Lab 5 – Information Retrieval

**Part I. Term Weighting**

Suppose that we have a collection of one million documents and that the TF (term frequency) data for the first three documents are shown in Figure 1. In addition, the DF (document frequency) values for four terms from them are shown in Table 2.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Doc1 | Doc2 | Doc3 |
| Car | 27 | 4 | 24 |
| Auto | 3 | 33 | 0 |
| Insurance | 0 | 33 | 29 |
| Best | 14 | 0 | 17 |

Figure 1. Table of **TF** values

|  |  |  |  |
| --- | --- | --- | --- |
|  | DF | N | idf­ = log10(N/DF) |
| Car | 10,000 | 1,000,000 | log10(1,000,000/10,000)= 2 |
| Auto | 10,000 | 1,000,000 | log10(1,000,000/10,000)= 2 |
| Insurance | 1,000 | 1,000,000 | log10(1,000,000/1,000)= 3 |
| Best | 100,000 | 1,000,000 | log10(1,000,000/100,000)= 1 |

Figure 2. Table of **DF** values

Example: log10(1000/10) = log10(100) = log10(102) = 2

|  |  |
| --- | --- |
|  | Doc1 |
| Car | 2\*27= 54 |
| Auto | 2\*3= 6 |
| Insurance | 3\*0= 0 |
| Best | 1\*14= 14 |

Figure 3. Table of **TF\*idf** values

1) (24 points) Calculate the terms’ idf values and their TF\*idf values for Doc1.

2) (6 points) Explain why terms should be given different weights (i.e. why some terms are more informative than others and should be weighted higher). Use the terms in this exercise as examples.

The terms should be weighted because if what I am looking for is a good car insurance document then there should be more weight placed in a document that has the word “insurance” a few times then the word “Best” many times since there are many things that could be referred to as the best

|  |
| --- |
| **Part II. PageRank for Web Search Ranking** |

**Given the following nodes (pages) and links, calculate the pages’ PageRank scores, i.e., R values.**

**Using PageRank formula:**

****

**with damping factor d = 0.2.**

where ***p*** denotes the node being considered and ***pi*** is one of the nodes that link to node ***p***. For example, if three nodes X, Y, and Z link to A, then the PageRank score of A: *R(A) = d/T + (1-d) \* [R(X)/C(X) + R(Y)/C(Y) + R(Z)/C(Z)]*.

1. (2 points) Count the total number of nodes.

T = 4

1. (16 points) Collect basic degree information about the nodes (pages).

|  |  |  |
| --- | --- | --- |
| **Node** | **In-degree** | **Out-degree = C(p)** |
| A | 1 | C(A) = 2 |
| B | 3 | C(B) =0 |
| C | 1 | C(C) = 2 |
| D | 1 | C(D) = 2 |

1. **Step 1.** (2 points) Initialize all nodes’ PageRank values (all R values) with value 1.

|  |  |
| --- | --- |
| **Node** | **Step 1 value** |
| A | R(A) = 1 |
| B | R(B) = 1 |
| C | R(C) =1 |
| D | R(D) =1 |

1. **Step 2.** (8 points) Recalculate R values using values from step 1. Use the above PageRank formula. **Please provide calculation details. Make sure any decimal values use five places after the decimal point.**

|  |  |
| --- | --- |
| **Node** | **Step 2 value** |
| A | R(A) = .45 |
| B | R(B) = 1.25 |
| C | R(C) =.45 |
| D | R(D) =.45 |

1. **Step 3.** (8 points) Recalculate R values using values from step 2. **Please provide calculation details.**

**Make sure any decimal values use five places after the decimal point.**

|  |  |
| --- | --- |
| **Node** | **Step 3 value** |
| A | R(A) = .23 |
| B | R(B) = .59 |
| C | R(C) =.23 |
| D | R(D) =.23 |

1. (4 points) Compare R values from step 3 with the nodes’ in-degrees. What do you find?

The more nodes’ in-degrees the higher the R value

**What to Turn In**

Please finish all questions in both Part I and Part II. Be sure to fill out all highlighted blanks.

For Part II, please provide calculation steps and details (you may require to add an additional page to this answer sheet if needed). **Make sure any decimal values use five places after the decimal point.**