Big Data Analytics Assignment I
Due Date: 23:55, 8th October, 2020

September 24, 2020

1 Text Processing

Given the email dataset emailtexts, process the texts according to the following steps.
i) Build the vocabulary base, remove stop words, and implement word stemming
ii) compute the tf-idf value for each word in the vocabulary

2 Inverted Index

Transfer all the texts into three-element tuples (docID, termID, weight), order the tuples according to termID and construct the inverted index for each word in the vocabulary.

3 Ranking

Given two queries: *"home run", "ESPN hockey"*, return the top 100 documents for each query based on:
i) boolean model
ii) vector space model.
iii) BM25 model.

4 PageRank Implementation (5 points)

Assume the directed graph $G = (V, E)$ has n nodes (numbered 1, 2, . . . , n) and m edges, all nodes have positive out-degree, and $M = [Mji]_{n\times n}$ is a an $n \times n$ matrix as defined in class such that for any $i, j \in [1, n]$:

$$M_{ji} = \begin{cases} \frac{1}{\deg(i)} & \text{if } (i \to j) \in E \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Here, $\deg(i)$ is the number of outgoing edges of node i in G. If there are multiple edges in the same direction between two nodes, treat them as a single edge. By the definition

of PageRank, assuming $1-\beta$ to be the **teleport** probability, and denoting the PageRank vector by the column vector $\mathbf{r}$, we have the following equation:

$$\mathbf{r} = (1-\beta)\frac{\mathbf{1}}{n} + \beta M\mathbf{r}, \tag{2}$$

where $\mathbf{1}$ is the $n \times 1$ vector with all entries equal to 1. Based on this equation, the iterative procedure to compute PageRank works as follows:

- Initialize: $\mathbf{r}^0 = \frac{1}{n}\mathbf{1}$
- For t from 1 to k, iterate: $\mathbf{r}^{(t)} = (1-\beta)\frac{\mathbf{1}}{n} + \beta M\mathbf{r}^{(t-1)}$

Given a graph `graph-small.txt`, run the aforementioned iterative process for 40 iterations (assuming β=0.8) and obtain the PageRank vector r. Compute the following:

- List the top 5 node ids with the highest PageRank scores.
- List the bottom 5 node ids with the lowest PageRank scores.

Note: You have to implement the PageRank iterative process by yourself without any third-party packages.

5 Submission Guidelines

You can use any programming language to do the assignments. You can use open source codes to help you do the stop words removal, word stemming, but you have to finish other tasks through your own original programming.

The submission should include:

1) the source codes

2) the processed results in ReadMe.txt