	https://www.nbcnews.com/mach/video/-a-wrinkle- in-time-is-probably-right-about-multiple- dimensions-1182239299525
https://www.nbcnews.com/mach/scie nce/here-s-what-will-happen-tesla- spacex-shot-space-ncna845566	https://www.nbcnews.com/mach/video/astronaut- scott-kelly-s-genes-show-long-term-changes-after-a- year-in-space-1186826819794
https://www.nbcnews.com/mach/scie nce/month-space-pictures-cosmic- penguin-playing-mars-ncsl852081	https://www.nbcnews.com/mach/video/this- futuristic-jet-pack-that-lets-users-fly-up-to-10-000- feet-in-the-sky- 1189436995946?playlist=mmlsnnd_mach
https://www.nbcnews.com/mach/amp/ncna846446	https://www.nbcnews.com/mach/video/driverless- cars-can-now-be-tested-in-california-without-a- safety-driver-1172721219646
https://www.nbcnews.com/mach/scie nce/how-spacex-aims-bring-high- speed-internet-whole-world- ncna853766	https://www.nbcnews.com/news/world/un- approves-watered-down-new-sanctions-against- north-korea-n800511
https://www.nbcnews.com/mach/science/talking-robotic-head-will-join-astronauts-iss-ncna854821	https://www.nbcnews.com/news/world/prince- mohammed-bin-salman-aims-rebrand-saudi-arabia- n817201

Syrian War	
Lucene	PageRank
https://www.nbcnews.com/storyline/syrias-children/syrian-civil-war-photo-photographer-s-rescuegoes-viral-n748146	https://www.nbcnews.com/news/mideast/syria-s-eastern-ghouta-president-assad-visits-troops-front-line-n857801

https://www.nbcnews.com/news/mideast/syria-s-civil-war-has-been-raging-7-years-what-n849851	https://www.nbcnews.com/politics/white- house/saudi-crown-prince-if-iran-makes-nuclear-bomb- so-will-n857921
https://www.nbcnews.com/news/mideast/syrias-civil-war-has-been-raging-7-years-whats-behind-n849851	https://www.nbcnews.com/storyline/aleppos- children/aleppo-s-bana-al-abed-mother-take- questions-facebook-n700336
https://www.nbcnews.com/news/ other/barrel-bombs-just-another- symbol-brutality-syrian-civil-war- f2D11765640	https://www.nbcnews.com/news/north-korea/north-korea-uses-pirate-tactics-earn-millions-sanction-busting-n856921
https://www.nbcnews.com/slidesh ow/children-among-injured- besieged-damascus-suburb- n834526	https://www.nbcnews.com/storyline/isis-terror/isis-will-remain-threat-2018-experts-warn-n828146
https://www.nbcnews.com/slidesh ow/air-raids-cause-devastation- eastern-ghouta-region-syria- n834526	https://www.nbcnews.com/news/world/analysis- trump-s-america-first-foreign-policy-isolates-u-s- n833046
https://www.nbcnews.com/news/mideast/amp/syria-s-civil-war-has-been-raging-7-years-what-n849851	https://www.nbcnews.com/news/world/north-korea-s-kim-trump-meet-andrea-mitchell-weighs-risks-n855111
https://www.nbcnews.com/storyline/syrias-children/amp/syrian-civil-war-photo-photographer-srescue-goes-viral-n748146	https://www.nbcnews.com/news/world/putin-s-bravado-over-russian-nuclear-weapons-emboldened-trump-analysts-n852301
https://www.nbcnews.com/storyline/aleppos-children/aleppo-siegebombs-hit-hospital-rebel-held-syrian-city-nurse-n657841	http://www.nbcnews.com/id/51444762/ns/meet_the_press-transcripts/t/april-lindsey-graham-bill-richardson-michele-flournoy-mike-murphy-maggie-haberman-andrea-mitchell/
https://www.nbcnews.com/storyline/aleppos-children/aleppo-siegesyrian-kids-find-childhood-deepunderground-n622036	http://www.nbcnews.com/id/52898657/ns/meet_the_press-transcripts/t/august-john-lewis-cory-booker-bobby-jindal-raul-labrador-sheryl-wudunn-al-sharpton-david-brooks-doris-kearns-goodwin/

