HomeWork-4 Report

Methodology

Briefly describe the datasets (e.g., wt2g_inlinks) and tools used in the analysis. Explain the process of calculating PageRank, Hub, and Authority scores.

- The wt2g_inlinks dataset is given by assignment material, containing information about web pages (identified by URLs as its unique IDs) and their in-links. And I use Elasticsearch to retrieve in-links and out-links for pages from our merged index 'general_crawler_for_hw4'. The in-links and out-links are written to files in_link.txt and out_link.txt. And will be later read by my PageRank and HITS class. In the python file get_es.py, the tool used contains Elasticsearch7, numpy, math, os. In the Elasticsearch data fetching process, helpers.scan is used to efficiently iterate over large sets of documents.
- PageRank process starts from initialization: read in-links from a file (wt2g_inlinks.txt and our merged index) to build a graph of the web pages and their connections. Each page is initially assigned an equal PageRank value, essentially 1/N, where N is the total number of pages. The PageRank value of each page is distributed among its out-links. Sink nodes are pages with no out-links and they contribute their PageRank back to the system evenly across all pages. The new PageRank value would be calculated and updated using damping factor d as 0.85. Iterations continue until the change in PageRank values between iterations falls below a predefined threshold, indicating convergence.
- The HITS process starts from root set fetching, i use Elasticsearch to query for a root set of documents related to my specific query "Sino-Soviet split". Through the iterative process, the base set is expanded by adding pages that link to and from the pages in the current set. The expansion is limited by a parameter d to manage computational complexity and finally achieve 10000. Each page's authority score is calculated based on the sum of the hub scores of all pages that link to it. Each page's hub score is determined by the sum of the authority scores of all pages it links to. The iteration continues until the scores converge, meaning the changes between iterations are minimal.

Analysis

1. PageRank on WT2G_Inlinks Data

The page WT21-B37-76 has the highest Page Rank which is 0.0026945 and a large number of in-links, which is 2568, indicating a higher number of inlink pages may get a higher pagerank score. WT21-B37-75 follows with a PageRank score of 0.00153317 and 1704 in-links, showcasing a similar trend where a large volume of in-links correlates with a higher PageRank. WT25-B39-116 and WT23-B21-53 have fewer in-links, which is 169 and 198, but maintain high

PageRank scores (0.0014685 and 0.0013735). This may suggest that pages with fewer, potentially high-quality in-links can achieve high PageRank scores, likely due to receiving links from authoritative sources. And the number of out-links varies significantly, yet does not show a clear direct impact on PageRank scores. WT23-B39-340 has 396 out-links but still achieves a high PageRank score (0.0012411780918402), indicating that a high number of out-links doesn't necessarily dilute a page's importance if it also has significant in-link support. Sink pages like WT08-B18-400 who has 0 out-links still achieve notable PageRank scores that is 0.0011435, highlighting the algorithm's ability to redistribute PageRank effectively, ensuring that pages contributing valuable information are recognized regardless of their out-link count.

	Page	Page Rank	No. of Outlinks	No. of Inlinks
2	WT21-B37-76	0.0026944737779136	5	2568
3	WT21-B37-75	0.0015331787464243	1	1704
4	WT25-B39-116	0.0014685026346494	1	169
5	WT23-B21-53	0.0013735263702095	1	198
6	WT24-B26-10	0.0012761659886317	1	291
7	WT24-B40-171	0.0012455607092278	209	270
8	WT23-B39-340	0.0012411780918402	396	274
9	WT23-B37-134	0.0012051108089130	2	208
10	WT08-B18-400	0.0011435609558466	0	200
11	WT13-B06-284	0.0011248658902832	2	454
12	WT24-B26-46	0.0010850584092956	6	187
13	WT13-B06-273	0.0010447785198078	11	454
14	WT01-B18-225	0.0009884621682207	0	431
15	WT04-B27-720	0.0009365359592547	28	291
16	WT23-B19-156	0.0008942676845469	12	406
17	WT04-B30-12	0.0008164636719073	8	241
18	WT25-B15-307	0.0008043952604037	8	614
19	WT07-B18-256	0.0007753931779220	170	169
20	WT24-B26-2	0.0007713637156048	5	625

