Example (TD matrix, VSM model)

Let

$\boldsymbol{D} = \{\boldsymbol{D_1},\ldots,\boldsymbol{D_j},\ldots,\boldsymbol{D_m}\}$ denote the documents,

$\boldsymbol{T} = \{\boldsymbol{t_1},\ldots,\boldsymbol{t_i},\ldots,\boldsymbol{t_n}\}$ denote the terms constructed.

matrix TD = $(w_{ij})_{n\times m}$, $(i = 1,\ldots, n, j = 1,\ldots, m)$

where

$\boldsymbol{w_{ij}}$ the *weight* of term $\boldsymbol{t_i}$ in the document $\boldsymbol{D_j}$

$\boldsymbol{f_{ij}}$: the number of times term $\boldsymbol{t_i}$ occurs in document $\boldsymbol{D_j}$,

$\boldsymbol{m}$ is the number of documents

$\boldsymbol{F_i}$ is the number of documents where $\boldsymbol{t_i}$ occurs

$$TD := \begin{pmatrix} w_{1,1} & \vdots & \vdots & w_{1,j} & \vdots & \vdots & w_{1,m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{i,1} & \vdots & \vdots & w_{i,j} & \vdots & \vdots & w_{i,m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{n,1} & \vdots & \vdots & w_{n,j} & \vdots & \vdots & w_{n,m} \end{pmatrix}$$

Collection of book titles with the index terms

(n = 9 terms, m = 7 documents)

Documents:

D_1 = ″ Infant and Toddler First Aid ″

D_2 = ″ Babies and Children's Room (For your Home)″

D_3 = ″ Child Safety at Home ″

D_4 = ″Your Baby's Health and Safety: From Infant to Toddler ″

D_5 = ″ Baby Proofing Basics ″

D_6 = ″Your Guide to Easy Rust Proofing ″

D_7 = ″Beanie Babies Collectors Guide ″

Terms:

t_1 = ″Baby″

t_2 = ″Child″

t_3 = ″Guide″

t_4 = ″Health″

t_5 = ″Home″

t_6= ″Infant″

t_7 = ″Proofing″

t_8 = ″Safety″

t_9 = ″Toddler″

The *term-by-document* matrix

and the *term-by-query* are the following

using **binary weighting method**:

$$w_{ij} = \begin{cases} 1 & if \quad t_i \quad occurs\ in \quad D_j \\ 0 & otherwise \end{cases},$$

or **frequency weighting method**:

$$w_{ij} = f_{ij}.$$

e.g., f_{17}: the number of times term $\boldsymbol{t_1}$ *(Baby)* occurs in document $\boldsymbol{D_7}$ *(Beanie <u>Babies</u> Collectors <u>Guide</u>)*,

$$D := \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \qquad Q := \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

The *term-by-document* matrix
and the *term-by-query* are the following:

using *norm-tf*, length-normalized method:

$$w_{ij} = \frac{f_{ij}}{\sqrt{\sum_{k=1}^{n} f_{kj}^{2}}} .$$

D = ▮

$$q := \begin{pmatrix} 0 \\ 0.4472 \\ 0 \\ 0 \\ 0.4472 \\ 0.4472 \\ 0.4472 \\ 0.4472 \\ 0 \end{pmatrix}$$

Similarity measures

Cosine measure:

$$\sigma(\mathbf{w}_j, \mathbf{q}) = \frac{\sum_{i=1}^{n} w_{ij} q_i}{\sqrt{\sum_{i=1}^{n} w_{ij}^2 \cdot \sum_{i=1}^{n} q_i^2}}$$

Dice's coefficient:

$$\sigma(\mathbf{w}_j, \mathbf{q}) = 2 \cdot \frac{\sum_{i=1}^{n} w_{ij} q_i}{\sum_{i=1}^{n} (w_{ij} + q_i)}$$

Jaccard's coefficient:

$$\sigma(\mathbf{w}_j, \mathbf{q}) = \frac{\sum_{i=1}^{n} w_{ij} q_i}{\sum_{i=1}^{n} \left(\frac{w_{ij} + q_i}{2^{w_{ij} q_i}} \right)}$$

Doc.	**Similarity values (Rank)**		
	Cosine	**Jaccard**	**Dice**
D1	0.316 **(4.)**	0.092 **(4.)**	0.174 **(4.)**
D2	0.516 **(2.)**	0.142 **(2.)**	0.26 **(2.)**
D3	0.775 **(1.)**	0.224 **(1.)**	0.39 **(1.)**
D4	0.4 **(3.)**	0.094 **(3.)**	0.178 **(3.)**
D5	0.316 **(4.)**	0.092 **(4.)**	0.174 **(4.)**
D6	0.316 **(4.)**	0.092 **(4.)**	0.168 **(4.)**
D7	0	0	0

EXAMPLE (TD Matrix, VSM model)

Let the set T of index terms be

$T = \{t_1, t_2, t_3\} = \{$

t_1 = Bayes,

t_2 = probability,

t_3 = epistemology

$\}$.

Conceive the documents as sets of terms (together with their frequencies):

$D = \{D_1, D_2, D_3\}$, where

D_1 = { (Bayes, 1); (probability, 1); (epistemology, 0) }

D_2 = { (Bayes, 2); (probability, 1); (epistemology, 0) }

D_3 = { (Bayes, 3); (probability, 3); (epistemology, 3) }

Let the query Q be:

Q = {What is Bayesian epistemology? }

Let the query Q be (as a set of terms):

Q = {(t_1 = Bayes); (t_3 = epistemology) }

The *Term-by-Document matrix TD,*
and *term-by-query Q :*

$TD_{3\times3} = (w_{ij})$, where

- Using *Frequency Weighting Method*:

$$TD = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 3 \\ 0 & 0 & 3 \end{pmatrix} \quad Q = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

- Using *Binary Weigthing Method*

$$TD = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad Q = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

- Using *Binary Length Normalized Weighting Method*:

$$TD = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{3}}{3} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{3}}{3} \\ 0 & 0 & \frac{\sqrt{3}}{3} \end{pmatrix} \quad Q = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ 0 \\ \frac{\sqrt{2}}{2} \end{pmatrix}$$

Similarity values:

- *Dot Product:*

$Dot_1 = 0.5$; $Dot_2 = 0.5$; $Dot_3 = 0.816$

- *Cosine Measure:*

$Cosine_1 = 0.5$; $Cosine_2 = 0.5$; $Cosine_3 = 0.816$

- *Dice Measure:*

$Dice_1 = 0.354$; $Dice_2 = 0.354$; $Dice_3 = 0.519$

- *Jaccard Measure:*

$Jaccard_1 = 0.21$; $Jaccard_2 = 0.21$; $Jaccard_3 = 0.32$

EXAMPLE (different ranking order)

Let

$D = \{D_1, D_2, D_3, D_4, D_5, D_6\}$,

$T = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9\}$

Doc.	Similarity values **(Rank)**		
	Cosine	**Jaccard**	**Dice**
D1	0.676 **(3.)**	0.149 **(4.)**	0.276 **(4.)**
D2	0.845 **(1.)**	0.191 **(3.)**	0.347 **(2-3.)**
D3	0.632 **(4.)**	0.198 **(2.)**	0.347 **(2-3.)**
D4	0.775 **(2.)**	0.224 **(1.)**	0.39 **(1.)**
D5	0.258 **(6.)**	0.068 **(6.)**	0.13 **(6.)**
D6	0.316 **(5.)**	0.092 **(5.)**	0.173 **(5.)**