HomeWork-4 Report Template

Introduction

This section should provide a brief overview of the objectives of the assignment, including understanding how search algorithms like PageRank and HITS work, and how they assess the importance and relevance of web pages.

The objective of this assignment is to implement and understand two fundamental search algorithms: PageRank and HITS (Hypertext Induced Topic Search). Both algorithms aim to assess the importance and relevance of web pages in the context of a given set of interconnected pages.

1) PageRank is an algorithm developed by Google. The algorithm works by recursively calculating the probability of a user visiting a particular page by following links. Pages with higher PageRank are considered more important or relevant.

2) HITS is another algorithm used for ranking web pages. Unlike PageRank, HITS identifies two types of important pages: hubs and authorities. A hub is a page that links to many authorities, while an authority is a page that is linked by many hubs. The algorithm iteratively computes hub and authority scores for each page based on the links between them.

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.

1) Dataset
    a) wt2g_inlinks: This dataset contains information about the inbound links of web pages. Each line represents a web page followed by the web pages it links to. For example, "WT01-B01-2" links to "WT01-B01-6" and "WT01-B01-1".
2) Tools used
    a) Python, Elasticsearch, random, math library
3) Process of calculating PageRank:
    a) Initialization: Initialize the PageRank values for each page to an initial value, typically set to 1/N, where N is the total number of pages.
    b) Sink nodes: Identify and handle sink nodes (pages with no outbound links) separately to distribute their PageRank to all other pages.
    c) Iterative Calculation: Iteratively update the PageRank values for each page based on the PageRank values of the pages linking to it. The PageRank value for each page is calculated as a combination of its own importance and the importance of the pages linking to it.

    d) Convergence: Check for convergence by calculating the perplexity of the PageRank distribution. If the perplexity doesn't change significantly over iterations, the algorithm is considered converged.

