Web Search Engines — Problem Set 2

Himaja Rachakonda
N14633788
hr970

March 7, 2016

1 Problem 1

Given term-document matrix -

	Doc1	Doc2	Doc3	Doc4
Walrus	10	0	0	10
Carpenter	8	0	40	0
Bread	4	24	0	20
Butter	1	16	0	0

$$w(t,d) = \begin{cases} 1 + \log_2 f(t,d) & iff(t,d) > 0 \\ 0 & iff(t,d) = 0 \end{cases}$$

$$i(t) = 1 + \log_2(c/o(t))$$

$$\vec{d} = w(t,d) * i(t)$$

Calculating $f(t,d)$,$w(t,d)$, $\vec{d}$ for each of the terms given -

Walrus $o(t) = 2$, $c = 4$, $i(t) = 2$

	$f(t,d)$	$w(t,d)$	$\vec{d}$
Doc1	10	4.32	8.64
Doc2	0	0	0
Doc3	0	0	0
Doc4	10	4.32	8.64

Carpenter $o(t) = 2$, $c = 4$, $i(t) = 2$

	$f(t,d)$	$w(t,d)$	$\vec{d}$
Doc1	8	4	8
Doc2	0	0	0
Doc3	40	6.32	12.64
Doc4	0	0	0

Bread $o(t) = 3$, $c = 4$, $i(t) = \log_2(4/3)$

	$f(t,d)$	$w(t,d)$	$\vec{d}$
Doc1	4	3	5.656
Doc2	24	5.58	7.89
Doc3	0	0	0
Doc4	20	5.32	7.52

Butter $o(t) = 2$, $c = 4$, $i(t) = 1$

	$f(t,d)$	$w(t,d)$	$\vec{d}$
Doc1	1	1	2
Doc2	16	5	10
Doc3	0	0	0
Doc4	0	0	0

So, the Document vectors with each of these terms as a dimension is 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

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

1.1 Query - Document Rankings

Query - "Walrus" $\vec{q} =< 1, 0, 0, 0 >$

	$sim(\vec{d}, \vec{q})$	Rank
Doc1	0.654	2
Doc2	0	3
Doc3	0	3
Doc4	0.754	1

Query - "Walrus Carpenter" $\vec{q} =$ <0.707,0.707,0,0>

	$sim(\vec{d}, \vec{q})$	Rank
Doc1	0.89	1
Doc2	0	4
Doc3	0.707	2
Doc4	0.533	3

Query - "Walrus Bread Butter" $\vec{q} =$ <0.57,0,0.57,0.57>

	$sim(\vec{d}, \vec{q})$	Rank
Doc1	0.702	3
Doc2	0.800	2
Doc3	0	4
Doc4	0.803	1

2 Problem 2

2.1 Document Similarity

$sim(\vec{d_1}, \vec{d_2})$ = 0.427*0.619 = 0.264
$sim(\vec{d_1}, \vec{d_3})$ = 0.605 * 1 = 0.605
$sim(\vec{d_1}, \vec{d_4})$ = 0.654*0.754 + 0.427*0.656 = 0.773

2.2 Word Similarity

Doc1 $o(t) = 4$, $c = 4$, $i(t) = 1$

	$f(t,d)$	$w(t,d)$	$\vec{d}$
Walrus	10	4.32	4.32
Carpenter	8	4	4
Bread	4	3	3
Butter	1	1	1

Doc2 $o(t) = 2$, $c = 4$, $i(t) = 2$

	$f(t,d)$	$w(t,d)$	$\vec{d}$
Walrus	0	0	0
Carpenter	0	0	0
Bread	24	5.58	11.16
Butter	16	5	10

Doc3 $o(t) = 1$, $c = 4$, $i(t) = 3$

	$f(t,d)$	$w(t,d)$	$\vec{d}$
Walrus	0	0	0
Carpenter	40	6.32	18.96
Bread	0	0	0
Butter	0	0	0

Doc4 $o(t) = 2$, $c = 4$, $i(t) = 2$

	$f(t,d)$	$w(t,d)$	$\vec{d}$
Walrus	10	4.32	8.64
Carpenter	0	0	0
Bread	20	5.32	10.64
Butter	0	0	0

The cumulative word-document matrix 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

The normalized vector is as follows -

	Walrus	Carpenter	Bread	Butter
Doc1	0.447	0.206	0.184	0.01
Doc2	0	0	0.685	0.996
Doc3	0	0.978	0	0
Doc4	0.014	0	0.653	0

