HomeWork-4 Report

Karthik Chintamani Dileep

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.

PageRank is an algorithm that determines the importance of a web page based on the number and quality of pages linking to it. It operates on the principle that more important pages tend to receive more links from other important pages. By modeling the web as a graph and performing a random walk on this graph, PageRank can estimate the probability of a random web surfer visiting a particular page, which is used as a proxy for the page's importance.

HITS (Hyperlink-Induced Topic Search) is an algorithm that specifically focuses on identifying relevant and authoritative pages for a given topic or query. It distinguishes between two types of pages: "authorities" (pages with valuable content on the topic) and "hubs" (pages that link to many relevant authorities). HITS operates on the idea that good authorities are linked to by many good hubs, and good hubs link to many good authorities. Through an iterative process, HITS calculates hub and authority scores for pages, allowing it to identify the most relevant and authoritative pages for a particular topic.

Both algorithms assess the importance and relevance of web pages by analyzing the link structure of the web. PageRank uses the global link structure to estimate a page's overall importance, while HITS uses the local link structure around a topic to identify the most relevant and authoritative pages for that topic.

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.

Datasets:

- **climate-change** – crawled index in Elasticsearch.
- **wt2g_inlinks.txt**

Tools:

- **Elasticsearch Cloud**
- **Python**

Calculation of PageRank scores:

- All pages, inlinks, outlinks, and sink pages (pages with no outlinks) are retrieved from the Elasticsearch index and/or wt2g_inlinks dataset
- The PageRank scores are computed using the iterative method, considering the damping factor, sink pages, and the link structure (inlinks and outlinks).
- The top 500 pages are sorted by their PageRank scores and written to a file, along with their outlink and inlink counts.

Calculation of HITS (Hub and Authority) scores:

1. A root set of approximately 1000 pages is created by querying the climate change index in Elasticsearch using the "climate change" query and ranking the pages.
2. The root set is expanded by adding pages that the current set points to (outlinks) and pages that point to the current set (inlinks). The latter set is limited to a maximum of 200 random pages to control the size of the root set.
3. The expanded set (base set) is used to compute the HITS scores:
    - The authority and hub scores are initialized to 1 for all pages in the base set.
    - In each iteration, the authority score is updated by summing the hub scores of pages pointing to it, and the hub score is updated by summing the authority scores of pages it points to.
    - The scores are normalized after each iteration.
    - The process continues until convergence or a maximum number of iterations is reached.
4. The top 500 authority and hub pages are written to separate files, along with their respective scores.

Analysis

PageRank Analysis

PageRank on WT2G_Inlinks Data

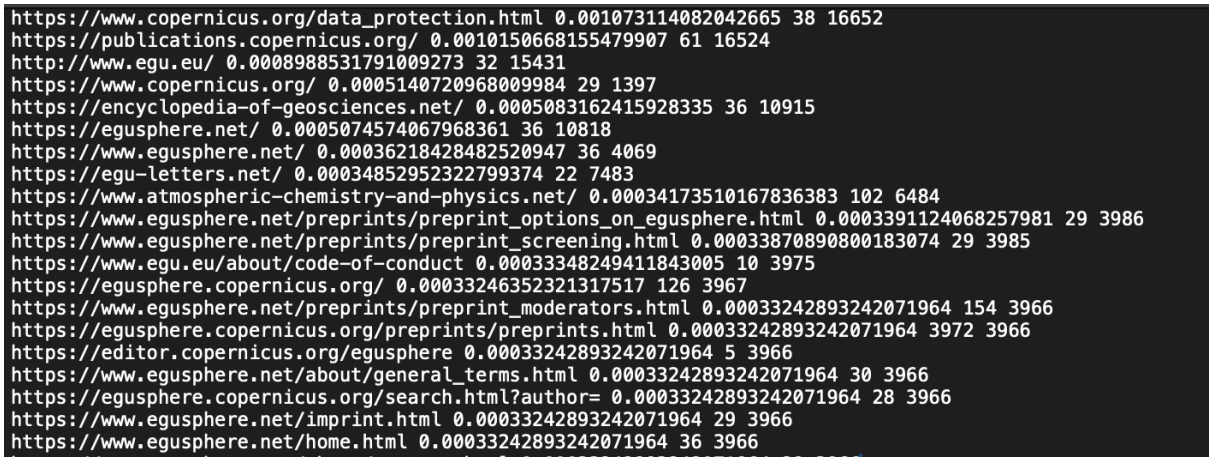
- Include a screenshot of the first 20 rows.

```
WT21-B37-76 0.002694470870551518 5 2568
WT21-B37-75 0.0015331771248564394 1 1704
WT25-B39-116 0.0014685088124070647 1 169
WT23-B21-53 0.0013735336119046905 1 198
WT24-B26-10 0.001276215242790007 1 291
WT24-B40-171 0.0012452589803631758 209 270
WT23-B39-340 0.0012428610383833519 395 274
WT23-B37-134 0.0012054274132022166 2 207
WT08-B18-400 0.0011447764346453053 0 990
WT13-B06-284 0.0011365503403142805 2 454
WT13-B06-273 0.0010549175456506178 11 452
WT01-B18-225 0.0009553812179194965 0 1137
WT04-B27-720 0.0009409558792930838 27 291
WT24-B26-46 0.0008622309394783624 3 179
WT23-B19-156 0.000825066916738247 12 364
WT04-B30-12 0.0008166209010213502 8 241
WT25-B15-307 0.0007972229995179634 8 605
WT07-B18-256 0.0007750022279615515 169 169
WT24-B40-167 0.0007076055482559692 152 153
WT14-B03-220 0.0006988599396070423 162 163
```