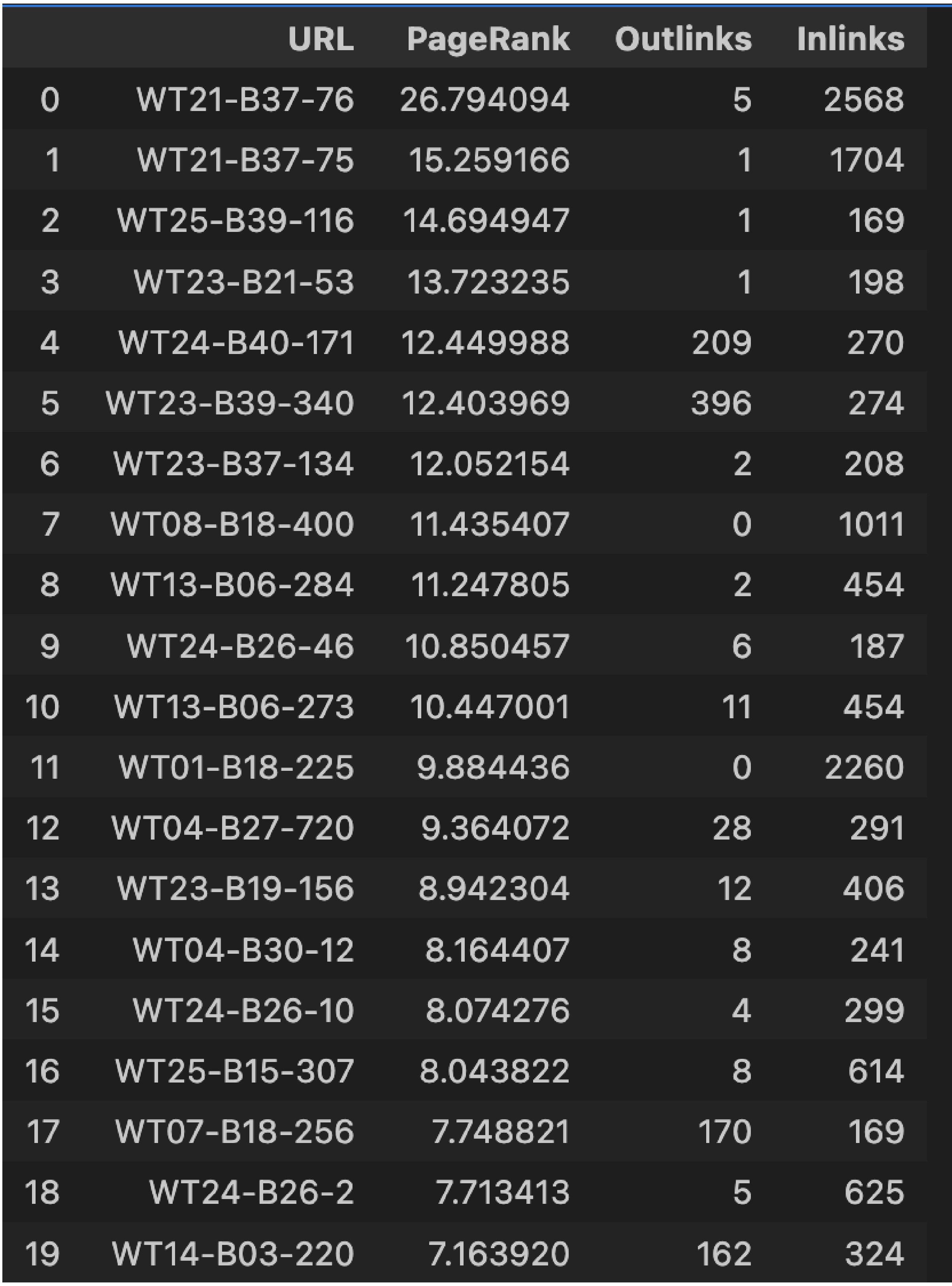
4) Process of calculating Hub and Authority scores (HITS algorithm):
    a) Initialization: Create a root set of web pages, often by querying a search engine for pages relevant to a specific topic.
    b) Update Set: Expand the root set by including additional pages linked to and from the pages in the root set. This helps in capturing a broader set of relevant pages.
    c) Iterative Computation: Iteratively compute the hub and authority scores for each page in the root set. Hub scores represent the quality of outbound links from a page, while authority scores represent the quality of inbound links to a page.
    d) Normalization: Normalize the hub and authority scores to ensure that they represent relative importance rather than absolute values.
    e) Convergence: Repeat the iterative computation process until convergence criteria are met, such as a certain number of iterations or when the scores stabilize.

Analysis

PageRank Analysis

PageRank on WT2G_Inlinks Data

- Include a screenshot of the first 20 rows.



	URL	PageRank	Outlinks	Inlinks
0	WT21-B37-76	26.794094	5	2568
1	WT21-B37-75	15.259166	1	1704
2	WT25-B39-116	14.694947	1	169
3	WT23-B21-53	13.723235	1	198
4	WT24-B40-171	12.449988	209	270
5	WT23-B39-340	12.403969	396	274
6	WT23-B37-134	12.052154	2	208
7	WT08-B18-400	11.435407	0	1011
8	WT13-B06-284	11.247805	2	454
9	WT24-B26-46	10.850457	6	187
10	WT13-B06-273	10.447001	11	454
11	WT01-B18-225	9.884436	0	2260
12	WT04-B27-720	9.364072	28	291
13	WT23-B19-156	8.942304	12	406
14	WT04-B30-12	8.164407	8	241
15	WT24-B26-10	8.074276	4	299
16	WT25-B15-307	8.043822	8	614
17	WT07-B18-256	7.748821	170	169
18	WT24-B26-2	7.713413	5	625
19	WT14-B03-220	7.163920	162	324

