Creating a link collection using Screaming Frog

This project involves collection of hyperlinks using the Screaming Frog SEO. The collected hyperlinks are to be imported in Gephi for further analysis.

First, you need to create the link collection. Use the Wikipedia page of IIT Kanpur https://en.wikipedia.org/wiki/IIT_Kanpur as the seed page. Collect the new web addresses by collecting links mentioned in the IIT Kanpur Wikipedia page, and reuse them to collect more web addresses via the new seed lists.

You may use the following steps to create your link collection.

1. Save the URL of IIT Kanpur Wikipedia page in a .csv file. Now upload this file by selecting 'List' option in 'Mode' menu bar at the top of the Screaming Frog SEO.
2. Run the crawler to find all the links that the Wikipedia page points to.
3. Select the 'Internal' tab, select all web addresses in the top window. Now select the 'Outlinks' in the bottom window. This shows all outgoing links corresponding to the web addresses mentioned in the Internal tab.
4. Select 'Export' tab at the top of the bottom window.
5. Download the file as .xlsx file. It should contain approximately 366,350 rows. You can give it an appropriate name, for example, wiki_iitk_step1.xlsx
6. The extracted file contains more information than we need. Such as, Anchor Text, Status code, Path Type etc. You are encouraged to look for their definition in the Screaming Frog user guide. However, for this project, we need 'From' and 'To' columns.
7. Copy the 'To' column and remove the duplicate entries. You will use this column with unique entries as the next seed set. Let's give it an appropriate name seeds_step2.csv. This should result in approximately 129,000 links.
8. Upload seeds_step2.csv and repeat steps 1-7 until the crawling is complete for the first 1990 links in seeds_step2.csv. **Note:** The free version of Screaming Frog is limited to crawling 500 URLs. Therefore, the seed list of 1990 nodes must be divided into shorter lists. You may limit your list to 499 URLs in order to avoid crawl limit.

You now have created a Web graph of reasonable size for this project. The data from the top window contains many other details of a node such as, Status Code, Word Count, Readability Index etc. We will skip these details in this project.

Data cleaning and preparation

Collect all wiki_iitk_step files in a single Excel file. This file has two columns: Source (From) and Target (To). You may remove protocol (https) and www texts from each link. Also, remove the extremely long web addresses. Usually, they are links to apps, payment gateways or bots. Finally, remove the duplicate edges.

Import in Gephi

After completing above steps you should have an excel file with two columns, Source and Target, and only one copy of each link. Save this file in the .csv format to import it in Gephi.

Report details

Including the following information in your report.

1. Import the graph and include its snapshot in the report using the Yifan Hu layout.
2. Report the number of edges and nodes in the graph.
3. Report the top-10 nodes based on: (a) in-degree, (b) out-degree.
4. Run the PageRank algorithm and report the top 10 nodes based on this metric.
5. Adjust the size of each node based on its PageRank value, and include the corresponding graph in the report.
6. Run the modularity algorithm and color the nodes based on their modularity class.
7. Include a snapshot of the different clusters identified in the previous step.
8. Filter the largest cluster identifed by the modularity algorithm and export it. How many nodes and edges are present in the graph?
9. Visualize the graph in the previous step. This graph should include the label of each node. What could be a possible explanation for grouping these nodes in the same cluster? Why are these nodes highly interrelated?
10. Finally, mention the contribution of each team member.