Solutions for Homework 2

IIR Book:

Exercise 1.2 (0.5')

Consider these documents:

Doc 1 breakthrough drug for schizophrenia

Doc 2 new schizophrenia drug

Doc 3 new approach for treatment of schizophrenia

Doc 4 new hopes for schizophrenia patients

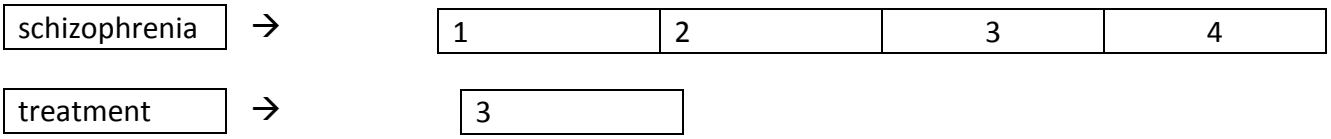
a. Draw the term-document incidence matrix for this document collection.
b. Draw the inverted index representation for this collection, as in Figure 1.3 (page 6).

a. Term-document incidence matrix

	Doc1	Doc2	Doc3	Doc4
approach	0	0	1	0
breakthrough	1	0	0	0
drug	1	1	0	0
for	1	0	1	1
hopes	0	0	0	1
new	0	1	1	1
of	0	0	1	0
patients	0	0	0	1
schizophrenia	1	1	1	1
treatment	0	0	1	0

b. inverted index representation for this collection (change the order between "hopes" and "for")

approach	→	3
breakthrough	→	1
drug	→	1 → 2
for	→	1 → 3 → 4
hopes	→	4
new	→	2 → 3 → 4
of	→	3
patients	→	4



Exercise 1.3 (0.5')

For the document collection shown in Exercise 1.2, what are the returned results for these queries:

a. schizophrenia AND drug

b. for AND NOT(drug OR approach)

a. Doc1, Doc 2

b. Doc 4

Exercise 4.11 (1')

Apply MapReduce to the problem of counting how often each term occurs in a set of files. Specify map and reduce operations for this task. Write down an example along the lines of Figure 4.6. (should follow the example in Figure 4.6).

Method 1:

Schema:

map: input -> list(k, v)

reduce: (k, list(v)) -> output

Instantiation of the schema for term counting

map: a set of files -> list(term, 1)

reduce: <(term1, 1), (term2, 1), (term3, 1)…> -> list(term, total count)

Example for term counting

map: d1: I hear, I forget. d2: I see, I remember. -> <I, 1> <hear, 1><I, 1><forget 1> <I, 1><see, 1><I, 1><remember 1>

reduce: <I,(1,1,1,1)><hear, 1><forget, 1><see, 1><remember, 1> -> <I, 4><hear, 1><forget, 1><see, 1><remember, 1>

Method 2:

Schema:

map: input -> list(k, v)

reduce: (k, list(v)) -> output

Instantiation of the schema for term counting

map: a set of files -> list(term, count in one file)

reduce: <(term1, count1), (term2, count2), (term3, count3)…> -> list(term, total count)

Example for term counting

map: d1: I hear, I forget. d2: I see, I remember. -> <I, 2> <hear, 1><forget 1> <I, 2><see, 1><remember 1>

reduce: <I,(2, 2)><hear, 1><forget, 1><see, 1><remember, 1> ->
<I, 4><hear, 1><forget, 1><see, 1><remember, 1>

Exercise 6.10 (0.5')

Consider the table of term frequencies for 3 documents denoted Doc1, Doc2, Doc3 in Figure 6.9. Compute the tf-idf weights for the terms car, auto, insurance, best, for eachdocument, using the idf values from Figure 6.8.

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

▶ Figure 6.9 Table of tf values for Exercise 6.10.

term	df_t	idf_t
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

▶ Figure 6.8 Example of idf values. Here we give the idf's of terms with various frequencies in the Reuters collection of 806,791 documents.