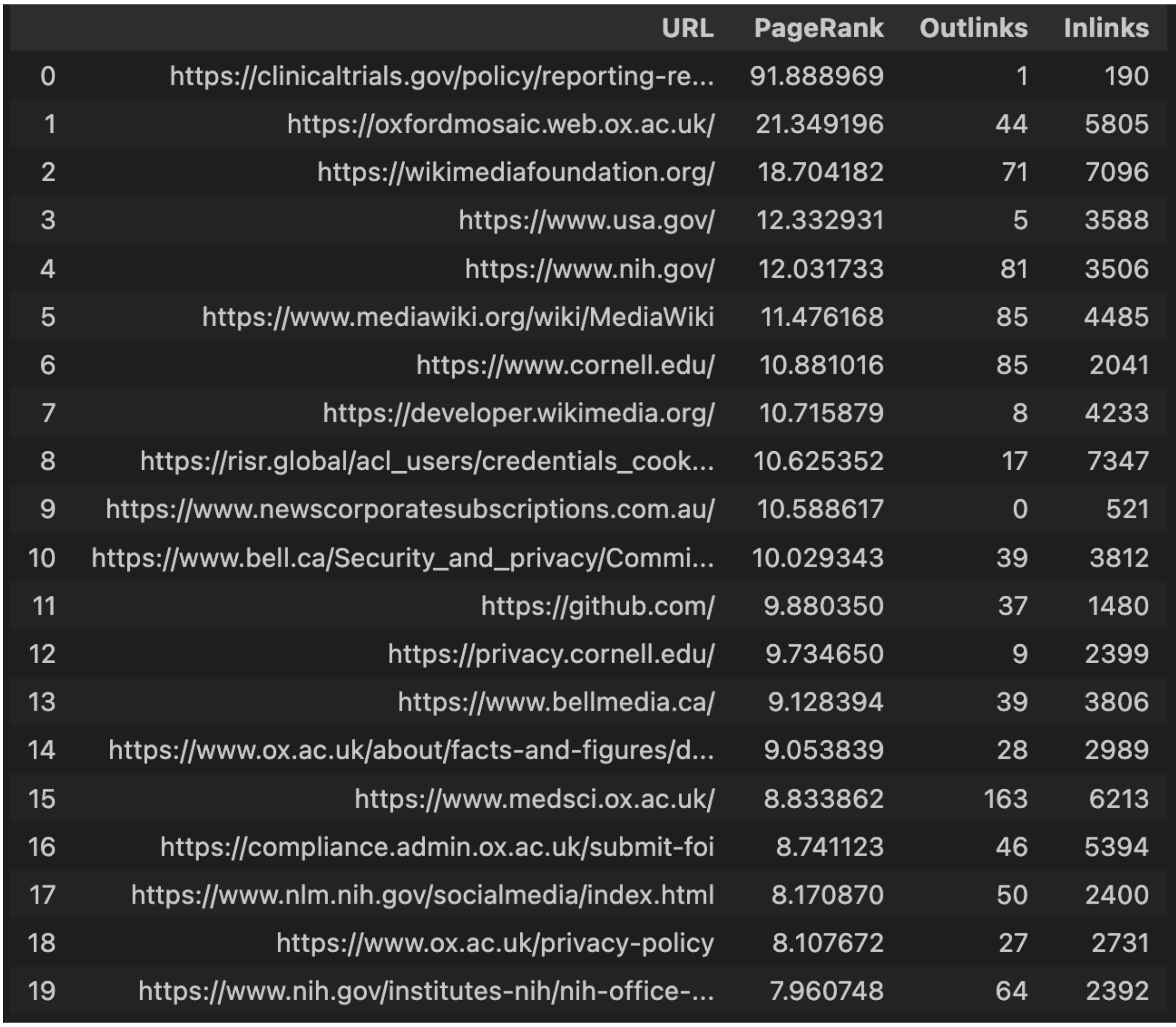
- Discuss any notable patterns or anomalies observed in the PageRank scores compared to inlink counts.
    - High PageRank with Low Inlinks:
        - Some pages exhibit high PageRank scores despite having relatively low numbers of inlinks. For example, "WT21-B37-76" has a PageRank of 26.794094 with only 2568 inlinks. This suggests that although the

page doesn't receive a large number of incoming links, the quality and importance of those incoming links are significant enough to contribute to its high PageRank.