- Discuss any notable patterns or anomalies observed in the PageRank scores compared to inlink counts.
    - There are several pages with relatively high PageRank scores but low inlink counts. There are pages with relatively low PageRank scores but high inlink counts. There seems to be a general positive correlation between PageRank scores and inlink counts.

PageRank on Merged Data

- Include a screenshot of the first 20 rows.



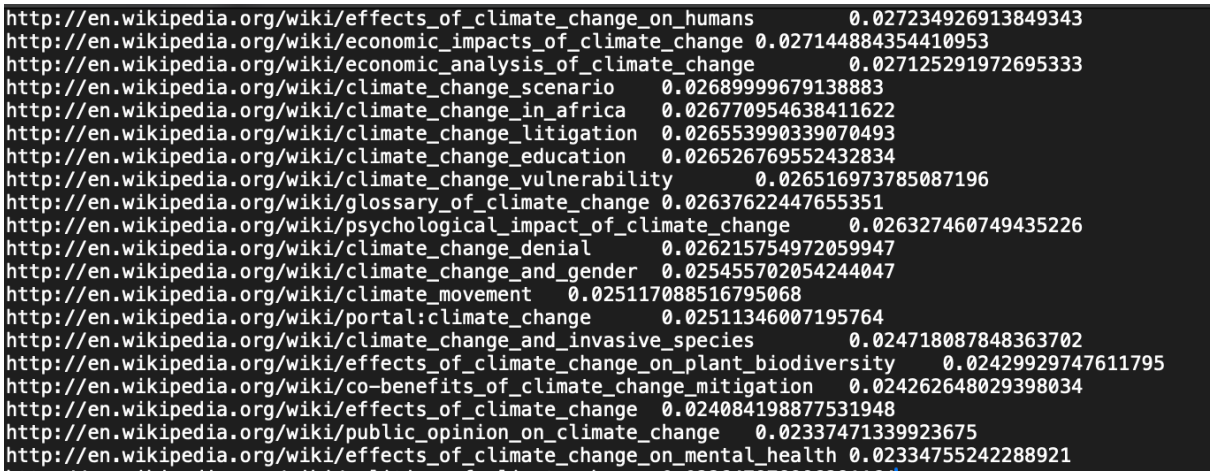
```
https://www.copernicus.org/data_protection.html 0.001073114082042665 38 16652
https://publications.copernicus.org/ 0.0010150668155479907 61 16524
http://www.egu.eu/ 0.0008988531791009273 32 15431
https://www.copernicus.org/ 0.0005140720968009984 29 1397
https://encyclopedia-of-geosciences.net/ 0.0005083162415928335 36 10915
https://egusphere.net/ 0.0005074574067968361 36 10818
https://www.egusphere.net/ 0.00036218428482520947 36 4069
https://egu-letters.net/ 0.00034852952322799374 22 7483
https://www.atmospheric-chemistry-and-physics.net/ 0.00034173510167836383 102 6484
https://www.egusphere.net/preprints/preprint_options_on_egusphere.html 0.0003391124068257981 29 3986
https://www.egusphere.net/preprints/preprint_screening.html 0.0003387089080018307 4 29 3985
https://www.egu.eu/about/code-of-conduct 0.00033348249411843005 10 3975
https://egusphere.copernicus.org/ 0.0003324635232131751 7 126 3967
https://www.egusphere.net/preprints/preprint_moderators.html 0.00033242893242071964 154 3966
https://egusphere.copernicus.org/preprints/preprints.html 0.00033242893242071964 3972 3966
https://editor.copernicus.org/egusphere 0.00033242893242071964 5 3966
https://www.egusphere.net/about/general_terms.html 0.00033242893242071964 30 3966
https://egusphere.copernicus.org/search.html?author= 0.00033242893242071964 28 3966
https://www.egusphere.net/imprint.html 0.00033242893242071964 29 3966
https://www.egusphere.net/home.html 0.00033242893242071964 36 3966
```