Solution 1(raw term frequency weighting)

	Doc1	Doc2	Doc3
car	44.55	6.6	39.6
auto	6.24	68.64	0
insurance	0	53.46	46.98
best	21	0	25.5

Solution 2(log term frequency weighting)

	Doc1	Doc2	Doc3
car	4.0118	2.6434	3.9273
auto	3.0724	5.2385	0
insurance	0	4.08	3.9891
best	3.2192	0	3.3457

Exercise 6.17 (1')

With term weights as computed in Exercise 6.15, rank the three documents by computed score for the query car insurance, for each of the following cases of term

weighting in the query:
1. The weight of a term is 1 if present in the query, 0 otherwise.
2. Euclidean normalized idf.

Solution 1:
Normalize the raw tf-idf weights computed in Ex. 6.10, we get

	Doc1	Doc2	Doc3
car	0.8974	0.0756	0.5953
auto	0.1257	0.7867	0
insurance	0	0.6127	0.7062
best	0.4230	0	0.3833

1. The query is q = [1, 0, 1, 0]

score(q, doc1)= 0.8974, score(q, doc2) = 0.6883, score(q, doc3) = 1.3015
Ranking: doc3, doc1, doc2
2. The query is q = [1.62, 0, 1.61, 0]

score(q, doc1) = 1.4807, score(q, doc2) = 1.1174, score(q, doc3) = 2.1262
Ranking: doc3, doc1, doc2

Solution 2:
Normalize the log tf-idf weights computed in Ex. 6.10, we get

	Doc1	Doc2	Doc3
car	0.6696	0.3699	0.6022
auto	0.5128	0.7330	0
insurance	0	0.5709	0.6117
best	0.5375	0	0.5130

1. The query is q = [1, 0, 1, 0]

score(q, doc1)= 0.6696, score(q, doc2) = 0.9408, score(q, doc3) = 1.2139
Ranking: doc3, doc2, doc1
2. The query is q = [1.62, 0, 1.61, 0]

score(q, doc1) = 1.1048, score(q, doc2) = 1.5351, score(q, doc3) = 1.9846
Ranking: doc3, doc2, doc1

Chapter 5 of MMDS Textbook:

5.1.1 (0.5')

The transition matrix for the graph is:

$$\begin{bmatrix} 1/3 & 1/2 & 0 \\ 1/3 & 0 & 1/2 \\ 1/3 & 1/2 & 1/2 \end{bmatrix}$$

By equation method ($M\lambda = \lambda$), we get the result $\lambda = \left[\frac{3}{13}, \frac{4}{13}, \frac{6}{13}\right]^T$

By iteration method, we get the following list*:

$$\begin{bmatrix}0.3333\\0.3333\\0.3333\end{bmatrix},\begin{bmatrix}0.2777\\0.2777\\0.4444\end{bmatrix},\begin{bmatrix}0.2314\\0.3148\\0.4537\end{bmatrix},\begin{bmatrix}0.2345\\0.3040\\0.4614\end{bmatrix},\begin{bmatrix}0.2301\\0.3088\\0.4609\end{bmatrix}\ldots\ldots\begin{bmatrix}0.2307\\0.3076\\0.4615\end{bmatrix}$$

5.1.2 (0.5')

The iteration process is :

$$\begin{aligned}\mathrm{v}' &= \beta\mathrm{M}\mathrm{v} + (1-\beta)\mathrm{e}/\mathrm{n}\\ &= \begin{bmatrix}4/15 & 2/5 & 0\\ 4/15 & 0 & 2/5\\ 4/15 & 2/5 & 2/5\end{bmatrix}\mathrm{v} + \begin{bmatrix}1/15\\1/15\\1/15\end{bmatrix}\end{aligned}$$

If we solve this equation directly, we get $\mathrm{v} = \left[\frac{7}{27}, \frac{25}{81}, \frac{35}{81}\right]^{\mathrm{T}}$

