INFORMATION RETRIEVAL

HOMEWORK - 2

Problem 1

Document 1 : You say goodbye, I say hello

term-count will be

you - 1

say - 1

goodbye - 1

i - 1

hello - 1

Document 2 : You say stop, I say go

you - 1

say - 2

stop - 1

i - 1

go - 1

Document 3 : Hello, hello, you say goodbye

hello - 2

you - 1

say - 1

goodbye - 1

Document 4 : I say yes, you say no

i - 1

say - 2

yes - 1

you - 1

no - 1

Q1 : say goodbye

say - 1

goodbye - 1

Q2 : you hello

you - 1

hello - 1

(a) Binary Term Matrix.

	you	say	hello	i	stop	goodbye	yes	no	go
Document 1	1	1	1	1	0	1	0	0	0
Document 2	1	1	0	1	1	0	0	0	1
Document 3	1	1	1	0	0	1	0	0	0
Document 4	1	1	0	1	0	0	1	1	0

(b) Raw Term Frequency

	you	say	hello	i	stop	goodbye	yes	no	go
Document 1	1	2	1	1	0	1	0	0	0
Document 2	1	2	0	1	1	0	0	0	1
Document 3	1	1	2	0	0	1	0	0	0
Document 4	1	2	0	1	0	0	1	1	0

c) Normalized Term Frequency

	you	say	hello	i	stop	goodbye	yes	no	go
Document 1	0.16	0.33	0.16	0.16	0	0.16	0	0	0
Document 2	0.16	0.33	0	0.16	0.16	0	0	0	0.16
Document 3	0.2	0.2	0.4	0	0	0.2	0	0	0
Document 4	0.16	0.33	0	0.16	0	0	0.16	0.16	0

d)

Inverse document frequency (idf)

$$= \left[\ln\left(N/(n_j+1)\right)+1\right]$$

Document 1 :

① For term 'you' :

$$idf = \left[\ln\left(N/(n_j+1)\right)+1\right]$$

$$= \left[\ln\left(4/(4+1)\right)+1\right]$$

$$= \left[\ln\left(\frac{4}{5}\right)+1\right] = -0.22+1$$

$$= \quad = 0.77$$

~~idf('you') = 0.47~~ idf('you') = 0.77

$$tf(\text{'you'}) = 1$$

$$tf\text{-}idf(\text{'you'}) = tf_{D1}(\text{'you'}) * idf_{D1}(\text{'you'})$$

$$= 1 * 0.7768$$

$$\boxed{tf\text{-}idf(\text{'you'})_{D1} = 0.7768}$$

2) $$idf(\text{'say'}) = \ln(4/5)+1 = 0.77$$

$$\boxed{idf(\text{'say'}) = 0.77}$$

$$tf\text{-}idf(\text{'say'}) = tf(\text{'say'}) * idf(\text{'say'})$$

$$\boxed{tf\text{-}idf(\text{'say'}) = \begin{matrix} 2 * 0.77 \\ 1.54 \end{matrix}}$$

3) $$idf(\text{'hello'}) = 1.28$$

$$tf\text{-}idf(\text{'hello'}) = tf(\text{'hello'}) * idf(\text{'hello'})$$

$$\boxed{tf\text{-}idf(\text{'hello'}) = \begin{matrix} 1 * 1.28 \\ 1.28 \end{matrix}}$$

4) $idf('i') = \ln(4/4)+1$

[illegible]

$$\boxed{idf('i') = 1}$$

$$tf\text{-}idf('i') = tf('i') * idf('i')$$

$$= 1 * 1$$

$$\boxed{tf\text{-}idf('i') = 1}$$

5) $idf('stop') = \ln(4/2)+1$

$$= 1.69$$

$$tf\text{-}idf('stop') = 0 * 1.69$$

$$\boxed{tf\text{-}idf('stop') = 0}$$

6) $idf('goodbye') = \ln(4/3)+1 \quad =$

$$= 1.28$$

$$tf\text{-}idf('goodbye') = 1 * 1.28$$

$$\boxed{tf\text{-}idf('goodbye') = 1.28}$$

7) $idf(\text{'yes'}) = \ln(4/2) + 1 = 1.69$

$tf\text{-}idf = 0 * 1.69 = 0$

$$\boxed{tf\text{-}idf = 0}$$

8) $idf(\text{'no'}) = \ln(4/2) + 1 = 1.69$

$tf\text{-}idf = 0 * 1.69$

$$\boxed{tf\text{-}idf(\text{'no'}) = 0}$$

9) $idf(\text{'go'}) = \ln(4/2) + 1 = 1.69$

$tf\text{-}idf(\text{'go'}) = 0 * 1.69$

$$\boxed{tf\text{-}idf(\text{'go'}) = 0}$$

[redacted]

[redacted]

d) tf-idf weights

	you	say	hello	i	stop	goodbye	yes	no	go
Document 1	0·77	1·54	1·28	1	0	1·28	0	0	0
Document 2	0·77	1·54	0	1	1·69	0	0	0	1·69
Document 3	0·77	0·77	2·56	0	0	1·28	0	0	0
Document 4	0·77	1·54	0	1	0	0	1·69	1·69	0

Problem 2

Q1 : Say goodbye

Q2 : you hello.

