

## 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 = $\log_{10}(N/DF)$
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:  $\log_{10}(1000/10) = \log_{10}(100) = \log_{10}(10^2) = 2$

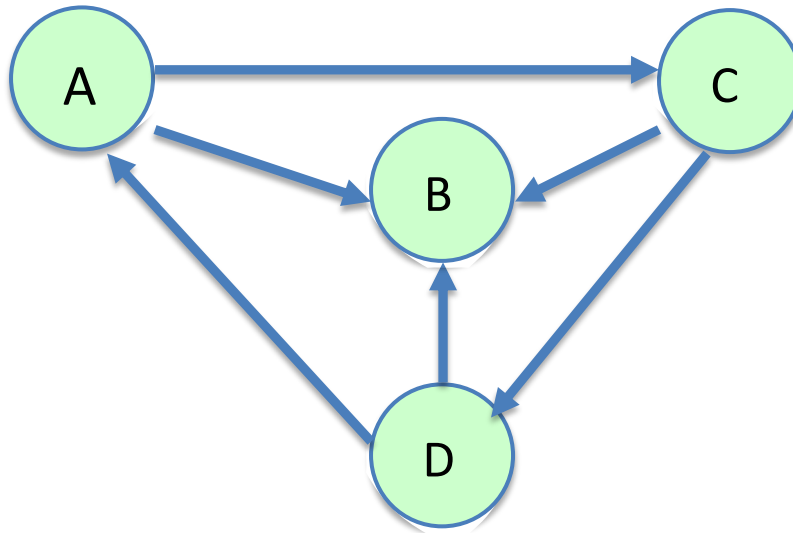
	Doc1
Car	54
Auto	6
Insurance	0
Best	14

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.  
 Different Terms should be given different weights because not all terms are as informative as others . The more informative terms should be weighted higher because they are rare and specific terms.

## 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:

$$R(p) = d \cdot \frac{1}{T} + (1 - d) \cdot \sum_{i=1}^k \frac{R(p_i)}{C(p_i)}$$

with damping factor  $d = 0.2$ .

where  $p$  denotes the node being considered and  $p_i$  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.

$T = 4$

2. (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

3. **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	$R(D) = 1$

4. **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	$R(A) = .2(1/4) + .8(1/2) = .45000$
B	$R(B) = .2(1/4) + .8(.45/2 + 1/2 + 1/2) = 1.03000$
C	$R(C) = .2(1/4) + .8(.45/2) = .23000$
D	$R(D) = .2(1/4) + .8(.23/2) = .14200$

5. **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	$R(A) = .2(1/4) + .8(.23/2) = .10680$
B	$R(B) = .2(1/4) + .8(.1068/2 + .142/2 + .23/2) = .24152$
C	$R(C) = .2(1/4) + .8(.24152/2) = .09272$
D	$R(D) = .2(1/4) + .8(.09272/2) = .08709$

6. (4 points) Compare R values from step 3 with the nodes' in-degrees. What do you find?

The values get pretty proportional to the nodes' in-degrees as the calculations continue and approach a stable 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.**