IR Project 1

Run the script PageRank.py

The scripts reads two files:

1. **LinkGraph**
   This file contains the graph:
   A D E F
   B A F
   C A B D
   D B C
   E B C D F
   F A B D

2. **wt2g_inlinks.txt**

Running the script will generate the following files in the same folder:

1. **PageRankSixNodes**
   This file contains: PageRank values for each of the six vertices from the LinkGraph file after 1, 10, and 100 iterations of the PageRank algorithm.
2. **PageRankWT2g**
   This file contains: Top 50 pages with high page ranks
3. **perplexity_values**
   This file contains the list of perplexity value on each iteration until the perplexity value has converged.
   Convergence: Getting same perplexity value for the unit's place for 4 iterations
4. **Top_inlinks_count**
   This file contains: Top 50 pages with high in link count.

Analysis on Top 10 page rank and in links for each page with high page rank and high in links:

For analysis of page rank and in link counts, two separate files have been composed:

analysis.sh: will prompt for a document id and check if that document has been in linked by another document id whose page rank is in the Top 50. You can quit the program by typing in "quit"

Ananlysis.py: For a page, displays the in link count, out link count, sink node or not, the in links to the page, in link and out link count for each in link and whether the page has out link to itself or not. Also, I am using the save state of in link, out link, sink node and pages to avoid computation for each test I execute.

a. **WT21-B37-76** has the highest page rank: 0.00267941 and the highest number of in links: 2568.
   This page should be important because it has the highest number of in links and we view links as information about the popularity of a web page.
   Further analysis:
   Applying analysis.sh, we find if a page is linked by some page which has a page rank in the top 50.

For this page:
WT21-B37-75 – second highest on page rank list and has 1704 inlinks

Additionally, the lemur interface tells us that the page is an experimental home page to *The Economist.* So, we can understand that when the user types "economist" in the query, it might want look at this page.

b. **WT21-B37-75** has a page rank: 0.00152592 and in links: 1704
This page also has very high number of in links and can be considered as important/interesting
This page has an in link from WT21-B37-76, which we just analyzed above. Thus, these two pages increase their in links by linking to each other and thus, their page rank value are also so high.

The lemur interface tells us that this page contains the "Copyright Notice" for the *The Economist.*
So, there will be a link to this page from most of the pages or articles of a newspaper near the bottom of the page. But, it might not be that interesting to some user who will visit the online newspaper to read some article.

c. **WT25-B39-116** has a page rank: 0.00146949 and in links: 169
Even though, this page so less page ranks, why would it have the 3rd highest page rank?
The reason for this is, most of the pages this page has an in link from, do not have any in links of their own.
Thus, to reach the in links of this page we would require teleportation.
Plus, the in links to this page have only one out link, i.e., to **WT25-B39-116.**
And, most importantly this page has only one out link- to itself. So, it is more likely to move to itself than teleport.
Thus, even though the page might not be that interesting to the user, it's page rank has been increased through the links to it.

d. **WT01-B18-225** has 2260 in links and a page rank of 0.00098844, rank 13.
On the lemur interface, this page is the online library of drug policy and it might be of importance to the user.
But, even though it has third highest number of in links, its page rank is not in the top 10 because most of its in links are duplicate. The page that link to it have tried to increase its page rank by linking multiple times to it. To give more importance to the page, the linking could have used some more sophisticated methods like anchor text or some other weighting factor.

e. **WT23-B21-53** has a high page rank of 0.00137232 and only 198 in links
The lemur interface shows a list of web development members for this page.
Thus, this page might not be very important to some user.
But, the reason it has a high page rank is that it has only one out link- to itself. So, it is more likely to move to itself than teleport.