Similairty of word bread with other words -

$sim(\vec{bread}, \vec{walrus})$ = 0.447*0.184 + 0.014 * 0.653 = 0.091
$sim(\vec{bread}, \vec{Carpenter})$ = 0.206 * 0.184 = 0.037
$sim(\vec{bread}, \vec{Butter})$ = 0.184*0.099 + 0.685 * 0.996 = 0.700

3 Problem 3

3.1 Property A: Invariance under irrelevant words

This Property does not hold as the document vectors can be different for both 'd' and 'e' even when query vector is same. The similarity measure is given by the formulae -

$$sim(\vec{d}, \vec{q}) = \frac{\vec{d} \cdot \vec{q}}{||d|| ||q||}$$

Here, the document vector may contain different weights to other words even the weight of the query term may be the same for both the documents.

Example : Let the query term be "search" for the following the term-document matrix. The term search has the same $f(t, d)$ for the documents D1 and D4 -

	Doc1	Doc2	Doc3	Doc4
search	8	0	0	8
web	16	0	16	0
engines	32	32	32	32
class	64	64	0	0

After computing the $w(t, d)$ and $i(t)$ the normalized term-document matrix with tf-idf values is as follows -

	Doc1	Doc2	Doc3	Doc4
search	0.4	0	0	0.8
web	0.50	0	0.85	0
engines	0.30	0.39	0.51	0.6
class	0.70	0.92	0	0

The query vector for the term "search" will be $\vec{q}$ = <1,0,0,0>. The similarity of $\vec{d_1}$ and $\vec{d_2}$ with $\vec{q}$ will be - $sim(\vec{d_1}, \vec{q})$ = 0.4 and $sim(\vec{d_2}, \vec{q})$ = 0.8

Therefore, we can see that this property does not hold true.

3.2 Property B: Invariance under scaling

This property holds good for the ranking algorithm in Problem 1. The inverse document frequency 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 $1 + \log(c/o(t))$

Example: Consider the followinf term-document matrix with the terms in $f(t, Doc1) = 2 * f(t, Doc3)$

	Doc1	Doc2	Doc3	Doc4
search	8	0	16	0
web	16	0	32	0
engines	32	8	64	2
class	16	0	32	0

After computing $w(t,d)$ and $i(t)$ we have the following normalized tf-idf matrix -

	Doc1	Doc2	Doc3	Doc4
search	0.46	0	0.48	0
web	0.57	0	0.57	0
engines	0.34	1	0.34	1
class	0.57	0	0.574	0

3.3 Property C: Order invariance under Collection

This property does not always hold true as the ranking depends on tf-idf formulation and inverse document frequency. This depends on the two constants -

1. The number of documents in each collection (c)
2. The number of documents in which each term is found. $(o(t))$

if the constant $c/o(t)$ increases, the ranking may be higher, and lower otherwise.

4 Problem 4

4.1 $N = 9,\ e = 0.3,\ f = 1 - e \Rightarrow f = 1 - 0.3 \Rightarrow f = 0.7,$ $E = (e/N) \Rightarrow E = 0.033$

$$\begin{array}{rcl}
A & = & 0.033 + 0.7(0) \\
B & = & 0.033 + 0.7(A/4 + C/3) \\
C & = & 0.033 + 0.7(A/4 + I/2 + B/2) \\
D & = & 0.033 + 0.7(A/4 + H/1) \\
E & = & 0.033 + 0.7(A/4 + B/2 + C/3 + F/2 + D/2) \\
F & = & 0.033 + 0.7(C/3 + E/2) \\
G & = & 0.033 + 0.7(D/2) \\
H & = & 0.033 + 0.7(E/2 + G/1 + I/2) \\
I & = & 0.033 + 0.7(F/2)
\end{array}$$

4.2 Page Rank computation

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

To solve these system of equations, we represent these in the form $\vec{c} = B\vec{p}$ Therefore the system of equations can be represented as:

$$\begin{aligned} A &= 0.033 \\ -0.175A + B - 0.233C &= 0.033 \\ -0.175A - 0.35B + C - 0.35I &= 0.033 \\ -0.175A + D - 0.7H &= 0.033 \\ -0.175A - 0.35B - 0.233C - 0.35D + E - 0.35F &= 0.033 \\ -0.233C - 0.35E + F &= 0.033 \\ -0.35D + G &= 0.033 \\ -0.35E - 0.7G + H - 0.35I &= 0.033 \\ -0.35F + I &= 0.033 \end{aligned}$$

```
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;
0,0,0.233,0,0.35,0,0,0,0;
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];

b = eye(9) - q;

c = ones(9,1);

c = c * 0.033;

p = b \ c;

p =

    0.0330
    0.0586
```

```
0.0849
0.1686
0.1784
0.1152
0.0920
0.1855
0.0733
```

