Homework 4 Report

Name: Rohith Chandra Kandambeth

Introduction

There are three main objectives in this assignment.
1. To use page rank algorithm on the wt2g_inlinks data and get the top 500 pages
2. To use page rank algorithm on the crawled inlinks data and get the top 500 pages
3. Compute hits, with hubs and authority score on the root set (used to form a base set) for the specific topic.

The PageRank algorithm starts by assigning an initial rank value to all pages in the web graph, equally distributing 'rank' among them. It then iteratively updates the rank of each page to simulate a 'random surfer' who randomly follows links, but occasionally jumps to a random page. In each iteration, we first account for 'sink' pages, which don't link anywhere, redistributing their rank value across all pages—this ensures that the rank is conserved within the system. Each page's new rank is composed of a small fraction that comes from the random jumps (teleportation), ensuring every page gets a baseline rank value, plus a share of the rank from pages that link to it, proportionally to their rank and the number of links they have. Pages that are linked by high-ranked pages or many pages gain more rank. This process repeats until the ranks stabilize and stop changing significantly, which means we've reached convergence, and the pages' ranks are now a fair representation of their relative importance and authority in the web.

For the HITS algorithm, we start by selecting a root set of about 1,000 documents that are most relevant to our query, using an information retrieval function an Elasticsearch search. We expand this root set by adding pages linked from and to each page in the set. If a page links to more than 200 other pages, we add only a random selection of 200 to keep our set manageable. This process is repeated two or three times until we have a base set of around 10,000 pages. Then we calculate each page's HITS scores, starting both their authority and hub scores at 1. Authority scores increase with more in-links from high-scoring hubs, while hub scores increase with more out-links to high-scoring authorities. We continually update these scores and normalize them after each iteration until they stabilize, revealing which pages are the strongest authorities and hubs on our topic

Methodology

The wt2g_inlinks dataset is likely a collection of web pages along with their in-link information. There are 183811 documents/webpages.

The other dataset used is the data crawled for topics - West African Ebola epidemic, H1N1 Swine Flu pandemic, COVID 19.

PageRank is calculated by assigning an initial rank to all pages (P) equally, and then repeatedly refining that rank through iterations. Each page (p) starts with a rank of 1/N, where N is the total number of pages. During each iteration, we deal with 'sink' pages (S), which are those without out-links, by redistributing their rank (sinkPR) evenly across all pages. Each page's rank is then updated by adding a baseline teleportation rank of (1-d)/N, where d is the

damping factor, typically 0.85, plus a share of the sinkPR, plus the weighted rank contributions from pages that link to it (M(p)), with each contributing page (q) passing on a portion of its rank divided by its number of out-links (L(q)). This iterative process continues until changes in rank stabilize, indicating convergence, at which point the PageRank values represent the relative importance or likelihood of visiting each page in a random web surfing scenario.

Hubs and authority scores are calculated using an iterative process called the HITS algorithm. Initially, each page is given a hub and an authority score of 1. In each iteration, a page's authority score is updated to be the sum of the hub scores of all pages that link to it, reflecting the intuition that a good authority is a page that is linked by good hubs. Conversely, a page's hub score is updated to be the sum of the authority scores of all pages it links to, since a good hub is a page that links to good authorities. After each iteration, both hub and authority scores are normalized to prevent runaway scores. This process repeats until the scores converge. The resulting scores identify the most reputable pages (authorities) and the best directories or resource lists (hubs) within a network of web pages

