***CSCI 572 Report NBC News***

***Vibhanshu Sharma***

***USC ID: 9698-3656-82***

To complete the assignment, I installed Ubuntu virtual box in my system. Then, I installed and setup Solr in the virtual box. Once it was setup, I started the solr server and created a new core, naming it – myexample. Then I made the changes in the solrconfig.xml and managed-schema file as was mentioned in the assignment. Once, all the changes were made, I was ready to index html pages from my news website, i.e., NBC news.

After indexing the webpages, I created a user interface, where user can enter his queries and get the top ten results for that query, just like google search. This search result ranking was from Lucene. In order to get search result ranking according to Page Rank Algorithm, we had to compute page rank scores using page rank process.

In order to compute page rank, I first created a web graph structure and extracted the outgoing links from the downloaded html files provided. With the help of example code given in the assignment, I created edgelist from the downloaded files. Then, I gave this file as an input to my networkx in python. Using the below parameters, I computed the page rank.

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

After computing the page rank scores, I placed this file in the data folder of the core that I had defined. Again, I made changes in the solrconfig.xml and manages-schema file as mentioned in the assignment.

Some pages have higher page rank values because the number of incoming and outgoing links in that web page are more than others pages. This depicts that the page with higher page rank score has more links, in the form of outgoing links, of the query entered by the user.

**The overlap for NBC news is zero.**

Comparing the results of the two-ranking algorithm, solr and page rank, I got the following results:

1. **Brexit**

**Solr PageRank**

035b6cbb-904d-4dcb-acc1-b39b55555dd9.html 80f37b5a-18a2-479d-b9e3-b04c5263429b.html

2c45d221-46c9-4cf3-920d-c3a87485dfb8.html 4d722b23-249a-4773-a9f2-b950798ff056.html

61238700-526c-4e3f-9f71-69d152cc0d87.html 8b1859e6-33c0-4ea4-801e-122154a17e3d.html

cacc07c9-689c-4d38-819f-bbc797bdc7f6.html e5e0f461-2307-4097-b38f-a34090659562.html

dc3ce7bc-56b0-4dd5-b2de-f421d8167c5c.html 84b60d2a-7d3e-45a9-87ef-250c6e3d6e3f.html

46550da7-0f82-42bc-840b-e533e1d10a74.html b7eeeeae-9225-4137-8f00-99c599ec6f02.html

44659258-cfcb-4edb-ba77-6018c26bae37.html 638518c2-3c83-44df-b22a-f1f0d15f111a.html

4af864b4-e543-4ccb-b5ed-aff04ac2ba0c.html 549d660a-4317-431a-bc39-f57e1900dea8.html

ea483d6c-53c2-476f-948d-97d6cd0cb394.html 8f5c84dc-1513-406d-8a92-5a87d930e3b4.html

2f73e6ee-2725-4f3c-874e-bb13fc5c2ad7.html e4584bca-b012-4444-8164-c14ae8b517d1.html

1. **NASDAQ**

**Solr PageRank**

619371a9-8b61-4931-9f10-2aa323789d72.html 5b4f205d-1956-4b8b-adc0-abfb86883f81.html

f77ce5da-0913-44a7-b8ef-a1357a0dae66.html dd47892e-dd51-4f18-abeb-bdd1a3870a19.html

0ece91f9-a3bb-43c8-8f94-e1fa30daa331.html ae0e9652-7749-4207-b5d1-03d0149b4ee6.html

ffddfb44-e65b-4fb6-9a2f-630262352dfe.html 35ac79f4-2a62-4992-b0b7-c778298e0b52.html

92b21a9a-d9c3-4275-ab92-2c17c6ef3522.html 8b1859e6-33c0-4ea4-801e-122154a17e3d.html

6931aa4a-3275-42d4-8075-73ee419a1d4e.html 5125c2cc-e0cf-4b73-a8ca-0f9727b48aba.html

92f16f45-b494-4494-9c6e-6eb420bf96b0.html ea81a411-c9da-4e60-9e56-6e3a048838df.html

801de5e3-3c70-46f5-9854-20909c366d8b.html e749e1c1-9f57-4f89-ab7e-370d9e84b099.html

