Introduction to Data Science – CAP 5771

HW7 - Report

Name: Sai Vempali

UFID: 16141381

PageRank Calculation:

The formula $PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$ mentioned in the instructions has been used to calculate page rank for each of the Wikipedia pages. Three jobs were used. The first job has a structure like the WordCount example and finds the total number of unique pages in the set. The output for reducer after this job is <null, N> where totalCount gives the total number of pages. The second job does the page rank calculation. For calculation of page rank, since we have the list of outgoing links for each page the weight for each node is calculated suitably. If it is the first iteration, then a rank of 1/N where N is the total number of pages is allocated to each node. Then the reducer gets this input and it computes the total score for each page based on the set of incoming nodes using the above formula and returns the value along with the degree which is the number of outgoing links. Also the entire list of links for a page are also returned for future use. To distinguish between a line containing only the list of links and the one with rank and number of outgoing links a "$" sign was introduced at the beginning of lists containing only the links. The is used in the reducer to identify lines with list of links. In the reducer the score for each node is computed and added to the previous score. This gives the new score and finally the formula is applied to generate the final score. This score is returned with the entire list of links followed by it. So, output after the page rank calculation reducer is like <key, rank score, list of outgoing links>. Then the sorter mapper takes the input from the page rank calculation job and its output is as follows: <rank score, page title> and the sort reducer does the sorting and displays all pages with score >= 5/N in the form <page title, rank score>

Difficulties faced during this lab:

1. Handling the output files correctly by renaming and moving them.
2. Understanding page rank and finding out the logic for page rank calculation.
3. Giving the input correctly for each iteration and setting the input and output paths correctly for each iteration.