CSE 494/598
Lecture-4: Correlation Analysis

LYDIA MANIKONDA
HTTP://WWW.PUBLIC.ASU.EDU/~LMANIKON/

**Content adapted from last year’s slides**
Announcements

• Project-1 Due: February 12\textsuperscript{th} 2016
  • Analysis report: Before 4 pm (Hard copy)
  • Code: Before 11.59 pm (Through email: cse494s16@gmail.com )

<table>
<thead>
<tr>
<th>Important Dates</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>02/12/2016</td>
<td>Project part – 1</td>
</tr>
<tr>
<td>03/04/2016</td>
<td>Midterm – 1</td>
</tr>
<tr>
<td>03/18/2016</td>
<td>Project part – 2</td>
</tr>
<tr>
<td>04/22/2016</td>
<td>Project part – 3</td>
</tr>
<tr>
<td>05/02/2016 or 05/04/2016</td>
<td>Final Exam</td>
</tr>
</tbody>
</table>
Today

• Correlation analysis

• Review of Linear Algebra
So many things can go wrong

Reasons that ideal effectiveness is hard to achieve:

• Document representation loses information
• Users’ inability to describe queries precisely
• Similarity function used is not good enough
• Importance/weight of a term in representing a document and query may be inaccurate
• Same term may have multiple meanings and different terms may have similar meanings
Improving Vector Space Ranking

1. **Relevance Feedback** – Improves the query quality

2. **Correlation Analysis** – Looks at correlation between keywords (and thus effectively computes a thesaurus based on the word occurrence in the documents) to do query elaboration

3. **Principal Component Analysis** (also seen as Latent Semantic Indexing) – Subsumes correlation analysis and performs dimensionality reduction
Correlation/Co-occurrence Analysis

Co-occurrence Analysis

• Terms that are related to terms in the original query may be added to the query
• Two terms are related if they have high co-occurrence in documents
Correlation

- There are 1,000 documents in my corpus
- Keyword $k_1$ is in 50% of them
- Keyword $k_2$ is in 50% of them

- Are $k_1$ and $k_2$
  - Positively correlated?
  - Negatively correlated?
  - Not correlated?

- Assuming that the fraction of documents that have $k_1$ and $k_2$ together be $x\%$

  If $x >> 25\%$ +ve correlation
  $x << 25\%$ -ve correlation
  $x \sim 25\%$ independent
Correlation/Co-occurrence Analysis

Co-occurrence Analysis

• Terms that are related to terms in the original query may be added to the query
• Two terms are related if they have high co-occurrence in documents

Let there be $n$ number of documents

• $n_1$ and $n_2$ be the # of documents containing terms $t_1$ and $t_2$
• $m$ be the # of documents having both $t_1$ and $t_2$

If $t_1$ and $t_2$ are independent

$$n \cdot \frac{n_1}{n} \cdot \frac{n_2}{n} \approx m$$

>> if

Inversely correlated

If $t_1$ and $t_2$ are correlated

$$n \cdot \frac{n_1}{n} \cdot \frac{n_2}{n} \ll m$$
Terms and Documents as mutually dependent vectors

Instead of doc-doc similarity, we can compute term-term distance.

• If terms are independent, the term-term similarity matrix should be diagonal (*)
  • If it is not diagonal, we use the correlations to add related terms to the query
  • But can also ask the question “Are there independent dimensions which define the space where terms & docs are vectors?”

(*) Note that $ij^{th}$ element in the term-term matrix is the dot product of $i^{th}$ term vector and $j^{th}$ term vector

<table>
<thead>
<tr>
<th>Term</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>User</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>System</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Human</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Computer</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Response</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Time</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EPS</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Survey</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Trees</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Graph</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Minors</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Association Clusters

• Let $t-d$ be the term-document matrix
  • For the full corpus (global)
  • For the documents in the set of initial results (local)
  • Sometimes stems are used instead of terms

• Correlation matrix $C = t-d \times t-d^T$

$$C_{uv} = \sum_{dj} f_{tu,dj} \times f_{tv,dj}$$

$$S_{uv} = C_{uv} \quad \text{Un-normalized Association Matrix}$$

$$S_{uv} = \frac{C_{uv}}{C_{uu}+C_{vv}-C_{uv}} \quad \text{Normalized Association Matrix}$$

$N^{th}$-Association Cluster for a term $t_u$ is the set of terms $t_v$ such that $S_{uv}$ are the $n$ largest values among $S_{u1}, S_{u2}, \ldots, S_{uk}$
**Example**

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
<th>$d_4$</th>
<th>$d_5$</th>
<th>$d_6$</th>
<th>$d_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_1$</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$K_2$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>$K_3$</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Correlation Matrix**

