Homework 4 Report

Introduction

The objective of this assignment was to delve into the intricacies of ranking algorithms, such as PageRank and HITS. These algorithms play a crucial role in search engines helping rank the value, relevance and importance of web pages so that the engine can present the most optimal and relevant search results.

PageRank, initially designed by Larry Page and Sergey Brin, operates on the principle that not all pages are equal and that some pages, due to the number and quality of links to it from other pages, carry more weight than others. Using the analogy of citation analysis, it views links as ‘votes’, where a page connected to by highly ranked pages is considered to be important. The PageRank of a page is computed iteratively using the ranks of its voters and a damping factor to model the likelihood of clicking on links versus landing on a page randomly. This approach ensures that PageRank not only considers the number of inlinks as a metric but also evaluates the quality of those links.

HITS, on the other hand, introduces the concepts of 'hubs' and 'authorities'. Authorities are pages that offer valuable information on a specific topic while hubs are pages that link to various authority pages, acting as a guide to online content. The HITS algorithm works by giving each page an authority and a hub score and then iteratively adjusting these scores based on the structure of the web graph. A high authority score suggests a page with valuable information on a particular subject whereas a high hub score signifies a page that serves as an excellent directory to high-quality content.

Methodology

In this assignment, we utilised datasets such as the 'wt2g_inlinks’ as well as our self made group merged indexes. Tools used included ES for index operations (link graph, search, etc.), python packages numpy, random, and pickle for help implementing, matrix algebra for computation, etc. The process included first extracting the link graphs, both from the merged ES index and the WT2G text file, then initialising the base values for pagerank and hub/authority scores, and finally iteratively computing PR, hub and authority scores.

Analysis

PageRank Analysis

PageRank on WT2G_Inlinks Data

```
URL: WT21-B37-76, PageRank: 0.002708250818794516, Inlinks: 2568, Outlinks: 5
URL: WT21-B37-75, PageRank: 0.00154074154018171 3, Inlinks: 1704, Outlinks: 1
URL: WT25-B39-116, PageRank: 0.0014619660419097143, Inlinks: 169, Outlinks: 1
URL: WT23-B21-53, PageRank: 0.0013654331445348085, Inlinks: 198, Outlinks: 1
URL: WT23-B39-340, PageRank: 0.0012620236297305218, Inlinks: 274, Outlinks: 395
URL: WT24-B40-171, PageRank: 0.0012599558592240028, Inlinks: 270, Outlinks: 209
URL: WT24-B26-10, PageRank: 0.0012490363797455478, Inlinks: 291, Outlinks: 1
URL: WT23-B37-134, PageRank: 0.0012002110974870083, Inlinks: 207, Outlinks: 2
URL: WT08-B18-400, PageRank: 0.0011467540543661802, Inlinks: 990, Outlinks: 0
URL: WT13-B06-284, PageRank: 0.001137035091226585, Inlinks: 454, Outlinks: 2
URL: WT13-B06-273, PageRank: 0.0010553785387760605, Inlinks: 452, Outlinks: 11
URL: WT01-B18-225, PageRank: 0.0009574415958034686, Inlinks: 1137, Outlinks: 0
URL: WT04-B27-720, PageRank: 0.0009440099419502785, Inlinks: 291, Outlinks: 27
URL: WT24-B26-46, PageRank: 0.0008620976667652068, Inlinks: 179, Outlinks: 3
URL: WT23-B19-156, PageRank: 0.0008225724115733233, Inlinks: 364, Outlinks: 12
URL: WT04-B30-12, PageRank: 0.0008171951417106377, Inlinks: 241, Outlinks: 8
URL: WT25-B15-307, PageRank: 0.000798535543328242, Inlinks: 605, Outlinks: 8
URL: WT07-B18-256, PageRank: 0.000786703164160688, Inlinks: 169, Outlinks: 169
URL: WT24-B40-167, PageRank: 0.0007164695678037854, Inlinks: 153, Outlinks: 152
URL: WT14-B03-220, PageRank: 0.0007102162441563615, Inlinks: 163, Outlinks: 162
```

- Although the first two entries both have a high inlink count, looking further down the list we can see some pages with low inlink counts ranking higher than pages with higher inlink counts. This is because the PageRank of a page is not solely based on the number of inlinks but also on the quality/importance of those inlinks. Therefore, pages with fewer inlinks but from high-quality pages can rank higher than pages with lots of inlinks but only from low-quality pages.