2. PageRank for Index

Page	Page Rank	No. of Outlinks	No. of Inlinks
c45a066a545333a9fd8a65e27521f4c13b3bf575b5586e28721bf5acc975dd8e	0.0000567459018664	64	1108
6f7b3f9dd587cbf7dc9f64aa922f33402ca948b2bce7ced9c6cc21bd73def95e	0.0000543736880759	73	1107
503b2b944245a3ebefefeff7362afdeb53ec83554e9218da1968a129d7be45f0	0.0000543362922535	26	1106
90ed511a2189046b32b9d9753f3c08dfef6218955c8371006e8e47b72c2d7a83	0.0000522876781804	3	1106
ca5255cafe03e1362bf9a7f97d4b527eae913b89e9cd6a80e4437a3ccf50ba87	0.0000515344721645	21	1088
ae8607ed972d7f94ca3143cdc60fcf7a5c9209bc6a9f3a2b3ccf1ae614445ba3	0.0000510576916741	17	1086
da979fc55fc6150f2732648ae0e215de8321f63d54ff06b68f7bc6264aedae7b	0.0000507355494097	15	1086
d7abae127e9916f2a51c5ac5a23a6bd9d69b700b7a9404b1e4798deb8e6e6855	0.0000505025572355	42	1085
ac52f5e0852fed4b282a7a0085bd612dfb71ffa21521f89d8077a4e41fb093cb	0.0000443427038855	47	1086
b636e16f0ac5b49c82e0fae51a8760e7d7b0c0546f12a9494bbb5e0c78dcfd41	0.0000441364994945	40	1082
799fd01f4cbb1f5b6a3a685738d23d725f23c1a85ba6eb8f41975fe1b228bb5a	0.0000292461499936	15	989
e68d73d017f9132f6240876dbd474c9c845ae860a5920089321c04f8c4625fcd	0.0000288237321498	40	282
23c1013aa950839ff8b903e1461275ce4682eba3d52dc21707c409b6a1325d72	0.0000272388639522	59	272
d74b839c182c0b6308aa22682ed3d79de14794c63e6edbf94f0370c55773101b	0.0000238477591356	119	235
c20774b63d04db782a061ce0b6400826b510a79122e9e8a6a68d1596ea0ac5c9	0.0000233820472786	0	407
afb71f174b7aeb0f80185ae9ceefbbda7ad3dae4bf4f3fcb9f7482f9821a603a	0.0000222763579526	4	271
f1f15afcbe09528365b6584bb319465c3394aeb35f0008ba2e0e32a7c8a799c1	0.0000218196097581	121	233
16e1688aee907d6b6e7605d32dc812c2b6f82c8edbe13d4a789543b9374d1ae6	0.0000189219941505	5	270
050fea02bd9ebccade5104fc5e8b6fe44e2d3e451292b0c4afa728678d13e6f8	0.0000185632076293	0	38

The PageRank scores in the merged dataset are significantly lower across the board compared to the scores observed in the WT2G_Inlinks dataset. Since we have a 180000 urls dataset, this could indicate a larger dataset with more pages, leading to a more diluted distribution of PageRank scores due to the increased competition for inbound links. Similar to the WT2G_Inlinks data, pages with a higher number of in-links tend to have higher PageRank scores, affirming the principle that pages well-cited by others are deemed more important. The relationship between the number of out-links and PageRank scores does not display a clear pattern, suggesting that while the algorithm distributes PageRank through out-links, the number of out-links a page has does not directly impact its PageRank score in a linear fashion. In both datasets, a higher number of in-links generally correlates with higher PageRank scores, reinforcing the fundamental concept of PageRank where pages that are more frequently linked are considered more authoritative.

3. Hub Scores

top_500_hubs.txt

```
00a78d500abb018f0b35f5a60b62096a7d9e27b6e5350be91cb473afa22da7f5    0.0803932354487455
f129c414f075a7ba648a5bbd9356dd3a2bd358c1d6842ea5e90e6b25cb1bafdc    0.08005704345548476
e88870527f354adcddddeed2ed2a8c06d7d3022b528e75a910e20dc9230a21d6    0.07933765906785394
11f619b6bac764470988f21ffc678675ffae545f93b35686de5bc0f1a86aa3f5    0.07904482273240057
13d1ea9f6f515c418da4e8887980241487 65f7002d95289464fc89e8a52f244e    0.07755847571859048
14f0d8e51baaa229a723cb86a0259d27395c7597f184511692661308b28eff52    0.0772080452601315
06e150b2bb5e8fcc87c9fa49cd591f65a2e46dada09a9218f994b31399f5670c    0.07609904722681313
f29ab9ab25db662612ce320e4f8309e456d4a65eaad1b3f49cfcc584e6a4b138    0.0759008650793278
63caa7693254b441d7e534250e467e61931e14078a139fbd117d40c87ebdc23f    0.07500306425145675
85d41589a183d2fc466f1e32bde014a7ae7a58dbbcc56cd6f46471df3217d615    0.07483569318549219
4925cd3542748b3e31f88a23d46a535f105355964416872a62cb20d8a881e576    0.0748329284833426
24d485cd5593eae907325684947936d566ffe6dc857f214bec7e9425662756ef    0.07472396221088158
209768417b8801967d006f9307d8fc7307276e5e86968a753bd0dc2c329b094c    0.07460997333164884
2edfd05258f7db9441a6f4597e2fb3782535495a17db1a0eb919adc673c966ad    0.07450403876700142
d97629a38eab8d27b5b64561aa6f60c3de87fe365e98185a0e174f970f63bba0    0.07432303771474276
676b5785195d048cdff1fcf254a75925e94e37205ff0874fa789db47ff6a13a1    0.07429337225106393
db3321fc887f76e23ad9711ed768839f89fc16bca335f9ef293537beee581edd    0.0742018918952357
9c07946607018c790d264a6285b250dc7e78e8995b3aee58500cda91a7102545    0.07402563890989387
be573f309a7da520447b8d41bfb682388f7184c54bb31561f72b5f0743f55d29    0.07397714276968624
5be20630f074032e5e2434881b159a5e1eff04833a5135a281cf74e874268782    0.0736575907098614
```

