Web Search Engines – Problem Set 2

Narasimman Sairam (ns3184/N13296703)

Problem 1

Given the term document matrix:

Calculating f(t,d), w(t,d), vector d

Term: Walrus

O(t) = 2; c = 4; i(t) = 1 + log(4/2) = 2

	f(t,d)	w(t,d)	Vector d
Doc1	10	4.32	8.64
Doc2	0	0	0
Doc3	0	0	0
Doc4	10	4.32	8.64

Term: Carpenter

O(t) = 2; c = 4; i(t) = 1 + log(4/2) = 2

	f(t,d)	w(t,d)	Vector d
Doc1	8	4	8
Doc2	0	0	0
Doc3	40	6.32	12.64
Doc4	0	0	0

Term: bread

O(t) = 3; c = 4; i(t) = 1 + log(4/3) = 1.414

	f(t,d)	w(t,d)	Vector d
Doc1	4	3	5.656
Doc2	24	5.58	7.89
Doc3	0	0	0
Doc4	20	5.32	7.52

Term: butter

O(t) = 2; c = 4; i(t) = 1 + log(4/2) = 2

	f(t,d)	w(t,d)	Vector d
Doc1	1	1	2
Doc2	16	5	10
Doc3	0	0	0
Doc4	0	0	0

So, the document vectors are as follows:

	Doc1	Doc2	Doc3	Doc4
Walrus	8.64	0	0	8.64
Carpenter	8	0	12.64	0
Bread	5.65	7.89	0	7.52
Butter	2	10	0	0

Therefore, the normalized document vector is as follows:

	Doc1	Doc2	Doc3	Doc4
Walrus	0.654	0	0	0.754
Carpenter	0.605	0	1	0
Bread	0.427	0.619	0	0.656
Butter	0.151	0.785	0	0

Query:

1. “Walrus” - q<1,0,0,0>

	Sim(d,q)	Rank
Doc1	0.654	2
Doc2	0	3
Doc3	0	4
Doc4	0.754	1

2. “Walrus carpenter” - q<0.707,0.707,0,0>

	Sim(d,q)	Rank
Doc1	0.890	1
Doc2	0	4
Doc3	0.707	2
Doc4	0.533	3

3. “walrus bread butter” - q<0.57,0,0.57,0.57>

	Sim(d,q)	Rank
Doc1	0.702	3
Doc2	0.800	2
Doc3	0	4
Doc4	0.803	1

Problem 2:

A.

Sim(d1,d2) = 0.427*0.619 = 0.264
Sim(d1,d3) = 0.605*1 = 0.605
Sim(d1,d4) = 0.654*0.754 + 0.427*0.656 = 0.773

B.

Doc 1 : o(t) = 4; c = 4; i(t) = 1

	f(t,d)	w(t,d)	Vector d
Walrus	10	4.32	4.32
Carpenter	8	4	4
Bread	4	3	3
Butter	1	1	1

Doc 2 : o(t) = 2; c = 4; i(t) = 2

	f(t,d)	w(t,d)	Vector d
Walrus	0	0	0
Carpenter	0	0	0
Bread	24	5.58	11.16
Butter	16	5	10

Doc 3 : o(t) = 1; c = 4; i(t) = 3

	f(t,d)	w(t,d)	Vector d
Walrus	0	0	0
Carpenter	40	6.32	18.96
Bread	0	0	0
Butter	0	0	0

Doc 4 : o(t) = 2; c = 4; i(t) = 2

	f(t,d)	w(t,d)	Vector d
Walrus	10	4.32	8.64
Carpenter	0	0	0
Bread	20	5.32	10.64
Butter	0	0	0

So, the word vector is as follows:

	Walrus	Carpenter	Bread	Butter
Doc1	4.32	4	3	1
Doc2	0	0	11.16	10

Doc3	0	18.96	0	0
Doc4	8.64	0	10.64	0

So, the normalized vector is as follows:

	Walrus	Carpenter	Bread	Butter
Doc1	0.447	0.206	0.184	0.099
Doc2	0	0	0.685	0.996
Doc3	0	0.978	0	0
Doc4	0.014	0	0.653	0

sim(“bread”, “walrus”) = 0.447*0.184 + 0.014 * 0.653 = 0.091
sim(“bread”, “carpenter”) = 0.206 * 0.184 = 0.037
sim(“bread”, “butter”) =).184*0.099 + 0.685 * 0.996 = 0.700

Problem 3:

A. Invariance under irrelevant words

This property does not hold true because the document vectors can be different for both d and e while the query vector is same.

The calculation of the document vector depends on the number of terms and the length of the document as well. Just because f(t,d) = f(t,e) does not mean that the document vectors are same.

Since document vectors can be different, the similarity is a cross product of doc vector and query vector. Hence, the similarity values will be different.