14aa1fbe-c131-43d9-a22e-bbe7f05e9a87.html d9fadc48-237a-47af-8232-70901c4608be.html

47353bda-6fe2-4cf8-8431-0643863e4001.html c68a6adc-92ea-4d39-ab4b-2b84adc25e8f.html

1. **NBA**

**Solr PageRank**

2d79f0ea-4ccf-4bd7-be75-26d64a67fee8.html 5b4f205d-1956-4b8b-adc0-abfb86883f81.html

67dbcb60-1e63-42d0-bd4d-4a988236c493.html 963f0e3b-dff7-4c77-b532-ad2bd3357bef.html

f2d01e8b-5f47-4a93-a2db-dc09b5b6aeb6.html bfca1cda-95c9-4ed2-85ac-2a710e534e08.html

23dd7a9a-0edc-429f-8a17-76f6769e9528.html 824be3ec-0389-4e12-b6e7-5ccec84d445b.html

0eb00e96-d3fb-469d-a91c-6b6ca4d724b2.html 66156dd0-3230-4a6c-a952-9351b7889df8.html

8923f191-028a-44ad-b09c-e1c599b1cd28.html fba95525-7331-4160-a3c3-d758ea0c198c.html

24afcc01-1665-4e74-9b1a-87a343ad3e4d.html 1f9ac5e4-d876-4327-8e40-79e2fbf0f3af.html

b8a4e7d0-1070-484f-bc35-315f04ebeaff.html b2e82089-acab-4ae5-afe8-fcb792251047.html

5a786e10-ccac-4e7a-9e1e-58647b285d12.html e1ee34a6-a961-4ffb-b5b2-0873f9163641.html

e32397a2-a09f-4135-bd67-46e9ece82bb9.html de5dc4e0-4eb6-4bc3-bbcb-775723dfa7e3.html

1. **Snapchat**

**Solr PageRank**

d7f88191-453c-4cc1-8cad-25c9f2541bd0.html 2c456792-7f88-4b20-8128-0843660a0267.html

af7031ce-74be-4b88-81f4-84056b26206b.html b510c49b-c9ad-4459-aecb-c6bfe2221389.html

a10ab0ed-e53d-4b4f-8508-cfbd1577ad32.html a52e857a-3c6b-4ddf-850b-551989619f9b.html

a4c6131e-e962-4d9b-8808-649ca9030de7.html 283132e5-1845-45b6-9501-baaf90a3120f.html

ac902771-644d-4027-93ee-39fdb5ff8a4b.html 7d6b8eca-a7e0-42c9-b7de-ecc9a3cc85f7.html

a431a5d1-ec3c-4e7d-9f29-faa727fa76d8.html 3316704a-8283-435f-a5a3-117cd8fd8413.html

e0c17f48-c22b-4399-996a-faab4ea3c087.html 9077b0d9-b9ce-43b8-85dc-4781efd29b2d.html

e63b1d46-2084-4634-8b59-76c7487ab9a0.html 9fb2754a-7428-49d1-b790-2da89abe1e3a.html

e7b0bcc8-260b-438d-8950-41ec392e8905.html 03cfade2-ebb4-4398-aa04-baf671d9bd21.html

85f874d4-1527-49d9-9386-ac088edd6177.html 9ee7c022-1d37-4a21-94dd-06098d242521.html

1. **Illegal Immigration**

**Solr PageRank**

2c42b257-1f91-4eae-8cc5-0f41effbee29.html 5b4f205d-1956-4b8b-adc0-abfb86883f81.html

e3b635f2-487a-42c8-bfe3-97a1afa6b44d.html 5730614b-e0d9-4b59-b376-d581341ab3a1.html

25b8a702-17f0-4c4f-9d9b-8c347de2b066.html 7179303f-100e-4015-b10f-e8616189b6a4.html

d27d0384-616a-4588-af26-a21e0b5be5d1.html 5cfd8710-4f3a-491d-b1bf-36eff9d319f9.html

70298026-cb94-4740-81aa-a5ce9c45b38e.html 7729f519-cae1-4770-91ee-3ded2fa84222.html

e1bfb199-b5e3-4413-9b88-eea733a94ba2.html fc906fed-5d11-4fc2-af7d-1eb6b7b7e472.html

a51bc697-8d5f-41f0-8cfd-0c103cecdca5.html 9163cdba-a2d9-4bc9-8bf5-f74a97147543.html