By iteration method, we get the following iteration list*:

$$\begin{bmatrix}0.3333\\0.3333\\0.3333\end{bmatrix},\begin{bmatrix}0.2888\\0.2888\\0.4222\end{bmatrix},\begin{bmatrix}0.2592\\0.3125\\0.4281\end{bmatrix},\begin{bmatrix}0.2608\\0.3070\\0.4320\end{bmatrix},\begin{bmatrix}0.2590\\0.3090\\0.4318\end{bmatrix}\ldots\ldots\begin{bmatrix}0.2592\\0.3086\\0.4320\end{bmatrix}$$

5.1.3 (1')

The matrix A for the graph is:

$$\beta\begin{bmatrix}0 & \frac{1}{\mathrm{n}} & \cdots & \frac{1}{\mathrm{n}} & 0\\ \frac{1}{\mathrm{n}} & 0 & \cdots & \frac{1}{\mathrm{n}} & 0\\ \vdots & \vdots & \ddots & \vdots & 0\\ \frac{1}{\mathrm{n}} & \frac{1}{\mathrm{n}} & \frac{1}{\mathrm{n}} & 0 & 0\\ \frac{1}{\mathrm{n}} & \frac{1}{\mathrm{n}} & \frac{1}{\mathrm{n}} & \frac{1}{\mathrm{n}} & 0\end{bmatrix} + (1-\beta)\begin{bmatrix}\frac{1}{n+1} & \cdots & \frac{1}{n+1} & \frac{1}{(n+1)(1-\beta)}\\ \vdots & \ddots & \vdots & \vdots\\ \frac{1}{n+1} & \cdots & \frac{1}{n+1} & \frac{1}{(n+1)(1-\beta)}\end{bmatrix}$$

Easy to see that all nodes in the clique have the same PageRank value, so we suppose vector v to be [x, x, …, x, y]T, where x is PageRank of node in the clique, and y represents the additional node outside the clique. Then we get the equation

$$\begin{cases}\mathrm{x} = \dfrac{\beta x(n-1)}{n} + \dfrac{nx(1-\beta)}{n+1} + \dfrac{y}{n+1}\\ \mathrm{n}*\mathrm{x} + \mathrm{y} = 1\end{cases}$$

Solve the equation, we get

$$\begin{cases}\mathrm{x} = \dfrac{\mathrm{n}}{\mathrm{n}^2 + \mathrm{n} + \beta}\\ \mathrm{y} = \dfrac{\mathrm{n}+\beta}{\mathrm{n}^2 + \mathrm{n} + \beta}\end{cases}$$

And the corresponding vector will be given.

5.1.6 (0.5')

There will be only one node, the head node with a self-direction, and PageRank for this node is 1. PageRank for all the remaining nodes will be 1/2.

5.2.2 (1')

a)

Source	Degree	Destinations
A	3	B, C, D
B	2	A, D
C	1	E
D	2	B, C

b)

Source	Degree	Destinations
a	3	a, b, c
b	2	a, c
c	2	b, c

5.2.3 (1')

The four-node graph is divided into four 2-by-2 blocks (M11, M12, M21, M22).

M11:

Source	Degree	Destinations
A	3	B
B	2	A

M12:

Source	Degree	Destinations
D	2	B

M21:

Source	Degree	Destinations
A	3	C, D
B	2	D

M22:

Source	Degree	Destinations
D	2	C

5.3.1 (1')

The transition matrix of Figure 5.15 is:

$$\begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

Suppose β is 0.8

a)

$$\mathrm{v}' = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} \mathrm{v} + \begin{bmatrix} 1/5 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

If we solve this equation directly, we get $\mathrm{v} = [\frac{3}{7}, \frac{4}{21}, \frac{4}{21}, \frac{4}{21}]^{\mathrm{T}}$