Analysis

PageRank Analysis

PageRank on WT2G_Inlinks Data

wt2g_pagerank.txt

Page	Page Rank	No. of Outlinks	No. of Inlinks
WT21-B37-76	0.0026944708785777444	5	2568
WT21-B37-75	0.0015331771293438034	1	1704
WT25-B39-116	0.0014685087868547102	1	169
WT23-B21-53	0.0013735335821988344	1	198
WT24-B26-10	0.001276215008801935	1	291
WT24-B40-171	0.0012452591223336598	209	270
WT23-B39-340	0.0012428612869828874	395	274
WT23-B37-134	0.0012054273922617123	2	207
WT08-B18-400	0.0011447764367003175	0	990
WT13-B06-284	0.001136550377992955	2	454
WT13-B06-273	0.0010549175801714342	11	452
WT01-B18-225	0.0009553812196934016	0	1137
WT04-B27-720	0.000940955907280795	27	291
WT24-B26-46	0.0008622309745390188	3	179
WT23-B19-156	0.0008250669463171077	12	364
WT04-B30-12	0.0008166209191956031	8	241
WT25-B15-307	0.0007972230008250924	8	605
WT07-B18-256	0.0007750024108448045	169	169
WT24-B40-167	0.0007076056547892882	152	153
WT14-B03-220	0.0006988600464577245	162	163
WT18-B31-240	0.0006942192732102163	31	259
WT14-B03-227	0.0006852823072029095	147	148
WT04-B40-202	0.0006846752449326029	36	322
WT08-B19-222	0.0006495321309790612	1	1041
WT23-B20-363	0.0006396307474757167	193	181
WT27-B28-203	0.0006270789092354628	2	589
WT13-B39-295	0.0006215355451936358	19	443
WT13-B15-160	0.0006198581863076683	0	484

.

A notable pattern is that some pages have high PageRank scores despite a low count of inlinks, suggesting these inlinks are from highly authoritative sources. Conversely, pages with many inlinks but lower PageRank scores might be linked from less authoritative sites. Additionally, sink pages with no outlinks can disproportionately retain PageRank, influencing the overall distribution of scores. Even though it has fewer inlinks than some other pages, it likely has inlinks from pages with high PageRank themselves. Those high-quality "votes" boost

its own PageRank score. The algorithm doesn't directly check any features of the page itself to determine importance

PageRank on Merged Data

Page	Page Rank	No. of Outlinks	No. of Inlinks
https://wikimediafoundation.org/	0.0030544770738605 98	54	7096
https://www.mediawiki.org/wiki/MediaWiki	0.0021888858035014835	80	4485
https://developer.wikimedia.org/	0.002080621304738688	2	4233
https://support.apple.com/?cid=gn-ols-home-hp-tab	0.001998266388381666	21	838
https://clinicaltrials.gov/policy/reporting-requir	0.001996959077386701	1	190
https://oxfordmosaic.web.ox.ac.uk/	0.001415092417026366	29	5805
https://apps.apple.com/us/app/apple-store/id375380	0.001173776232657226	17	114
https://github.com/	0.0011597987047498508	21	1480
https://proquest.libguides.com/termsofuse	0.0010454282279149833	1	260
https://www.nih.gov/	0.0010352003666634993	53	3506
https://shop.slate.com/collections/holiday-gift-gu	0.0010348153798556492	1	2
https://en.wikipedia.org/wiki/Pandemic	0.001018775902224508	1098	3024
https://www.usa.gov/	0.0009637495141655669	24	3588
https://wikimediafoundation.org/our-work/wikimedia	0.0009597355869752258	26	116
https://www.bell.ca/Security_and_privacy/Commitmen	0.0009512268169778663	14	3812
https://www.cornell.edu/	0.0009008034138420843	71	2041
https://www.gnu.org/licenses/gpl-3.0.html	0.0008892826998997395	2	4
https://www.nlm.nih.gov/	0.0008636252275703621	14	1549
https://en.wikipedia.org/wiki/Main_Page	0.0007891145856137053	34	2890
https://knowledge.exlibrisgroup.com/	0.0007342981018919921	17	642
https://en.wikipedia.org/wiki/Help:Contents	0.0007305776861150904	26	2392
https://github.blog/	0.00072221891595626	106	1518
https://knowledge.exlibrisgroup.com/TERMS_OF_USE	0.0007160980049307191	3	621
https://knowledge.exlibrisgroup.com/Cross-Product/	0.000715166461003263	4	619
https://privacy.cornell.edu/	0.0007139744898959616	3	2399
https://www.alumni.ox.ac.uk/	0.0007032460503681074	14	1666
https://www.githubstatus.com/	0.0007020940082008358	15	1495
https://automattic.com/	0.0006911269729690046	8	1956
https://en.wikipedia.org/wiki/Help:Introduction	0.0006547700850516802	14	2349
https://www.nih.gov/institutes-nih/nih-office-dire	0.0006336319995539928	41	2392
https://en.wikipedia.org/wiki/Wikipedia:Contact_us	0.0006330980679334794	23	2336
https://identi.ca/jimfulner	0.0006252195470295262	1	10
https://en.wikipedia.org/wiki/Wikipedia:About	0.0006100293530990573	29	2329