PageRank on Merged Data

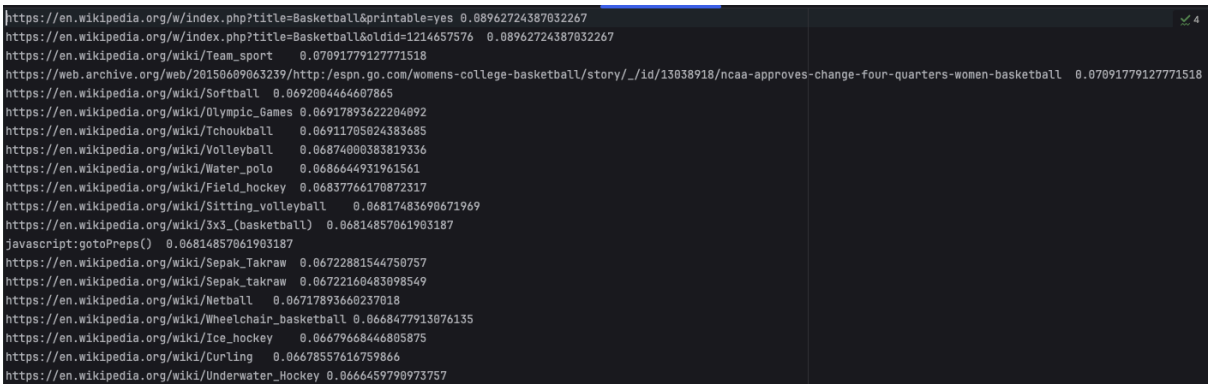
```
URL: http://www.theguardian.com/football/womensfootball, PageRank: 0.009410507609609832, Inlinks: 30, Outlinks: 2
URL: http://www.theguardian.com/soccer, PageRank: 0.007561241608302206, Inlinks: 174, Outlinks: 1
URL: http://www.theguardian.com/football, PageRank: 0.005490008045610901, Inlinks: 131, Outlinks: 3
URL: https://www.latinbasket.com/Argentina/basketball.aspx, PageRank: 0.004845345637236852, Inlinks: 7186, Outlinks: 5
URL: https://www.latinbasket.com/Aruba/basketball.aspx, PageRank: 0.004845345637236852, Inlinks: 7186, Outlinks: 7
URL: https://www.mediawiki.org/, PageRank: 0.004197722021720867, Inlinks: 7185, Outlinks: 6
URL: https://wikimediafoundation.org/, PageRank: 0.00393476389837136, Inlinks: 7185, Outlinks: 4
URL: http://www.bbc.co.uk/sport/football, PageRank: 0.0038816879098204637, Inlinks: 77, Outlinks: 6
URL: https://www.latinbasket.com/Antigua/basketball.aspx, PageRank: 0.0037237360617237, Inlinks: 7181, Outlinks: 1
URL: http://en.wikipedia.org/wiki/Portal:Association_football, PageRank: 0.002973522018171433, Inlinks: 762, Outlinks: 29
URL: https://en.wikipedia.org/wiki/Help:Contents, PageRank: 0.0026570993820024475, Inlinks: 5563, Outlinks: 31
URL: https://en.wikipedia.org/wiki/Main_Page, PageRank: 0.0025818046392375974, Inlinks: 6003, Outlinks: 27
URL: http://commons.wikimedia.org/wiki/Category:The_Football_Association, PageRank: 0.0024036287466172117, Inlinks: 63, Outlinks: 55
URL: http://www.bbc.co.uk/sport/american-football, PageRank: 0.0023227923065776658, Inlinks: 76, Outlinks: 1
URL: http://commons.wikimedia.org/wiki/Category:Football, PageRank: 0.002082261418125679, Inlinks: 56, Outlinks: 50
URL: http://www.bbc.com/sport/football, PageRank: 0.002073374720479579, Inlinks: 35, Outlinks: 1
URL: http://en.wikipedia.org/wiki/Women%27s_association_football, PageRank: 0.0020713746955277196, Inlinks: 539, Outlinks: 108
URL: http://www.bbc.com/sport/american-football, PageRank: 0.0020695503727551252, Inlinks: 34, Outlinks: 1
URL: http://en.wikipedia.org/wiki/Category:Football_at_multi-sport_events, PageRank: 0.0019893999083947164, Inlinks: 57, Outlinks: 49
URL: http://www.theguardian.com/football/football, PageRank: 0.0019455316079694375, Inlinks: 69, Outlinks: 3
```