For Example:

	Doc1	Doc2	Doc3	Doc4
Walrus	2	0	0	2
Carpenter	4	0	4	0
Bread	8	8	8	8
Butter	16	16	0	0

If we calculate the similarity of “Walrus” query with document 1 and document 2, we will get different values because the tf-idf values will be different.

B. Invariance under scaling

This property holds good for the ranking algorithm in Problem 1.

The idf reduces the weight of the vector which occurs frequently in all the documents or the complete collection. Even though a higher weight is given to the dimensions of the more verbose document; they are penalized by a factor of 1 + log(c/o(t))

So, the documents with more words that are repeating k times.

	Doc1	Doc2	Doc3	Doc4
Walrus	2	4	0	2
Carpenter	4	8	4	0
Bread	8	16	8	8
Butter	16	32	0	0

Here, when we calculate the similarity of “walrus” for document 1 and document 2, we would get the same value. This is also because the normalization values will be of the ratio ‘k’. Thus, we would get the same similarity.

C. **Order invariance under collection**

This property does not hold true for the following reasons:

a. The number of documents in each collection might be different
b. The number of documents in which each of the term is found will be different. O(t).

As these factors depends on the number of documents in the collection and the term frequency of the documents, the tf-idf calculation will result in different values and hence the ranking of a document in collection c may not be the same in a different collection d.

Problem 4

A.

N = 9; e = 0.3 ; f = 0.7

E = e / N = 0.3/9 = 0.033

A = 0.033 + 0.7 (0)
B = 0.033 + 0.7 (A/4 + C/3)
C = 0.033 + 0.7 (A/4 + B/2 + I/2)
D = 0.033 + 0.7 (A/4 + H)
E = 0.033 + 0.7 (A/4 + B/2 + C/3 + D/2 + F/2)
F = 0.033 + 0.7 (C/3 + E/2)
G = 0.033 + 0.7 (D/2)
H = 0.033 + 0.7 (E/2 + G + I/2)
I = 0.033 + 0.7 (F/2)

B. Page rank computation:

$$Q = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.175 & 0 & 0.233 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.175 & 0.35 & 0 & 0 & 0 & 0 & 0 & 0 & 0.35 \\ 0.175 & 0 & 0 & 0 & 0 & 0 & 0 & 0.7 & 0 \\ 0.175 & 0.35 & 0.233 & 0.35 & 0 & 0.35 & 0 & 0 & 0 \\ 0 & 0 & 0.233 & 0 & 0.35 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.35 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.35 & 0 & 0.7 & 0 & 0.35 \\ 0 & 0 & 0 & 0 & 0 & 0.35 & 0 & 0 & 0 \end{pmatrix}$$

```
>> a = zeros(9,1)

a =

     0
     0
     0
     0
     0
     0
     0
     0
     0

>> c = 0.033 * ones(9,1)

c =

    0.0330
    0.0330
    0.0330
    0.0330
    0.0330
    0.0330
    0.0330
    0.0330
    0.0330

>> q =
[0,0,0,0,0,0,0,0,0;0.175,0,0.233,0,0,0,0,0,0;0.175,0.35,0,0,0,0,0,0,0.
35;0.175,0,0,0,0,0,0,0.7,0;0.175,0.35,0.233,0.35,0,0.35,0,0,0;0,0,0.23
3,0,0.35,0,0,0,0;0,0,0,0.35,0,0,0,0,0;0,0,0,0,0.35,0,0.7,0,0.35;0,0,0,
0,0,0.35,0,0,0]


>> a=c+q*a
```

```
a =

    0.0330
    0.0330
    0.0330
    0.0330
    0.0330
    0.0330
    0.0330
    0.0330
    0.0330

>> a=c+q*a

a =

    0.0330
    0.0465
    0.0619
    0.0619
    0.0811
    0.0522
    0.0445
    0.0792
    0.0445
```

After 10 iterations;

```
a =

    0.0330
    0.0586
    0.0849
    0.1685
    0.1783
    0.1152
    0.0919
    0.1853
    0.0733
```

5. For e = 0.99

N = 9 ; e = 0.99 ; f = 0.01

E = e / N = 0.99/9 = 0.11

A = 0.11 + 0.01 (0)

B = 0.11 + 0.01 (A/4 + C/3)

