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

<table>
<thead>
<tr>
<th>Term</th>
<th>Doc1: 0</th>
<th>Doc2: 0</th>
<th>Doc3: 1</th>
<th>Doc4: 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>approach</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>breakthrough</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>drug</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>for</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>hopes</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>new</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>of</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>patients</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>schizophrenia</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>treatment</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

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> <forget 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> 
reduce: <I, 2><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 each document, using the idf values from Figure 6.8.

Solution 1 (raw term frequency weighting)

<table>
<thead>
<tr>
<th>term</th>
<th>Doc1</th>
<th>Doc2</th>
<th>Doc3</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>44.55</td>
<td>6.6</td>
<td>39.6</td>
</tr>
<tr>
<td>auto</td>
<td>6.24</td>
<td>68.64</td>
<td>0</td>
</tr>
<tr>
<td>insurance</td>
<td>0</td>
<td>53.46</td>
<td>46.98</td>
</tr>
<tr>
<td>best</td>
<td>21</td>
<td>0</td>
<td>25.5</td>
</tr>
</tbody>
</table>

Solution 2 (log term frequency weighting)

<table>
<thead>
<tr>
<th>term</th>
<th>Doc1</th>
<th>Doc2</th>
<th>Doc3</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>4.0118</td>
<td>2.6434</td>
<td>3.9273</td>
</tr>
<tr>
<td>auto</td>
<td>3.0724</td>
<td>5.2385</td>
<td>0</td>
</tr>
<tr>
<td>insurance</td>
<td>0</td>
<td>4.08</td>
<td>3.9891</td>
</tr>
<tr>
<td>best</td>
<td>3.2192</td>
<td>0</td>
<td>3.3457</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Doc1</th>
<th>Doc2</th>
<th>Doc3</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>0.8974</td>
<td>0.0756</td>
<td>0.5953</td>
</tr>
<tr>
<td>auto</td>
<td>0.1257</td>
<td>0.7867</td>
<td>0</td>
</tr>
<tr>
<td>insurance</td>
<td>0</td>
<td>0.6127</td>
<td>0.7062</td>
</tr>
<tr>
<td>best</td>
<td>0.4230</td>
<td>0</td>
<td>0.3833</td>
</tr>
</tbody>
</table>

1. The query is \(q = [1, 0, 1, 0]\)
\[\text{score}(q, \text{doc1}) = 0.8974, \text{score}(q, \text{doc2}) = 0.6883, \text{score}(q, \text{doc3}) = 1.3015\]
Ranking: doc3, doc1, doc2
2. The query is \(q = [1.62, 0, 1.61, 0]\)
\[\text{score}(q, \text{doc1}) = 1.4807, \text{score}(q, \text{doc2}) = 1.1174, \text{score}(q, \text{doc3}) = 2.1262\]
Ranking: doc3, doc1, doc2

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

<table>
<thead>
<tr>
<th></th>
<th>Doc1</th>
<th>Doc2</th>
<th>Doc3</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>0.6696</td>
<td>0.3699</td>
<td>0.6022</td>
</tr>
<tr>
<td>auto</td>
<td>0.5128</td>
<td>0.7330</td>
<td>0</td>
</tr>
<tr>
<td>insurance</td>
<td>0</td>
<td>0.5709</td>
<td>0.6117</td>
</tr>
<tr>
<td>best</td>
<td>0.5375</td>
<td>0</td>
<td>0.5130</td>
</tr>
</tbody>
</table>

1. The query is \(q = [1, 0, 1, 0]\)
\[\text{score}(q, \text{doc1}) = 0.6696, \text{score}(q, \text{doc2}) = 0.9408, \text{score}(q, \text{doc3}) = 1.2139\]
Ranking: doc3, doc2, doc1
2. The query is \(q = [1.62, 0, 1.61, 0]\)
\[\text{score}(q, \text{doc1}) = 1.1048, \text{score}(q, \text{doc2}) = 1.5351, \text{score}(q, \text{doc3}) = 1.9846\]
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
\[
\lambda = \begin{bmatrix}
\frac{3}{13}, \\
\frac{4}{13}, \\
\frac{6}{13}
\end{bmatrix}^T
\]
By iteration method, we get the following list*:
\[
\begin{bmatrix}
0.3333 & 0.2777 & 0.2314 & 0.2345 & 0.2301 & \ldots & 0.2307 \\
0.3333 & 0.2777 & 0.3148 & 0.3040 & 0.3088 & \ldots & 0.3076 \\
0.3333 & 0.4444 & 0.4537 & 0.4614 & 0.4609 & \ldots & 0.4615 \\
\end{bmatrix}
\]