- Highlight any differences observed when comparing the PageRank scores from the WT2G_Inlinks data to the merged data set.
    - Top PageRank scores more clustered together in the merged data set.
    - Higher concentration of pages with relatively high inlink counts in the merged data set.
    - Differences likely due to the topic-specific nature and targeted crawling process of the merged data set.

HITS Analysis

Hub Scores

- Include a screenshot of the first 20 rows.

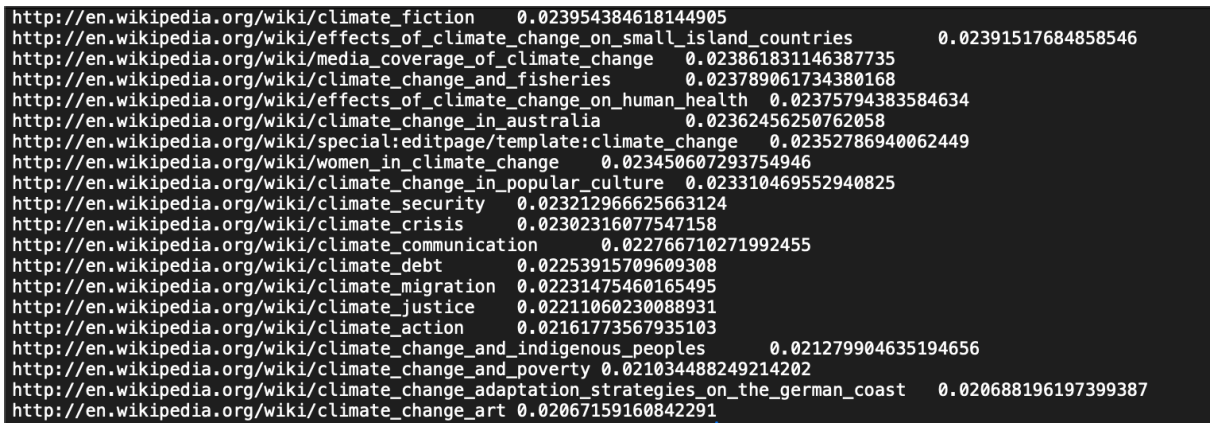


```
http://en.wikipedia.org/wiki/effects_of_climate_change_on_humans        0.027234926913849343
http://en.wikipedia.org/wiki/economic_impacts_of_climate_change 0.027144884354410953
http://en.wikipedia.org/wiki/economic_analysis_of_climate_change        0.027125291972695333
http://en.wikipedia.org/wiki/climate_change_scenario    0.02689999679138883
http://en.wikipedia.org/wiki/climate_change_in_africa   0.026770954638411622
http://en.wikipedia.org/wiki/climate_change_litigation  0.026553990339070493
http://en.wikipedia.org/wiki/climate_change_education   0.026526769552432834
http://en.wikipedia.org/wiki/climate_change_vulnerability       0.026516973785087196
http://en.wikipedia.org/wiki/glossary_of_climate_change 0.02637622447655351
http://en.wikipedia.org/wiki/psychological_impact_of_climate_change     0.026327460749435226
http://en.wikipedia.org/wiki/climate_change_denial      0.026215754972059947
http://en.wikipedia.org/wiki/climate_change_and_gender  0.025455702054244047
http://en.wikipedia.org/wiki/climate_movement   0.025117088516795068
http://en.wikipedia.org/wiki/portal:climate_change      0.02511346007195764
http://en.wikipedia.org/wiki/climate_change_and_invasive_species        0.024718087848363702
http://en.wikipedia.org/wiki/effects_of_climate_change_on_plant_biodiversity    0.02429929747611795
http://en.wikipedia.org/wiki/co-benefits_of_climate_change_mitigation   0.024262648029398034
http://en.wikipedia.org/wiki/effects_of_climate_change  0.024084198877531948
http://en.wikipedia.org/wiki/public_opinion_on_climate_change   0.02337471339923675
http://en.wikipedia.org/wiki/effects_of_climate_change_on_mental_health 0.02334755242288921
```

- Provide insights into the relationship between Hub scores and the structure of the web graph.
    - High Hub scores indicate pages with many outgoing links, acting as central nodes.
    - Hub pages often represent broad topics or categories within a domain.
    - Distribution of links from Hubs shapes the connectivity and information flow.
    - Influential Hubs can serve as entry points for exploring a topic or domain.
    - Clusters of high-scoring Hubs may indicate sub-communities or related topics.

