|  |  |  |
| --- | --- | --- |
| **Query term** | **No. of overlaps between Lucene(Boolean Vector Space model and PageRank)** | **Matched Document IDs** |
| Brexit | 1 | 82a97ada-a5e6-454d-9c27-7197c64a358e.html |
| NASDAQ | 2 | |  | | --- | | 3ab1b080-82be-4e05-9ffe-4960303b3935.html | | 2c2f4a33-b2e2-4655-b583-22d9267a203b.html | |
| NBA | 0 | NA |
| Snapchat | 1 | 22d862ac-4869-4dde-bcd1-8ec35dc8042b.html |
| Illegal Immigration | 0 | NA |
| Donald Trump | 0 | NA |
| Russia | 0 | NA |
| NASA | 0 | NA |
| **Total No. of overlaps** | **4** | NA |

**1.Overlap Table**

**2.Overlap Graph**

**3a.Top 10 Results in Lucene and Page Rank(Brexit)**

|  |  |
| --- | --- |
| **Lucene** | **Page Rank** |
| 25d8f6d1-808b-42c5-9b1e-074f3d211634.html | 69231edd-ceee-401c-9691-762281d40ffd.html |
| c399d9ce-4ea4-4895-98b7-87b35f8e98b7.html | 45cbcda5-6acd-4db0-a8a2-717f2f0657f2.html |
| a7c5a9f5-ab12-41fa-9ed5-8ee34607aa9d.html | 3c06d6d6-6f2c-4363-a1fe-c8a248a0f763.html |
| 82a97ada-a5e6-454d-9c27-7197c64a358e.html | 3f5a90de-7fdd-42e2-850c-fa738afaeb1c.html |
| 190c4540-3bde-45a4-9c65-a4776c88eef8.html | 023f03ba-bcdb-4b92-89a1-db3d9dc26884.html |
| a0d7199b-67b2-4c19-a59e-afc9c070fb26.html | 854f8c49-f98b-4d4a-80da-c0fa3bdc2f05.html |
| 0764af1f-3d42-469b-a1e3-ba8c9754776a.html | 82a97ada-a5e6-454d-9c27-7197c64a358e.html |
| a66a8982-1735-4493-88ef-80f341989b98.html | 55f5c603-f395-45fa-9092-dc359bf6d146.html |
| 9db10abd-2419-4a6a-884b-cd0a428131db.html | 613e64d2-e4e2-4f7e-8c75-58ec6afc9bc7.html |
| a83f07b6-4a67-45b5-8c48-4ef43608e847.html | 613e64d2-e4e2-4f7e-8c75-58ec6afc9bc7.html |

**3b.Top 10 Results in Lucene and Page Rank(NASDAQ)**

|  |  |
| --- | --- |
| **Lucene** | **Page Rank** |
| 3ab1b080-82be-4e05-9ffe-4960303b3935.html | 023f03ba-bcdb-4b92-89a1-db3d9dc26884.html |
| fa3bb02c-3ef3-48d3-9ea0-f2c19e2368af.html | 3da5a28a-2348-472d-976f-75013fdc8408.html |
| cf21932a-4697-4974-9656-7df0df79bdc3.html | 37216d61-debc-42cd-a00a-f902038e10ee.html |
| 28f22131-7932-47fb-a873-f0719eeefbc1.html | 2c2f4a33-b2e2-4655-b583-22d9267a203b.html |
| d21b454c-bf80-41e7-99fb-0603d4bb5e4a.html | 3ab1b080-82be-4e05-9ffe-4960303b3935.html |
| 8db99298-4581-4092-b0a9-b9e0f0b4cedf.html | c9827c54-ee1b-46c8-90c9-0c95674cd1bf.html |
| b1266016-5c61-4817-ba31-f26190d14835.html | f6af0e78-c278-4c1d-abec-405f346eb3af.html |
| 2c2f4a33-b2e2-4655-b583-22d9267a203b.html | dc8f4fdb-cab9-4ec0-bf25-39a76d1e737d.html |
| 340df3a6-6208-4e65-ab63-66ed859a5022.html | 3a2f8010-d934-48dc-86ff-38e30feeef92.html |
| a0509cdc-57c8-4ae4-9681-66ec05087c3f.html | 73ca4bab-47b7-464b-9174-c77aa7cf1416.html |