297635ec-dda2-40ca-85b9-632114a1760f.html 7ebdc5ca-a505-48cb-8fc1-ae8dad969d37.html

a63ae513-fe56-4540-a839-70f9c8e05251.html 221cbe33-d9d3-4489-be90-dc13d25465c5.html

9f255dda-8527-4068-96e6-bd4593471bde.html 5940a162-2ec8-41aa-93ac-2c1aee48ad3e.html

1. **Donald Trump**

**Solr PageRank**

33eb2dbc-3efa-44d5-87ff-12ae64ffcbe0.html 5b4f205d-1956-4b8b-adc0-abfb86883f81.html

8e7351b7-6999-427a-bf70-7381b52bf199.html 7179303f-100e-4015-b10f-e8616189b6a4.html

76589e3e-97ef-4f4c-a643-3bfb1420c71d.html 5cfd8710-4f3a-491d-b1bf-36eff9d319f9.html

ff8826e8-5174-40e0-b999-4efc22adaef5.html 7729f519-cae1-4770-91ee-3ded2fa84222.html

56317d55-8606-475f-9659-2a1fdd48a589.html 47daadab-0d94-4087-bab6-9ee86a4a42fb.html

2517d47e-70bb-458d-940f-425878123bde.html fc906fed-5d11-4fc2-af7d-1eb6b7b7e472.html

03ed641c-8d2d-4a60-8f06-6ddd9ba18d66.html 9163cdba-a2d9-4bc9-8bf5-f74a97147543.html

61e38367-e9db-4129-b625-7c7859277a30.html 7ebdc5ca-a505-48cb-8fc1-ae8dad969d37.html

06d66823-dbed-43a7-a437-07a2a8a8f658.html 963f0e3b-dff7-4c77-b532-ad2bd3357bef.html

c8300a0c-3490-4e0b-ad21-7ddc5e16bd45.html fe1875da-6d82-4128-881c-d252d4e571ab.html

1. **Russia**

**Solr PageRank**

b7e88093-ed7b-4751-9503-af2cc8ea814a.html 5b4f205d-1956-4b8b-adc0-abfb86883f81.html

36474563-085a-4c2f-9b34-caa1bf29c106.html 7729f519-cae1-4770-91ee-3ded2fa84222.html

70fdd57d-67cb-4c4f-a5b0-b6d699637ea5.html 47daadab-0d94-4087-bab6-9ee86a4a42fb.html

28cdb474-34f0-48c4-9900-c423aeed1eef.html 9163cdba-a2d9-4bc9-8bf5-f74a97147543.html

3f5af4c4-69cc-4d6c-a727-bc03df2ac247.html fe1875da-6d82-4128-881c-d252d4e571ab.html

ad3958bb-6d89-4d2d-b767-ba8819a03985.html 221cbe33-d9d3-4489-be90-dc13d25465c5.html

4c821d32-d588-4661-8bf5-931b2cfb29b1.html b9ec6c29-626c-4fbe-a78a-41c24a87f337.html

f0463bcb-cbdb-4819-ab09-5702b6ea6fea.html c43fd018-bc79-4a7d-a1a4-b706aedda959.html

6f4cc215-2f21-45da-8011-3fec118e9b7d.html a86559ae-4237-414e-bb68-79d516a65a03.html

9521c97d-265b-48a2-9b2f-51bd93cef532.html 8e161071-7fea-4ce2-bea2-f0b999b7b230.html

1. **NASA**

**Solr PageRank**

8c987f61-5620-4ae3-814e-bce029b6baaa.html 5cfd8710-4f3a-491d-b1bf-36eff9d319f9.html

5cfc90c7-8949-4cb3-a6b4-82755eb5ba5f.html ae71e879-abab-4f04-b305-903553a0dbf9.html

b17a8d1f-3f8f-4c74-897e-03a98a48d317.html 2eeecec2-6eeb-47ed-8f9f-85387cf3028c.html

8f5f6963-5d57-4e01-ad09-f8d910ccee11.html e14b087a-bd31-4437-9c31-c3b1f4f7c693.html