- $11$ $4$ $6$
- $4$ $34$ $11$
- $6$ $11$ $26$

**Normalized Correlation Matrix**

- $1.0$ $0.097$ $0.193$
- $0.097$ $1.0$ $0.224$
- $0.193$ $0.224$ $1.0$

$$S_{12} = \frac{(s_{12})^4}{(s_{11})^4 + (s_{22})^4 - (s_{12})^4}$$

1\textsuperscript{th} Association Cluster for $K_2$ is $K_3$
On the Database/Statistics Example

<table>
<thead>
<tr>
<th></th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>t6</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>24</td>
<td>21</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>d2</td>
<td>32</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d3</td>
<td>12</td>
<td>16</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>d4</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d5</td>
<td>43</td>
<td>31</td>
<td>20</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>d6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>d7</td>
<td>0</td>
<td>0</td>
<td>32</td>
<td>0</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>d8</td>
<td>0</td>
<td>22</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d9</td>
<td>1</td>
<td>0</td>
<td>34</td>
<td>27</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>d10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>4</td>
<td>23</td>
</tr>
</tbody>
</table>

t1= database
t2=SQL
t3=index
t4=regression
t5=likelihood
t6=linear

3679 2391 1308 238 302 273
2391 1807 953 0 123 63
1308 953 536 32 87 27
238 0 32 3277 1584 1573
302 123 87 1584 972 887
273 63 27 1573 887 1423

database is most related to SQL and second most related to index

<table>
<thead>
<tr>
<th></th>
<th>1.0000</th>
<th>0.7725</th>
<th>0.4499</th>
<th>0.0354</th>
<th>0.0694</th>
<th>0.0565</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7725</td>
<td>1.0000</td>
<td>0.6856</td>
<td>0</td>
<td>0.0463</td>
<td>0.0199</td>
<td></td>
</tr>
<tr>
<td>0.4499</td>
<td>0.6856</td>
<td>1.0000</td>
<td>0.0085</td>
<td>0.0612</td>
<td>0.0140</td>
<td></td>
</tr>
<tr>
<td>0.0354</td>
<td>0</td>
<td>0.0085</td>
<td>1.0000</td>
<td>0.5944</td>
<td>0.5030</td>
<td></td>
</tr>
<tr>
<td>0.0694</td>
<td>0.0463</td>
<td>0.0612</td>
<td>0.5944</td>
<td>1.0000</td>
<td>0.5882</td>
<td></td>
</tr>
<tr>
<td>0.0565</td>
<td>0.0199</td>
<td>0.0140</td>
<td>0.5030</td>
<td>0.5882</td>
<td>1.0000</td>
<td></td>
</tr>
</tbody>
</table>

Can also do this on query logs instead of documents
Scalar Clusters

• Consider the normalized association matrix $S$

• The “association vector” $A_u$ of term $u$ is $(S_{u1}, S_{u2}, ... S_{uk})$

• To measure neighborhood-induced correlation between terms: Take the cosine-theta between the association vectors of terms $u$ and $v$

$$S_{uv} = \frac{A_u \cdot A_v}{|A_u| \times |A_v|}$$

Even if terms $u$ and $v$ have low correlations, they may be transitively correlated (e.g. a term $w$ has high correlation with $u$ and $v$).

$N^{th}$-Association Cluster for a term $t_u$ is the set of terms $t_v$ such that $S_{uv}$ are the $n$ largest values among $S_{u1}, S_{u2}, ..., S_{uk}$
Example

<table>
<thead>
<tr>
<th></th>
<th>d₁</th>
<th>d₂</th>
<th>d₃</th>
<th>d₄</th>
<th>d₅</th>
<th>d₆</th>
<th>d₇</th>
</tr>
</thead>
<tbody>
<tr>
<td>K₁</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>K₂</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>K₃</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Normalized Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>d₁</th>
<th>d₂</th>
<th>d₃</th>
<th>d₄</th>
<th>d₅</th>
<th>d₆</th>
<th>d₇</th>
</tr>
</thead>
<tbody>
<tr>
<td>d₁</td>
<td>1.0</td>
<td>0.097</td>
<td>0.193</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d₂</td>
<td>0.097</td>
<td>1.0</td>
<td>0.224</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d₃</td>
<td>0.193</td>
<td>0.224</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Scalar (Neighborhood) Cluster Matrix

<table>
<thead>
<tr>
<th></th>
<th>d₁</th>
<th>d₂</th>
<th>d₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>d₁</td>
<td>1.0</td>
<td>0.226</td>
<td>0.383</td>
</tr>
<tr>
<td>d₂</td>
<td>0.226</td>
<td>1.0</td>
<td>0.435</td>
</tr>
<tr>
<td>d₃</td>
<td>0.383</td>
<td>0.435</td>
<td>1.0</td>
</tr>
</tbody>
</table>