**3c.Top 10 Results in Lucene and Page Rank(NBA)**

|  |  |
| --- | --- |
| **Lucene** | **Page Rank** |
| 27f86ea4-de92-4e46-b783-afbe480fdef1.html | 5328816a-ad71-44fb-8ff5-c453601efb37.html |
| 24148f94-b278-4338-8a63-88ee8965f526.html | b11d4fa2-5703-4783-b292-a8ae419d8ed8.html |
| 978adc0c-e519-4bcc-ac9f-b2312840dd5f.html | 3d18dd47-efe7-48d5-856a-bca3a24741d0.html |
| 3cab90e1-350c-4e1b-b656-eda814119503.html | bd34721b-2227-4e0d-8d92-09ea2b5862ee.html |
| 07179d47-270b-49fc-bb49-745cc636b018.html | 1a532cf7-7c39-48e6-9e37-5aa8b7a8d1ab.html |
| 7a9616ec-422d-46c8-a2ef-07b87819a572.html | 192cf384-d794-456b-85c0-29b5d8dec242.html |
| 069c60f9-8554-48b8-82fb-10fe1d8ea718.html | 0f7ae22a-3016-4ef6-a9af-cb953da216af.html |
| e9c31554-28c7-4cc6-ab2b-7ca7b15c6cb1.html | 04d0cd22-eace-4527-a766-7f568360adcb.html |
| 731106cf-3bd2-43c1-bd93-7e6cff99f0ea.html | da38403f-06a3-4557-88c0-dd541b3e8663.html |
| 7ae8781b-dc71-45c1-90be-38a9fc5d96d3.html | cf2cd074-7bc7-459f-ba70-e09d937251fc.html |

**3d.Top 10 Results in Lucene and Page Rank(Snapchat)**

|  |  |
| --- | --- |
| **Lucene** | **Page Rank** |
| 0cb21b42-d443-4664-8a99-15c2eb20e81f.html | 56d42531-5dc7-443e-abad-75827000c98f.html |
| 010f3ed6-10de-4a7e-a90c-eee36411d529.html | 0422de08-73b9-4639-a2c3-88b45c655a05.html |
| 4e8275ec-fe49-417b-b584-fdaaf26f916c.html | 2d381114-d720-4ac8-8559-4a91608c7b15.html |
| 80263883-5064-4ded-b4e5-1c267e39e174.html | 6365d874-cf31-4e1a-a18d-d34a349e3e72.html |
| fe8d4dd1-8bc2-4b05-942c-3ad90c398fc3.html | c9449668-4c16-4686-9554-ee90d7a3c56a.html |
| 2ae8fd83-0fde-4106-965f-b227f2231150.html | 22d862ac-4869-4dde-bcd1-8ec35dc8042b.html |
| de5bb543-684d-4c6b-a45e-39bcf0d11bd6.html | 024f82b6-4616-4048-9f33-bbed10822799.html |
| 22d862ac-4869-4dde-bcd1-8ec35dc8042b.html | d4e31447-3a8c-49a7-a9b5-64d87af30197.html |
| f9eecd1c-46a8-46ae-b996-ece18dca0177.html | 9f3ca530-d309-4b0a-a510-40b5cf2f9bb8.html |
| d5eb6923-506a-401d-be71-26ffb4828534.html | c7b3498e-a592-45bf-b310-03537fdb64ed.html |

**3e.Top 10 Results in Lucene and Page Rank(Illegal Immigration)**

|  |  |
| --- | --- |
| **Lucene** | **Page Rank** |
| 1f5c661e-72aa-4aa1-942d-89ad708ffda7.html | b35af3b6-d161-4a9e-ac3f-be5920b7188d.html |
| a6afdbe5-4b86-44aa-b9c6-c9077115b96b.html | 326bb6d5-fb9f-4ca0-82d5-75f5baa7f894.html |
| eb9d3a79-ddef-4884-9da3-7ceeffd79c8b.html | dd432b45-2219-4c2a-8bf1-bb8834e13584.html |
| ba06f123-c441-425c-a4e9-2d5bb4019751.html | f69f6fdc-6018-4bfe-bc62-3f141cb6973c.html |
| 3e8c1075-0926-408c-99f6-9bbbe39e474a.html | 137d3d61-85d5-49a9-bfdc-84328c5460ba.html |
| da39a88a-c347-4b00-91f7-28984ae50a1f.html | 8f22fc06-cc35-448e-b9d4-b02af62dad6f.html |
| 6ea43c16-5e37-47c9-bb46-7a2b7e273add.html | 0222a4b6-1476-4616-bdf3-21e9bcfd8f3b.html |
| 1e99f0be-9e3f-42c5-bb7f-e42eeeda555b.html | e256b2de-82e5-4efb-8eef-ea306f54046f.html |
| c97fb364-b2f8-4718-af0d-6b993ae74435.html | bd8c7738-daa4-48af-ae60-9ac9ff3b3b3f.html |
| 84bc1c7c-d0f3-40d3-851c-8efc62ecba28.html | 83addeae-2d90-4776-aa46-f17082bad944.html |