links > top_500_hubs.txt

```
https://en.wikipedia.org/wiki/A-35_anti-ballistic_missile_system    0.0803932354487455
http://en.wikipedia.org/wiki/Hyacinth_(plant)    0.08005704345548476
f73387f8895e5202c1e887f72aa33c89280320a2a8cb846650ab9c2c21d3c25b    0.07933765906785394
http://en.wikipedia.org/wiki/People%27s_Liberation_Party_(Turkey)    0.07904482273240057
https://en.wikipedia.org/wiki/Revolutionary_wave    0.07755847571859048
http://www.prometej.ba/clanak/povijest/sukob-jugoslavije-s-informbiroom-sukob-tita-i-staljina-ii-2096    0.0772080452601315
https://en.wikipedia.org/wiki/Andr%C3%A9s_Nin    0.07609904722681313
http://en.wikipedia.org/wiki/Lavrentiy_Beria    0.0759008650793278
http://en.wikipedia.org/wiki/Battle_of_Vukovar    0.07500306425145675
https://www.marxists.org/history/erol/ncm-5/pol-pot-sumup.htm    0.07483569318549219
https://en.wikipedia.org/wiki/Chilean_Communist_Party_(Proletarian_Action)    0.0748329284833426
https://sh.wikipedia.org/wiki/Kulturna_revolucija    0.07472396221088158
7c8fae84e7a44331a09d8ad74b1bf2ad307fc21a7d0457f9645dcc10552cdbd9    0.07460997333164884
88e896f30f7778231486f9d35549e50dd8f0ba93d74b1d59308637f7ea50c64c    0.07450403876700142
http://en.wikipedia.org/wiki/Political_views_of_Joseph_Stalin    0.07432303771474276
https://en.wikipedia.org/wiki/Moscow_Summit_(1972)    0.07429337225106393
https://es.wikipedia.org/wiki/Tesis_de_abril    0.07755847571859048
https://es.wikipedia.org/wiki/Planta_(arquitectura) 0.0772080452601315
60831ac93cda93b01ffe0b0318d75ba5f0c3d7daaa0fcc7f1d0240bb1d3ce1ad    0.07609904722681313
http://dle.rae.es/srv/fetch 0.0759008650793278
```

The HITS algorithm provides a HUB score to help us measure the value and content of a website based on its ability to link to authoritative resources. As the top 20 highest scoring sites are linking to information from various authoritative sites, such as Wikipedia, which is a key node for users to visit. Some pages, even if they don't have a lot of links overall, can accidentally score well if they link to one of several highly authoritative pages.

