**Homework 4 – Comparing Search Engine Ranking Algorithms**

**Steps executed during the assignment:**

Step 1: Basic Installation:

* Installed Ubuntu on my Windows machine using Oracle VM VirtualBox.
* Installed Apache Solr software and checked if Solr Admin was working.
* Downloaded and extracted all the crawled files from google drive for LA times**.**

Step 2: Indexing the HTML files:

* Started the Solr server using **bin/solr start**.
* 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 all the text content can be mapped properly.
* The Solr Admin page showed that all the files were indexed.
* I queried in the Solr Admin page to find the relevant results.
* This made me sure that Solr is working fine.

Step 3: Apache, PHP installation:

* I installed Apache web server and PHP in Ubuntu and started with the Sample PHP code.
* I cloned the GitHub repository for solr-php-client on my system and made necessary changes in the sample code.
* This indicates that a client is designed in PHP to run queries, retrieve result and display them.

Step 4: Generating the Web Graph:

* I wrote a Java program using JSoup library. It first created a mapping csv which has both the mapping csv files and I created fileUrlMap and urlFileMap data structures using the combined mapping file.
* Using this I extracted the outgoing links of each web page by traversing the directory to create the edgeList.txt file.
* This is the web graph or the directed graph which contains web page files as vertices and edge between two indicates a link between them.

Step 5: PageRank Computation:

* I executed python code to compute the page rank values using the NetworkX library.
* The python file 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**.

* Finally, the program creates external\_pageRankFile.txt having format <document\_id>=<page\_rank\_score>.

Step 6: Further changes:

* I placed the external\_pageRankFile.txt which is used as external page rank file in Solr in **solr-6.2.1/server/solr/myexample/data** and made necessary changes to managed-schema by adding fieldType and field parameters.
* I modified the solrconfig.xml file to include listeners as per the instructions.
* I reloaded the core on the Solr admin page. I queried in Solr Admin Page by placing “pageRankFile desc” in sort parameter of the query to get the results based on pageRankFile in descending order.

Step 7: Visual representation using PHP:

* I modified the PHP code to handle the external page rank algorithm.
* This was achieved by adding radio buttons for Solr default configuration and Page rank. For page rank I added additional parameters ‘sort’ => ‘pageRankFile desc’ before sending the query request to Solr.
* I changed the way the results are displayed for each query and I am displaying the title, link and clickable links to the actual news article featured on LATimes.
* 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.

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

1. **Brexit:**

|  |  |
| --- | --- |
| **Default Solr Method** | **External Page rank method** |
| ﻿39a7634a-199b-45a7-8ec0-4cd68eb36a1c.html | ﻿290668ae-cfb3-4ea7-9089-e714251d8493.html |
| ﻿ebd91d9c-1e87-4c3f-8e77-2b665c3824e3.html | ﻿8ef10e81-9666-4028-a247-9e236dbfc159.html |
| ﻿78ebd410-efa6-4a45-a4c4-b5eaba56dfde.html | ﻿ae385852-64f2-40e9-8d87-7973ad6a4587.html |
| ﻿19793351-63e0-4dea-8ced-1a56ea4c70f1.html | ﻿dc1e389b-0024-4465-95fd-9a3cbfbcdb9e.html |
| ﻿e1d8349d-303b-4650-b62f-9da84d7a9498.html | ﻿22bf3ac1-7fd1-41de-aac6-e784e35d0f7c.html |
| ﻿bf11ea70-7d9a-458d-881b-48be6f3deb51.html | ﻿c910bfd4-500d-48e0-a9c0-563c8fb3b683.html |
| ﻿c4e94f55-e2e5-4704-af20-cda2bcfe8a1f.html | ﻿bf11ea70-7d9a-458d-881b-48be6f3deb51.html |
| ﻿217d7d7e-ebf8-4384-89c2-7139d2aa2c77.html | ﻿ea2f70bf-4804-46cb-b837-ffff4e99a7aa.html |
| ﻿b41f1263-0f37-4140-afaa-9eb34e0f00b2.html | ﻿070739fd-7a3b-4632-8d66-67ed7cc484b8.html |
| ﻿7557772c-8c0d-4c8d-a220-d9188a47989f.html | ﻿39f138a5-53be-4ce0-9536-b48d6af324e5.html |

1. **NASDAQ:**

|  |  |
| --- | --- |
| **Default Solr Method** | **External Page rank method** |
| ﻿f995e404-aab6-42be-a09a-bd0c34ee90b0.html | ﻿04ef426b-7f8d-45a1-bb22-cf72a547b205.html |
| ﻿20edb790-b152-4263-b7ee-fe7c77222e91.html | ﻿d4baaf9e-9ebf-4856-b30d-5f8260524703.html |
| ﻿93895cfa-48e5-422a-966c-d57a655fe39b.html | ﻿ac1e02bd-2049-4a12-93a0-a5c4e456e7e2.html |
| ﻿0d21ac16-9668-47a3-96e2-75ec89d62d31.html | ﻿5819d959-1ac8-4837-a02b-b85350117b9d.html |
| ﻿7d66758d-8e2f-4d55-a8a4-1bc430716d6d.html | ﻿3c3168d0-874a-4629-ae96-a766cfaf63f3.html |
| ﻿27fe4596-0136-45fe-890e-3f08bb1f34aa.html | ﻿0d21ac16-9668-47a3-96e2-75ec89d62d31.html |
| ﻿cc2a6a1a-5286-43c0-8463-c7e1da1c1af7.html | ﻿63f2f8ee-f456-4c2d-bcf7-e8e65194e838.html |
| ﻿ed5c7512-063a-4431-83b1-a39bc045aaa1.html | ﻿844ddf56-b0fd-47f8-9506-ab3c13739be3.html |
| ﻿f947eaef-8d81-4687-8ace-791dbff5e1e3.html | ﻿c255d484-1285-46f1-be29-8fb2a2ce2c13.html |
| ﻿3c3168d0-874a-4629-ae96-a766cfaf63f3.html | ﻿09638dfd-e647-4e5d-8ce4-d0b906188aad.html |