- Low PageRank with High Inlinks:
    - Conversely, there are instances where pages have a high number of inlinks but relatively low PageRank scores. For instance, "WT01-B18-225" has 2260 inlinks but a PageRank of only 9.884436. This could indicate that while the page receives many incoming links, they may be from low-quality or less relevant sources, impacting its overall PageRank.
- PageRank Discrepancies:
    - There are instances where the PageRank scores seem disproportionate to the number of inlinks or outlinks. For instance, "WT25-B39-116" has a relatively high PageRank of 14.694 with 169 inbound links, which appears to be relatively low compared to other pages with similar PageRank scores.

PageRank on Merged Data

- Include a screenshot of the first 20 rows.

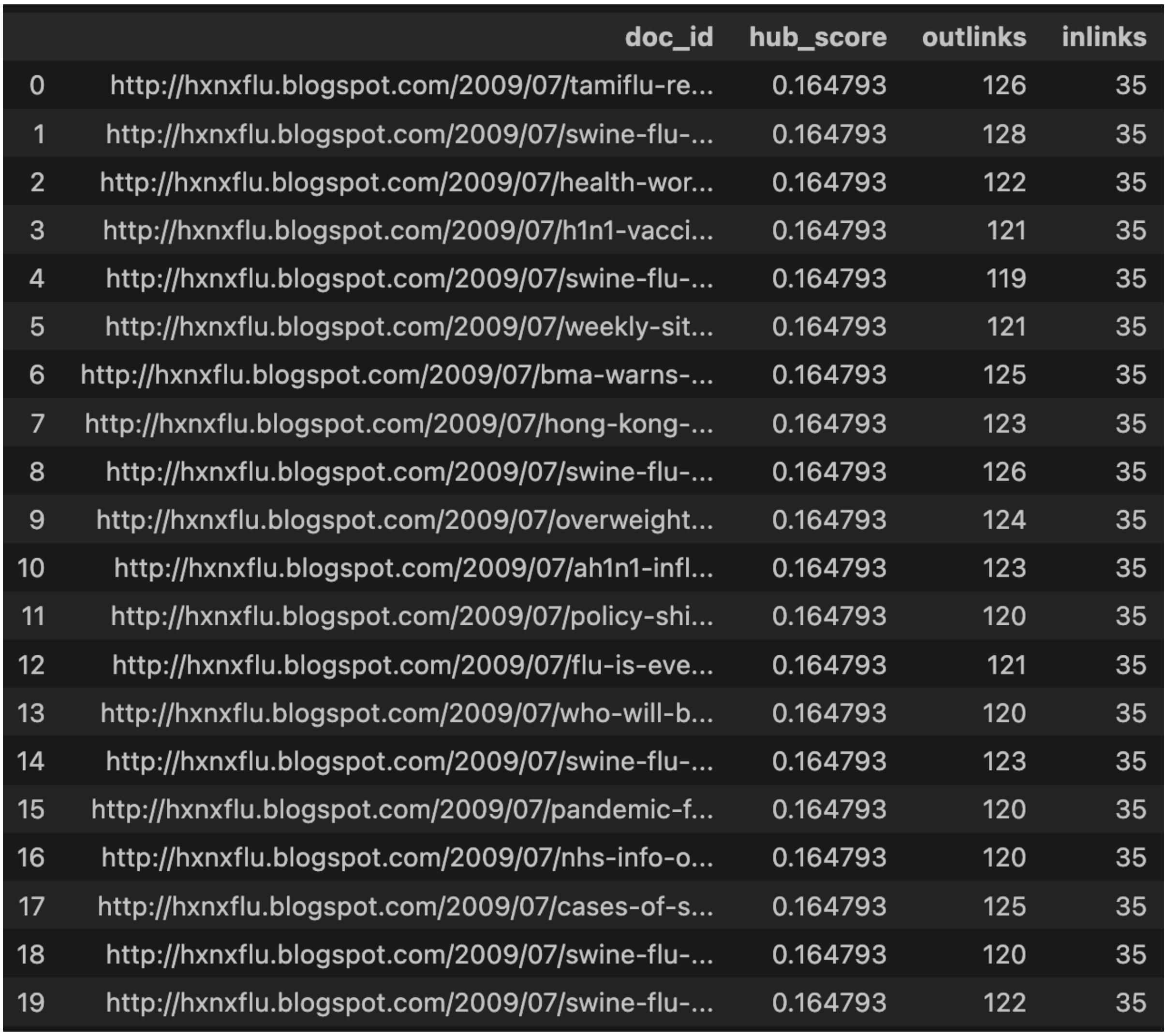


	URL	PageRank	Outlinks	Inlinks
0	https://clinicaltrials.gov/policy/reporting-re...	91.888969	1	190
1	https://oxfordmosaic.web.ox.ac.uk/	21.349196	44	5805
2	https://wikimediafoundation.org/	18.704182	71	7096
3	https://www.usa.gov/	12.332931	5	3588
4	https://www.nih.gov/	12.031733	81	3506
5	https://www.mediawiki.org/wiki/MediaWiki	11.476168	85	4485
6	https://www.cornell.edu/	10.881016	85	2041
