1

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 | **TF\*idf** =27\*2=54 |
| Auto | **TF\*idf** =3\*2=6 |
| Insurance | **TF\*idf** =0\*3=0 |
| Best | **TF\*idf** =14\*1=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.

Rare terms are more informative to searches and therefore should be given larger weight than frequent terms. In this example it is seen that “insurance” is multiplied by 3 since it a rarer term within the documents, compared to “car” which is multiplied by only 2 since it is a more common word within the documents. Moreover, the weight vs frequency of a term balances out considering that tf\*idf values increase with the rarity of a term within a document and also with the increase in frequency of a term in the entire collection.

|  |
| --- |
| **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 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) =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) = (d/T)+(1-d)(R(D)/C(D))=(.2)(.25)+(.8)(.5)=0.45000 |
| B | R(B) = (d/T)+(1-d)(R(D)/C(D)+(R(C)/C(C) )+(R(A)/C(A))= (.2)(.25)+ (.8)(.45+.45+.45)=1.03000 |
| C | R(C) =(d/T)+(1-d)((R(A)/C(A))=(.2)(.25)+(.8\*.45)(.5)=0.23000 |
| D | R(D) =(d/T)+(1-d)((R(C)/C(C) )= (.2)(.25)+(.23)(.5)=0.14200 |

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) = (d/T)+(1-d)(R(D)/C(D))=(.2)(.25)+(.8)(0.142000/2)=0.10680 |
| B | R(B) = (d/T)+(1-d)(R(D)/C(D)+(R(C)/C(C) )+(R(A)/C(A)= (.2)(.25)+(.8)((.10680/2)+(.23000/2)+(.142000/2))=0.24152 |
| C | R(C) =(d/T)+(1-d)((R(A)/C(A))= (.2)(.25)+(.8)(..10680/2)=0.09272 |
| D | R(D) =(d/T)+(1-d)((R(C)/C(C) )= (.2)(.25)+(.8)(.09272/2)=0.087088 |

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

Upon comparison, it can be seen that R(B) having the only in-degree of 3, has the highest R value in step 3. A is the next highest, followed by C, and then D with the lowest of the 3. This is very interesting given that all 3 of these values have the same in-degree of 1.

**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.**