00e08ef4-472e-42ca-b9b9-cb30153d933c.html 9caa74aa-d911-4d83-ae85-8a548c7f36c9.html

2a2bc981-3a07-42f4-88d3-c9672bdc5771.html 5dadf35f-aa29-491e-bc3e-55bea2a067e2.html

abbe6c8a-775d-411f-93cc-eb6627d756e4.html ff343d0c-f054-414f-adf3-6af34ad48b92.html

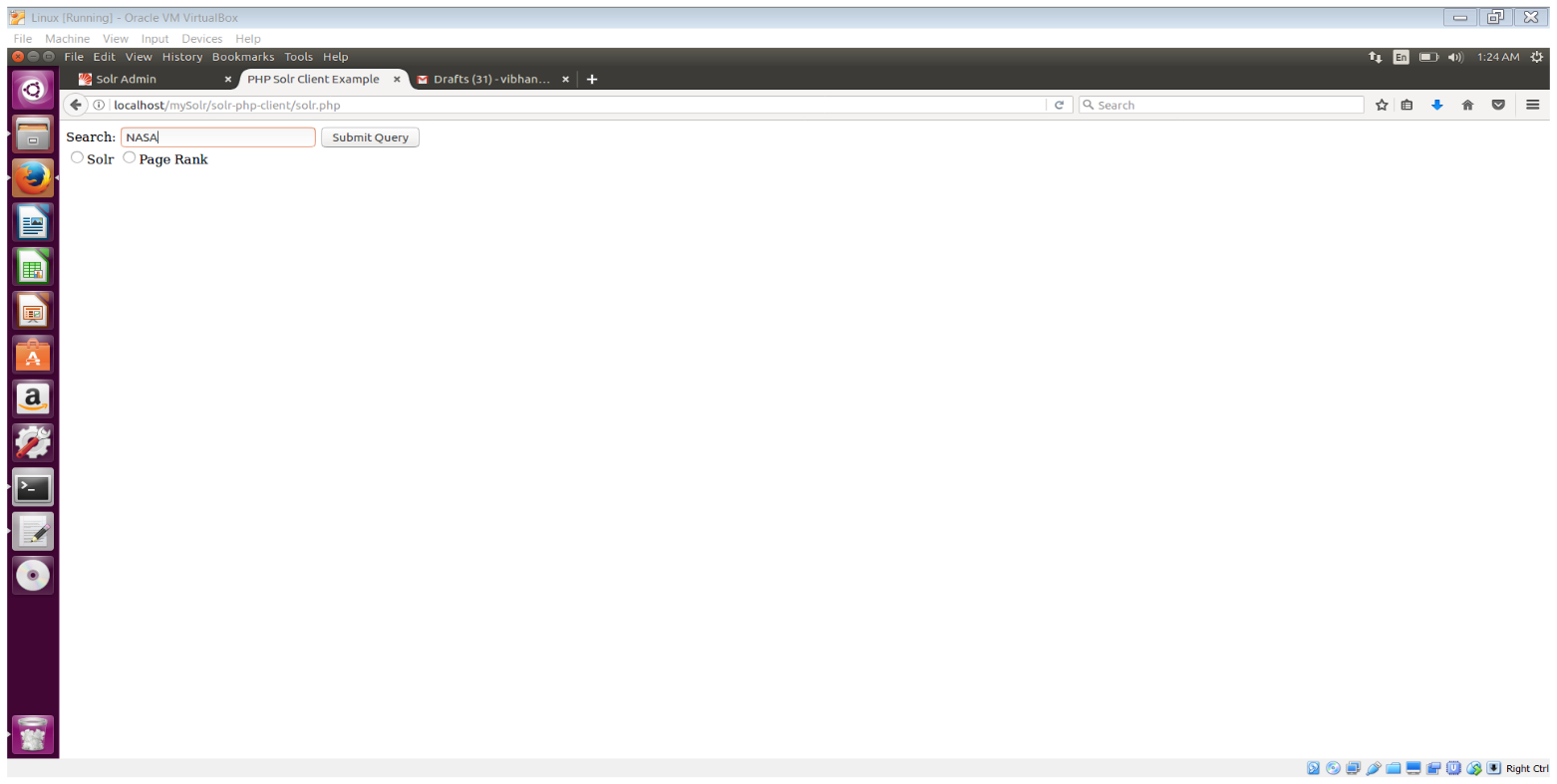
8826e497-6e0f-4384-9869-bc3e8ca0f08d.html bb189ad6-dcd8-4a14-8a10-d7d02e727415.html

f890b005-a5d5-4b0e-885b-93d263a65aca.html dd47892e-dd51-4f18-abeb-bdd1a3870a19.html

