# Project 1 Page Ranks and analysis

**Qn 1**:

Page Rank values for six vertices after 1, 10, 100 iterations:

|  |  |  |
| --- | --- | --- |
| **Iteration 1**  A 0.24975  B 0.11963  C 0.14329  D 0.13146  E 0.21427  F 0.14329 | **Iteration 10**  A 0.25218  B 0.13948  C 0.15145  D 0.11889  E 0.18702  F 0.15145 | **Iteration 100**  A 0.25213  B 0.13931  C 0.15131  D 0.11891  E 0.18705  F 0.15131 |

**Pls find the code for the first question in Qn1.py**

Qn 2:

Perplexity Values with Shannon entropy

|  |  |  |
| --- | --- | --- |
| 183810.999998  79669.9231957  86267.6741024  72260.3536067  75132.4076593  68932.6029131  71197.8334108  67782.5377845  69379.5774141  67383.7075589  68477.8018834  67207.1847963  68004.1538837  67138.9553795  67708.2593908  67131.6639346  67524.4769137  67132.1110911  67413.7101219 | 67138.8498145  67339.825439  67149.7850062  67290.830658  67158.762079  67259.2257454  67166.0293807  67237.7802261  67172.3205073  67223.127439  67177.1443729  67213.3194519  67180.7511381  67206.5975774  67183.5491211  67201.9331033  67185.6323888  67198.7420931  67187.1593955 | 67196.5257185  67188.2983336  67194.9793761  67189.1321605  67193.9036297  67189.740378  67193.1507383  67190.1847357  67192.621374  67190.5078833  67192.250355  67190.7423803  67191.9887961  67190.912902  67191.804427  67191.036087  67191.6743361  67191.1253299 |

**Pls find the code for the second question in Qn2.py**

Qn. 3:

The Document IDs of the top 50 pages sorted by PageRank and their PageRank

**Top 50 pages by page rank:**

|  |  |
| --- | --- |
| WT21-B37-76 0.00267941005759  WT21-B37-75 0.00152591699385  WT25-B39-116 0.00146949352594  WT23-B21-53 0.00137232204363  WT24-B40-171 0.0012450717567  WT23-B39-340 0.00124048970549  WT23-B37-134 0.00120521440914  WT08-B18-400 0.0011435408633  WT13-B06-284 0.00112478163921  WT24-B26-46 0.00108504628436  WT13-B06-273 0.00104470114963  WT01-B18-225 0.000988443749006  WT04-B27-720 0.000936408657488  WT23-B19-156 0.000894231205525  WT04-B30-12 0.000816440873075  WT24-B26-10 0.000807421967764  WT25-B15-307 0.000804382292729  WT07-B18-256 0.000774948428349  WT24-B26-2 0.00077134157658  WT14-B03-220 0.000716442619324  WT24-B40-167 0.00070750883028  WT14-B03-227 0.00068502362319  WT18-B31-240 0.000660190199768  WT04-B40-202 0.000658703995483  WT08-B19-222 0.000643432387733 | WT27-B28-203 0.000627001437232  WT13-B15-160 0.0006212965578  WT13-B39-295 0.000616960634898  WT12-B30-56 0.000602289689649  WT10-B02-288 0.000575993082918  WT22-B38-403 0.00057457221106  WT14-B36-337 0.000558256735308  WT27-B34-57 0.000555500110617  WT23-B20-363 0.000550862343293  WT23-B01-40 0.000550421871083  WT27-B32-30 0.000549770818687  WT21-B40-37 0.000548180008165  WT21-B35-155 0.000540052161788  WT08-B08-60 0.000535637027043  WT04-B22-268 0.000532768123456  WT14-B02-400 0.000532566194076  WT18-B14-66 0.00053205951422  WT23-B27-31 0.000525616292471  WT23-B38-120 0.000520927838604  WT06-B35-151 0.000520104124682  WT06-B14-69 0.00051906658947  WT06-B35-161 0.000518209396918  WT10-B33-300 0.00051680665582  WT14-B36-336 0.000515459986569  WT14-B36-335 0.000515459986569 |

The Document IDs of the top 50 pages sorted by in-link count and their in-link counts