Q1 :

Say	1
goodbye	1

Q2 :

you	1
hello	1

a) Binary Term Matrix

	Say	goodbye	you	hello.
Q1	1	1	0	0
Q2	0	0	1	1

b) Raw Term Frequency.

	Say	good bye	you	hello.
Q1	1	1	0	0
Q2	0	0	1	1

Problem

Normalized Term Frequency

c)

	Say	goodbye	you	hello.
Q1	0.5	0.5	0	0
Q2	0	0	0.5	0.5

d) tf-idf weight.

	Say	you.	good bye	hello
Q1	0.77	0	1.28	0
Q2	0.	0.77	0	1.28

Q1 : tf - idf('Say') = 0.77 * 1 =0.77

Q1: tf - idf ('you') = tf ('you') * idf('you')
= 0 * 0.77 = 0.

Q1 : tf-idf ('goodbye') = 1.28 * 1 = 1.28

Q1 : tf-idf ('hello') = tf('hello') * idf('hello')
= 0 * 1.28 = 0.

Q2 : tf-idf('Say') = 0.77 * 0 = 0.
tf-idf('you') = 0.77 * 1 = 0.77
tf-idf('goodbye') = 1.28 * 0 = 0

tf-idf ('hello')
= 1.28 * 1
= 1.28.

Problem 3

(a) Inner Product

term 'Say' in Q1 = 0.77 * 1.54

+

term 'goodbye' in Q1 = 1.28 * 1.28.

in Document 1 = 2.82

$\boxed{Q1 \,\&\, D1 = 2.82}$

term 'you' & 'hello' in Q1 = 0

term 'Say' & 'goodbye' in Q2 = 0

term 'you' & 'hello' in Q2 = 0.77 * 0.77

+

in Document 2 0 * 1.28

= 0.59.

$\boxed{Q2 \,\&\, D2 = 0.59}$

b) **Cosine similarity**

$$\text{Cos}(d_i, d_j) = \frac{\sum_{k=1}^{n} w_{ik} w_{jk}}{|d_i||d_j|}$$

$|d_1|$ = Sum of squares of all tf-idf weights for Document 1

$$= \sqrt{(0.77)^2 + (1.54)^2 + (1.28)^2 + (1)^2 + (0)^2 + (1.28)^2 + 0^2 + 0^2 + 0^2}$$

$$= 2.705$$

$|d_2| = 3.123$ $\quad$ $|q_1| = 1.49$

$|d_3| = 3.066$ $\quad$ $|q_2| = 1.49.$

$|d_4| = 3.123$

$$\text{Cos}(d_1, q_1) = \frac{\text{Inner Product}}{|d_1||q_1|} = \frac{2.82}{2.705 \times 1.49}$$

$$= \frac{2.82}{4.03} = \boxed{0.699.}$$

Problem

$$\boxed{\cos(d_1, q_1) = 0.699}$$

$$\cos(d_2, q_1) = \frac{\text{Inner Product}}{|d_2||q_1|} = \frac{0.998}{3.12 * 1.49}$$

$$\boxed{\cos(d_2, q_1) = 0.21}$$

$$\cos(d_3, q_1) = \frac{\text{Inner Product}}{|d_3|\,|q_1|} = \frac{2.24}{3.066 * 1.49}$$

$$\boxed{\cos(d_3, q_1) = 0.49}$$

$$\cos(d_4, q_1) = \frac{\text{Inner Product}}{|d_4|\,|q_1|} = \frac{1.21}{3.123 * 1.49}$$

$$\boxed{\cos(d_4, q_1) = 0.24}$$

$$\boxed{\begin{aligned} \cos(d_1, q_2) &= 0.567 \\ \cos(d_2, q_2) &= 0.13 \\ \cos(d_3, q_2) &= 0.70 \\ \cos(d_4, q_2) &= 0.13 \end{aligned}}$$

Problem 4

For query q, we will Rank documents based on Cosine Similarity.

		Cos(di, qi)
Q1	D1	0.699
	D2	0.21
	D3	0.49
	D4	0.24

$\boxed{\text{Rank} = D_1, D_3, D_4, D_2}$.

Q2	D1	= 0.56
	D2	= 0.13
	D3	= 0.70
	D4	= 0.13

$\boxed{\text{Rank} = D_3, D_1, D_4, D_2}$

Observing The Ranking achieved here, we can say that, for query Q_1 and Q_2 Documents D_1 and D_3 were much relevant compared to D_2 and D_4