In comparing the WT2G_Inlinks PageRank data to a merged dataset, we might observe that pages with fewer inlinks can have high PageRank scores, likely due to links from high-authority sites, while others with many inlinks don't rank as highly, possibly because of links from less authoritative sources. Pages with a balanced mix of inlinks and outlinks, like github.com, often maintain strong PageRank scores, indicating a robust web presence. The discrepancies between datasets can arise from the unique linking structures and the varying authority of the pages within each dataset, with a merged dataset potentially introducing new link patterns that could shift PageRank scores.

HITS Analysis

Hub Scores

https://www.bmj.com/company/bmj-tag/ 0.051559668340021216
https://www.bmj.com/company/anti-racism-at-bmj-2/ 0.05155828909157505
https://www.bmj.com/company/apha2021/ 0.05155818279332216
https://www.bmj.com/company/work-at-bmj-today/ 0.05155818130879347
https://www.bmj.com/company/global-health-ii/climate-change-and-infectious-diseases/ 0.05155818126301736
https://www.bmj.com/company/newsroom/the-world-can-learn-from-chinas-response-to-the-pandemic-say-experts/ 0.051558181199064725
https://www.bmj.com/company/newsroom/male-scientists-frame-their-research-findings-more-positively-than-women/ 0.051557850714690455
https://www.bmj.com/company/newsroom/the-nhs-paid-private-hospitals-2bn-in-the-pandemic-but-some-treated-more-private-patients-than-nhs-ones/ 0.051557850714690455
https://www.bmj.com/company/newsroom/significant-racial-disparities-in-care-of-heart-patients-during-first-pandemic-wave/ 0.051557850714690455
https://www.bmj.com/company/global-health-ii/making-the-most-of-your-time/ 0.051557850714690455
https://www.bmj.com/company/newsroom/mass-covid-testing-at-uk-universities-is-haphazard-and-unscientific-finds-bmj-investigation/ 0.051557850714690455
https://www.bmj.com/company/bmj-resource-centre/competitions/ 0.051557850714690455
https://www.bmj.com/company/newsroom/low-blood-folate-may-be-linked-to-heightened-dementia-and-death-risks-in-older-people/ 0.051557850714690455
https://www.bmj.com/company/newsroom/celebrity-tweets-likely-shaped-us-negative-public-opinion-of-covid-19-pandemic/ 0.051557850714690455
https://www.bmj.com/company/newsroom/new-editor-in-chief-for-bmj-open-sport-exercise-medicine/ 0.051557850714690455
https://www.bmj.com/company/newsroom/acute-psychotic-illness-triggered-by-brexit-referendum/ 0.051557850714690455
https://www.bmj.com/company/newsroom/limited-supply-may-scupper-proposals-to-use-antimalarials-to-ward-off-covid-19/ 0.051557850714690455
https://www.bmj.com/company/newsroom/maintaining-healthy-lifestyle-might-prevent-up-to-60-of-inflammatory-bowel-disease-cases/ 0.051557850714690455
https://www.bmj.com/company/newsroom/new-study-updates-evidence-on-rare-heart-condition-after-covid-vaccination/ 0.051557850714690455