1. **NBA:**

|  |  |
| --- | --- |
| **Default Solr Method** | **External Page rank method** |
| ﻿235fb710-8da8-4885-9115-5f24248eec58.html | ﻿29f4eafd-c724-483b-9d14-4d91ba7b1ee3.html |
| ﻿7c4bb32e-0a0b-4cb2-877f-1ced31d851c1.html | ﻿394613f1-aa68-4c1d-8bbf-c675ae7ce405.html |
| ﻿fc1f7fd5-590f-4359-aedf-76367f755d7e.html | ﻿9b36e7df-415f-4075-a852-54eb27c41496.html |
| ﻿6903a2a2-fa00-48be-ba01-a053024290fb.html | ﻿edd774a2-0bf3-46f4-9461-c440fd34252a.html |
| ﻿ccb9332d-0beb-4b9a-9b35-1e8c87ecc839.html | ﻿6803a01c-70b1-49ec-9601-838d03738bbb.html |
| ﻿df946ab6-0270-4a70-a45e-36b0d4831cfb.html | ﻿d9bbe0d3-ad56-4909-a8df-bb1097a671fd.html |
| ﻿fef09156-5317-40e7-9558-180fda9263c5.html | ﻿4695b2e5-9f45-49b1-a7ad-a281d9bc9ce6.html |
| ﻿58ae6c05-3859-4c26-b4b8-c309ea452a25.html | ﻿ded80a6a-8d03-4d09-81b6-21d0a3e7e1a1.html |
| ﻿165ac121-2d93-4dd9-953e-c406f3e79400.html | ﻿2d872a24-3734-428c-aca2-79ad1e0d6782.html |
| ﻿fbdb30f6-2f9a-4403-9a95-c642b69b26a6.html | ﻿cef5d430-0a3d-4f15-a38e-2d767f8cd235.html |

1. **Snapchat:**

|  |  |
| --- | --- |
| **Default Solr Method** | **External Page rank method** |
| ﻿cf629de5-20f3-4848-9f8e-6e928f992f20.html | ﻿9f0adc35-3c5e-402a-ba97-31e66dd0e404.html |
| ﻿902d28e2-cc30-49fe-a5d4-94b19190767f.html | ﻿22abf67d-f856-4017-ae8e-8561a0c7b4df.html |
| ﻿6f9e308a-df4a-4dde-8293-708da8d320ce.html | ﻿ab0e6aa2-dacd-437f-b82e-eedf3266f5d4.html |
| ﻿80c0a32f-be08-4b03-892a-b34a0bcceff9.html | ﻿9ff22e03-d130-4253-8302-b947747a6018.html |
| ﻿b7a52afa-c808-4a7f-8e24-1593fd7c3ee1.html | ﻿04ef426b-7f8d-45a1-bb22-cf72a547b205.html |
| ﻿337ae7a8-d0a2-49a4-916d-4f5be1e19611.html | ﻿d4baaf9e-9ebf-4856-b30d-5f8260524703.html |
| ﻿ba0abffa-2bda-4240-bca9-4fbd170f6653.html | ﻿81047177-86f7-4ee5-b608-98b843e4f280.html |
| ﻿4ff9771c-80be-45bf-907a-d53e47f37f5d.html | ﻿a0f12977-4071-4fb6-b00f-aba6f4b708e2.html |
| ﻿116517ef-8a7a-4db3-abe8-32026c3c2e2e.html | ﻿8e5754e0-2a03-4402-91a0-bdf0caa1406d.html |
| ﻿9f0adc35-3c5e-402a-ba97-31e66dd0e404.html | ﻿ac1e02bd-2049-4a12-93a0-a5c4e456e7e2.html |

1. **Illegal Immigration:**

|  |  |
| --- | --- |
| **Default Solr Method** | **External Page rank method** |
| ﻿954518a3-8853-45dc-acf4-af55fb1423c9.html | ﻿993fa1e7-9c43-44f3-8d2f-d3d6bc38f14b.html |
| ﻿5b8f344c-aadd-401a-9993-1482f88f5b24.html | ﻿c25d7b3d-99e6-4a4e-8759-b94533e7954c.html |
| ﻿bb756b23-3831-4610-9c4f-0a6071b3c0a3.html | ﻿fe38df1d-dfd5-44a2-b066-e352ba826537.html |
| ﻿3ec0aedf-0ba0-47eb-9b71-1149c687f4cc.html | ﻿e7c05a32-22fa-4279-b075-9d01b12d5e08.html |
| ﻿14fbfaf1-4d3c-4de2-9ff7-e24f6acf8c48.html | ﻿61c32d0a-0745-4c1a-974e-b7511b589463.html |
| ﻿67c4b468-7e04-4549-8e02-309a99ddfd3f.html | ﻿60caf73c-35b2-459f-8d2a-4b078e3d74ca.html |
| ﻿f5642f9b-9be9-4a27-8af3-2308ed7573c7.html | ﻿9e9c4b23-7004-4002-b168-77ad20844e62.html |
| ﻿0563ff55-62b3-4dc2-acec-16ff18b88715.html | ﻿8cbc59d4-0cfb-4803-9b79-a1794ee7e2e0.html |
| ﻿67b30c42-bcb8-4bc3-a05a-8bcd6a851c6b.html | ﻿4a548688-2b2e-4b9b-b047-229ea7c3d23f.html |
| ﻿98738e6b-5982-4fa0-81ea-7ff60f179f42.html | ﻿c8e7da7b-e863-48af-98de-2ab8755f6279.html |