**3f.Top 10 Results in Lucene and Page Rank(Donald Trump)**

|  |  |
| --- | --- |
| **Lucene** | **Page Rank** |
| 0c395e07-8d48-48f1-8a09-ad062be76428.html | c84ddeb3-a200-421a-aa27-952bbc5d66cc.html |
| 32ddda93-cecc-46ed-b860-3226ec717c4a.html | e894e860-bfe9-4bdf-b622-9911c9438a22.html |
| 7fbc9283-38ac-4e2c-a7d6-a18592d4f7d0.html | bcff6ff4-a751-41f7-b121-1583879d9c2b.html |
| 8ae6718c-2f8a-4c38-bde2-402bdc63724e.html | a52b85a8-d9cc-4798-b39d-2e7564bf3e0c.html |
| f336c49d-da11-4b72-99a0-6a8dffa91285.html | bc5e5e3e-62fd-47d5-a5c2-32f219608f67.html |
| fbe923b2-3b6f-481d-bf2d-0c11580542c3.html | 3a9091eb-89a3-4b4e-a9c8-76677c9961be.html |
| dfd4c799-a803-4185-9362-b6d4c41606e5.html | b35af3b6-d161-4a9e-ac3f-be5920b7188d.html |
| 163950c6-89ab-404c-adf4-ebcbd715b77c.html | 244a18e8-8a8a-4e00-bdb3-7f35c600b92a.html |
| b5f15b40-5ccd-4aea-ab63-49750c4a3f54.html | fff879a1-b3cf-4eca-a93f-183bbd5ce73d.html |
| 746c9239-649d-4536-9ea7-f3d4d2458987.html | b5ab432d-a7dc-44a7-b532-1291975e6ee1.html |

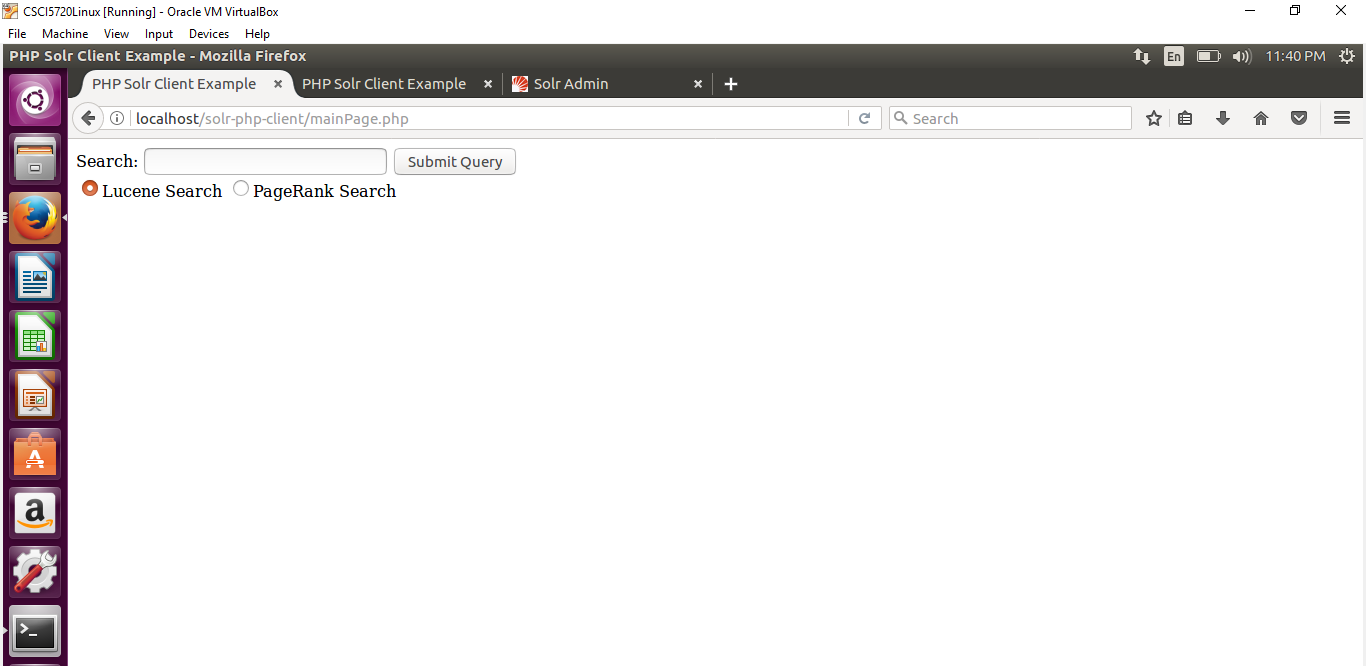
**3g.Top 10 Results in Lucene and Page Rank(Russia)**

|  |  |
| --- | --- |
| **Lucene** | **Page Rank** |
| 3e52fc77-5e3e-4bf3-86a2-3d4394d18e41.html | c84ddeb3-a200-421a-aa27-952bbc5d66cc.html |
| 20f8b522-47af-4cf8-81b3-e7b48065f8bd.html | b5ab432d-a7dc-44a7-b532-1291975e6ee1.html |
| 9c971824-8243-4288-8dbf-ce7054375b16.html | 326bb6d5-fb9f-4ca0-82d5-75f5baa7f894.html |
| 996ebb98-e611-47c5-99e9-f01c18567cfb.html | a8aac8c6-9520-4ed4-92ca-413a94704f35.html |
| 662beacd-433a-4743-9df4-42e85b95fa6d.html | ff81de16-4b1c-43f6-a5c1-40b0c3ff4ccd.html |
| 194471ca-94c5-4c1e-b11e-688a4bc4bb9b.html | c1fe061c-463a-4c22-9f00-89fa89c2c316.html |
| 36033678-07c9-4645-bc45-fa47af1d13c6.html | 60b148c3-be19-421c-9028-cce49caf9607.html |
| 0cf12bd0-e0b5-41d1-b4f9-10ee970c3e46.html | 7eb11b9e-9375-4a30-8634-e6c9226fb6b3.html |
| 38b8cc92-d40b-4407-8dd5-1a5102d0444c.html | 5c7ff6a2-effa-46f6-bb42-b4b3c526dcd7.html |
| 6958dbf9-427f-4286-9205-85b127ae24e7.html | 62ad6bf1-4597-4007-b42e-8004bf03e686.html |