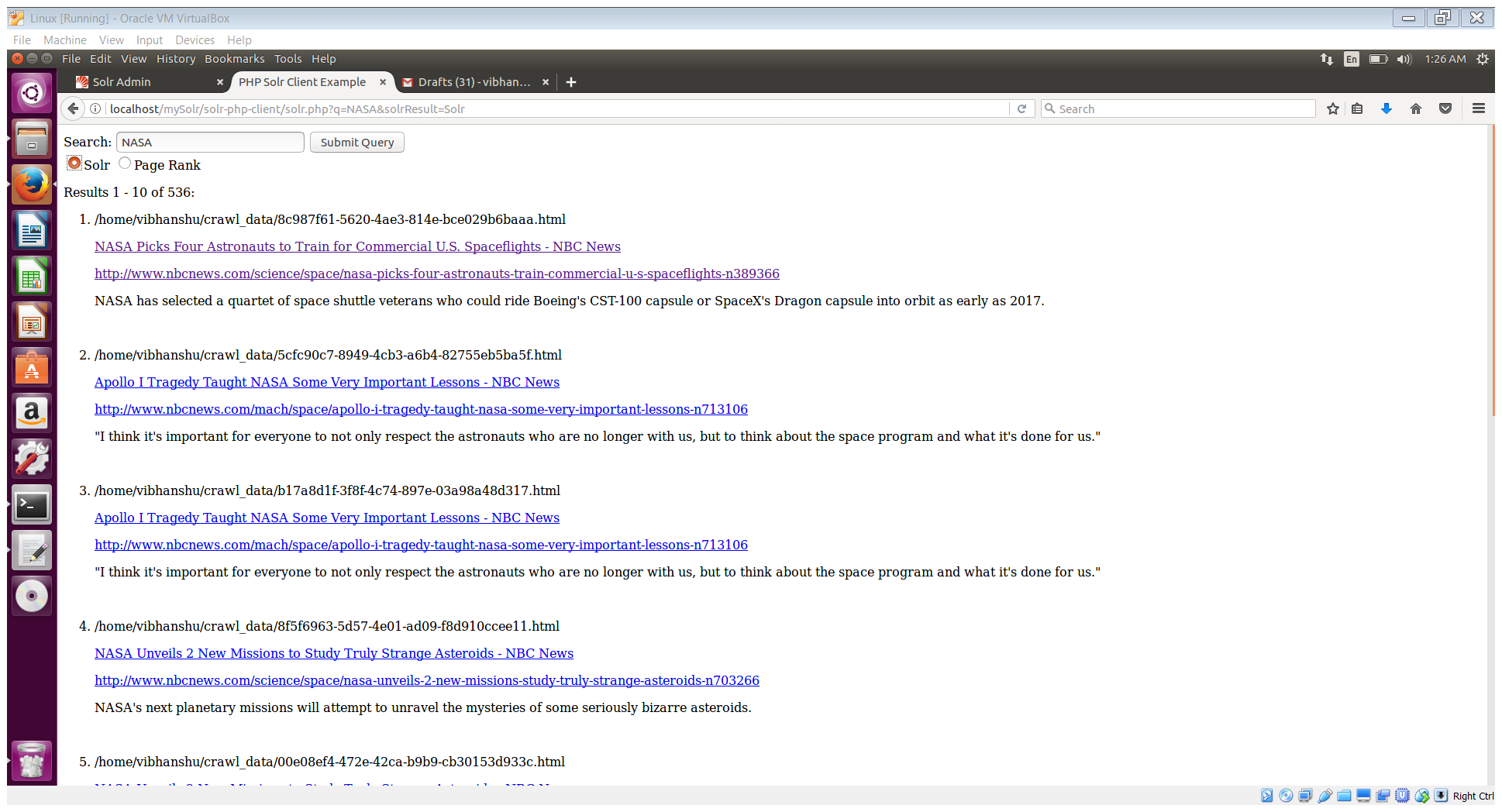
29caabc4-cf76-4bd2-915f-538681dd41f4.html ae0e9652-7749-4207-b5d1-03d0149b4ee6.html