1³ Scalar Cluster for K₂ is still K₃
On the Database/Statistics Example

## Scalar Clusters

<table>
<thead>
<tr>
<th></th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>t6</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>24</td>
<td>21</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>d2</td>
<td>32</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>d3</td>
<td>12</td>
<td>16</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>d4</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>d5</td>
<td>43</td>
<td>31</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>d6</td>
<td>2</td>
<td>0</td>
<td>18</td>
<td>7</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>d7</td>
<td>0</td>
<td>1</td>
<td>32</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d8</td>
<td>3</td>
<td>0</td>
<td>22</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d9</td>
<td>1</td>
<td>0</td>
<td>34</td>
<td>27</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>d10</td>
<td>6</td>
<td>0</td>
<td>17</td>
<td>4</td>
<td>23</td>
<td>0</td>
</tr>
</tbody>
</table>

- t1 = database
- t2 = SQL
- t3 = index
- t4 = regression
- t5 = likelihood
- t6 = linear

Notice that index became much closer to database

## Association Clusters

<table>
<thead>
<tr>
<th></th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>t6</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>3679</td>
<td>2391</td>
<td>1308</td>
<td>238</td>
<td>302</td>
<td>273</td>
</tr>
<tr>
<td>d2</td>
<td>2391</td>
<td>1807</td>
<td>953</td>
<td>0</td>
<td>123</td>
<td>63</td>
</tr>
<tr>
<td>d3</td>
<td>1308</td>
<td>953</td>
<td>536</td>
<td>32</td>
<td>87</td>
<td>27</td>
</tr>
<tr>
<td>d4</td>
<td>238</td>
<td>0</td>
<td>32</td>
<td>3277</td>
<td>1584</td>
<td>1573</td>
</tr>
<tr>
<td>d5</td>
<td>302</td>
<td>123</td>
<td>87</td>
<td>1584</td>
<td>972</td>
<td>887</td>
</tr>
<tr>
<td>d6</td>
<td>273</td>
<td>63</td>
<td>27</td>
<td>1573</td>
<td>887</td>
<td>1423</td>
</tr>
</tbody>
</table>

- database is most related to SQL and second most related to index

Can also do this on query logs instead of documents
Ways of doing correlation analysis

ANALYZE DOCUMENT CORPUS

- Global thesaurus construction
  - Do the analysis on the entire document corpus

- Local (query-specific) thesaurus construction
  - Do the analysis only over the top-k documents with high vector similarity to the current query
    - There may be query specific correlations that are drowned out in full corpus

ANALYZE QUERY LOGS

- Or analyze query logs (in as much as you actually have users asking queries — i.e., people are actually using your retrieval system)
  - Instead of doc-term matrix, you will start with query-term matrix

“cold start” problem
You have query logs only if you have users.
Example – Correlation analysis for recommendations
Connection to Collaborative Filtering

If you think of

• “documents” or “queries” as users
• “terms” as items that the users have bought/liked
• The correlation clusters tell us what other items are owned/liked by people who liked the item you have
• Collaborative Filtering
Collaborative Filtering
Collaborative Filtering

User-based filtering

Item-based filtering
Terms and Documents as mutually dependent vectors

Instead of doc-doc similarity, we can compute term-term distance.

• If terms are independent, the term-term similarity matrix should be diagonal (*)
  • If it is not diagonal, we use the correlations to add related terms to the query
  • But can also ask the question “Are there independent dimensions which define the space where terms & docs are vectors?”

(*) Note that $i^{th}$ element in the term-term matrix is the dot product of $i^{th}$ term vector and $j^{th}$ term vector
Beyond Correlation Analysis

Latent Semantic Indexing
Linear Algebra review
Vectors and Matrices

• A vector is a set of numbers
• A matrix is a set of vectors
A vector in space

- In space, a vector can be shown as an arrow
  - starting point is the origin
  - ending point are the values of the vector
Properties of vectors

- Its “size” $|v|$
- the 2-norm of the vector
  $$|(1, 2, 3)| = \sqrt{1^2 + 2^2 + 3^2}$$
- A unit vector is a vector of size 1.
Operations on vectors

- Addition, Subtraction – easy.
  \[(1, 2, 3) + (100, 200, 300) = (101, 202, 303)\]

- Dot product
  \[(1, 2, 3) \cdot (100, 200, 300) = (100, 400, 900)\]

- Multiplication – with other matrices.

- Division – not defined.
Matrices

- No intuitive representation in space
- Addition / Subtraction – easy
- Multiplication – matrix multiplication
  - Not commutative
- Division – not defined
  - If the matrix is
    - a square matrix
    - Invertible
  - then take inverse and multiply
Matrix Multiplication can be seen as computing vector dot products.