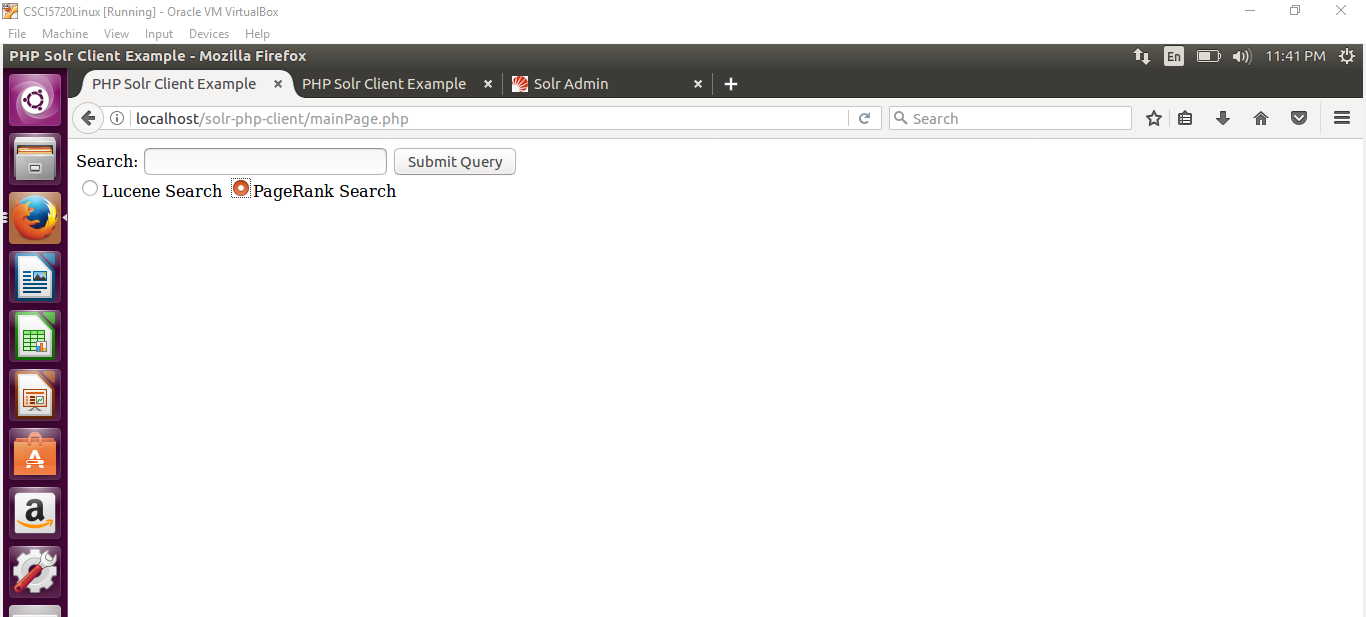
**3h.Top 10 Results in Lucene and Page Rank(NASA)**

|  |  |
| --- | --- |
| **Lucene** | **Page Rank** |
| 9f3aae9f-6ac8-4e0f-bcb6-b3bad0da5a3d.html | c84ddeb3-a200-421a-aa27-952bbc5d66cc.html |
| 9b8bc114-5bfb-4a07-a5dc-4e789db9beb7.html | 2030b587-8eba-460a-87f3-2303e3390857.html |
| ab2070cb-1bf7-4d98-8685-f3929ba2a285.html | e894e860-bfe9-4bdf-b622-9911c9438a22.html |
| 74b37142-1822-4ccb-81ea-44b24d900e86.html | a52b85a8-d9cc-4798-b39d-2e7564bf3e0c.html |
| 7f0e7647-2139-4a2e-9252-6ccd1b56d21c.html | b5ab432d-a7dc-44a7-b532-1291975e6ee1.html |
| 47172b45-3f17-40a9-a706-d7a3c5999c63.html | 326bb6d5-fb9f-4ca0-82d5-75f5baa7f894.html |
| f8057905-1f21-4578-a0e4-ca6676950f5c.html | dd432b45-2219-4c2a-8bf1-bb8834e13584.html |
| 72f68214-c043-4753-b77c-36acab37f8cd.html | a8aac8c6-9520-4ed4-92ca-413a94704f35.html |
| 7a99b487-01bc-4ef8-834d-f16745a41709.html | 435f99ca-31cf-4c07-982e-cb250415eb9e.html |
| 082915f1-5c7b-4fd5-98a8-314c033e6af0.html | 43397f6a-c173-4552-b5c8-82b25c357d29.html |