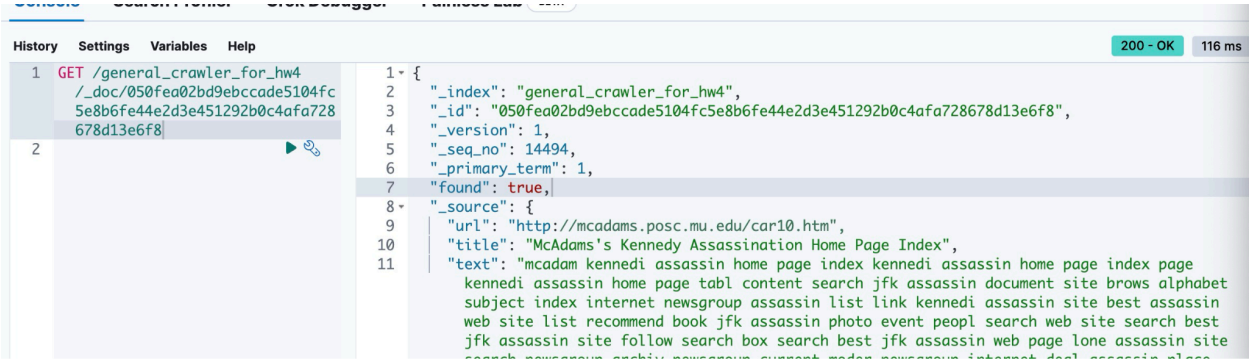
4. Authority Scores

1	5e360268d41ff3817a33874e5575ab924c6b86c33382505d91d0e47a1cd17b90	0.06588461743972976
2	4f839238a76ee46fd15ce5ba764019500360ce81557a9a96025eb79e37917173	0.06564821302284721
3	d08f02b8bff83fd4b85312330689fca1f93b76ae214425b1da17cdf537325d08	0.06500704132725946
4	33d499ae0b7b1b72d16bf7ca2b5aed3933a6a5a5ce09769ba231c79f1f6070b8	0.06496540041679635
5	3b0960742ae92cf0f96f4ca0543e585ce1f06cdf21b892979f3280052a0f8e30	0.06488172886404979
6	03569ed0e142d3a3a48686d3bf75bb66ae779bd301c10350f3ad770f9e9b390f	0.06487491258148068
7	eadd5a6c6bf055702f171f26429b1a20520ebc3997e157f3fcd3b12d9448afc4	0.06478835001570084
8	6d4d0f95611622a7d7625a2d1a891cf938e72a492afa60f8c175cae0a21c8e91	0.06460399702397229
9	23c200d62cf850d95f3c23fccfb04808ad89a43e9fc36d0d17e2d820a5973bd2	0.06459335037264441
10	1c0579520416011e1edf9a2600db37d0a2e8e34a8ce1c0156541f1b807b80f3c	0.06456682968611997
11	24de9e844517c7f68695d9099ed5fa9c8f02137ab8acdeabfb14c6f61a92ca8e	0.06452760647643101
12	dc95312468f7baf93172044601b1f74392b922fccbda758cadbddb39b200330d	0.06451276400848509
13	49a71f99bf9e9c15b4e04d4521284892f1198b2bc12c3313c3f841d08bf49089	0.06448782275744697
14	2d8231488893fa61d8edc78d1d32f78f82f437ebfbafb6378f2f4e9d5f6de26a	0.06448571041194828
15	b21901cc4b90cefa0a6f1a2de6576e9c18afe623df10cda9e94bb81031febf74	0.06446984479454011
16	e8360b0005934eb1660ae64cc972de230c7540f76553a6812fe36f83914ee4e7	0.06446590773251085
17	d43eccfdf25af8de0f726fd92b2c980460accf313ce7d921d01cb4a8958bcb68	0.06443610623455537
18	7d41bb8a67e06f238861bdb042b81320279c70888aa375d2fb0c4b3e6a8c9240	0.06441782256982649
19	c10e290991b9b63d989a9fce76bc3e0b77a760b07bc8bc72129319c8643df331	0.06440932079788055
20	425f16cb05d753f96a29daf8665d9b1df24dd4edda79ad6d13f8baf5d51618f6	0.0644032923428019