https://www.bmj.com/company/newsroom/mid-life-moderate-vigorous-physical-activity-quota-associated-with-brain-power/ 0.051557850714690455
https://www.bmj.com/company/advertising-sponsorship-2/for-advertisers-and-sponsor/ 0.051557850714690455
https://www.bmj.com/company/legal-information/ 0.051557850714690455
https://www.bmj.com/company/newsroom/media-reports-of-celebrity-suicide-linked-to-increased-suicide-rates/ 0.051557850714690455
https://www.bmj.com/company/newsroom/most-health-claims-on-infant-formula-products-seem-to-have-little-or-no-supporting-evidence/ 0.051557850714690455
https://www.bmj.com/company/newsroom/key-healthcare-and-tech-companies-pledge-to-decarbonise-nhs-supply-chain-by-2045/ 0.051557850714690455
https://www.bmj.com/company/our-products/ 0.051557850714690455
https://www.bmj.com/company/newsroom/menstrual-discs-may-be-best-for-heavy-monthly-blood-flow/ 0.051557850714690455
https://www.bmj.com/company/bmj-revenue-sources/ 0.051557850714690455
https://www.bmj.com/company/bmj-resource-centre/bmj-case-reports-resources/bmj-case-reports-user-guides/ 0.051557850714690455
https://www.bmj.com/company/newsroom/new-study-shows-covid-19-virus-persists-longer-and-peaks-later-in-the-respiratory-tissue-of-patients-with-severe-disease/ 0.051557850714690455
https://www.bmj.com/company/newsroom/many-antibiotic-courses-for-common-infections-not-in-line-with-guidelines/ 0.051557850714690455
https://www.bmj.com/company/advertisers-sponsors-signup-alerts/ 0.051557850714690455
https://www.bmj.com/company/who-we-are/values/ 0.051557850714690455
https://www.bmj.com/company/newsroom/major-surgery-associated-with-small-long-term-decline-in-brain-functioning/ 0.051557850714690455
https://www.bmj.com/company/newsroom/moderate-egg-intake-not-associated-with-cardiovascular-disease-risk/ 0.051557850714690455
https://www.bmj.com/company/newsroom/rise-in-childhood-short-sightedness-may-be-linked-to-pandemic-suggests-hong-kong-study/ 0.051557850714690455
https://www.bmj.com/company/newsroom/disrupted-access-to-healthcare-during-pandemic-linked-to-avoidable-hospital-admissions/ 0.051557850714690455
https://www.bmj.com/company/newsroom/midlife-chronic-conditions-linked-to-increased-dementia-risk-later-in-life/ 0.051557850714690455
https://www.bmj.com/company/bmj-resource-centre/bmj-best-practice-resource-centre/comorbidities_manager/getting-started-guide/ 0.051557850714690455
https://www.bmj.com/company/bmj-resource-centre/research-to-publication/curriculum-guides/ 0.051557850714690455
https://www.bmj.com/company/newsroom/metal-pins-no-better-than-traditional-plaster-cast-for-a-broken-wrist/ 0.051557850714690455
https://www.bmj.com/company/newsroom/mediterranean-diet-promotes-gut-bacteria-linked-to-healthy-ageing-in-older-people/ 0.051557850714690455
https://www.bmj.com/company/newsroom/labelling-foods-with-amount-of-physical-activity-needed-to-burn-off-calories-linked-to-healthier-choices/ 0.051557850714690455

