**CSC 395**

**Practice Problem: Vector Space Model**

**Date: Sept 24, 2019**

Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Given the following collection:

D1: class Grinnell classroom test

D2: dog Grinnell test

D3: test time test material

D4: material test

**Step 01:**

Determine the tf and idf of the vocabulary.

**Vocabulary = (class, classroom, dog, Grinnell, material, test, time)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Raw term frequency** | **class** | **classroom** | **dog** | **Grinnell** | **material** | **test** | **time** |
| **D1** | **1** | **1** | **0** | **1** | **0** | **1** | **0** |
| **D2** | **0** | **0** | **1** | **1** | **0** | **1** | **0** |
| **D3** | **0** | **0** | **0** | **0** | **1** | **2** | **1** |
| **D4** | **0** | **0** | **0** | **0** | **1** | **1** | **0** |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **class** | **classroom** | **dog** | **Grinnell** | **material** | **test** | **time** |
| **IDF-score** | **log(4/1)=0.602** | **log(4/1)** | **log(4/1)** | **log(4/2)** | **log(4/2)** | **log(4/4)** | **log(4/1)** |

**Vector Space Representation with:tf (normalized) \* IDF**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| normalized term frequency | class | classroom | dog | Grinnell | material | test | time |
| D1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| D2 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| D3 | 0 | 0 | 0 | 0 | 0.5+0.5/2 | 0.5+0.5 \*2/2 | 0.5+0.5/2 |
| D4 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |

Step 02:

Represent the documents as vectors.

D1 = **<0.602, 0.602,0,0.301,0,0,0>**

D2 = **<0,0,.602,0.301,0,0,0>**

D3 = **<0,0,0,0,0.22575,0,0.1505>**

D4 = **<0,0,0,0,0.301,0,0>**

Step 03:

Represent the given query q = test Grinnell Grinnell as a vector.

q = <0,0,0,2,1,0, 0> [raw]

q = <0,0,0,2/2\*0.301,0, 1/2\*0,0> [normalized tf \* IDF]

q = <0,0,0, 0.301,0,0,0>

Step 04:

Apply cosine similarity between each document and the query. Show your result.

/ )

Do similar computation for the other documents.

Step 05:

Return the retrieved documents in some order of relevance.