LA Rams	
Lucene	PageRank
https://www.nbcnews.com/news/sports/how-history-informs-colin-kaepernick-s-national-anthem-protest-n639366	https://www.nbcnews.com/news/sports/sports-broadcasting-legend-dick-enberg-dies-82-n832076
https://www.nbcnews.com/news/sports/amp/how-history-informs-colin-kaepernick-s-national-anthem-protest-n639366	https://www.nbcnews.com/news/sports/how-history-informs-colin-kaepernick-s-national-anthem-protest-n639366
https://www.nbcnews.com/news/sports/amp/sports-broadcasting-legend-dick-enbergdies-82-n832076	https://www.nbcnews.com/news/sports/amp/sports-broadcasting-legend-dick-enbergdies-82-n832076
https://www.nbcnews.com/news/sports/sports-broadcasting-legend-dick-enberg-dies-82-n832076	https://www.nbcnews.com/news/sports/amp/how-history-informs-colin-kaepernick-s-national-anthem-protest-n639366

Black Panther	
Lucene	PageRank
https://www.nbcnews.com/news/nbcblk/black-panther-s-wakanda-sheds-light-black-excellence-n844801	https://www.nbcnews.com/think
https://www.nbcnews.com/news/nbcblk/black-panther-sets-fandango-record-top-preselling-marvel-movie-n836631	https://www.nbcnews.com/better/pop- culture/girl-scout-cookies-ranked-better- worst-ncna737381
https://www.nbcnews.com/news/nbcblk/black-panther-novelist-journalism-sci-fi-bring-hidden-light-n848011	https://www.nbcnews.com/pop- culture/movies/black-panther-tops-box- office-fifth-straight-weekend-equaling-avatar- n857781
https://www.nbcnews.com/tech/internet/pl ot-destroy-black-panther-ratings-thwarted- facebook-n844226	https://www.nbcnews.com/think/hot-take
https://www.nbcnews.com/pop- culture/movies/black-panther-blows-away- box-office-192-million-weekend-n849246	https://www.nbcnews.com/think/thought- experiment

https://www.nbcnews.com/news/nbcblk/black-panther-black-comic-creators-spentyears-breaking-down-barriers-n848461	https://www.nbcnews.com/pop- culture/celebrity/celebrity-deaths-2017- looking-back-famous-figures-we-lost-n831216
https://www.nbcnews.com/news/nbcblk/black-panther-actress-janeshia-adams-ginyard-talks-living-stuntwoman-life-n847646	https://www.nbcnews.com/better/pop- culture/pros-cons-america-s-favorite-diet- plans-ncna854396
https://www.nbcnews.com/think/opinion/defeat-killmonger-black-panther-holds-lessons-israel-ncna850831	https://www.nbcnews.com/pop- culture/awards/golden-globes-2018-metoo- casts-shadow-over-celebration-movies-tv- n835451
https://www.nbcnews.com/news/amp/ncna 848011	https://www.nbcnews.com/better/pop- culture/anti-diet-movement-leading-us- astray-ncna854856
https://www.nbcnews.com/news/amp/ncna 847706	https://www.nbcnews.com/storyline/confede rate-flag-furor/house-would-ban- confederate-flags-va-cemetery-flagpoles- n576906

Austin Bombings	
Lucene	PageRank
https://www.nbcnews.com/news/us- news/police-identify-teen-killed-austin- package-bombing-investigation-grows- n856371	https://www.nbcnews.com/better
https://www.nbcnews.com/news/us- news/deadly-austin-bombings-were-meant- send-message-police-chief-says-n857786	https://www.nbcnews.com/mach
https://www.nbcnews.com/news/us- news/amp/deadly-austin-bombings-were- meant-send-message-police-chief-says- n857786	https://www.nbcnews.com/leftfield
https://www.nbcnews.com/widget/video- embed/1184733251954	https://www.nbcnews.com/news/us- news/nbc-affiliates-n19981

