**Homework assignment Set #2**

**Instructions:** This document is generated with Microsoft Word MacOS version 16.16.6. Please contact the instructor if you have technical difficulty to read the problems. Type the answers into empty space below each problem, using additional pages if needed. Please create the figures using your preferred plotting tool, e.g., Excel, PyPlot, and paste the them to where is suitable. Please do not make any changes to the original problem! You may insert equations using the Insert->Equation module if it is available, but this is not obligated. Alternatively, you may use Latex software such as TexShop to input your solutions without typing the narratives of original problems.

After finishing the problems, please rename the file by appending your first name after the current file name. For example, set#-jian.docx (or set#-jian.pdf if you use Latex), and upload it to the box folder I shared with you by **noon on Tuesday, April 16th, 2019.**

Please honor the **academic integrity** in the syllabus. Violators, if confirmed, will be graded as 0.

In this exercise, you will create a simplified Lucene index. To get partial credit in case of miscalculations, please give detailed solutions.

Given the following documents:

D1: You say “goodbye”, I say “hello, hello, hello”

D2: You say stop, I say go.

D3: “Hello, hello, hello,” you say “goodbye”.

D4: I say yes, you say no.

**Problem 1: Build the inverted index for the documents.**

1. Dictionary file, e.g.,

|  |  |
| --- | --- |
| **Term** | **DocFreq** |
| Hello | 2 |
| I | 3 |

1. Posting file (terms are implicit), e.g.,

|  |  |
| --- | --- |
| **Doc#** | **Frequency** |
| 1 | 3 |
| 3 | 3 |

1. Posting file (terms are implicit from dictionary file, use absolute position of terms in the document), e.g.,

|  |  |  |  |
| --- | --- | --- | --- |
| **D1** | **D2** | **D3** | **D4** |
| 6,7,8 | 0 | 1,2,3 | 0 |
| 4 | 4 | 0 | 1 |

1. For a given query:  
   Q: say goodbye,  
   describe the process to search the inverted index.

**Problem 2: Based on Problem 1**

* 1. Estimate the total size of the inverted index file in bytes. Numbers and characters are counted as 4 bytes. Strings are counted as the number of characters multiplied by 4 bytes. For example, the size of string “hello” is 5\*4 = 20 bytes.
  2. Compare the results from 2a. to the total size of the documents in bytes.

**Problem 3: Given the following people (nodes) and relations (edges), analyze the network.**

*Nodes*:

P1: Tom, P2: Jean, P3: Nat, P4: Steven, P5: Vijay, P6: Mary, P7: Oscar, P8: Michelle, P9: Jian, P10: Clark

*Edges*:

P1 -> P4

P2 -> P1, P2 -> P4

P3 -> P2

P4 <-> P10

P5 -> P4, P5 -> P1

P6 -> P7, P6 -> P8, P6 -> P9

P7 -> P8

P10 -> P6

P10 -> P9

1. Draw the network as a graph and a matrix.
2. Calculate the following metrics:
   1. In degree of each node
   2. Out degree of each node

**Problem 4: Based on the condition in Problem 3, calculate the PageRank scores for each node in the social network given the following modified directed connection.** We suggest you use either:   
<https://webworkshop.net/seo-tools/pagerank_calculator>

Or the excel sheet that comes with this homework set.

1. Use initial score of 1 for all nodes. Calculate 10 iterations of the PageRank scores.
2. Change the initial score to 0.15. Calculate 10 iterations of the PageRank scores.
3. Change the initial score to 1.5. Calculate 10 iterations of the PageRank scores.
4. Compare the convergence behavior for the three initial values.