1. **Donald Trump**

|  |  |
| --- | --- |
| **Default Solr Method** | **External Page rank method** |
| ﻿2ae102d4-ba21-4d63-b504-0487aa97c4f3.html | ﻿993fa1e7-9c43-44f3-8d2f-d3d6bc38f14b.html |
| ﻿2d785261-e13f-4f0f-85ff-c68cff09b677.html | ﻿2240d649-3b7d-4dc7-93ea-566bbd3f8b75.html |
| ﻿0a86bd95-589c-49db-9ae2-71debb730423.html | ﻿c25d7b3d-99e6-4a4e-8759-b94533e7954c.html |
| ﻿d001b304-0be8-4861-a480-fd8dfd13aa51.html | ﻿31e478d9-cd70-44f9-bf89-1d7ce1e14bbd.html |
| ﻿ebd9fe36-4ca4-444f-8239-0d1311bee488.html | ﻿fe38df1d-dfd5-44a2-b066-e352ba826537.html |
| ﻿eb190bc8-e613-4d75-b2f4-f015d2c804fb.html | ﻿9c719802-2578-4c4f-86d5-c0438420b77a.html |
| ﻿a5e1f044-a0e1-4186-a654-d47138f5ce52.html | ﻿da4f2f9f-bcce-4be3-a3a2-557a83caf5d2.html |
| ﻿8de52abf-1243-4bc8-89c6-bd33953a5030.html | ﻿e7c05a32-22fa-4279-b075-9d01b12d5e08.html |
| ﻿5eeb8087-1012-49bc-b612-2adb623b919d.html | ﻿8f64c94b-a869-438f-be6f-def1e1b459c3.html |
| ﻿f1713486-3d58-44ba-8965-a78ef320d35b.html | ﻿61c32d0a-0745-4c1a-974e-b7511b589463.html |