.

The consistent hub scores across various BMJ web pages indicate that these pages may serve similarly as conduits to authoritative and relevant content, both within the BMJ's ecosystem and beyond. This uniformity could be due to the initial setup before the iterative process of the HITS algorithm fully differentiates the pages based on the strength and relevance of their outbound links. In a typical web graph, strong hubs are pages that reference many authoritative sources, suggesting that BMJ's site architecture might be designed to evenly distribute this role among its pages. As the HITS algorithm refines these scores through iterations, we would expect to see a variation develop where certain pages become recognized as more central hubs based on the quality and number of their links to recognized authorities.

Authority Scores

```
https://www.bmj.com/company/your-privacy/       0.04953772256565161
https://journals.bmj.com/       0.04950214891282642
https://www.bmj.com/company/legal-information/  0.04949488001409312
https://authors.bmj.com/        0.04949488001409312
https://www.bmj.com/company/legal-information/accessibility/    0.04949329699453884
https://www.bmj.com/company/openaccess/ 0.04949329699453884
https://www.bmj.com/company/    0.04948846469647341
https://www.bmj.com/company/americas/librarian-hub/edi-policy/  0.04948783247929487
https://www.bmj.com/company/americas/meet-the-team/     0.04948783247929487
https://www.bmj.com/company/benefits/   0.04948783247929487
https://www.bmj.com/company/americas/rights-licensing-and-permissions/  0.04948783247929487
https://www.bmj.com/company/americas/education/ 0.04948783247929487
https://www.bmj.com/company/americas/contact-us/        0.04948783247929487
https://www.bmj.com/company/americas/who-we-are-americas/       0.04948783247929487
https://www.bmj.com/company/americas/ebm-resources/     0.04948783247929487
https://www.bmj.com/company/americas/pharma-relevant-journals/  0.04948783247929487
https://www.bmj.com/company/americas/librarian-hub/faq/ 0.04948783247929487
https://www.bmj.com/company/americas/conferences-meetings/      0.04948783247929487
https://www.bmj.com/company/americas/impact-analytics/  0.04948783247929487
https://www.bmj.com/company/americas/advertising-and-sponsors/  0.04948783247929487
https://www.bmj.com/company/americas/   0.04946439778229687
https://www.bmj.com/company/americas/getpublished/      0.049459039988774434
https://www.bmj.com/company/americas/librarian-hub/product-brochures-and-videos/        0.04943497116491986
https://www.bmj.com/company/americas/state-of-the-art-clinical-reviews/ 0.049223681951844775
https://www.bmj.com/company/americas/librarian-hub/webinars/    0.04919520721661801
https://www.bmj.com/company/americas/librarian-hub/the-bmj/     0.0491079373535226
https://www.bmj.com/company/open-access-for-librarians/ 0.049102440803204736
https://www.bmj.com/company/americas/librarian-hub/admin-latam/ 0.04910012328595136
https://www.bmj.com/company/americas/librarian-hub/admin-brazil/        0.04910012328595136
https://www.bmj.com/company/americas/librarian-hub/admin-canada/        0.04910012328595136
https://www.bmj.com/company/americas/librarian-hub/admin-us/    0.04910012328595136
https://www.bmj.com/company/americas/librarian-hub/ijgc/        0.04909949372032798
https://www.bmj.com/company/americas/librarian-hub/bmj-neurology-neurosurgery-psychiatry/       0.04909949372032798
https://www.bmj.com/company/americas/librarian-hub/bmj-annals-rheumatic-diseases/       0.04909949372032798
https://www.bmj.com/company/americas/librarian-hub/bmj-evidence-based/  0.04909854961329791
https://recruiter.bmj.com/      0.04909314662172245
https://www.bmj.com/company/americas/librarian-hub/promo-latam/ 0.04909010606555967
https://www.bmj.com/company/americas/librarian-hub/promo-us/    0.04909010606555967
https://www.bmj.com/company/americas/librarian-hub/promo-ca/    0.04909010606555967
https://www.bmj.com/company/americas/librarian-hub/promo-brasil/        0.04909010606555967
https://www.bmj.com/company/americas/librarian-hub/case-reports/        0.04909010598010329
https://www.bmj.com/company/americas/librarian-hub/bmj-quality-safety/  0.04909010598010329
https://www.bmj.com/company/americas/librarian-hub/bmj-emergency-medical-journal/       0.04909010598010329
https://www.bmj.com/company/americas/librarian-hub/bmj-sports-medicine/ 0.04909010598010329
https://www.bmj.com/company/americas/librarian-hub/rapm/        0.04909010598010329
https://www.bmj.com/company/americas/librarian-hub/bmj-journal-epidemiology-community-health/   0.04909010598010329
https://www.bmj.com/company/americas/librarian-hub/casereports-uscorporate-bmj/ 0.04908821792841559
https://www.bmj.com/company/americas/librarian-hub/bmj-journals-archive/        0.049088217799047956
https://www.bmj.com/company/americas/high-impact-journals/      0.049088217799047956
```

Discuss how Authority scores compare with PageRank and Hub scores, and what this implies about the web pages' importance or relevance.

Case Study: PageRank vs. Inlink Count

Out of top 50 the below pages have fewer inlinks but higher pagerank
'WT13-B39-321' has 52 inlinks
'WT06-B14-69' has 57 inlinks
'WT23-B38-87' has one inlinks
We can look at the example of 'WT23-B38-87' having one inlink WT23-B37-134 having a pagerank of 8. Eventhough inlinks are low since the pagerank of the inlinks are high it gives a higher pagerank.