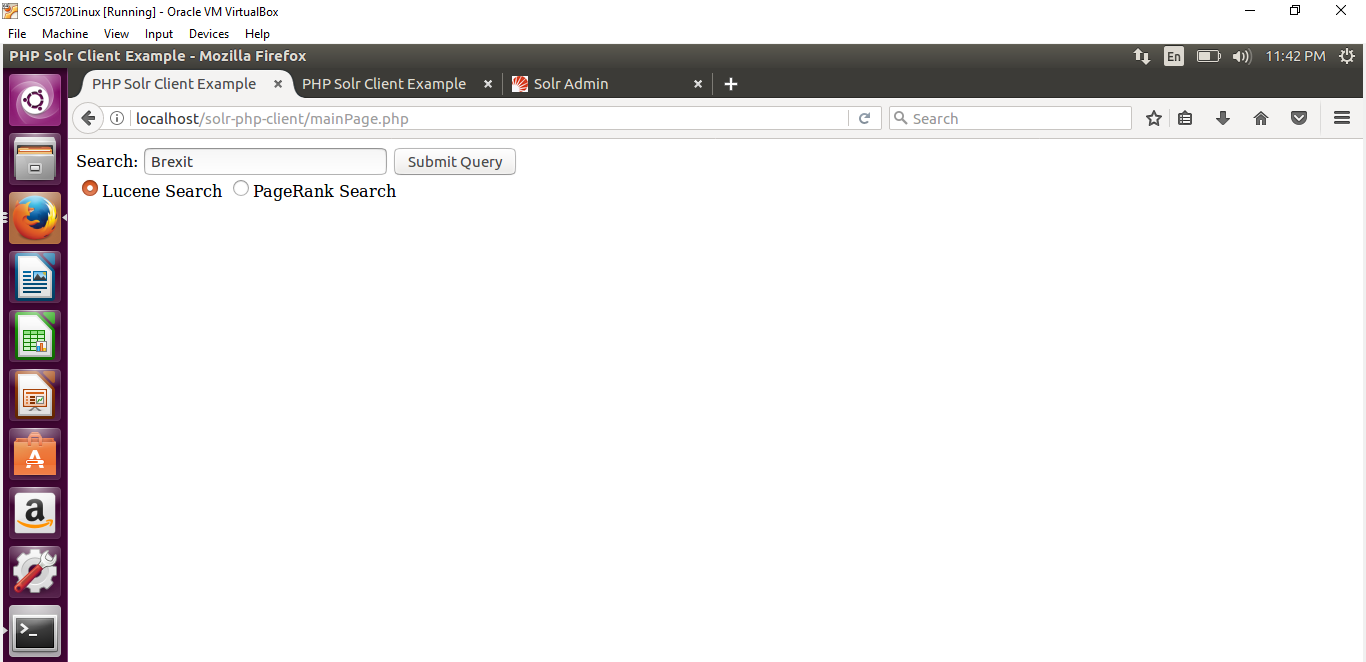
**Screenshot-1:Initial Screen-Lucene**

****

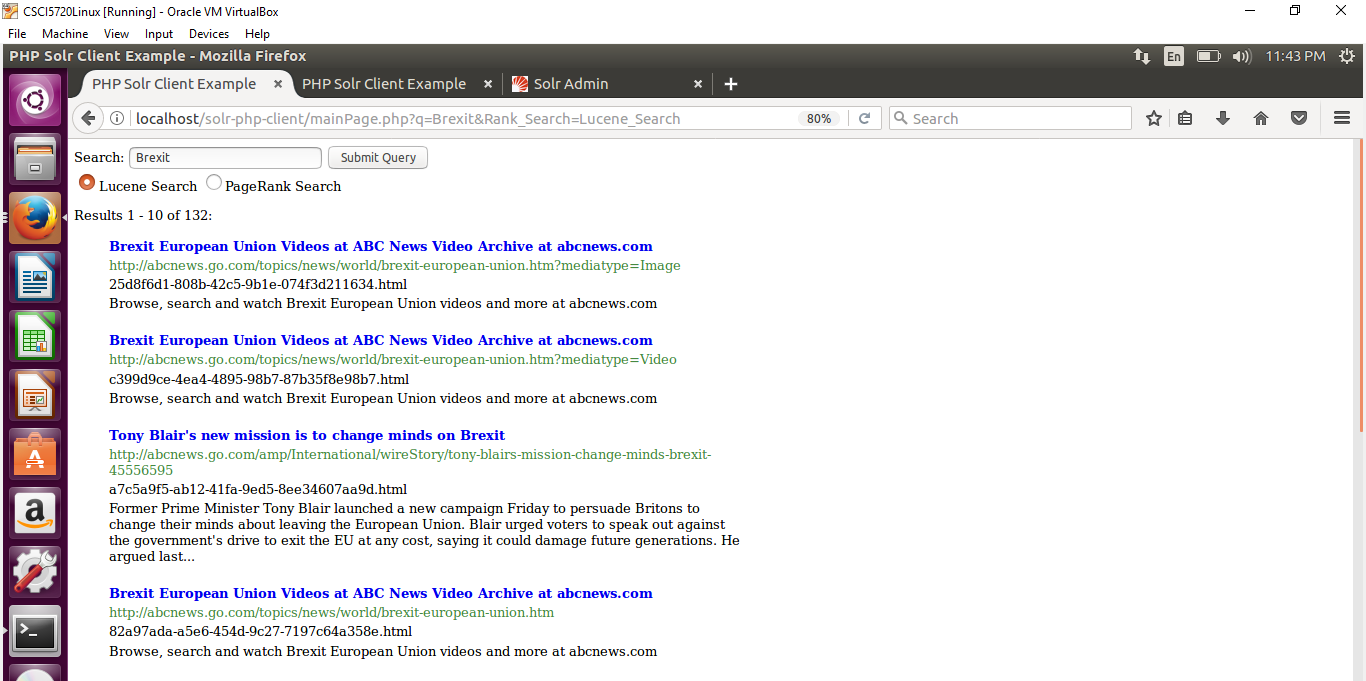
**Screenshot-2:Initial PageRank Screen**