**Top 50 pages by in-link count and the in-link count:**

|  |  |  |
| --- | --- | --- |
| WT21-B37-76 2568  WT18-B29-37 2269  WT01-B18-225 2260  WT23-B27-29 1940  WT21-B37-75 1704  WT27-B34-57 1257  WT27-B32-30 1255  WT08-B19-222 1041  WT08-B18-400 1011  WT10-B36-88 946  WT10-B36-90 943  WT10-B36-103 939  WT10-B36-89 896  WT21-B40-447 779  WT18-B28-345 728  WT12-B40-248 686  WT24-B26-2 625 | WT25-B15-307 614  WT27-B28-203 598  WT18-B40-82 576  WT21-B37-71 560  WT22-B38-403 544  WT08-B01-173 539  WT13-B15-160 484  WT23-B30-88 478  WT18-B29-36 477  WT27-B28-177 470  WT13-B06-284 454  WT13-B06-273 454  WT07-B02-55 449  WT13-B39-295 443  WT17-B34-499 442  WT17-B34-500 436  WT24-B04-192 430 | WT14-B36-337 417  WT17-B34-505 410  WT10-B33-300 409  WT23-B19-156 406  WT23-B31-215 402  WT17-B34-503 402  WT23-B23-51 400  WT08-B11-28 396  WT23-B12-215 388  WT23-B01-107 384  WT23-B30-105 380  WT17-B34-506 376  WT17-B34-504 374  WT17-B34-498 374  WT14-B36-323 373  WT07-B23-234 371 |

Qn. 4.

Top 10 pages

|  |  |
| --- | --- |
| **WT21-B37-76**  **WT21-B37-75**  **WT25-B39-116**  **WT23-B21-53**  **WT24-B40-171**  **WT23-B39-340**  **WT23-B37-134**  **WT08-B18-400**  **WT13-B06-284**  **WT24-B26-46** | **0.00267941005759**  **0.00152591699385**  **0.00146949352594**  **0.00137232204363**  **0.0012450717567**  **0.00124048970549**  **0.00120521440914**  **0.0011435408633**  **0.00112478163921**  **0.00108504628436** |

Figure 1

|  |  |  |  |
| --- | --- | --- | --- |
| **Pages** | **Total No. of inlinks** | **Ranking** | **Description** |
| WT21-B37-76 | 2568 | 1 | The Economist Homepage |
| WT21-B37-75 | 1704 | 5 | The Economist copyright notice |
| WT25-B39-116 | 169 | 280 | Security Assurance Requirement. |
| WT23-B21-53 | 198 | 178 | Infor. about the web development team |
| WT24-B40-171 | 270 | 106 | Evening news on every week night. |
| WT23-B39-340 | 274 | 102 | Financial Information of the companies |
| WT23-B37-134 | 208 | 162 | Health :Disclaimer on Material |
| WT08-B18-400 | 1011 | 9 | General Disclaimer |
| WT13-B06-284 | 454 | 29 | WebMaster Information. |
| WT24-B26-46 | 187 | 226 | Milton's HomePage, Department of Psychiatry |

Figure 2

**Brief Analysis:**

**WT21-B37-76:** has the highest page rank, as indicated by the table above, that is **0.00267941005759** and it is ‘The Economist Home Page’. There is one sink node, **WT08-B18-400 - 1011** (General Disclaimer, almost every page has a disclaimer and hence most of the inlinks are present for this page). Pages from **WT25-B39-116** to **WT24-B26-46** indicate that inlinks are not the sole factor on which the page rank depends. This can be seen from the fact that pages with higher inlinks are ranked lower than the once having a higher inlinkFrom above Figure 2, let me present my analysis on how each of the documents fare on being interesting,

**WT21-B37-76:** has the highest page rank, as indicated by the table above, that is **0.00267941005759** and it is ‘The Economist Home Page’. Counting the number of inlinks, I find that it has the highest number of inlinks. Every inlink to this page in turn also have inlinks. Interpreted in a different way, there is no page connecting to this page which is a sink node. It is particularly interesting as the search for recent events would hopefully lead to this page being in the list of returned queries. In addition, users would like to know the changes in economic policies and how that affects their daily lives. Hence this could be the reason for its popularity.