One, more interesting thing which was observed was that many pages that link to our page have in links who also link to our page. Example: WT23-B21-229 -> WT23-B21-54 -> WT23-B21-53, WT23-B21-229 -> WT23-B21-53
Also, there are big loops in the web graph: WT23-B21-189 -> WT23-B21-53, WT23-B21-189 -> WT23-B21-51 -> WT23-B21-53, WT23-B21-189 -> WT23-B21-55 -> WT23-B21-197 -> WT23-B21-53. This, is an example of a link farm where the index of the search engine gets spammed.

f. **WT24-B40-171** has a page rank: 0.00124507 but has only 270 in links.
Like before, for **WT25-B39-116,** the pages that this page has in links from have only one out link, i.e. to this page.
Also, most of the in links of this page have one in link themselves and guess what, it from our page: **WT24-B40-171 or another page which links to our page.** It's a cyclic loop.
Further, this page has a loop on itself, but its effect is normalized by the other 208 out links that it has.
Lemur interface tells us that this page is an online archive and thus will not be of much importance to the user.

g. **WT23-B39-340** has a page rank of 0.00124049, 7th on our page rank list and has a low in link count of 274.
The Lemur interface tells us that this page has links to financial reports of different companies. This page would be important to some users.
But, the reason it has a high page rank with low number of in links is that, most of the pages that link to this page have only one out link; hence our page gets the full share of page rank from all its in links. And further analysis, show that many pages that link to our page have in links from our page too. Hence this forms a cyclic loop like we saw before.

h. **WT23-B37-134** has a page rank of 0.00120521, 8th on our page rank list has a low in link count of 208.
Lemur interface shows that this page contains copy right and disclaimer information, something that the user might not be interested in.
The reason that it has a good page rank, is that the page that link to it have either one out link (to our page) or out links to pages that link to our page. Thus, our page is getting good page rank share directly or indirectly.

i. **WT08-B18-400** has a page rank of 0.00114354, 9th on our page rank list and also has a high in link count of 1011
The lemur interface shows that this page is the home of The Toronto-Dominion Bank and yes, this page would be important for a user query.

j. **WT13-B06-284** has a page rank of 0.00112478, 10th on our page rank list and has good number on in links too : 454 (in our top 50 list for in link counts)
This page contains people information of the Florida Program.

This page would be of importance for a user query, but it has a good page rank largely because the page has two out links- one to itself (self-loop) and another to WT13-B06-273, which has out link to our page WT13-B06-284 (cyclic loop).
And WT13-B06-273 also has 454 in links and good page rank: 0.0010447. But, our page beats it because of the self-loop and a one more cyclic loop.

k. **WT18-B29-37** has a very high number of in links 2269, but its page rank is not even in the top 50.
The reason for this is that most of the in links to the page are duplicates but the number duplicate in links will have that much high dividing factor while calculating the page rank.
Plus most of the pages that link to our page have themselves 1 to 2 in links. So, the probability of being on an in link of our page is also less.

This page contains Environmental News for states. This page might be important to a user query. So to increase its page rank sophisticated methods like weighting should be used to provide it a better page rank than just provide multiple links to the page.

l. **WT23-B27-29** has high number of in links 1940, but its page rank is not is the top 50.
It's a home page for SportsGate, something that would be relevant to the user query, but its page rank is no high because most of the in links to it are duplicates and many in links to this page have only one in link themselves.

m. **WT27-B34-57** has 1257 in links and a page rank of 0.0005555
Page of the Skeptics Society Message board. This page would be of some importance to the user. Its page rank is low because most of the in links are duplicates. So, having more in links won't have a big effect since that many number of out link would be a dividing factor while calculating the page rank.

n. **WT27-B32-30** has 1255 in links and a page rank of 0.00054977
This page is some faq page and would be of little importance to a user query.
These pages have a good page rank in the top 50, but they have so many in links just because many of the in links are duplicates. Also, this page has an in link from WT27-B34-57 discussed above.

o. **WT08-B19-222** has 1041 in links and a page rank of 0.00064343
This is the page of the terms and conditions of the **Toronto-Dominion Bank.** This page won't be of that much importance to the user. The reason for the high page rank is because of so many distinct in links, but it is not that high, because most of the pages that link to this page have themselves only one in link and the in link of the in link is only one. Also, some of the in links to this page are sink nodes which can be reached only through teleportation. So, the page rank of this page will not be in the top 10.

p. **WT10-B36-88** has 946 in links 10th on our list of in links, but its page rank is not even in the top 50.

The main reason for so many in links and insignificant page rank is that most of its in links are duplicates who are either sink nodes or have 1 or two in links, Thus the page rank will be insignificant for our page.

This page is a site of the gay, lesbian, bisexual and transgender community. This page would be f interest to many users belonging to this community to communicate, find business listings, etc. So, its page rank should have been increased by sophisticated measures like weighting factor, anchor text, etc.