****

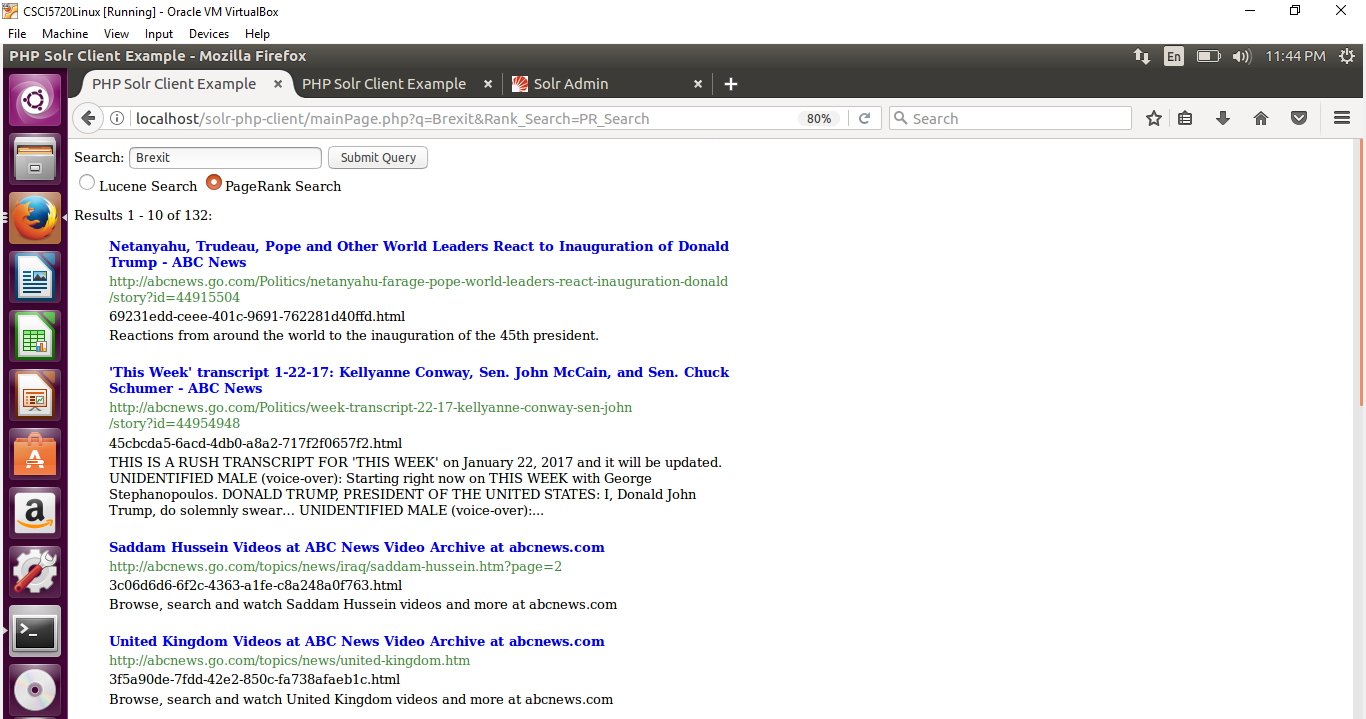
**Screenshot:3 Search term in Query box**