1. **Russia**

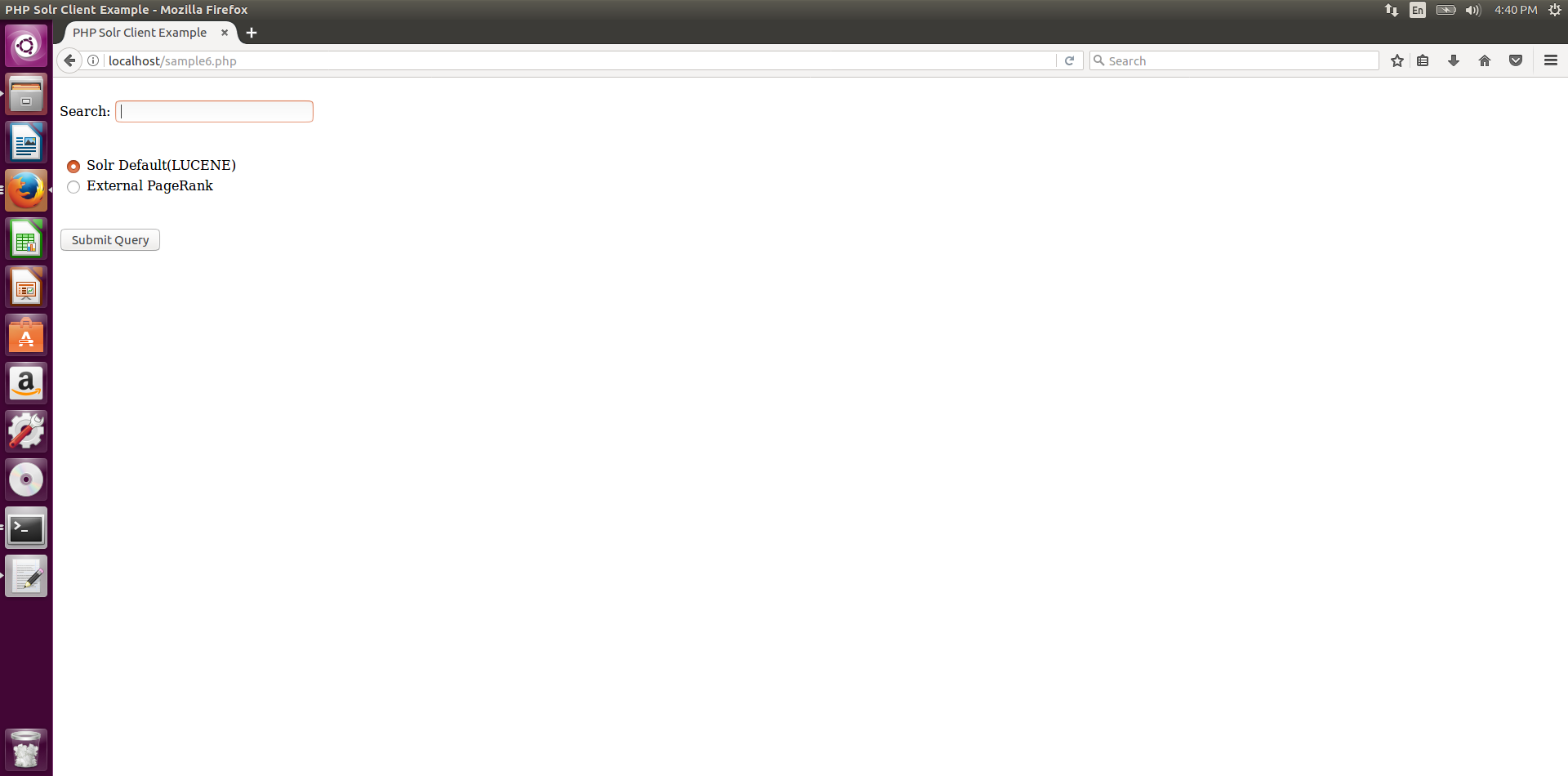
|  |  |
| --- | --- |
| **Default Solr Method** | **External Page rank method** |
| ﻿0102ac62-fdc1-477b-8b85-5f9684890d85.html | ﻿993fa1e7-9c43-44f3-8d2f-d3d6bc38f14b.html |
| ﻿e1fb04ab-3b6e-45d0-a1ae-cc66731790fc.html | ﻿2240d649-3b7d-4dc7-93ea-566bbd3f8b75.html |
| ﻿e0d69cec-d60d-46bc-ac64-7392ea6735a9.html | ﻿31e478d9-cd70-44f9-bf89-1d7ce1e14bbd.html |
| ﻿a26471ed-618e-47e2-950d-fcc0b90dc136.html | ﻿9c719802-2578-4c4f-86d5-c0438420b77a.html |
| ﻿f2217c74-8537-41a2-8367-419e92995bf7.html | ﻿da4f2f9f-bcce-4be3-a3a2-557a83caf5d2.html |
| ﻿8ed03175-70f5-4a66-bead-ec4dca7e6b7e.html | ﻿61c32d0a-0745-4c1a-974e-b7511b589463.html |
| ﻿03392469-11ef-4d2a-a048-4112355f0996.html | ﻿22abf67d-f856-4017-ae8e-8561a0c7b4df.html |
| ﻿5a692b2b-6a95-4c2d-93d6-6083e7f30afa.html | ﻿60caf73c-35b2-459f-8d2a-4b078e3d74ca.html |
| ﻿a394ec9e-fd3a-46bd-88b3-3c71f1a0f981.html | ﻿5d6a28c4-e602-40aa-ae96-45808ff3efc8.html |
| ﻿0a90f9f0-4024-4268-9fa0-4253e7c73c18.html | ﻿4b5022a1-e2e7-4474-9dc6-cda34c091b4a.html |