**WT21-B37-75, WT25-B39-116:** These pages are probably not something that hopefully would turn up in the list of returned queries as these are copyright notice, the security assurance requirement document

**WT23-B21-53**: provides information regarding the webmaster could be possibly linked by various documents. Users could possibly require that information for contact/complaints/comments purpose. Hence, even though it does not serve much purpose,

it has a higher PageRank. It is interesting to notice that all the inlinks are from the documents with document IDs commencing with WT23. Each and every document would have copyright information which might contain the link to the web development team’s page. Hence, this page even though not imparting much information like news, economy etc does have a higher page rank

**WT24-B40-171:** is the evening news on every weeknight. News pages normally have various links in-order to keep people updated with the most recent happenings. They are not only utilized to keep track of the most recent happenings but also to supplement present information. Hence, there would be many articles/blogs/web pages which link to this page. People require news and hence news is a very common query which makes this information interesting and important.

**WT23-B39-340:** is financial Information of the companies. This information should be important to many articles and also provide research material for them. Hence, many of the documents would be pointing to this page. People would like to keep an eye on companies that are sound and stable to invest or buy stocks. Hence this link is also of much importance

**WT23-B37-134:** is a disclaimer for all the materials provided by the Health Department of WA. All the documents having Document IDs starting with WT23 have a link to this page. As a general practice, articles and other academic material have a disclaimer. Hence, most of the pages would have a link to this disclaimer which adds up to the number of inlinks for the following page.

**WT08-B18-400:** is a **Sink Node** and a General Disclaimer is not as important like news or financial news. Yet it has a high pagerank and surprisingly a high number of inlinks. This can be attributed to the fact that most of the articles need a general disclaimer. Almost all the articles in the given set of documents have a link to this document. Given that this is a disclaimer page one does not expect this to link to another page and as expected there is no link out from this document.

**WT13-B06-284:** is the document containing the information of Director of web

development. There are 454 inlinks to the pages, all from the documents whose document

IDs are beginning with WT13. This could be because users may require the contact information for complaints/ comments/business offers.

**WT24-B26-46:** is a homepage of Dr. Milton who is a professor in the Department of Psychiatry at University of Michigan. All the documents, with the IDs starting with WT24 appear to belong to documents pertaining to Department of Psychiatry (for eg WT13-B06-284 contains the course information etc). Most of these pages have a link to Dr. Milton’s homepage.

According to multiple definitions of PageRank on the web:

*PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites. (wikipedia:PageRank)*

*PageRank is a probability distribution used to represent the likelihood that a person randomly*

*clicking on links will arrive at any particular page.(wikipedia:PageRank)*

In the case, of the second document (WT21-B37-75) and the third document (WT25-B39-116) the above definition of PageRank might be true. The second document has a high number of inlinks (1704) and ranks fifth when sorted according to the number of inlinks. Although, for the third document (WT25-B39-116) the number of inlinks is not that high (169, ranked at 280 when sorted according to the number of inlinks). This data (inlinks) can be seen in two ways, firstly by the number of inlinks within these ten documents i.e which document has an inlink from any other document amongst these ten documents.

*When one page links to another page, it is effectively increasing the importance for the other page. The more links are created for a page, the more important the page must be. Also, the*

*importance of the page that is linking to another page determines how important the link itself is. (discussed in class)*

**WT21-B37-76: WT21-B37-75, WT21-B37-75: WT21-B37-76**

As can be seen, the above answers why the second document is ranked higher. It has an inlink

from the first document which has the highest number of inlinks. This increases the likelihood

of a person arriving at this document.

To analyze why the third document has a higher PageRank, I looked at what links are linking to the third document. Are these links popular? I found the following:

WT25-B39-116 has two links that have high counts compared to other links. **WT25-B38-117:5 and WT25-B39-116:169.** There is one self link and there is only one document which has five inlinks (WT25-B38-117)

This link is a beginning of a book, more precisely, the first part titled ‘Introduction and General Model’. This seems an interesting page to appear in the results of a person’s search. Although the number of inlinks are not as high enough, which means the five documents linking to this document should have more inlinks. All the documents with their documents IDs deal with the subject matter of security assurance and there is a lot of research regarding this subject. Most of the materials keep reiterating to the beginning of the book, which has contents listed. These contents might contain a link to the subject requirement.