5.1.2 \( (0.5') \)
The iteration process is:
\[
v' = \beta M v + (1 - \beta) e / n
\]
\[
= \begin{bmatrix} 4/15 & 2/5 & 0 \\ 4/15 & 0 & 2/5 \\ 4/15 & 2/5 & 2/5 \end{bmatrix} v + \begin{bmatrix} 1/15 \\ 1/15 \\ 1/15 \end{bmatrix}
\]

If we solve this equation directly, we get \( v = \begin{bmatrix} 7/27 \\ 25/81 \\ 35/81 \end{bmatrix} \)

By iteration method, we get the following iteration list*:
\[
\begin{bmatrix}
0.3333 & 0.2888 & 0.2592 & 0.2608 & 0.2590 & \ldots & 0.2592 \\
0.3333 & 0.2888 & 0.3125 & 0.3070 & 0.3090 & \ldots & 0.3086 \\
0.3333 & 0.4222 & 0.4281 & 0.4320 & 0.4318 & \ldots & 0.4320 \\
\end{bmatrix}
\]

5.1.3 \( (1') \)
The matrix \( A \) for the graph is:
\[
\begin{bmatrix}
0 & 1/n & \ldots & 1/n & 0 \\
1/n & 0 & \ldots & 1/n & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
1 & 1 & \ldots & 1 & 0 \\
0 & n & \ldots & n & 0 \\
The matrix \( A \) for the graph is:
\[
\begin{bmatrix}
0 & 1/n & \ldots & 1/n & 0 \\
1/n & 0 & \ldots & 1/n & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
1 & 1 & \ldots & 1 & 0 \\
0 & n & \ldots & n & 0 \\
\end{bmatrix}
\]

Easy to see that all nodes in the clique have the same PageRank value, so we suppose vector \( v \) to be \([x, x, \ldots, x, y]^T\), where \(x\) is PageRank of node in the clique, and \(y\) represents the additional node outside the clique. Then we get the equation
\[
\begin{bmatrix}
x = \frac{\beta x(n - 1)}{n} + \frac{nx(1 - \beta)}{n + 1} + \frac{y}{n + 1} \\
\end{bmatrix}
\]
Solve the equation, we get
\[
\begin{bmatrix}
x = \frac{n}{n^2 + n + \beta} \\
y = \frac{n + \beta}{n^2 + n + \beta}
\end{bmatrix}
\]

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.
\section*{5.2.2 \ (1')}\ 
\textbf{a)}

\begin{tabular}{|c|c|c|}
\hline
Source & Degree & Destinations \\
\hline
A & 3 & B, C, D \\
B & 2 & A, D \\
C & 1 & E \\
D & 2 & B, C \\
\hline
\end{tabular}

\begin{tabular}{|c|c|c|}
\hline
\textbf{b)}

Source & Degree & Destinations \\
\hline
a & 3 & a, b, c \\
b & 2 & a, c \\
c & 2 & b, c \\
\hline
\end{tabular}

\section*{5.2.3 \ (1')}\ 
The four-node graph is divided into four 2-by-2 blocks (M11, M12, M21, M22).

