Information Retrieval

Dra. Mireya Paredes

Vector Space Model

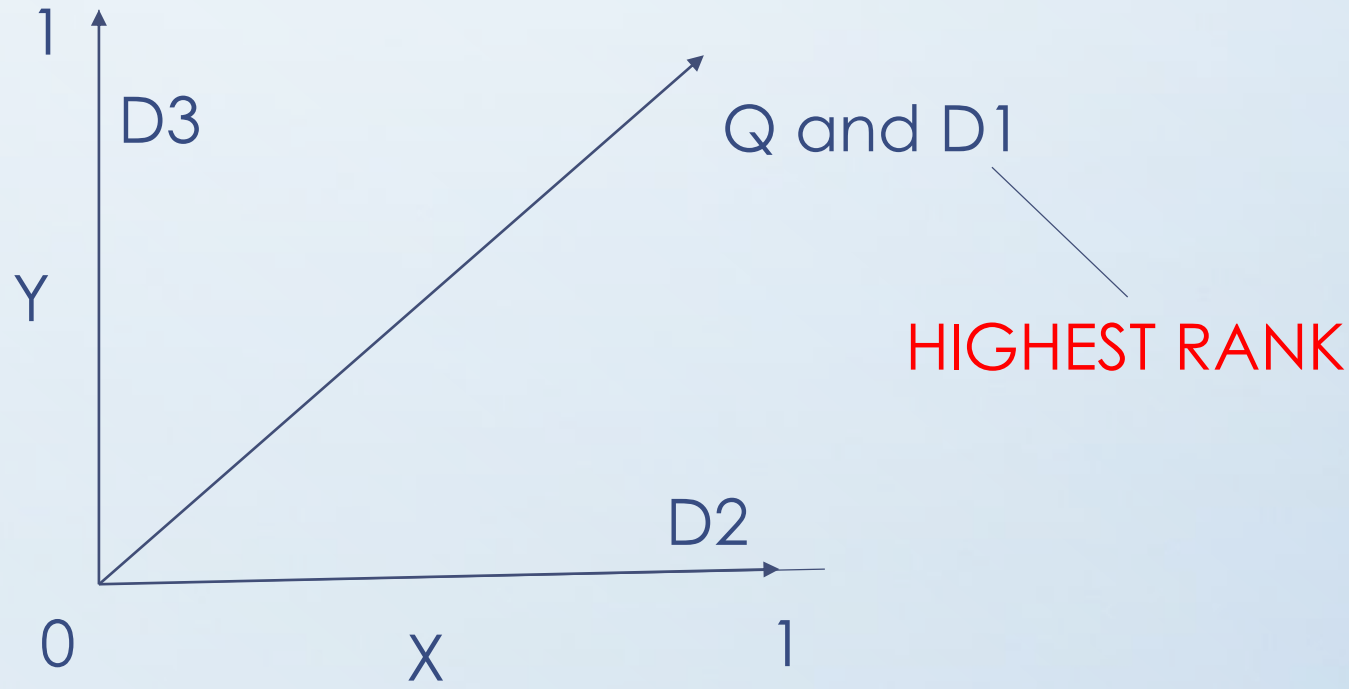
Computes a measure of similarity of defining a vector that represents each document, and a vector that represents the query.

- The **meaning of a document** is conveyed by the **words used**.

- Constructing a vector which represents the terms in the document and choosing a method of measuring the closeness of any two vectors.

Example of a tiny vector space model

D1 = A I = <1 1>
D2 = A = <1 0>
D3 = I = <0 1>
Q = A I = <1 1>



Measuring vectors closeness

Similarity coefficient
This could be computed as the distance from the query to the two vectors.

Inner product is usually the method used between the ***Query vector*** and the **Document vector**.

Assigning weight to TERMS

Index document Frequency (IDF)
per term

$$\mathrm{idf}_t = \log \frac{N}{\mathrm{df}_t}.$$

Total number of documents

Total number of documents term appears.

Q: “gold silver truck”
D1 = “Shipment of gold damaged in a fire”
D2 = “Delivery of silver arrived in a silver truck”
D3 = “Shipment of gold arrived in a truck”

IDF of the terms

T1→a	= 0	**N** = 3
T2→arrived	=0.176	
T3→damaged	=0.477	
T4→delivery	=0.477	
T5→fire	=0.477	
T6→gold	=0.176	
T7→in	=0	
T8→of	=0	
T9→silver	=0.477	
T10→shipment	=0.176	
T11→truck	=0.176	

Doc ID	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
D1	0	0	.477	0	.477	.176	0	0	0	.176	0
D2	0	.176	0	.477	0	0	0	0	.954	0	.176
D3	0	.176	0	0	0	.176	0	0	0	.176	.176
Q	0	0	0	0	0	.176	0	0	.477	0	.176

$$SC(Q, D) = (0)(0) + (0)(0) + (0)(0.477) + (0)(0) + (0)(0.477) + (0.176)(0.176)+(0)(0)+(0)(0)+(0.477)(0) + (0)(0.176) + (0.176)(0) + = (0.176)(0.176) = 0.031$$

Q: “today weather England”
D1 = “Today is such a good day”
D2 = “Suppose we are trying to predict weather”
D3 = “In England the weather is usually rainy”

1. Homework:
   On relevance, probabilistic indexing and information retrieval
   M. E. Maron and J.L. Kuhns

   To read the first 2 pages
   to write a paragraph explaining what is it about?