Below are the screenshots that describe whole flow from query page till the result page for both ranking methods:

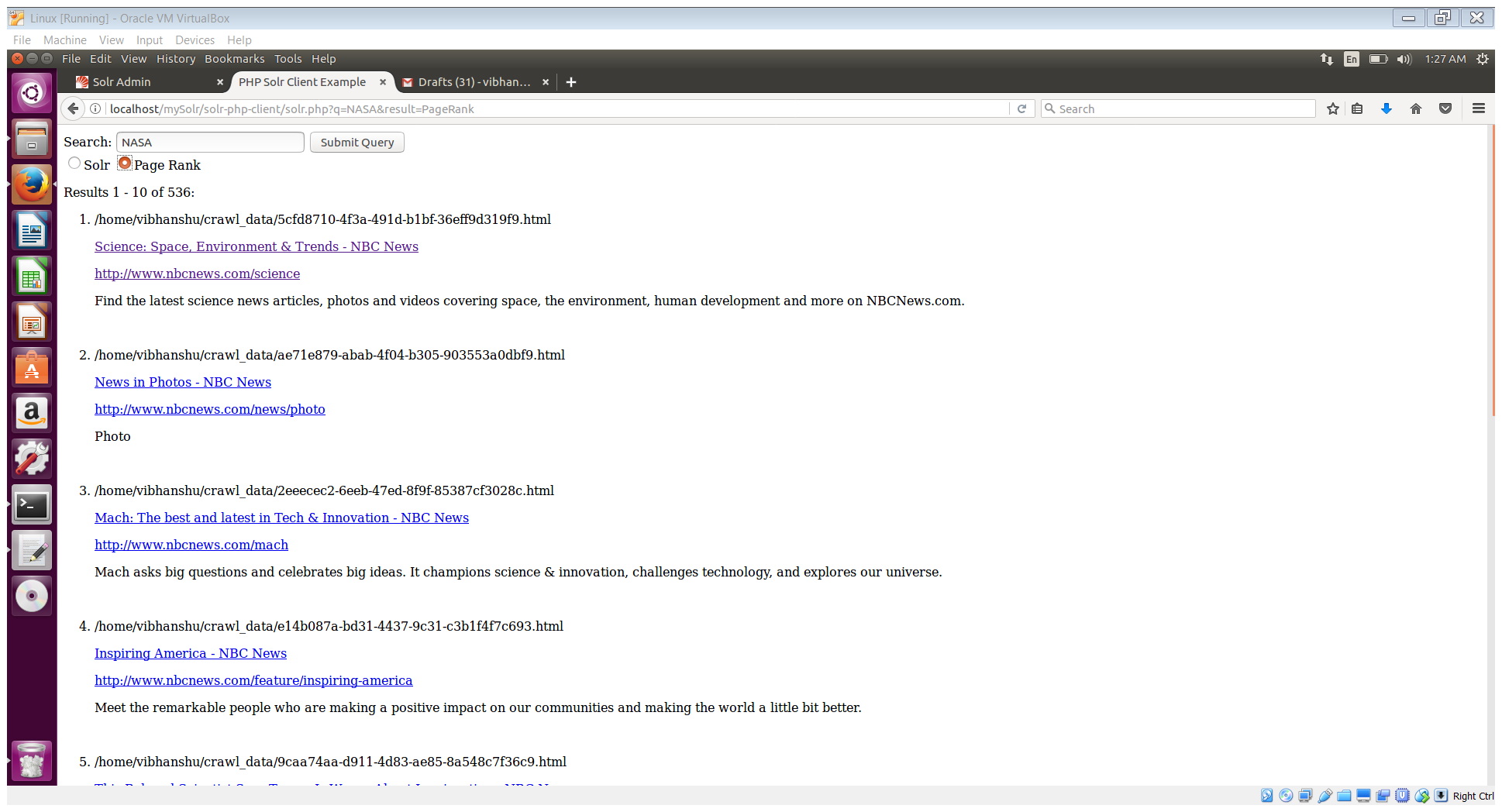
1. Query NASA entered in the search block.