7	https://developer.wikimedia.org/	10.715879	8	4233
8	https://risr.global/acl_users/credentials_cook...	10.625352	17	7347
9	https://www.newscorporatesubscriptions.com.au/	10.588617	0	521
10	https://www.bell.ca/Security_and_privacy/Commi...	10.029343	39	3812
11	https://github.com/	9.880350	37	1480
12	https://privacy.cornell.edu/	9.734650	9	2399
13	https://www.bellmedia.ca/	9.128394	39	3806
14	https://www.ox.ac.uk/about/facts-and-figures/d...	9.053839	28	2989
15	https://www.medsci.ox.ac.uk/	8.833862	163	6213
16	https://compliance.admin.ox.ac.uk/submit-foi	8.741123	46	5394
17	https://www.nlm.nih.gov/socialmedia/index.html	8.170870	50	2400
18	https://www.ox.ac.uk/privacy-policy	8.107672	27	2731
19	https://www.nih.gov/institutes-nih/nih-office-...	7.960748	64	2392

- Highlight any differences observed when comparing the PageRank scores from the WT2G_Inlinks data to the merged data set.
    - Pages in the merged dataset seem to have higher counts of outlinks and inlinks compared to the WT2G_Inlinks dataset. For instance, in the merged dataset, we see pages with up to 163 outbound links and 7347 inbound links, whereas the WT2G_Inlinks dataset had fewer links on average.