- The actual page rank values for the merged data are larger
- The WT2G_Inlinks data has more inlinks and outlinks compared to our merged data

- Some pages, despite having significantly fewer inlinks compared to other pages, still rank highly. This might indicate a strong endorsement from highly authoritative pages.

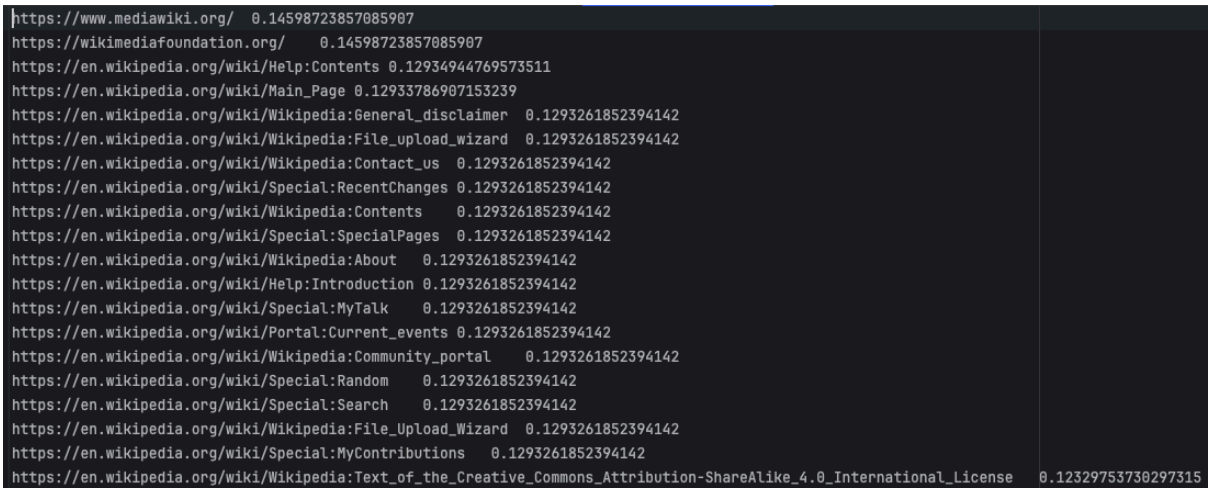
HITS Analysis

Hub Scores



- The relationship between Hub scores and the structure of the web graph revealed that pages serving as guides to authoritative content on specific topics are identified as valuable hubs.
- The top hubs are mostly wikipedia and showcase the algorithm's ability to identify pages that serve as central points for directing traffic to more authoritative pages.

Authority Scores



- Top authorities, like https://www.mediawiki.org/ and https://wikimediafoundation.org/, show the impact of inbound links from reputable hubs.
- The authority scores mirror the hub scores to some extent, emphasizing the reciprocal nature of web relationships where quality content attracts links from prominent hubs, thereby increasing the authority score.
- Both PageRank and HITS highlight the importance of link quality over sole link count. High PageRank and authority scores for pages with fewer but more authoritative inlinks suggest that the algorithms favor quality “votes”.

Case Study: PageRank vs. Inlink Count

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

- http://www.theguardian.com/football/womensfootball was the highest ranked page from the merged index although it only had 30 inlinks. Following this trail we can see that, the second and third highest ranked pages are http://www.theguardian.com/soccer and http://www.theguardian.com/football both of which link to http://www.theguardian.com/football/womensfootball and have high inlink counts. This showcases how even with few inlinks a page can rank highly if the pages that do link to it are valuable and important. Valuing quality over merely quantity.