Exercise : Vector Space Model[1]

Consider a collection made of the 4 following documents (one document per line)
d1. John gives a book to Mary
d2. John who reads a book loves Mary
d3. who does John think Mary love ?
d4. John thinks a book is a good gift

Question 1

These documents are pre-processed using a stop-list and a stemmer. The resulting index is built to allow to apply vector-based queries. Give a (graphical or textual) representation of this index.

Index :

term t	N/df_t	→	d1 : tf_{t,d_1}	d2 : tf_{t,d_2}	d3 : tf_{t,d_3}	d4 : tf_{t,d_4}
book	4/3	→	d1 :1	d2 :1	d4 :1	
gift	4/1	→	d4 :1			
give	4/1	→	d1 :1			
good	4/1	→	d4 :1			
John	4/4	→	d1 :1	d2 :1	d3 :1	d4 :1
love	4/2	→	d2 :1	d3 :1		
Mary	4/3	→	d1 :1	d2 :1	d3 :1	
read	4/1	→	d2 :1			
think	4/2	→	d3 :1	d4 :1		

Question 2

We now focus on 3 terms belonging to the dictionary, namely book, love and Mary. Compute the tf − idf-based vector representation for the 4 documents in the collection (these vectors are normalized using the euclidian normalization).

$$\vec{d1} = \begin{cases} (book) & \frac{1*\log(4/3)}{D1} \\ (love) & 0 \\ (Mary) & \frac{1*\log(4/3)}{D1} \end{cases}$$

$$\vec{d3} = \begin{cases} (book) & 0 \\ (love) & \frac{1*\log(4/2)}{D3} \\ (Mary) & \frac{1*\log(4/3)}{D3} \end{cases}$$

1 From http://www.sfs.uni-tuebingen.de/~parmenti/doc/final.pdf

$$\vec{d2}=\begin{cases}(book) & \frac{1*\log(4/3)}{D2}\\ (love) & \frac{1*\log(4/2)}{D2}\\ (Mary) & \frac{1*\log(4/3)}{D2}\end{cases}$$

$$\vec{d4}=\begin{cases}(book) & \frac{1*\log(4/3)}{D4}\\ (love) & 0\\ (Mary) & 0\end{cases}$$

by $\log_{10}$ to the base 10.

Where

$D1=\sqrt{(\log(4/3)^2+0+\log(4/3)^2)}=0.4068$ = 0.1766

$D2=\sqrt{(\log(4/3)^2+\log(4/2)^2+\log(4/3)^2)}=0.8037$ =

$D3=\sqrt{(0+\log(4/2)^2+0+\log(4/3)^2)}=0.7505$

$D4=\sqrt{(\log(4/3)^2+0+0)}=0.2877$

This is log to the base e natural log.

Question 3

Consider the query "love Mary". Give the results of a ranked retrieval for this query. What document is considered to be the most relevant ?

$$\vec{q}=\begin{cases}(book) & 0\\ (love) & 1\\ (Mary) & 1\end{cases}$$

Ranking:

1. $s(\vec{q},\vec{d3})=0+\log\frac{(2)}{D3}+\log\frac{(4/3)}{D3}=1.3069$
2. $s(\vec{q},\vec{d2})=0+\log\frac{(2)}{D2}+\log\frac{(4/3)}{D2}=1.2204$
3. $s(\vec{q},\vec{d1})=0+0+\log\frac{(4/3)}{D1}=0.7071$
4. $s(\vec{q},\vec{d4})=0+0+0=0$

0.0156096879