HITS Analysis

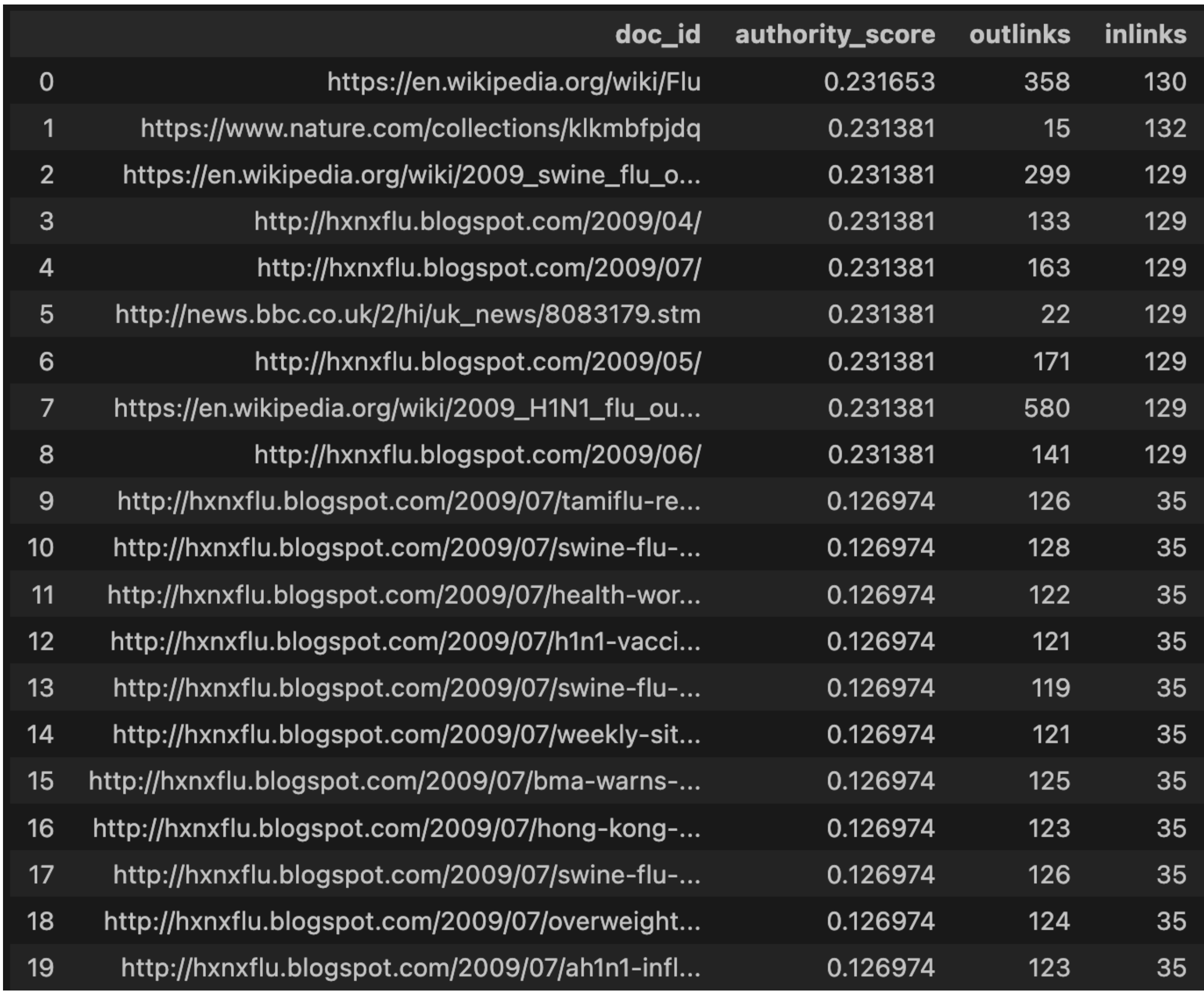
Hub Scores

- Include a screenshot of the first 20 rows.
    - 

	doc_id	hub_score	outlinks	inlinks
0	http://hxnxflu.blogspot.com/2009/07/tamiflu-re...	0.164793	126	35
1	http://hxnxflu.blogspot.com/2009/07/swine-flu-...	0.164793	128	35
2	http://hxnxflu.blogspot.com/2009/07/health-wor...	0.164793	122	35
3	http://hxnxflu.blogspot.com/2009/07/h1n1-vacci...	0.164793	121	35
4	http://hxnxflu.blogspot.com/2009/07/swine-flu-...	0.164793	119	35
5	http://hxnxflu.blogspot.com/2009/07/weekly-sit...	0.164793	121	35
6	http://hxnxflu.blogspot.com/2009/07/bma-warns-...	0.164793	125	35
7	http://hxnxflu.blogspot.com/2009/07/hong-kong-...	0.164793	123	35
8	http://hxnxflu.blogspot.com/2009/07/swine-flu-...	0.164793	126	35
9	http://hxnxflu.blogspot.com/2009/07/overweight...	0.164793	124	35
10	http://hxnxflu.blogspot.com/2009/07/ah1n1-infl...	0.164793	123	35
11	http://hxnxflu.blogspot.com/2009/07/policy-shi...	0.164793	120	35
12	http://hxnxflu.blogspot.com/2009/07/flu-is-eve...	0.164793	121	35
13	http://hxnxflu.blogspot.com/2009/07/who-will-b...	0.164793	120	35
14	http://hxnxflu.blogspot.com/2009/07/swine-flu-...	0.164793	123	35
15	http://hxnxflu.blogspot.com/2009/07/pandemic-f...	0.164793	120	35
16	http://hxnxflu.blogspot.com/2009/07/nhs-info-o...	0.164793	120	35
17	http://hxnxflu.blogspot.com/2009/07/cases-of-s...	0.164793	125	35
18	http://hxnxflu.blogspot.com/2009/07/swine-flu-...	0.164793	120	35
19	http://hxnxflu.blogspot.com/2009/07/swine-flu-...	0.164793	122	35

- Provide insights into the relationship between Hub scores and the structure of the web graph.