Authority Scores

- Include a screenshot of the first 20 rows.



```
http://en.wikipedia.org/wiki/climate_fiction    0.023954384618144905
http://en.wikipedia.org/wiki/effects_of_climate_change_on_small_island_countries        0.02391517684858546
http://en.wikipedia.org/wiki/media_coverage_of_climate_change   0.023861831146387735
http://en.wikipedia.org/wiki/climate_change_and_fisheries       0.023789061734380168
http://en.wikipedia.org/wiki/effects_of_climate_change_on_human_health  0.02375794383584634
http://en.wikipedia.org/wiki/climate_change_in_australia        0.02362456250762058
http://en.wikipedia.org/wiki/special:editpage/template:climate_change   0.02352786940062449
http://en.wikipedia.org/wiki/women_in_climate_change    0.023450607293754946
http://en.wikipedia.org/wiki/climate_change_in_popular_culture  0.023310469552940825
http://en.wikipedia.org/wiki/climate_security   0.023212966625663124
http://en.wikipedia.org/wiki/climate_crisis     0.02302316077547158
http://en.wikipedia.org/wiki/climate_communication      0.022766710271992455
http://en.wikipedia.org/wiki/climate_debt       0.02253915709609308
http://en.wikipedia.org/wiki/climate_migration  0.02231475460165495
http://en.wikipedia.org/wiki/climate_justice    0.02211060230088931
http://en.wikipedia.org/wiki/climate_action     0.02161773567935103
http://en.wikipedia.org/wiki/climate_change_and_indigenous_peoples      0.021279904635194656
http://en.wikipedia.org/wiki/climate_change_and_poverty 0.021034488249214202
http://en.wikipedia.org/wiki/climate_change_adaptation_strategies_on_the_german_coast   0.020688196197399387
http://en.wikipedia.org/wiki/climate_change_art 0.02067159160842291
```

- Discuss how Authority scores compare with PageRank and Hub scores, and what this implies about the web pages' importance or relevance.
    - Authority scores measure the content quality and relevance of a page for a given topic.
    - High Authority scores indicate authoritative and valuable content on the topic.
    - PageRank measures overall importance based on the link structure of the web.
    - Hub scores measure how well a page acts as a hub, linking to many relevant pages.
    - Pages with high Authority but low PageRank/Hub scores may have high-quality content but lack visibility or connectivity.
    - Pages with high PageRank/Hub scores but low Authority may be well-connected but lack authoritative content.
    - Ideally, relevant and authoritative pages should have high scores across all three metrics.
    - Authority scores highlight content expertise, while PageRank and Hub scores reflect structural importance.

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.
- Explain why these pages have a higher PageRank despite a smaller inlink count, focusing on the quality or authority of the incoming links.

https://www.nasa.gov/
PageRank: 0.00015827491882294977
Inlinks: 206
Outlinks: 1635

While this page has a relatively high number of inlinks (206), its very high PageRank of 0.00015827 suggests that many of those 206 inlinks are from highly authoritative and relevant pages themselves. NASA.gov, being a major scientific organization's website, likely receives many inbound links from other reputable .edu, .gov, and scientific domains, which could increase the passing of PageRank value to nasa.gov.

https://www.giss.nasa.gov/tools/panoply
PageRank: 9.554243539559107e-05

Inlinks: 56
Outlinks: 1161

This page has only 56 inlinks, which is relatively low, but has a decent PageRank of 9.554e-05. Looking at the inlinks, many are from other nasa.gov and giss.nasa.gov pages, which are likely high-authority and topically relevant domains. So while the raw inlink count is low, the quality and relevance of those inbound links from other NASA pages allows this page to accumulate a higher PageRank.

https://environment.yale.edu/centers/ycec
PageRank: 4.772446338288567e-05
Inlinks: 48

Outlinks: 607

With only 48 inlinks, this page's relatively high PageRank of 4.772e-05 may be influenced by inbound links from other authoritative yale.edu URLs, as well as potentially links from other reputable .edu domains related to environmental studies. The topic authority and academic standing of Yale could give weight to the inlinks.

So in summary, while these pages have lower inlink counts, their higher PageRank scores can likely be attributed to high-quality, relevant, and authoritative inbound links from sources like major scientific organizations, academic institutions, and topically focused domains related to the page's content. The PageRank algorithm places emphasis on such authoritative links over purely counting link volumes.