

Page Rank

Language: Python

Version: 3.5

Tool: PyCharm

Libraries:

1. Math
2. Datetime

Step:

1. Name the inlinks file from WT2g as “inlinks.txt”
2. Store the file in same directory as the PageRank.py file
3. Run the PageRank.py file

Note: The file displays the requested output and writes the output page rank values for each document in file “Output.txt” which is generated inside the same directory. The writing of the file might take time.

Q: Speculate why these documents have high PageRank values, i.e., why is it that these particular pages are linked to by (possibly) many other pages with (possibly) high PageRank values. Are all of these documents ones that users would likely want to see in response to an appropriate query? Give some examples of ones that are and ones that are not. For those that are not "interesting" documents, why might they have high PageRank values? How do the pages with high PageRank compare to the pages with many in-links? In short, give an analysis of the PageRank results you obtain.

Analysis:

- A high page rank value depends on a number of factors like the number of links pointing to a particular page, the relevant content of page, the update frequency of page etc.
- A user would like to see a page which has most recent contents with more relevant information.
- The high page rank document like for example WT21-B37-76 and WT21-B37-75 has also the highest inlink count. The pages have more pages pointing to these documents but the contents on page WT21-B37-76 looks more relevant and desired by user than WT21-B37-75.
- There are several links on page WT21-B37-76 like copyright link pointing to WT21-B37-75 which are irrelevant to the user but still has high page rank and inlink count. The reason could be unnecessary clicking of the copyright from page WT21-B37-76 increasing the page rank value for WT21-B37-75.
- Also the last modified date of WT21-B37-76 is recent than WT21-B37-75. Hence it has more recent content.
- A user will be more interested in checking the economist page rather than the copyright page which has less relevant or desired information.
- Page WT21-B37-75 has high inlink and page rank value but the information is less relevant as it is just a copyright page.

Conclusion: From above analysis it could be concluded that although a page has higher page rank and inlink count, it is not necessary that the information on the page will always be relevant to the user. A page containing a higher count of inlinks but quality inlinks i.e. pointing to other relevant pages is more interesting to user than a page with higher page rank and inlink count with less relevant contents.