https://www.nbcnews.com/news/us- news/amp/police-identify-teen-killed- austin-package-bombing-investigation- grows-n856371	https://www.nbcnews.com/think
https://www.nbcnews.com/news/us- news/austin-texas-edge-after-spate- package-bombs-kill-two-n855956	https://www.nbcnews.com/news/us- news/deadly-austin-bombings-were-meant- send-message-police-chief-says-n857786
https://www.nbcnews.com/news/us- news/amp/austin-texas-edge-after-spate- package-bombs-kill-two-n855956	https://www.nbcnews.com/news/us- news/veterans-hospital-killings-spotlight- problem-violence-against-health-care- workers-n855931
http://www.nbcnews.com/search/	https://www.nbcnews.com/news/us- news/cirque-du-soleil-star-yann-arnaud- plunges-death-during-florida-n857761
https://www.nbcnews.com/news/latino/celi a-cruz-be-remembered-grammy-lifetime- achievement-award-n516956	https://www.nbcnews.com/news/latino/mad uro-challenger-henri-falcon-unsettles- venezuela-s-presidential-vote-n857836
https://www.nbcnews.com/news/latino/cali fornia-affirmative-action-bill-sparks-fierce- debate-n87426	https://www.nbcnews.com/better/wellness

Paris Climate Deal	
Lucene	PageRank
https://www.nbcnews.com/storyline /trump-s-first-foreign-trip/trump- declines-endorsing-paris-climate- change-deal-g7-summit-will-n765406	https://www.nbcnews.com/tech/social-media/tech-companies-struggle-human-side-fake-news-n857426
https://www.nbcnews.com/storyline/trump-s-first-foreign-trip/amp/trump-declines-endorsing-paris-climate-change-deal-g7-summit-will-n765406	https://www.nbcnews.com/storyline/trump-s-first-foreign-trip/trump-message-during-overseas-trip-now-he-s-home-n765136

https://www.nbcnews.com/politics/white-house/trump-pulls-u-s-out-paris-climate-agreement-n767066	https://www.nbcnews.com/news/world/analysis-trump-s-america-first-foreign-policy-isolates-u-s-n833046
https://www.nbcnews.com/politics/ white-house/here-s-what-happens-if- trump-leaves-paris-climate- agreement-n766761	https://www.nbcnews.com/news/us-news/trump- un-north-korean-leader-suicide-mission-n802596
https://www.nbcnews.com/politics/politics-news/forget-trump-u-s-storms-ahead-climate-change-never-n801826	https://www.nbcnews.com/feature/meet-the-press- 24-7/meet-press-august-6-2017-n790001
https://www.nbcnews.com/meet- the-press/meet-press-june-4-2017- n768046	https://www.nbcnews.com/news/us-news/trump- un-north-korean-leader-suicide-mission- n802596?cid=eml_nbn_20170919
https://www.nbcnews.com/news/world/analysis-trump-s-america-first-foreign-policy-isolates-u-s-n833046	https://www.nbcnews.com/news/china/china-backs-boasts-new-era-patriotic-movies-rap-music-n843686
https://www.nbcnews.com/storyline/trump-s-first-foreign-trip/trump-end-first-foreign-trip-pledges-unwavering-support-u-s-n765446	http://www.nbcnews.com/id/53771922/ns/meet_the_press-transcripts/t/dec-charlayne-hunter-gault-tom-brokaw-jesse-jackson-richard-stengel-charles-ogletree-al-sharpton-katty-kay-paul-gigot-harry-smith/
https://www.nbcnews.com/storyline/trump-s-first-foreign-trip/g-7-summit-leaders-brace-clash-trump-trade-climate-n764971	https://www.nbcnews.com/mach/science/michio-kaku-sees-amazing-things-our-future-except-those-scary-ncna851226
https://www.nbcnews.com/storyline/trump-s-first-foreign-trip/amp/trump-end-first-foreign-trip-pledges-unwavering-support-u-s-n765446	https://www.nbcnews.com/storyline/trump-s-first-foreign-trip/donald-trump-nato-why-his-silence-article-5-big-deal-n765011

5. Overlap Table:

Query	Number of Overlaps
Bitcoin Blockchain	5
Stephen hawking	0
Falcon Heavy	0
Syrian War	0
LA Rams	4
Black Panther	0
Austin Bombings	1
Paris Climate Deal	1

Overlap Chart:

