# MILITARY INSTITUTE OF SCIENCE & TECHNOLOGY

## CSE 453 Data Mining Assignment # 1

**Submitted by:**

Group – 02

**Group Members:**


1. 201514178 – Capt Akib Zaman
2. 201714014 – Abdullah-al-Sheak Jaber
3. 201714018 – Ayon Roy
4. 201714024 – Md. Aqib Alfaz
5. 201714043 – Nafiz Imtiaz Khan
6.


**Submitted to:**

Col Siddharth Malik, SM

# Question

**Set 2**

***Tutorial A***

Web Search has its root in Information retrieval

- Briefly explain how is Information Retrieval carried out?
- Briefly explain the Vector space model for information retrieval?

Question: Briefly explain how is Information Retrieval carried out?

Answer: Information retrieval is the field of study that helps the user find needed information from a large collection of text documents. In traditional IR, basic information unit is considered as a document, and a large number of documents formed the text database. However, in the web, the documents are web pages. Anything user asks to the search engine is called user query and re Retrieving information simply means finding a set of information that is relevant to the user query. Based on the relevance score of an particular query, a ranking of the set of of documents is performed. Depending on the ranking of the documents, search engine shows the user the most relevant document. List of keywords, also known as terms, are most commonly used query format. Data retrieval from database and IR are different as databases are highly structured whereas, we donot have any query language, such as,

SQL for text retreival.

Although web search is the most important application of IR, it doesn't simply apply traditional IR models. It uses some IR results, but it also has its unique techniques and presents many new problems for IR research. Web pages are also quite different from conventional text documents. Web pages use hyperlinks, anchor text which are not contained by traditional documents. for seach ranking algorithms to perform well, these informations (hyperlinks) play a vital role. Also A Web page is semi-structured as it doesn't simply contain some paragraps but it also contains some additional metadata like (title, body). These informations need to be organized and pre in several structured bbcks. Among these structured blocks, some blocks are important and some are not. Effectively detecting these bbcks are major issue in an efficient web seach. Finally, spamming is a major issue on the web. If a page is very relevant but ranked very low, user will unlikely to see this. That is why ranking of some target pages are improved ~~where by smay~~ by spamming.

In technical ~~tems~~, terms. IR studies the ~~acqui~~ acquisition, organization, storage, retrieval and distribution of information. An architecture of an IR system in presented in figure 1.