****

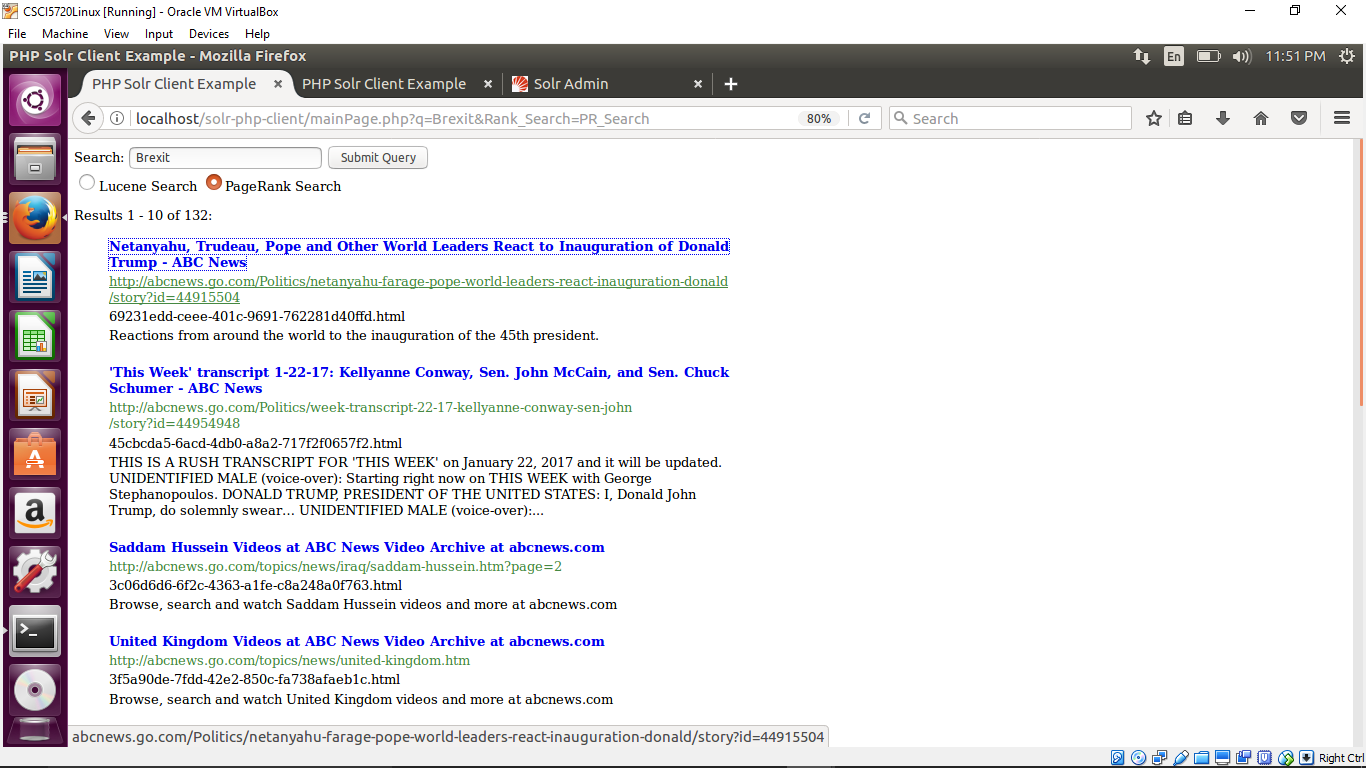
**Screenshot:4-Lucene Search Result**

****

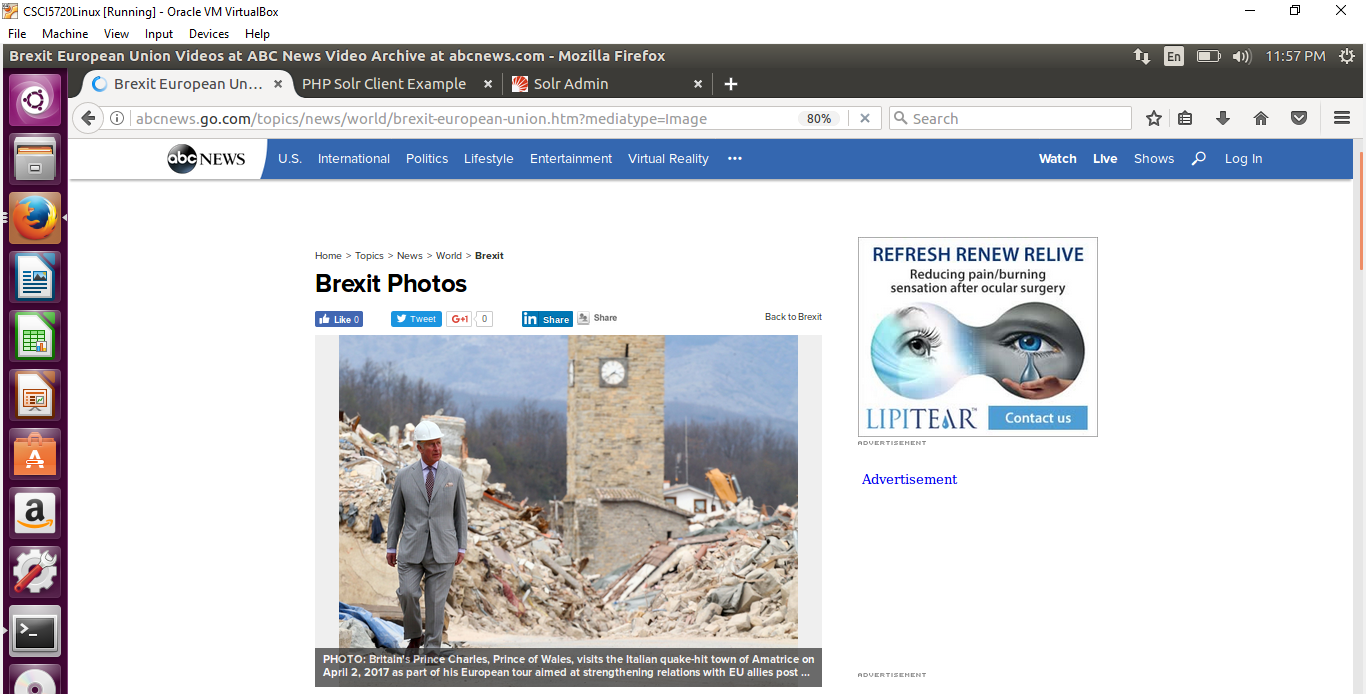
**Screenshot:5-PageRank Search Result**

****

**Screenshot:6-Click on link**

****

**Screenshot:7-Web page**

****

**Steps followed for comparing ranking algorithms**

**Step:1 - Installation of ubuntu 16.04 and solr-6.5.0**

The first step was to install ubuntu 16.04 and the source code download for solr-6.5.0 and

the installation of solr-6.5.0 on ubuntu

**Step:2 - Indexing crawl data's html pages using solr**

This step involved the following:

a.Starting and intitalizing solr using the bin/solr start command.

b.Creation of a new core on solr using the bin/post create command.

c.Once the core was created, Indexing was performed on html files provided in the mapping file by using the command:

bin/post –c abcNewsCore -filetypes html <path\_to\_crawl\_folder>

where abcNewsCore is the new core that was created.

The above mentioned steps would enable the user to query the indexed files using the solr admin portal

**Step:3 - Extracting the outgoing links from the downloaded crawl data**

--This step involved the creation of the edgeList.txt file using the crawl data and the mapping file provided using

the APIs available in the Jsoup library.

--Two hashmaps were used to store map encoded filenames to url and url to filename.

a.Using the Jsoup library, the outlinks for every file available in the mapping files were extracted and for each link an entry is made into the edgeList.txt file in the following format:

b.if a.html is the page and b.html is an outgoing link on a.html, then an entry a.html b.html would be made in the edgeList.txt file

**Step:4 - Constructing the networkX graph and computing the pagerank**

a.In this step, an open source utility networkX is used to computepagerank.

b. Using the read\_edgelist() api available as a part of networkX, a networkX graph G is constructed by feeding the

edgeList.txt file as input.

c.Using the pagerank function available as a part of networkX, the pagerank for all the files in the mapping is computed

and is stored in the external\_pageRankFile file. The parameters for the pagerank function are as follows:

G - Network produced from step b

alpha - 0.85

personalization - None

max\_iter - 30

tol - 1e-06

nstart - None

weight - 'weight'

dangling - None