5 Problem 5

5.1 $N = 9$, $e = 0.99$, $f = 1 - e \Rightarrow f = 1 - 0.99 \Rightarrow f = 0.01$, $E = (e/N) \Rightarrow E = 0.11$

$$\begin{array}{rcl}
A & = & 0.11 + 0.01(0) \\
B & = & 0.11 + 0.01(A/4 + C/3) \\
C & = & 0.11 + 0.01(A/4 + I/2 + B/2) \\
D & = & 0.11 + 0.01(A/4 + H/1) \\
E & = & 0.11 + 0.01(A/4 + B/2 + C/3 + F/2 + D/2) \\
F & = & 0.11 + 0.01(C/3 + E/2) \\
G & = & 0.11 + 0.01(D/2) \\
H & = & 0.11 + 0.01(E/2 + G/1 + I/2) \\
I & = & 0.11 + 0.01(F/2)
\end{array}$$

5.2 Page Rank Computaion

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

To solve these system of equations, we represent these in the form $\vec{c} = B\vec{p}$ Therefore the system of equations can be represented as:

$$\begin{array}{rcl}
A & = & 0.11 \\
-0.0025A + B - 0.0033C & = & 0.11 \\
-0.0025A - 0.005B + C - 0.005I & = & 0.11 \\
-0.0025A + D - 0.01H & = & 0.11 \\
-0.0025A - 0.005B - 0.0033C - 0.005D + E - 0.005F & = & 0.11 \\
-0.0033C - 0.005E + F & = & 0.11 \\
-0.005D + G & = & 0.11 \\
-0.005E - 0.01G + H - 0.005I & = & 0.11 \\
-0.005F + I & = & 0.11
\end{array}$$

```
q = [0,0,0,0,0,0,0,0,0;
0.0025,0,0.0033,0,0,0,0,0,0;
0.0025,0.005,0,0,0,0,0,0,0.005;
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];

b = eye(9) - q;

c = ones(9,1);

c = c * 0.11;

p = b \ c

p =

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

5.3 $N = 9$, $e = 0.01$, $f = 1 - e \Rightarrow f = 1 - 0.01 \Rightarrow f = 0.99$, $E = (e/N) \Rightarrow E = 0.001$

$$\begin{aligned}
A &= 0.001 + 0.99(0) \\
B &= 0.001 + 0.99(A/4 + C/3) \\
C &= 0.001 + 0.99(A/4 + I/2 + B/2) \\
D &= 0.001 + 0.99(A/4 + H/1) \\
E &= 0.001 + 0.99(A/4 + B/2 + C/3 + F/2 + D/2) \\
F &= 0.001 + 0.99(C/3 + E/2) \\
G &= 0.001 + 0.99(D/2) \\
H &= 0.001 + 0.99(E/2 + G/1 + I/2) \\
I &= 0.001 + 0.99(F/2)
\end{aligned}$$

5.4 Page Rank Computation

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2475 & 0 & 0.33 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2475 & 0.495 & 0 & 0 & 0 & 0 & 0 & 0 & 0.495 \\ 0.2475 & 0 & 0 & 0 & 0 & 0 & 0 & 0.99 & 0 \\ 0.2475 & 0.495 & 0.33 & 0.495 & 0 & 0.495 & 0 & 0 & 0 \\ 0 & 0 & 0.33 & 0 & 0.495 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.495 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.495 & 0 & 0.99 & 0 & 0.495 \\ 0 & 0 & 0 & 0 & 0 & 0.495 & 0 & 0 & 0 \end{bmatrix}$$

To solve these system of equations, we represent these in the form $\vec{c} = B\vec{p}$ Therefore the system of equations can be represented as:

$$\begin{aligned} A &= 0.001 \\ -0.2475A + B - 0.33C &= 0.001 \\ -0.2475A - 0.495B + C - 0.495I &= 0.001 \\ -0.2475A + D - 0.99H &= 0.001 \\ -0.2475A - 0.495B - 0.33C - 0.495D + E - 0.495F &= 0.001 \\ -0.33C - 0.495E + F &= 0.001 \\ -0.495D + G &= 0.001 \\ -0.495E - 0.99G + H - 0.495I &= 0.001 \\ -0.495F + I &= 0.001 \end{aligned}$$

```
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];

b = eye(9) - q;

c = ones(9,1);

c = c * 0.001;

p = b \ c

p =

    1.0000
   11.4695
```

```
 30.9758
215.7781
171.5447
 96.1366
107.8101
216.6975
 48.5876
```