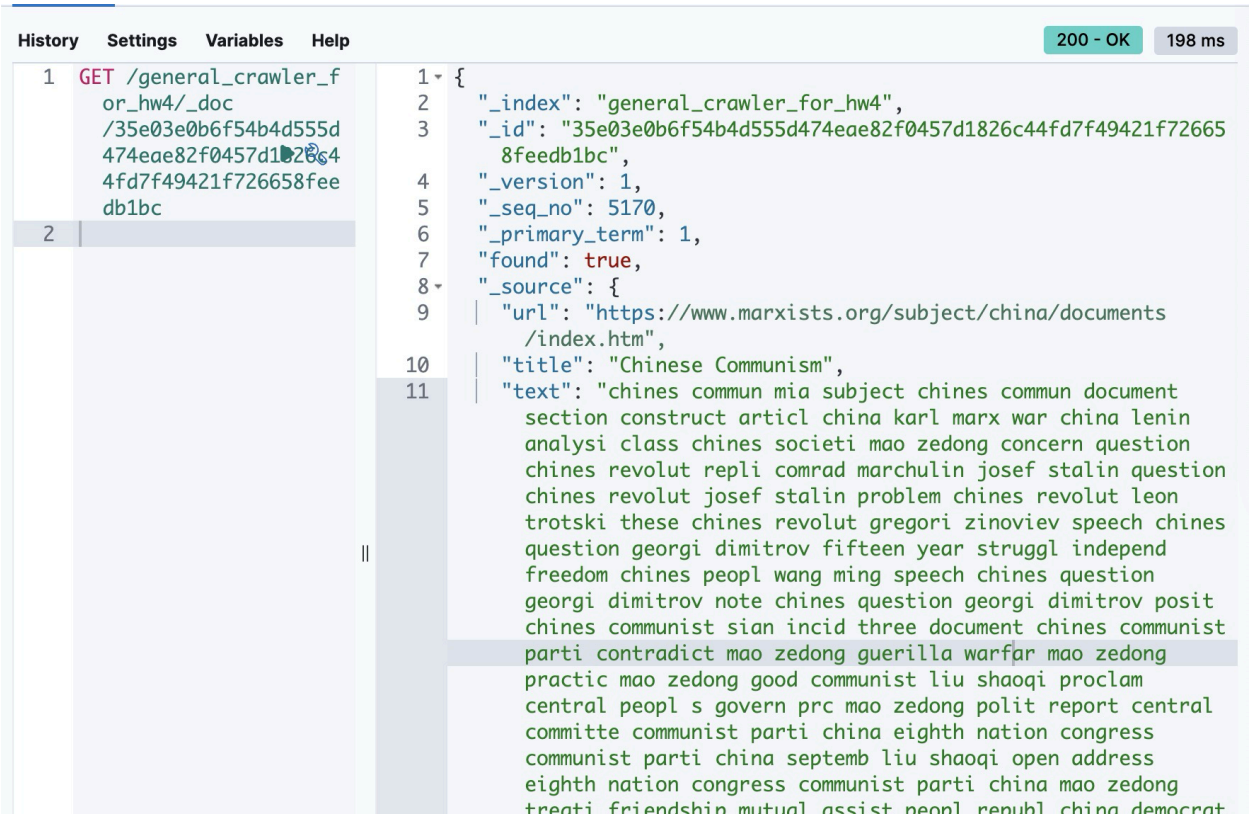
The url with the high hub score successfully links to the url with the high authority score we calculated earlier. a well structured web can distinguish well between the roles of hubs and authorities. pageRank measures the importance of a page based on its inbound links, but does not distinguish between the roles of these links, while the Authority score provides a more nuanced view. A high authority score in the top twenty figures listed here means that the page is not only popular or widely linked to, but is a key resource for its topic. A page with a high Authority score but an average PageRank may indicate that it is highly respected in a niche or specific community, rather than across the web. Conversely, a high score on both metrics indicates that the page is both authoritative on its topic and widely recognized in a variety of contexts. These data suggest that in a diverse ecosystem, both well-known and less visible pages can achieve a high level of authority within their domains, highlighting the richness and depth of the Web's information resources

Case Study: PageRank vs. Inlink Count

Id: 050fea02bd9ebccade5104fc5e8b6fe44e2d3e451292b0c4afa728678d13e6f8
Outlink count: 0, Inlink count: 38
This link only has 38 inlink counts but stays at NO.20 compared to these urls who have 1000+ inlink counts.
We can get the un hash url by:



"http://mcadams.posc.mu.edu/car10.htm"
The page receives inbound links from highly authoritative or reputable websites. The quality of inbound links significantly influences PageRank and Authority scores. If these linking sites are considered important or authoritative, they can pass a substantial amount of link equity to the page, boosting its score. The Authority score, in particular, could also reflect the page's content quality and relevance to specific topics. If the page is a definitive source of information on a subject that's well-cited across the web, it could achieve a high Authority score based on its inbound links, regardless of its own outbound links. The anchor text used in inbound links and the context in which those links are placed can also contribute to a higher score. If the inbound links use relevant, keyword-rich anchor text and come from contextually relevant pages, they can signal the page's importance to search engines and analysis algorithms.



“https://www.marxists.org/subject/china/documents/index.htm"

The website has only 26 in links and 36 out links. The site provides specialized content that might not be widely available elsewhere, especially documents related to a specific historical and political context. This unique and focused content can attract inbound links from academic, research, and interest-specific websites, which might be fewer in number but high in quality and relevance. Being part of "marxists.org", a domain that itself is a substantial repository of Marxist and socialist literature, the page benefits from the overall domain's authority and thematic relevance. The interconnected links within the site and to other parts of the domain can further reinforce its significance in the eyes of both users and algorithms. The outbound links from this page to 36 other pages, if those pages are also of high quality and relevance, can contribute positively to the site's perceived usefulness and authority. Outbound links to reputable sources can enhance a page's value as a resource hub for specific topics.