    - The Hub scores represent the quality and relevance of outbound links from a page within the web graph. In the provided data, all pages have identical Hub scores, suggesting that they are considered equally authoritative in terms of linking out to other pages. This indicates a uniform distribution of authority across the pages within the web graph, where each page is regarded as equally important in directing users to relevant content.

Authority Scores

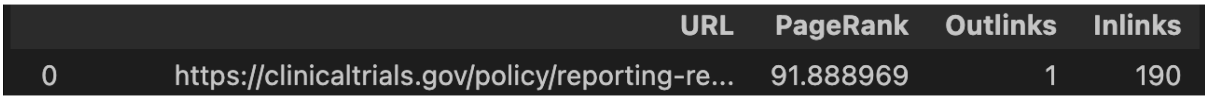
- Include a screenshot of the first 20 rows.
    - 

	doc_id	authority_score	outlinks	inlinks
0	https://en.wikipedia.org/wiki/Flu	0.231653	358	130
1	https://www.nature.com/collections/klkmbfpjdq	0.231381	15	132
2	https://en.wikipedia.org/wiki/2009_swine_flu_o...	0.231381	299	129
3	http://hxnxflu.blogspot.com/2009/04/	0.231381	133	129
4	http://hxnxflu.blogspot.com/2009/07/	0.231381	163	129
5	http://news.bbc.co.uk/2/hi/uk_news/8083179.stm	0.231381	22	129
6	http://hxnxflu.blogspot.com/2009/05/	0.231381	171	129
7	https://en.wikipedia.org/wiki/2009_H1N1_flu_ou...	0.231381	580	129
8	http://hxnxflu.blogspot.com/2009/06/	0.231381	141	129
9	http://hxnxflu.blogspot.com/2009/07/tamiflu-re...	0.126974	126	35
10	http://hxnxflu.blogspot.com/2009/07/swine-flu-...	0.126974	128	35
11	http://hxnxflu.blogspot.com/2009/07/health-wor...	0.126974	122	35
12	http://hxnxflu.blogspot.com/2009/07/h1n1-vacci...	0.126974	121	35
13	http://hxnxflu.blogspot.com/2009/07/swine-flu-...	0.126974	119	35
14	http://hxnxflu.blogspot.com/2009/07/weekly-sit...	0.126974	121	35
15	http://hxnxflu.blogspot.com/2009/07/bma-warns-...	0.126974	125	35
16	http://hxnxflu.blogspot.com/2009/07/hong-kong-...	0.126974	123	35
17	http://hxnxflu.blogspot.com/2009/07/swine-flu-...	0.126974	126	35
18	http://hxnxflu.blogspot.com/2009/07/overweight...	0.126974	124	35
19	http://hxnxflu.blogspot.com/2009/07/ah1n1-infl...	0.126974	123	35