C = 0.11 + 0.01 (A/4 + B/2 + I/2)

```
D = 0.11 + 0.01 (A/4 + H)
E = 0.11 + 0.01 (A/4 + B/2 + C/3 + D/2 + F/2)
F = 0.11 + 0.01 (C/3 + E/2)
G = 0.11 + 0.01 (D/2)
H = 0.11 + 0.01 (E/2 + G + I/2)
I = 0.11 + 0.01 (F/2)
```

$$Q = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.0025 & 0 & 0.0033 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.0025 & 0.005 & 0 & 0 & 0 & 0 & 0 & 0 & 0.005 \\ 0.0025 & 0 & 0 & 0 & 0 & 0 & 0 & 0.01 & 0 \\ 0.0025 & 0.005 & 0.0033 & 0.005 & 0 & 0.005 & 0 & 0 & 0 \\ 0 & 0 & 0.0033 & 0 & 0.005 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.005 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.005 & 0 & 0.01 & 0 & 0.005 \\ 0 & 0 & 0 & 0 & 0 & 0.005 & 0 & 0 & 0 \end{pmatrix}$$

```
>> a = zeros(9,1)

a =

     0
     0
     0
     0
     0
     0
     0
     0
     0

>> c = 0.11 * ones(9,1)

c =

    0.1100
    0.1100
    0.1100
    0.1100
    0.1100
    0.1100
    0.1100
    0.1100
    0.1100


>>  q =
[0,0,0,0,0,0,0,0,0;0.0025,0,0.0033,0,0,0,0,0,0;0.0025,0.005,0,0,0,0,0,0,
0,0.005;0.0025,0,0,0,0,0,0,0.01,0;0.0025,0.005,0.0033,0.005,0,0.005,0,
0,0;0,0,0.0033,0,0.005,0,0,0,0;0,0,0,0.005,0,0,0,0,0;0,0,0,0,0.005,0,0
.01,0,0.005;0,0,0,0,0,0.005,0,0,0]


>> a=c+q*a

a =

    0.1100
    0.1100
    0.1100
    0.1100
    0.1100
    0.1100
    0.1100
    0.1100
    0.1100

>> a=c+q*a

a =

    0.1100
    0.1106
    0.1114
    0.1114
    0.1123
    0.1109
    0.1106
    0.1122
    0.1106
```

After 10 iterations;

```
a =

    0.1100
    0.1106
    0.1114
    0.1114
    0.1123
    0.1109
    0.1106
    0.1122
    0.1106
```

For e = 0.01

N = 9 ; e = 0.01 ; f = 0.99

```
E = e / N = 0.01/9 = 0.001

A = 0.0011 + 0.99 (0)
B = 0.0011 + 0.99 (A/4 + C/3)
C = 0.0011 + 0.99 (A/4 + B/2 + I/2)
D = 0.0011 + 0.99 (A/4 + H)
E = 0.0011 + 0.99 (A/4 + B/2 + C/3 + D/2 + F/2)
F = 0.0011 + 0.99 (C/3 + E/2)
G = 0.0011 + 0.99 (D/2)
H = 0.0011 + 0.99 (E/2 + G + I/2)
I = 0.0011 + 0.99 (F/2)
```

$$Q \quad = \quad \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2475 & 0 & 0.33 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2475 & 0.495 & 0 & 0 & 0 & 0 & 0 & 0 & 0.495 \\ 0.2475 & 0 & 0 & 0 & 0 & 0 & 0 & .99 & 0 \\ 0.2475 & 0.495 & 0.33 & 0.495 & 0 & 0.495 & 0 & 0 & 0 \\ 0 & 0 & 0.33 & 0 & 0.495 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.495 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.495 & 0 & 0.99 & 0 & 0.495 \\ 0 & 0 & 0 & 0 & 0 & 0.495 & 0 & 0 & 0 \end{pmatrix}$$

```
>> a = zeros(9,1)

a =

     0
     0
     0
     0
     0
     0
     0
     0
     0

>> c = 0.0011 * ones(9,1)

c =

    0.0011
    0.0011
    0.0011
    0.0011
    0.0011
    0.0011
    0.0011
    0.0011
    0.0011

>>  q =
[0,0,0,0,0,0,0,0,0;0.2475,0,0.33,0,0,0,0,0,0;0.2475,0.495,0,0,0,0,0,0,
0.495;0.2475,0,0,0,0,0,0,0.99,0;0.2475,0.495,0.33,0.495,0,0.495,0,0,0;
0,0,0.33,0,0.495,0,0,0,0;0,0,0,0.495,0,0,0,0,0;0,0,0,0,0.495,0,0.99,0,
0.495;0,0,0,0,0,0.495,0,0,0]


>> a=c+q*a

a =

    0.0011
    0.0011
    0.0011
    0.0011
    0.0011
    0.0011
    0.0011
    0.0011
    0.0011

>> a=c+q*a

a =

    0.0011
    0.0017
    0.0025
    0.0025
    0.0034
    0.0020
    0.0016
    0.0033
    0.0016
```

After 10 iterations;

```
>> a=c+q*a

a =

    0.0011
    0.0044
    0.0094
    0.0481
    0.0405
    0.0234
    0.0239
```

```
0.0490
0.0123
```