# Vector space exercise

*We have a document database of 5 documents with the following content:*
Document 1(**d1**):"Information Retrieval Systems"
Document 2(**d2**):"Information Storage"
Document 3(**d3**):"Digital Speech Synthesis Systems"
Document 4(**d4**):"Speech Filtering"
Document 5(**d5**):"Speech Retrieval"

We want to retrieve the documents in that database that better match with my information need. For that, the query is **: Information Speech Filtering, Speech Retrieval.**

***Steps***

1. ***Frequency matrix***
2. ***Inverse document frequency:***
3. ***Query vector***
4. ***Similarity Function***

***1.-Frequency matrix: calculate the frequency of every term in each document***

| | Digital | Filtering | Information | Retrieval | Speech | Storage | Synthesis | Systems |
|---|---|---|---|---|---|---|---|---|
| d1 | | | | | | | | |
| d2 | | | | | | | | |
| d3 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| d4 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| d5 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| sum | 1 | 1 | | 2 | | 1 | 1 | 2 |

***2.-Inverse document frequency= log(number of documents/ frequency of the terms of all documents). IDF*** is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It helps to adjust for the fact that some words appear more frequently in general but there are not relevant

| TERM | DOC-FREQUENCY | IDF |
|---|---|---|
| Digital | 1 | log(5/1)=0.699 |
| Filtering | 1 | log(5/1)=0.699 |
| **Information** | | **log( )=** |
| Retrieval | 2 | log(5/2)=0.398 |
| **Speech** | | **log( )=** |
| Storage | 1 | log(5/1)=0.699 |
| Synthesis | 1 | log(5/1)=0.699 |
| **Systems** | | **log( / ) =** |

**3.- Calculate the matrix tf.idf = )multiply the frequency x the IDF of the term) and calculate the length of the vectors ( last column**)
**Length of d1=sqrt(_____^2+_____^2+_____^2)=**
**Length of d2=sqrt(_____^2+_____^2)=**
Length of d3=sqrt(0.699^2+0.222^2+0.699^2+0.398^2)=1.088
Length of d4=sqrt(0.699^2+0.222^2)=0.733
Length of d5=sqrt(0.398^2+0.222^2)=0.456

| | Digital | Filtering | Information | | Retrieval | Speech | Storage | Synthesis | Systems | **Length** |
|---|---|---|---|---|---|---|---|---|---|---|
| d1 | | | | | | | | | | |
| d2 | | | | | | | | | | |
| d3 | 1x0.699 | 0 | 0 | | 0 | 1x0.222 | 0 | 1x0.699 | 0.398 | **1.088** |
| d4 | 0 | 0.699 | 0 | | 0 | 0.222 | 0 | 0 | 0 | **0.733** |
| d5 | 0 | 0 | 0 | | 0.398 | 0.222 | 0 | 0 | 0 | **0.456** |

**4.-Query and query vector**
**The Query is**: **Information Speech Filtering, Speech Retrieval**
The maximum frequency of a term is ("Speech")=2
**Query vector**: frequency of the term/ max frequency of every term) X idf of the term

Length= sqrt (_____^2+_____^2+_____^2+_____^2)=_____

| | Digital | Filtering | Information | Retrieval | Speech | Storage | Synthesis | Systems | **Length** |
|---|---|---|---|---|---|---|---|---|---|
| q | | (1/2)*0.699=0.349 | | | | | | | |

**7.-Similarity Function: multiply the vector of the query by the vector of each document divided by the multiplication of its lengths**

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

cosSim(d1,q)=(_____*_____+_____*_____)/(_____*_____)=_____
cosSim(d2,q)=(0.398*0.199)/(0.501*0.804)=**0.197**
cosSim(d3,q)=(0.222*0.222)/(0.501*1.088)=**0.090**
cosSim(d4,q)=(0.222*0.222+0.699*0.349)/(0.501*0.733)=**0.799**
cosSim(d5,q)=(_____*_____+_____*_____)/(0.501*0.456)=_____

**the bigger the cosine the more similar de doc and the query**
**solution: the order of presentation is**

_____