- Discuss how Authority scores compare with PageRank and Hub scores, and what this implies about the web pages' importance or relevance.
    - The Authority scores reflect the perceived importance or relevance of web pages based on the quality and relevance of their inbound links. In the provided data, Authority scores appear to correlate with PageRank scores, indicating that pages with higher inbound link quality tend to have higher Authority scores. This suggests that these pages are considered more authoritative within the web graph.
    - Comparatively, Hub scores represent the quality and relevance of outbound links from a page. In this dataset, Hub scores are not provided, but if available, we could infer that pages with high Hub scores are influential in directing users to relevant content.
    - Overall, the alignment between Authority scores and PageRank suggests that the importance or relevance of a page is closely tied to both the quality of its inbound links and its overall connectivity within the web graph.

Case Study: PageRank vs. Inlink Count

Select a few pages that have a higher PageRank but a smaller inlink count. For each selected page:

- Identify and describe the other pages that point to them.
    - 

	URL	PageRank	Outlinks	Inlinks
0	https://clinicaltrials.gov/policy/reporting-re...	91.888969	1	190

        - https://clinicaltrials.gov:443/search?cond=%22Dementia%22&aggFilters=status%3Anot+rec+ava 0.08956820774416721
        - https://clinicaltrials.gov/study/NCT04192500 0.0826078987021847
        - https://clinicaltrials.gov/search?cond=%22Headache+Disorders%22&aggFilters=status:not%20rec 0.08373397399186093
        - https://clinicaltrials.gov:443/search?cond=Sjogren%27s+Syndrome&city= 0.08582920137713476
        - https://clinicaltrials.gov/study/NCT01142362 0.0826078987021847
        - https://clinicaltrials.gov/study/NCT03293498 0.0826078987021847
        - https://clinicaltrials.gov/study/NCT05426174 0.0826078987021847
        - https://clinicaltrials.gov/study/NCT05827978 0.0826078987021847
        - https://clinicaltrials.gov:443/search?cond=Rett+Syndrome&aggFilters=status%3Arec+act 0.08617622440054049
        - https://clinicaltrials.gov/study/NCT05155319 0.0826078987021847
        - https://clinicaltrials.gov/study/NCT05227001 0.0826078987021847
        - https://clinicaltrials.gov/search?cond=%22Adenoviridae+Infections%22&aggFilters=status:not%20rec 0.0836182932368123
        - https://clinicaltrials.gov/study/NCT03275389 0.0826078987021847
        - https://clinicaltrials.gov/study/NCT04622592 0.0826078987021847
        - https://clinicaltrials.gov/study/NCT00819013 0.0826078987021847
        - https://clinicaltrials.gov/study/NCT05284799 0.0826078987021847
        - https://clinicaltrials.gov/study/NCT05333289 0.0826078987021847
        - https://clinicaltrials.gov:443/search?cond=Bipolar+Disorder&city= 0.08622825833502684
        - https://clinicaltrials.gov:443/search?term=pulse+oximeters&city= 0.08493101532858965
        - https://clinicaltrials.gov/study/NCT05606965 0.0826078987021847
        - https://clinicaltrials.gov/study/NCT06125691 0.0826078987021847
- Explain why these pages have a higher PageRank despite a smaller inlink count, focusing on the quality or authority of the incoming links.
    - These pages likely have a higher PageRank despite a smaller inlink count because they receive inbound links from highly authoritative sources. PageRank considers not only the quantity but also the quality of incoming links.
    - In this case, the highest PageRank URL, "https://clinicaltrials.gov/policy/reporting-requirements," is linked to by other highly reputable and authoritative sources, such as clinical trial study pages and official government websites. These sources are likely considered authoritative in the context of clinical research and related topics.
    - Even though the inlink count may be smaller compared to other pages, the quality and authority of the incoming links play a significant role in determining the PageRank. Therefore, the presence of inbound links from trusted and relevant sources contributes to the higher PageRank of these pages.