1. **NASA**

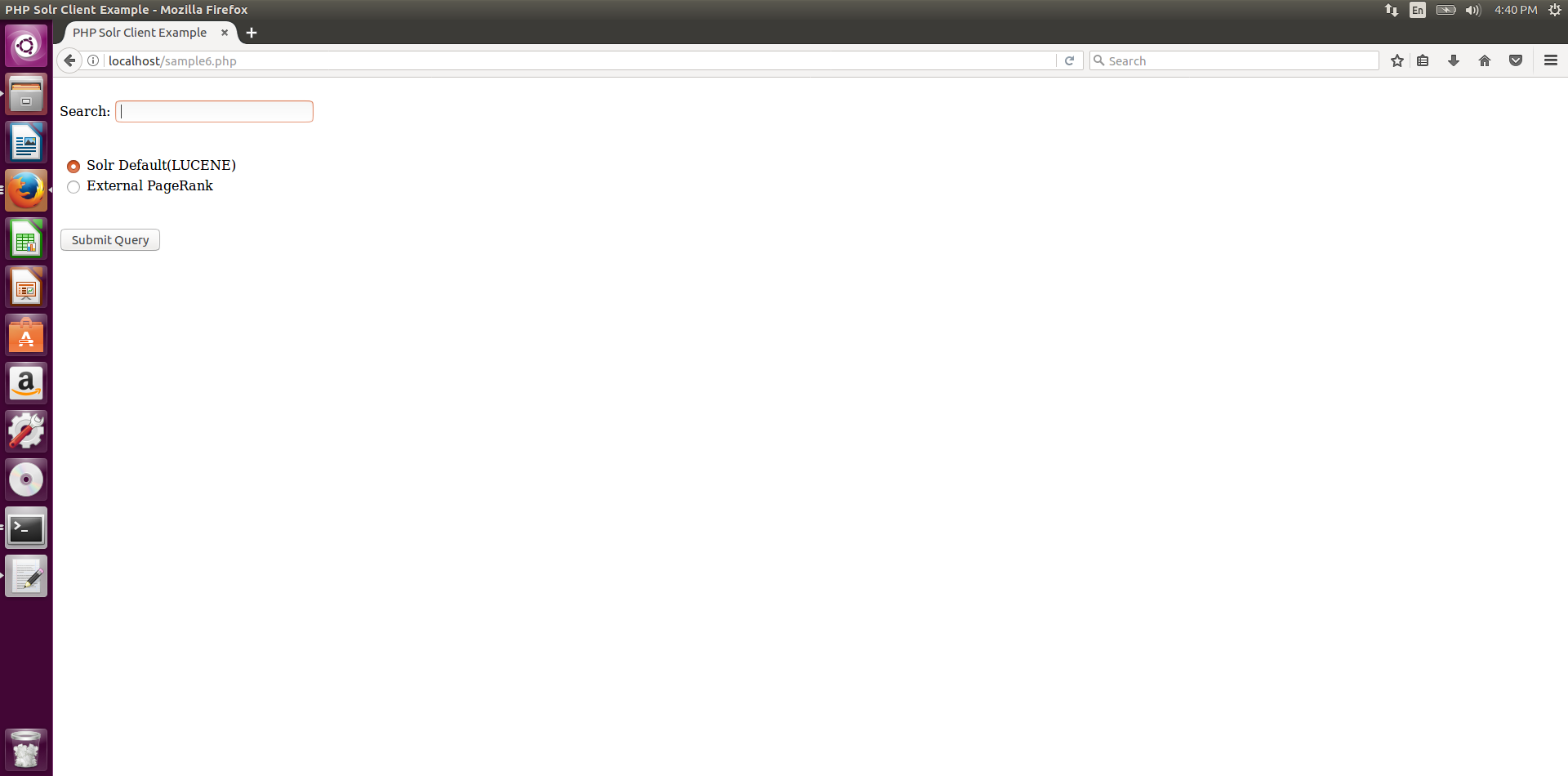
|  |  |
| --- | --- |
| **Default Solr Method** | **External Page rank method** |
| ﻿95d33041-9c33-442e-8875-710b5e1663d9.html | ﻿424ca584-ad82-4d76-a791-6eb9454bcc15.html |
| ﻿d5fd5eaf-ccee-49b5-98cd-7a78e311cf1f.html | ﻿bdd2ab43-3bfb-4689-b537-7bc2665818f6.html |
| ﻿bf592687-32ca-4c8d-a6f0-2e5f803ec589.html | ﻿2b61183b-8a99-45a6-812f-16d421aa6376.html |
| ﻿034e8eed-d942-4f5b-a43b-09e4bcf83524.html | ﻿14f80a4a-01bd-42ee-b127-6e6ee8cd3e26.html |
| ﻿224aa25d-0f1d-4bfb-acbd-e3f90534dd54.html | ﻿ee7c4767-b527-4098-acff-21bbb7c3ac2e.html |
| ﻿d52589c7-d44b-4415-8d8d-67ef50dd4f86.html | ﻿1af6e121-7d88-4e84-88b0-afaaae330ac8.html |
| ﻿51ffde63-d1d3-4df0-9344-f12a8a683990.html | ﻿e3309baa-2d31-4020-8108-406935906663.html |
| ﻿bd78fb70-b743-41d5-abf5-c4c9d7ccf339.html | ﻿142a6b05-7ecc-41e6-bca7-890ec071a921.html |
| ﻿9a145730-9575-42ca-8c65-cb765cf84012.html | ﻿37b5f92c-8b4c-4828-9a76-c7b826c8f39a.html |
| ﻿7356e678-e0d4-4af4-928e-6b438a331cfe.html | ﻿40a2750f-5e06-4892-a9a5-931d24066886.html |

Screenshots that describes the whole flow:

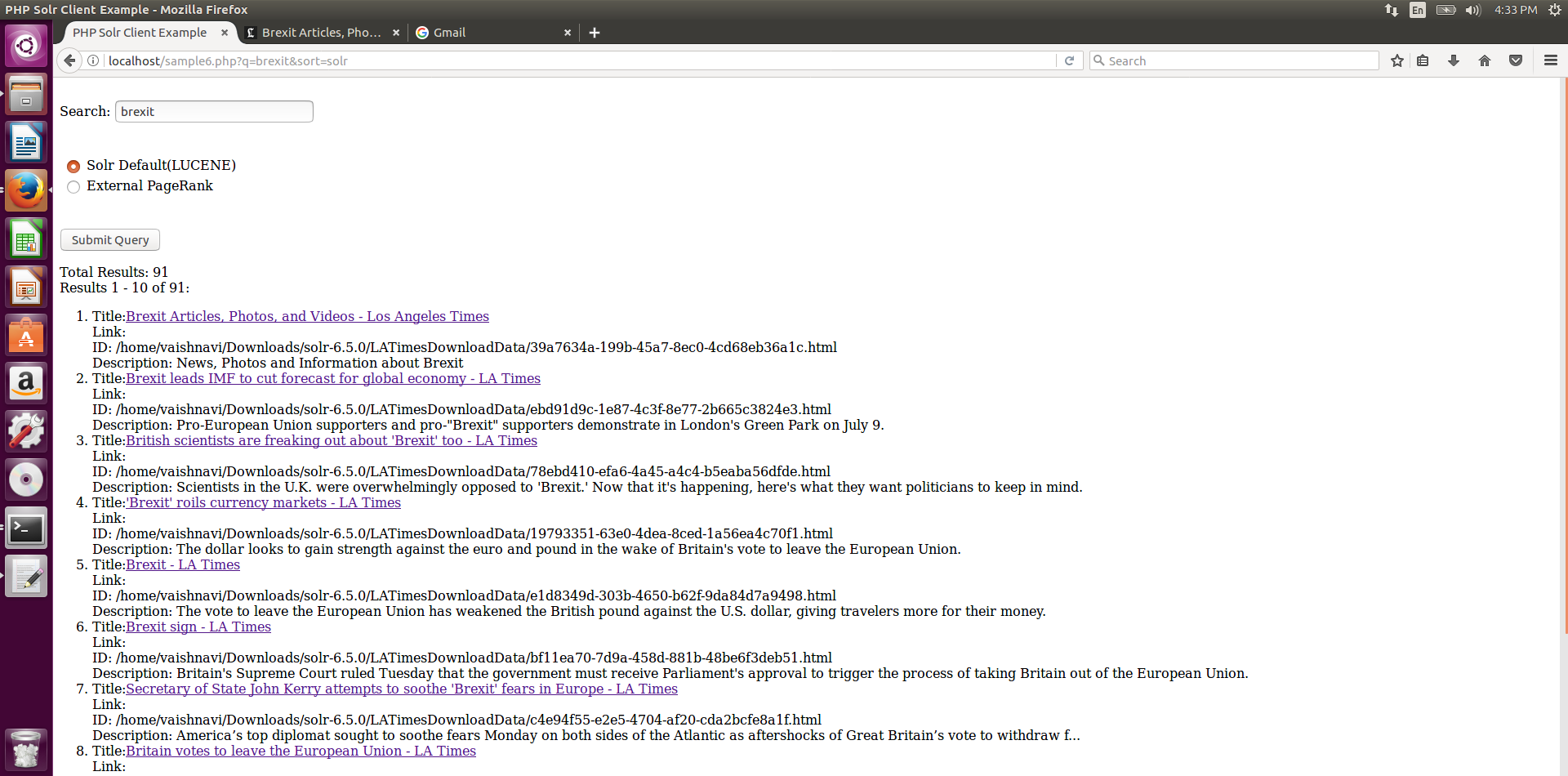
1. **Initial Page**

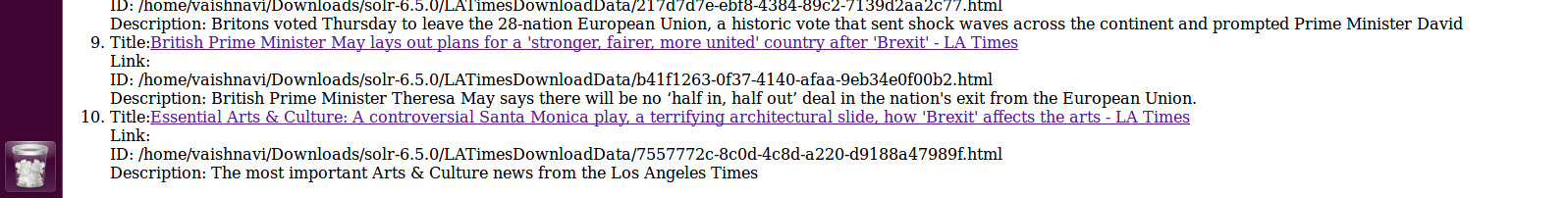


1. **Solr Method:**

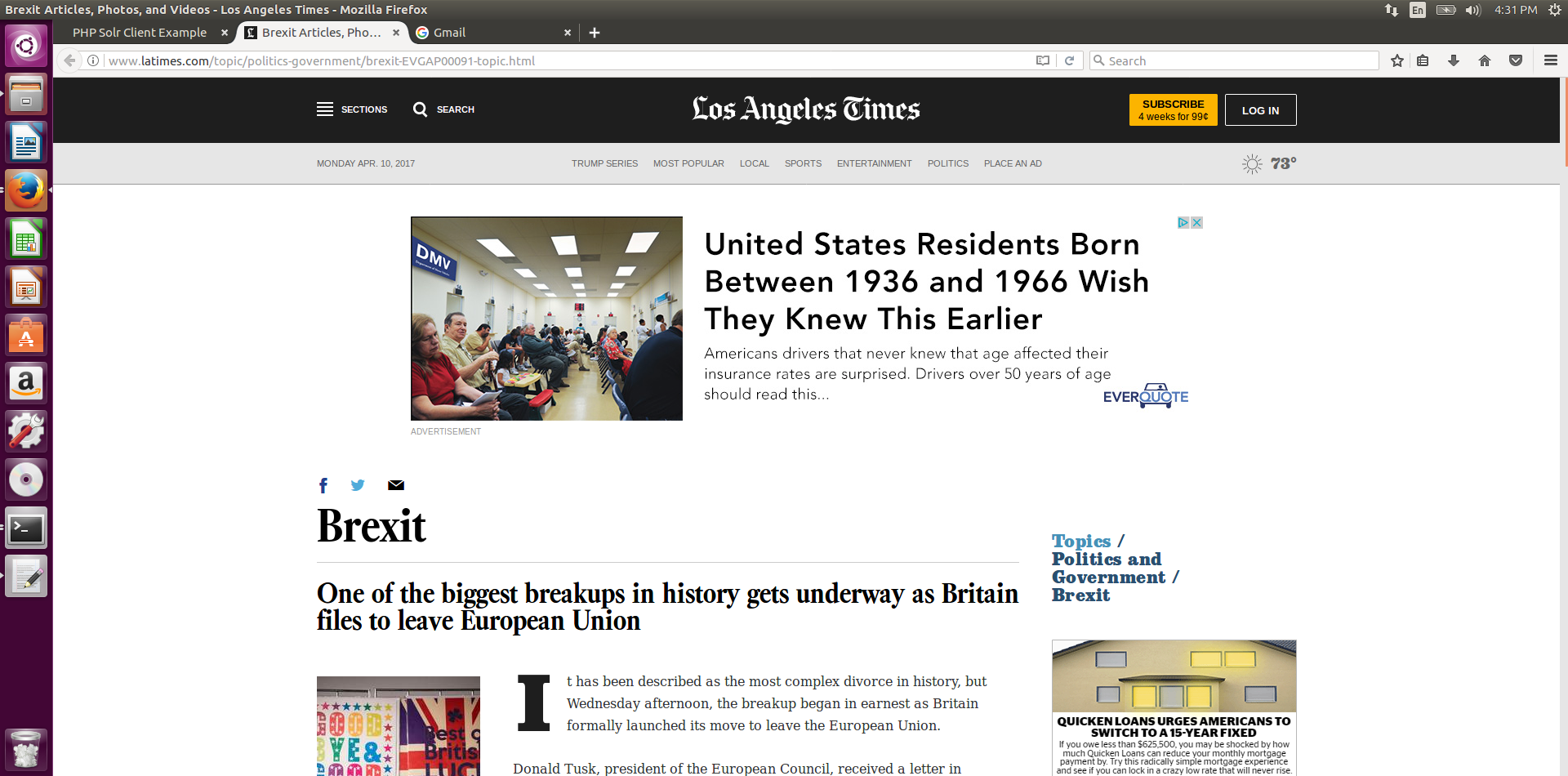


1. **Solr Method Results:**

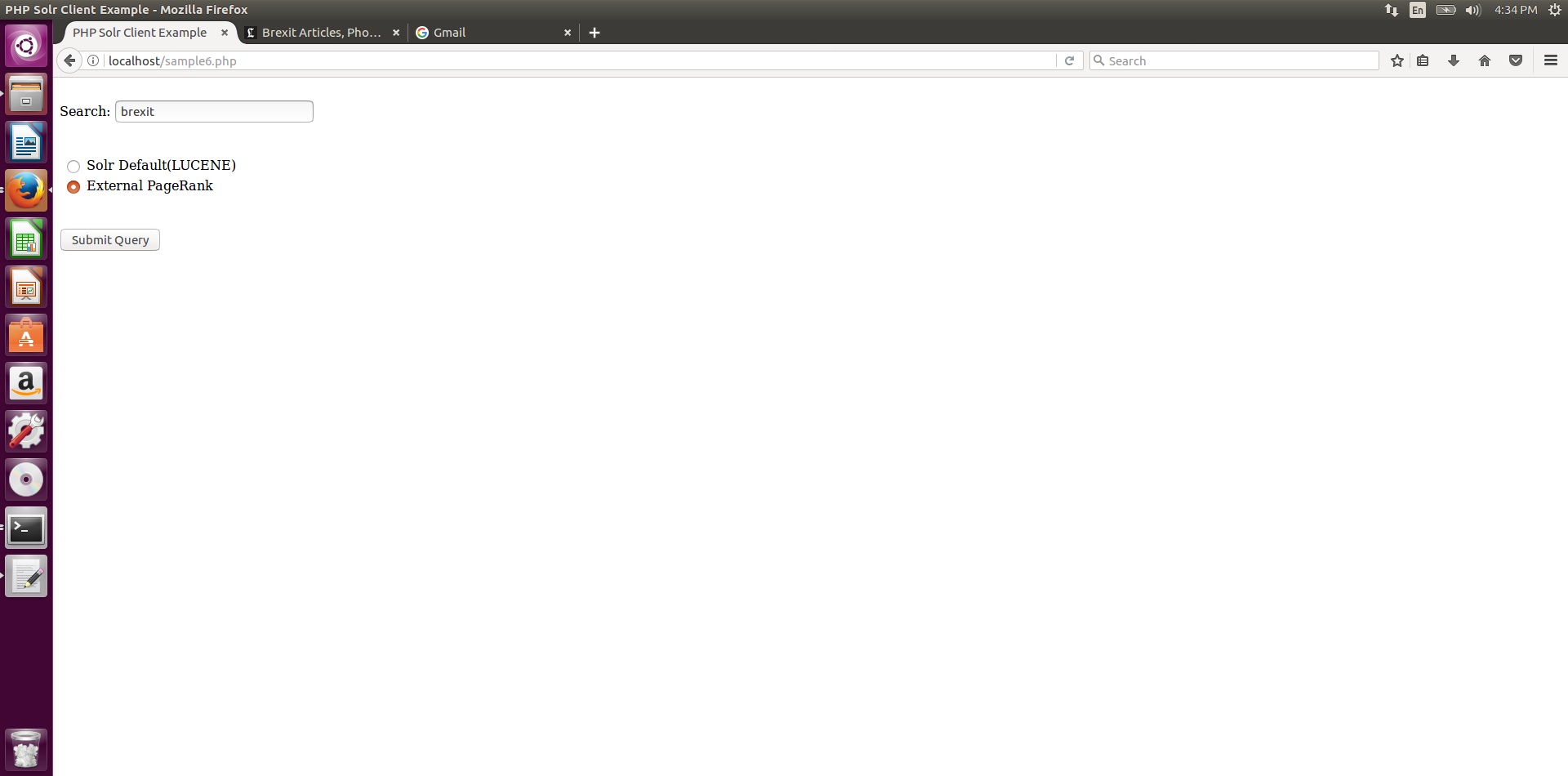