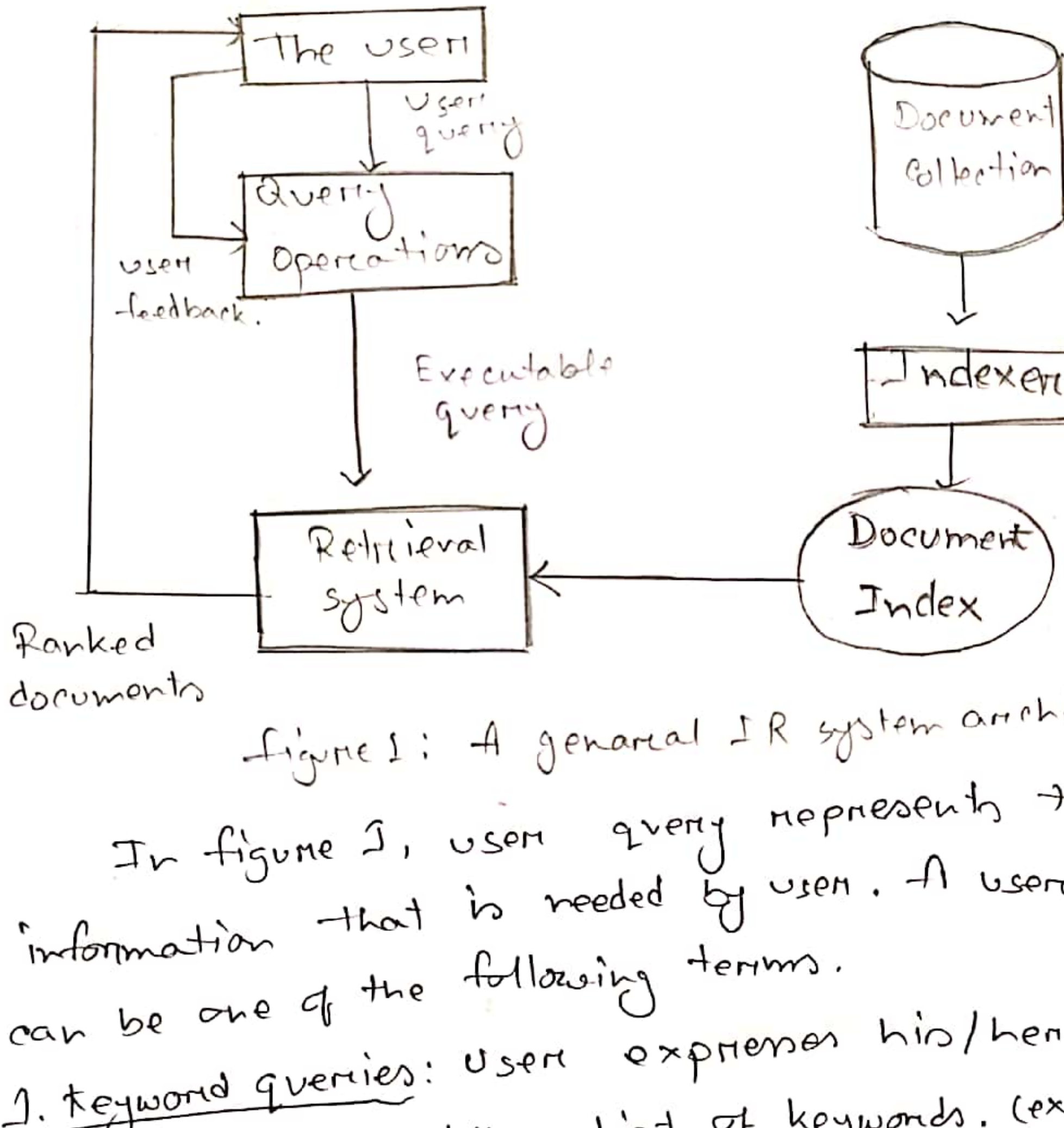



figure 1: A general IR system architecture.

In figure 1, user query represents the information that is needed by user. A user query can be one of the following terms.

1. Keyword queries: User expresses his/her information need with a list of keywords. (example: web mining)
2. Boolean queries: User may use boolean operators

in between the keywords to construct complex queries.

3. Phrase queries: User can search by a sequence of words (example: Data mining and applications)

4. Proximity queries: Proximity query is a relaxed version of phrase query and can be a combination of terms and phrases.

5. Full Document queries: User can directly ask for a specific document by providing URL.

6. Natural language questions: User can ask his/her query as natural language question.

The query operation module in the IR system architecture can range from very simple to very complex. It performs the pre-processing on the query, then pass it to retrieval system. Preprocessing on the query includes removal of stop-words, expanding of contractions, removal of noise, etc. this module transforms natural language queries into executable queries.

Indexer module in the system architecture indexes the original raw documents in some data structures to enable efficient retrieval. The result of this module is the document index. Finally, the retrieval system computes a relevance score for each indexed document to the query. Based on the relevance score, the documents are ranked and presented to the user. The Retrieval system does not match with user query with every, rather matches only a small subset of documents that contains only one query atleast one query term is first found based on index and sub-relevance score, user query is then compared with these small subset of documents. In this process, information retrieval is carried out.

② Briefly Explain How Vector Space Model work for information Retrieval.

## Vector Space Model

Vector Space Model or Term vector model is an algebraic model for representing text documents (and any object in general) as vectors of identifiers (such as index term). It is widely used in information filtering, information retrieval, indexing and relevancy rankings. Figure 1 shows a visulisation of Vector space Model.

Documents & Queries are represented as vectors.

$$d_j = (w_{j,j}, w_{2,j}, \ldots\ldots w_{t,j})$$

$$q = (w_{1,q}, w_{2,q}, \ldots\ldots w_{n,q})$$

Let us consider the issue of representation of documents in terms of the index terms $t_1, t_2, \ldots\ldots t_n$ be the terms used to represent documents. Corresponding to each term, $t_i$, suppose there exists a vector $t_i$ in the space. Without loss of generality, it is assumed that $t_i$s are the vectors of unit length. Now, suppose that each document $D_r$, $1 \le r \le m$, is a vector expressed in term of $t_i$s. Let the vector document $D_r$ be,

$$D_r = (a_{1r}, a_{2r}, \ldots\ldots a_{nr})$$

Where $a_{ir}$ are real numbers reflecting i in $D_r$.

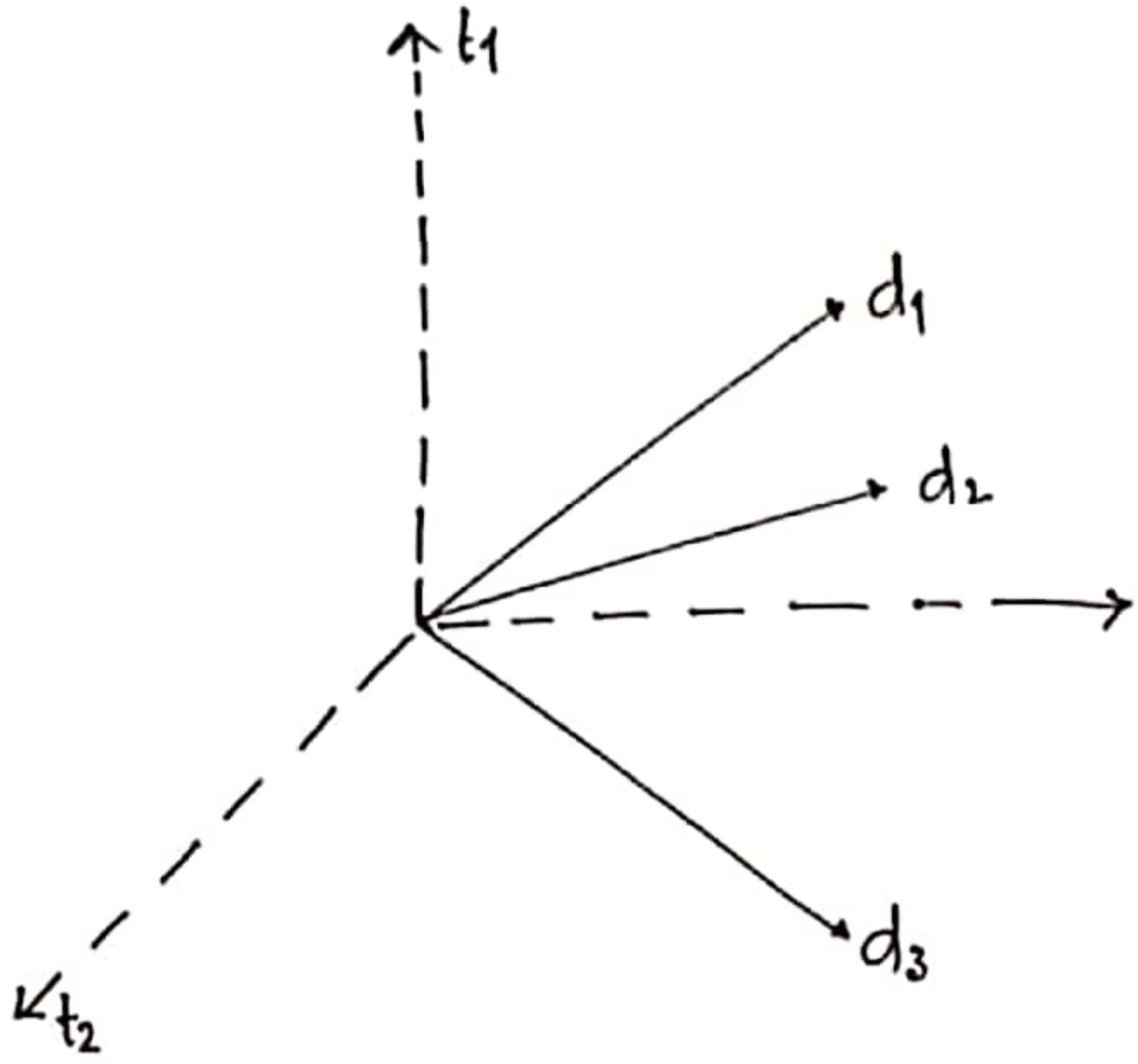
Since it is sufficient to restrict our scopes of disscussion to the subspace spanned by the term vectors, the ti s can be thought to be the generating set. Every vector in this subspace, and in particular all document vectors, are linear combinations of the term vectors. Thus, Dr can be equivalenty expressed as:

$$D_r = \sum_{i=1}^{n} a_{ir} t_i$$

The coefficients $a_{ir}$, for $1 \le i \le n$ and $1 \le r \le m$ are the components of Dr along the ti s.

Document similarity. is used in information retrieval to determine which document is more similar to a given query so this is one of the basic ideas in documentary tv or in search engines. Suppose, we have a query such as the vector Q which is shown in figure 2. Two documents $D_1$ and $D_2$ and we want to determine whether $d_1$ or $d_2$ is a better match the query Q.

As shown in figure 2, the documents and queries are represented in the same space. Now, we can use the angle between the vectors as a proxy for




