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 | 2 |
| Auto | 10,000 | 1,000,000 | 2 |
| Insurance | 1,000 | 1,000,000 | 3 |
| Best | 100,000 | 1,000,000 | 1 |

Figure 2. Table of **DF** values

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

|  |  |
| --- | --- |
|  | Doc1 |
| Car | 54 |
| Auto | 6 |
| Insurance | 0 |
| Best | 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.

Terms should be given different weights so that websites don’t try to “cheat” their way to being the first/most important website on the search engines. In this case, “best” and “car” are more likely to be on the internet in several spots, whereas “auto and “insurance” are going to appear less frequently on the web. This should give “auto” and “Insurance” more weight so that when someone says “car” multiple it is not outshining the more closely related website. The website with “car” on it multiple times would appear first in this case, but they could be talking about a race car driver and not the “auto” “insurance” that seems to be more important in this case.

|  |
| --- |
| **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 = 6 nodes

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) = 0.45000 |
| B | R(B) = 1.25000 |
| C | R(C) = 0.45000 |
| D | R(D) = 0.45000 |

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) = 0.23000 |
| B | R(B) = 0.59000 |
| C | R(C) = 0.23000 |
| D | R(D) = 0.23000 |

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) = 0.14200 |
| B | R(B) = 0.32600 |
| C | R(C) = 0.14200 |
| D | R(D) = 0.14200 |

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

Since node B has the highest in-degree and the lowest out degree this shows with it also having the highest R value in all the steps. Nodes A,B,&C all have an in-degree of 1. In Rv3 it is clear that they are the same.

**Calculation Steps/Details**

For my calculations I put the equation given above into excel. For example, my R(B3) value is equal to [=SUM((SUMPRODUCT(0.2/4))+(SUMPRODUCT((1-0.2)\*(SUM((SUMPRODUCT(E2/C2))+(SUMPRODUCT(E4/C4)+(SUMPRODUCT(E5/C5))))))))]. In order to start, I used 1 as my numerator when finding R1(p).

In this directory I have also added my work from the excel document.

**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.**