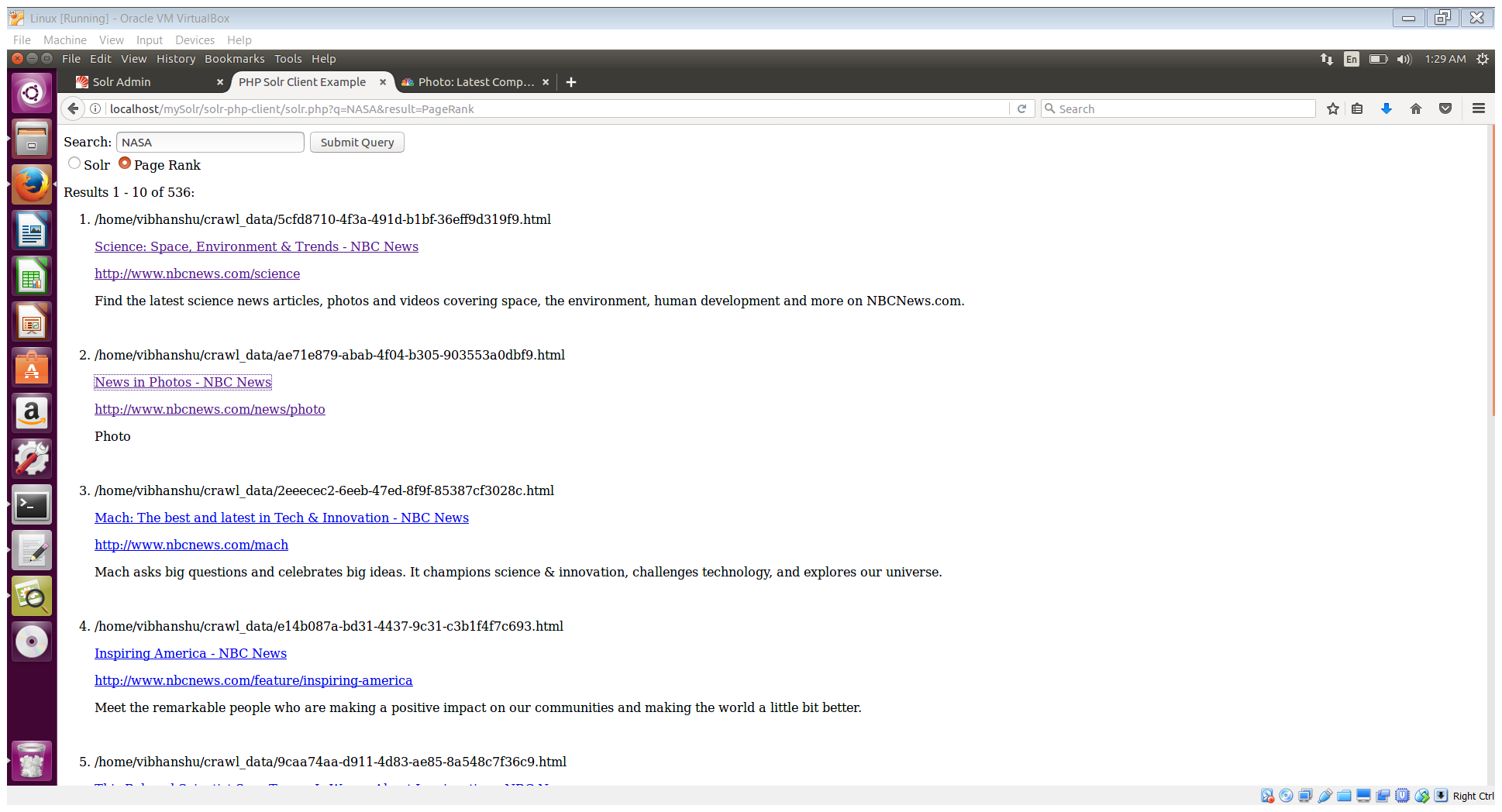
1. NASA Query submitted for Solr.



1. NASA Query submitted for Page Rank.



1. One of the URL (Second URL) Clicked.



1. The URL clicked opened in a new tab.