Suppose,
$t_1$ = Cat
$t_2$ = Dog
$t_3$ = Mouse

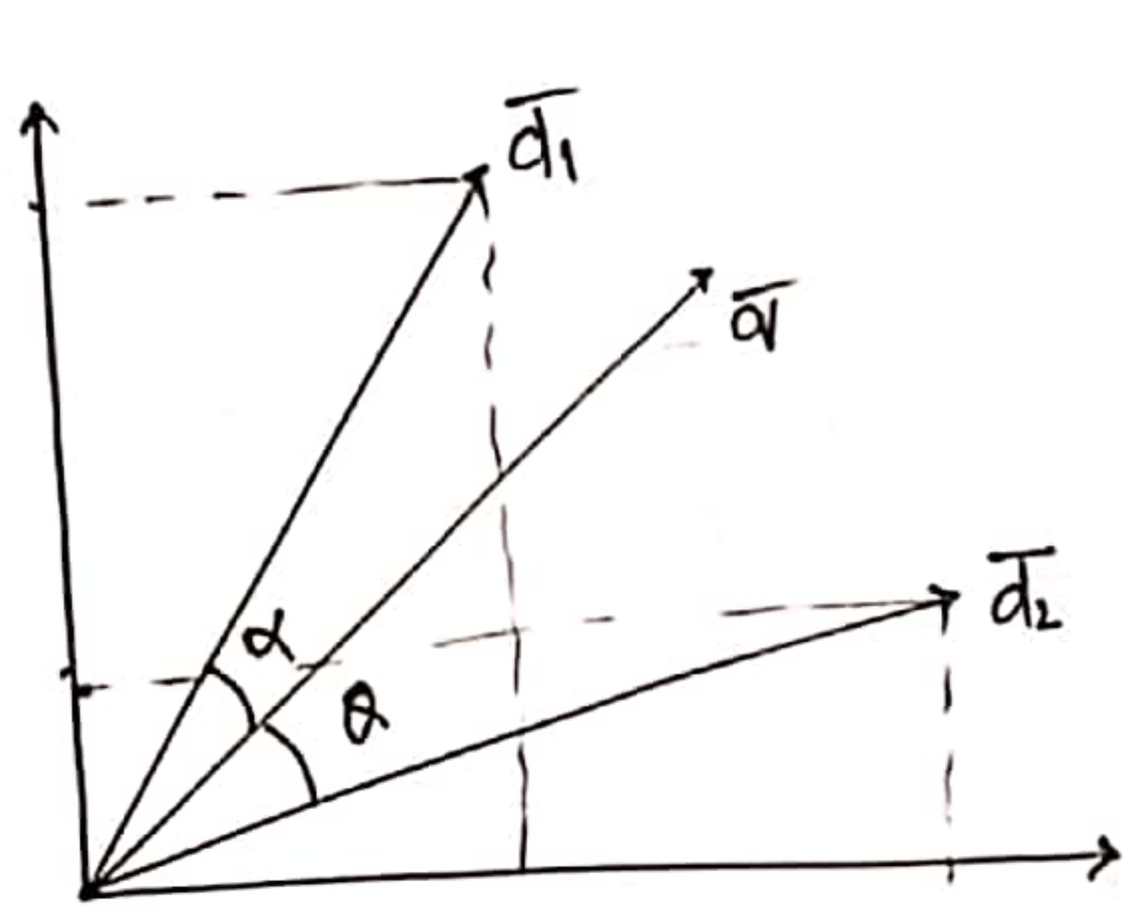
$d_1$ = "cat cat dog"
= <2,1,0>

$d_2$ = "dog dog cat mouse"
= <1,2,1>

$d_3$ = "cat dog"
= <1,1,0>

$t_1$, $t_2$, $t_3$ are the dimensions, where $d_1$, $d_2$, $d_3$ are documents.

Figure 1 : Vector Space Model




$D_1 = \vec{d_1}$
$D_2 = \vec{d_2}$
$q = \vec{q}$

similarity between $D_1$ an $q$ $\propto$ $\cos(\alpha)$
similarity beween $D_2$ and $q$ $\propto$ $\cos(\theta)$

Figure 2 : Document similarity using VSM

their similarity. Thus —

Similarity between $D_1$ and Q is proportional to the angle '$\alpha$' between them whereas $D_2$ and Q is proportional to angle '$\theta$' between them.

So, Cosine of the angles are considered. Thus. if cosine is smaller that means angle is smaller i.e similarity is larger and Viceversa.

Cosine measure is computed as the normalized dot product of two vectors.

$$\sigma(D,Q) = \frac{|D \cap Q|}{\sqrt{|D|\,|Q|}} = \frac{\Sigma(d_i q_i)}{\sqrt{\Sigma(d_i)^2}\ \sqrt{\Sigma(q_i)^2}}$$

A variant of cosine is Jaccard coefficient.

$$\sigma(D,Q) = \frac{|D \cap Q|}{|D \cup Q|}$$

We will look at an example of document similarity to understand easily.

## Example

Suppose, we have a document that is represented as Cat dog dog and cookie that represent as Cat dog mouse mouse. Thus ——

Dimensions are Cat, dog and mouse.

D = " Cat dog dog " = <1, 2, 0>

Q = " Cat dog mouse mouse " = <1, 1, 2>

Thus,

$$G(D,Q) = \frac{(1\times1) + (2\times1) + (0\times2)}{\sqrt{1^2+2^2+0^2}\ \sqrt{1^2+1^2+2^2}} = \frac{3}{\sqrt{5}\sqrt{6}} = 0.55$$

So, Q is 55% similiar to D.

As like document similarity, Vector space model is used in other information retrieval processes like filering, indexing etc.