By iteration method, we get the following list is:

$$\begin{bmatrix}1.0\\0.0\\0.0\\0.0\end{bmatrix}, \begin{bmatrix}0.2\\0.2666\\0.2666\\0.2666\end{bmatrix}, \begin{bmatrix}0.52\\0.16\\0.16\\0.16\end{bmatrix}, \begin{bmatrix}0.392\\0.2026\\0.2026\\0.2026\end{bmatrix}, \begin{bmatrix}0.4432\\0.1856\\0.1856\\0.1856\end{bmatrix}, \begin{bmatrix}0.4227\\0.1924\\0.1924\\0.1924\end{bmatrix} \cdots\cdots \begin{bmatrix}0.4285\\0.1904\\0.1904\\0.1904\end{bmatrix}$$

b)

$$\mathrm{v}' = \begin{bmatrix}0 & 2/5 & 4/5 & 0\\4/15 & 0 & 0 & 2/5\\4/15 & 0 & 0 & 2/5\\4/15 & 2/5 & 0 & 0\end{bmatrix}\mathrm{v} + \begin{bmatrix}1/10\\0\\1/10\\0\end{bmatrix}$$

If we solve this equation directly, we get $\mathrm{v} = [\frac{27}{70}, \frac{6}{35}, \frac{19}{70}, \frac{6}{35}]^{\mathrm{T}}$

By iteration method, we get the following list is:

$$\begin{bmatrix}0.5\\0.0\\0.5\\0.0\end{bmatrix}, \begin{bmatrix}\\\\\\\end{bmatrix}, \begin{bmatrix}0.5\\0.1333\\0.2333\\0.1333\end{bmatrix}, \begin{bmatrix}0.34\\0.1866\\0.2866\\0.1866\end{bmatrix}, \begin{bmatrix}0.404\\0.1653\\0.2653\\0.1653\end{bmatrix}, \begin{bmatrix}0.3784\\0.1738\\0.2738\\0.1738\end{bmatrix} \cdots\cdots \begin{bmatrix}0.3857\\0.1714\\0.2714\\0.1714\end{bmatrix}$$

5.5.1 (1')

The link matrix for Figure 5.1 is:

$$\mathrm{L} = \begin{bmatrix}0 & 1 & 1 & 1\\1 & 0 & 0 & 1\\1 & 0 & 0 & 0\\0 & 1 & 1 & 0\end{bmatrix}$$

Initialize a and h to be all 1's. given two equations

$$\mathrm{h} = \mathrm{La}$$
$$\mathrm{a} = \mathrm{L}^{\mathrm{T}}\mathrm{h}$$

By iteration method, we get the following list is:

$$\mathrm{h}: \begin{bmatrix}1\\1\\1\\1\end{bmatrix} \begin{bmatrix}1.0\\0.6666\\0.3333\\0.6666\end{bmatrix} \begin{bmatrix}1.0\\0.6666\\0.3333\\0.6666\end{bmatrix} \begin{bmatrix}1.0\\0.5333\\0.2\\0.6666\end{bmatrix} \begin{bmatrix}1.0\\0.5333\\0.2\\0.6666\end{bmatrix} \begin{bmatrix}1.0\\0.4657\\0.1506\\0.6849\end{bmatrix} \cdots\cdots \begin{bmatrix}1.0\\0.3919\\0.1027\\0.7108\end{bmatrix}$$

$$\mathrm{a}: \begin{bmatrix}1\\1\\1\\1\end{bmatrix} \begin{bmatrix}1\\1\\1\\1\end{bmatrix} \quad \begin{bmatrix}0.6\\1\\1\\1\end{bmatrix} \quad \begin{bmatrix}0.6\\1\\1\\1\end{bmatrix} \quad \begin{bmatrix}0.44\\1\\1\\0.92\end{bmatrix} \quad \begin{bmatrix}0.44\\1\\1\\0.92\end{bmatrix} \quad \cdots\cdots \begin{bmatrix}0.2891\\1.0\\1.0\\0.8136\end{bmatrix}$$