\textbf{M11:}

\begin{tabular}{|c|c|c|}
\hline
Source & Degree & Destinations \\
\hline
A & 3 & B \\
B & 2 & A \\
\hline
\end{tabular}

\textbf{M12:}

\begin{tabular}{|c|c|c|}
\hline
Source & Degree & Destinations \\
\hline
D & 2 & B \\
\hline
\end{tabular}

\textbf{M21:}

\begin{tabular}{|c|c|c|}
\hline
Source & Degree & Destinations \\
\hline
A & 3 & C, D \\
B & 2 & D \\
\hline
\end{tabular}

\textbf{M22:}

\begin{tabular}{|c|c|c|}
\hline
Source & Degree & Destinations \\
\hline
D & 2 & C \\
\hline
\end{tabular}

\section*{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 \( \beta \) is 0.8

\textbf{a)}

\[
v' = \begin{bmatrix}
0 & 2/5 & 4/5 & 0 \\
4/15 & 0 & 0 & 2/5 \\
4/15 & 0 & 0 & 2/5 \\
4/15 & 2/5 & 0 & 0
\end{bmatrix}
v + \begin{bmatrix}
1/5 \\
0 \\
0 \\
0
\end{bmatrix}
\]
If we solve this equation directly, we get \( \mathbf{v} = \begin{bmatrix} \frac{3}{7}, \frac{4}{21}, \frac{4}{21} \end{bmatrix}^T \)

By iteration method, we get the following list is:
\[
\begin{bmatrix}
1.0 & 0.2 & 0.52 & 0.392 & 0.4432 & 0.4227 & 0.4285 \\
0.0 & 0.2666 & 0.16 & 0.2026 & 0.1924 & 0.1924 & 0.1904 \\
0.0 & 0.2666 & 0.16 & 0.2026 & 0.1924 & 0.1924 & 0.1904 \\
0.0 & 0.2666 & 0.16 & 0.2026 & 0.1924 & 0.1924 & 0.1904
\end{bmatrix}
\]

b) \[
\mathbf{v}' = \begin{bmatrix}
0 & 2/5 & 4/5 & 0 \\
4/15 & 0 & 0 & 2/5 \\
4/15 & 0 & 0 & 2/5 \\
4/15 & 2/5 & 0 & 0
\end{bmatrix} \mathbf{v} + \begin{bmatrix} 1/10 \\ 0 \\ 0 \\ 0 \end{bmatrix}
\]

If we solve this equation directly, we get \( \mathbf{v} = \begin{bmatrix} \frac{27}{70}, \frac{6}{35}, \frac{19}{70}, \frac{6}{35} \end{bmatrix}^T \)

By iteration method, we get the following list is:
\[
\begin{bmatrix}
1.0 & 0.1 & 0.5 & 0.34 & 0.404 & 0.3784 & 0.3857 \\
0.0 & 0.2666 & 0.1333 & 0.1866 & 0.1653 & 0.1738 & 0.1714 \\
0.0 & 0.3666 & 0.2333 & 0.2866 & 0.2653 & 0.2738 & 0.2714 \\
0.0 & 0.2666 & 0.1333 & 0.1866 & 0.1653 & 0.1738 & 0.1714
\end{bmatrix}
\]

5.5.1 (1')
The link matrix for Figure 5.1 is:
\[
\mathbf{L} = \begin{bmatrix}
0 & 1 & 1 & 1 \\
1 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 \\
0 & 1 & 1 & 0
\end{bmatrix}
\]

Initialize \( \mathbf{a} \) and \( \mathbf{h} \) to be all 1’s, given two equations
\[
\mathbf{h} = \mathbf{L} \mathbf{a} \\
\mathbf{a} = \mathbf{L}^T \mathbf{h}
\]

By iteration method, we get the following list is:
\[
\begin{bmatrix}
1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\
0.6666 & 0.6666 & 0.5333 & 0.5333 & 0.4657 & 0.3919 \\
0.3333 & 0.3333 & 0.2 & 0.2 & 0.1506 & 0.1027 \\
0.6666 & 0.6666 & 0.6666 & 0.6666 & 0.6666 & 0.6666
\end{bmatrix}
\]
\[
\begin{bmatrix}
1.0 & 0.6 & 0.6 & 0.44 & 0.44 & 0.2891 \\
1.1 & 1 & 1 & 1 & 1 & 1 & 1.0 \\
1.1 & 1 & 1 & 1 & 1 & 1 & 1.0 \\
1.1 & 1 & 1 & 1 & 1 & 1 & 1.0
\end{bmatrix}
\]