**Step:5 - Reloading the index to facilitate pageRank**

a.The external\_pageRankFile created from the previous step is placed into the data folder of the core which was created

b. The following two lines were then added to the managed schema file

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

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

c.The following two lines were added to the solrconfig.xml

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

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

d.The core is then reloaded using the reload option available for the abcNewsCore in the solr admin portal

**Step:6 - Integrating the front end using PHP**

Apache and PHP are first installed on ubuntu and the github repository for solr-php-client was cloned into the

apache folder - /var/www/html

a.This step involves the design of the search box where the user can type in the query and the radio buttons to provide options for Lucene Search and PageRank based Search using the mainPage.php available inside the solr-php-clint

b.Once the query is submitted, an instance of the new solr search query is created by passing the following parameters

i. $query - the query string entered by the user

ii. $additionalParameters - If the search option was lucene, then this paramenter is not passed since it would be the default algorithm used by solr.If the search option is Page Rank, then this parameter consists of the field 'sort' with the value 'pageRankFile dec'

c.Once the search results are fetched using the indexed files, the top 10 results for each of the algorithm containing the id, url, title and

description is displayed with clickable links for url and title.

**Step:7 - Computing overlaps and generating bar graphs**

Based on the results produced by Lucene search and PageRank, the number of results which appear in both the searches are computed for all

the queries and are consolidated into a bar graph. The overlap parameter is the id which is unique for all the documents.

**Explanation for higher Page Rank**

The reason why certain pages have higher page rank is that the number of occurrence of those pages as outgoing links in the body of other downloaded crawl pages are high and the argument to the pagerank algorithm is the weight which is edge data key which is used to determine the rank of every document.