****

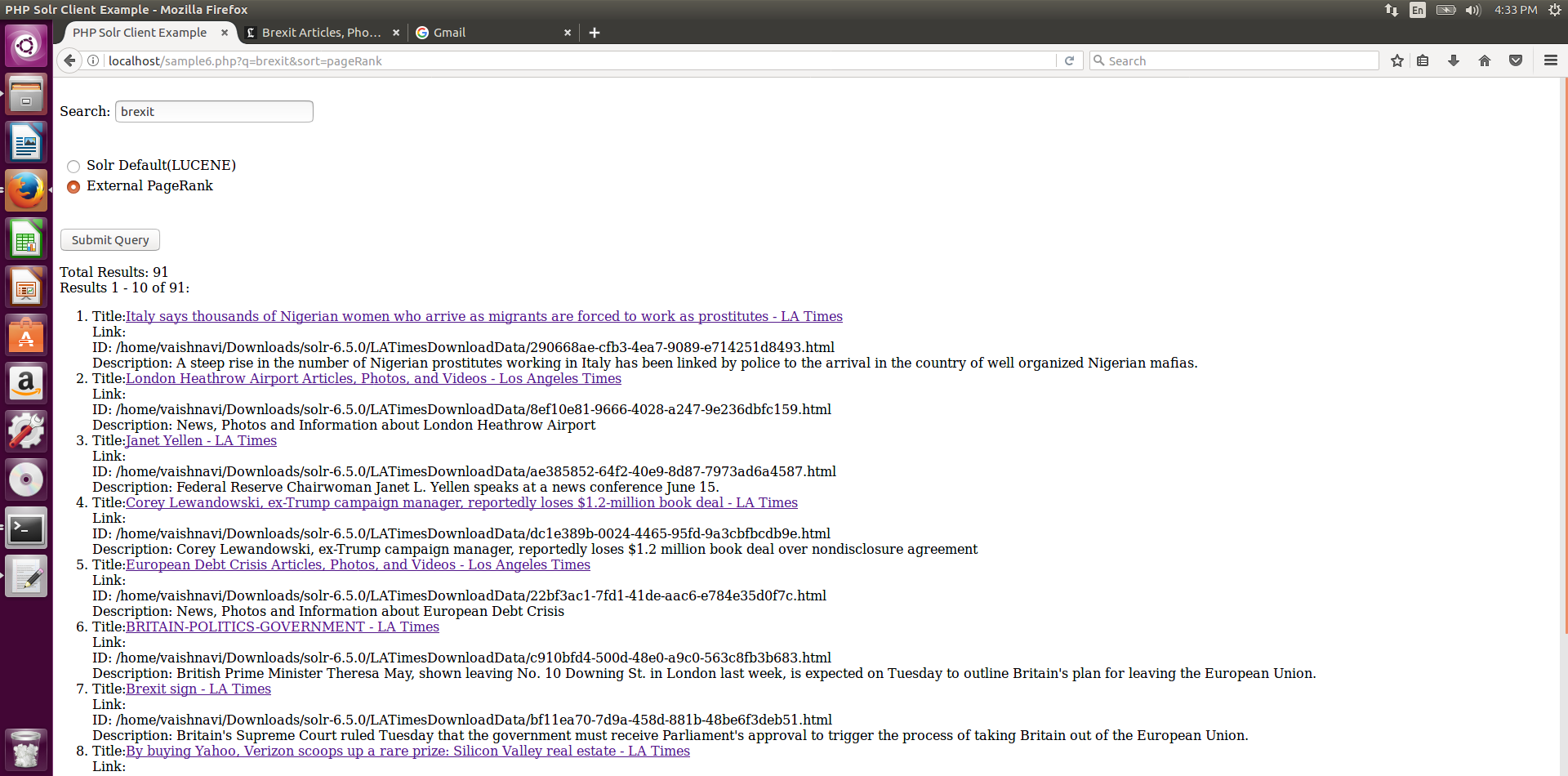
1. **Original Web Page Click in Solr Method:**

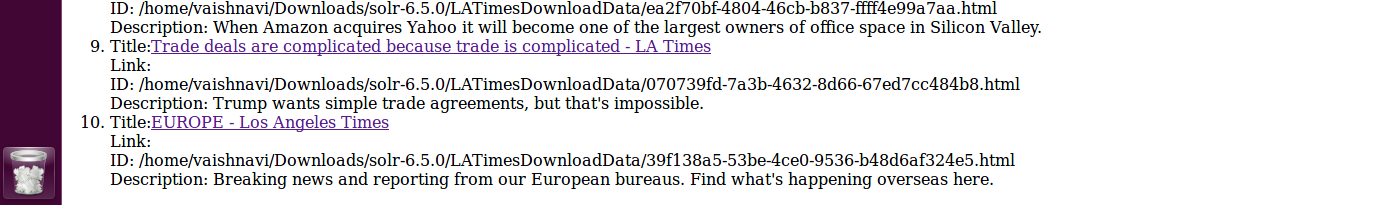


1. **External Page Rank Method:**



1. **External Page Rank Results:**



****

**Results Overlap**

|  |  |  |
| --- | --- | --- |
| **Query Number** | **Query** | **No. of Overlaps** |
| **1** | Brexit | 1 |
| **2** | NASDAQ | 1 |
| **3** | NBA | 0 |
| **4** | Snapchat | 0 |
| **5** | Illegal Immigration | 0 |
| **6** | Donald Trump | 0 |
| **7** | Russia | 0 |
| **8** | NASA | 0 |