# Web Information Retrieval

## Assignment 4

**Team Name : Gamma**

## Members

| Name | Matriculation Number |
|---|---|
| Mohammad Nizam Uddin | 216101140 |
| Md. Shohel Ahamad | 216203438 |
| MD Jakaria Nawaz | 216203442 |
| Shreya Chatterjee | 216100848 |

## 1 TF-IDF Calculation (10 Points) .

**Part-01**
**(1-4):**

Given,

Doc 1: “teach education school university education”
Doc 2: “education education campus teach teach”
Doc 3: “university university school teach learning”
Doc 4: “campus learning education learning”

**N=|Docs| = 4.**

| **Term** | **Term frequency** | | | | **inverse document frequency** | **DF** | **tf-idf** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Doc1** | **Doc 2** | **Doc3** | **Doc4** | **$(\mathrm{idf}_t = \log_{10} N/\mathrm{df}_t)$** | | **Doc 1** | **Doc 2** | **Doc 3** | **Doc 4** |
| teach | 1 | 2 | 1 | 0 | **$\log_{10}$ ( 4/3)=0.125** | **3** | 0.125 | 0.25 | 0.125 | 0 |
| education | 1 | 2 | 0 | 1 | **$\log_{10}$ ( 4/3)=0.125** | **3** | 0.125 | 0.25 | 0 | 0.125 |
| school | 1 | 0 | 1 | 0 | **$\log_{10}$ ( 4/2)=0.30** | **2** | 0.30 | 0 | 0.30 | 0 |
| university | 1 | 0 | 2 | 0 | **$\log_{10}$ ( 4/3)=0.125** | **2** | 0.125 | 0 | 0.25 | 0 |
| campus | 0 | 1 | 0 | 1 | **$\log_{10}$ ( 4/2)=0.30** | **1** | 0 | 0.30 | 0 | 0.30 |
| learning | 0 | 0 | 1 | 2 | **$\log_{10}$ ( 4/2)=0.30** | **2** | 0 | 0 | 0.30 | 0.60 |

**Part-02**

**(1-3):**

Given ,

Doc 1: “teach education school university education”
Doc 2: “education education campus teach teach”
Doc 3: “university university school teach learning”
Doc 4: “campus learning education learning”

Query: “teach teach education campus”.

| **Term** | **TF** | | | | | **inverse document frequency** | **tf-idf** |
|---|---|---|---|---|---|---|---|
| | **Doc1** | **Doc2** | **Doc3** | **Doc4** | Query | **(idf$_t$ = log$_{10}$ $N$/df$_t$)** | Query |
| campus | 0 | 1 | 0 | 1 | 1 | **log$_{10}$ ( 5/3)=0.22** | 0.22 |
| education | 1 | 2 | 0 | 1 | 1 | **log$_{10}$ ( 5/4)=0.10** | 0.10 |
| learning | 0 | 0 | 1 | 2 | 0 | **log$_{10}$ ( 5/2)=0.40** | 0 |
| school | 1 | 0 | 1 | 0 | 0 | **log$_{10}$ ( 5/2)=0.40** | 0 |
| teach | 1 | 2 | 1 | 0 | 2 | **log$_{10}$ (5/4)=0.10** | 0.20 |
| university | 1 | 0 | 2 | 0 | 0 | **log$_{10}$ ( 5/2)=0.40** | 0 |

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\bullet\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\bullet\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

Cosine Similarity (Query, Doc1) = { (3)/(√4.√6)}=0.61

Cosine Similarity (Query, Doc2) = {(1+2+4)/(√9.√6)}=0.95

Cosine Similarity (Query, Doc3) = {(2)/(√7.√6)}=0.31

Cosine Similarity (Query, Doc4) = {(2)/(√6.√6)}=0.33

**# The result of the query: Doc2**

**2 Relevance Feedback (6 Points)**

Consider the Rocchio method for relevance feedback as discussed in the lecture.
1. In Rocchio's algorithm, what weight setting for $\alpha/\beta/\gamma$ does a "Find pages like this one"-search correspond to?

**Ans:**

"Find pages like this one" tends to disregard the query and no negative judgments are considered.

Therefore, $\alpha = \gamma = 0$. Which implies $\beta = 1$.

2. Give three reasons why relevance feedback has been little used in web search.

**Ans:**

1. Relevance Feedback slows down returning results as it is needed to run two subsequent queries, the second of which is slower to compute than the first. Waiting is not a good experience for users.

2. Relevance Feedback is primarily used to increment recall, but users are more interested about the precision of the top few results.

3. Relevance Feedback is one way of handle the alternate ways to express an idea, but indexing anchor text is a better way to solve this problem.

**Reference:**
https://www.cs.helsinki.fi/group/doremi/courses/ir08/harj3.txt