- Given a matrix $R$, you can consider each row of $M$ as a vector.
- Thus $R = [r_1 
  r_2
  ..
  r_k]$
- Now Given another matrix $S$ whose column vectors are $s_1... s_k$, the $ij^{th}$ element in $R*S$ is $r_i.s_j$
- Is the dot product
- As a corollary, if a matrix $M$ is orthogonal—i.e., its row vectors are all orthogonal to each other, then $M*M'$ or $M'*M$ will both be diagonal matrices.
Some identities / properties

• Transpose of a matrix

\[
\begin{bmatrix}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 9
\end{bmatrix}^T =
\begin{bmatrix}
1 & 4 & 7 \\
2 & 5 & 8 \\
3 & 6 & 9
\end{bmatrix}
\]

• Determinant of a matrix

\[
\begin{vmatrix}
5 & 6 \\
8 & 10
\end{vmatrix} = 2
\]

\[
\begin{vmatrix}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 10
\end{vmatrix} = -3
\]

• $A A^{-1} = I$
What happens when you multiply a matrix by a vector?

The vector scales and rotates.

\[
\begin{bmatrix}
3 & -1 \\
1 & 3 \\
\end{bmatrix} \times \begin{bmatrix}
4 \\
3 \\
\end{bmatrix} = \begin{bmatrix}
9 \\
13 \\
\end{bmatrix}
\]
Example 1 – only rotation

\[
\begin{bmatrix}
\cos(45^\circ) & -\sin(45^\circ) \\
\sin(45^\circ) & \cos(45^\circ)
\end{bmatrix}
\times \begin{bmatrix}
1 \\
0
\end{bmatrix}
= \begin{bmatrix}
1/\sqrt{2} \\
1/\sqrt{2}
\end{bmatrix}
\]
Example 2 – only scaling

\[
\begin{bmatrix}
5 & 0 \\
0 & 5
\end{bmatrix} \times \begin{bmatrix}
3 \\
4
\end{bmatrix} = \begin{bmatrix}
15 \\
20
\end{bmatrix}
\]
Example 3 - both

\[
\begin{bmatrix}
2 \cos(30^\circ) & -2 \sin(30^\circ) \\
2 \sin(30^\circ) & 2 \cos(30^\circ)
\end{bmatrix} \times \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 3.93 \\ 9.20 \end{bmatrix}
\]
So a matrix is a bunch of numbers that tells us how to rotate and scale vectors

- Special matrices: Unit matrix
- Special matrices: Rotation matrix
- Special matrices: Scaling matrix
Can we make some general statements about a matrix?

• Given any matrix $M$, can we make some statements about how it affects vectors?
• Start with any vector. Multiply it over and over and over with a matrix. What happens?
Eigen vectors

- There are some vectors which don’t change direction on multiplication with a matrix.
- They are called **Eigen vectors**.

\[
\begin{bmatrix}
2 & 3 \\
3 & 2
\end{bmatrix}
\times
\begin{bmatrix}
1 \\
1
\end{bmatrix}
=
\begin{bmatrix}
5 \\
5
\end{bmatrix}
\]

- However, the matrix does manage to scale them. The factor it scales them by, are called the ‘Eigen values’.
How to find Eigen values

• Let’s assume one of the eigen vectors is $v$
• Then $A v = \lambda v$, where $\lambda$ is the eigenvalue.
• Transpose it $(A - \lambda I)v = 0$
• Theorem: if $v$ is not zero, then the above equation can only be true if the determinant of $(A - \lambda I)$ is zero. [proof: see Characteristic polynomial in Wikipedia]
• Use this fact to find values of $\lambda$
\[(A - \lambda I)v = 0\]

\[\Rightarrow |A - \lambda I| = 0\]

\[\Rightarrow \begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0\]

\[\Rightarrow \begin{vmatrix} 2 - \lambda & 3 \\ 3 & 2 - \lambda \end{vmatrix} = 0\]

\[\Rightarrow (2 - \lambda)^2 - 9 = 0\]

\[\Rightarrow \lambda = 5, -1\]
How to find Eigen Vectors

• Substitute $\lambda$ back into the equation $(A - \lambda I)v=0$

• Try to find $v$

• You will get two equations in two variables – but: there is a problem, the two equations are identical

$$\begin{pmatrix} 2 & 3 \\ 3 & 2 \end{pmatrix} - 5 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$$

$$\Rightarrow \begin{pmatrix} -3 & 3 \\ 3 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$$

$$\Rightarrow -3x_1 + 3x_2 = 0$$

and, $3x_1 - 3x_2 = 0$
Eigen vectors – the missing equation

• One equation, two variables

• Use the additional constraint that the Eigen vector is a unit vector (length 1)

• $x_1^2 + x_2^2 = 1$

• Using the $x_1 = x_2$ we found from the previous slide, we have the Eigen vector

\[
\begin{bmatrix}
\frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}}
\end{bmatrix}
\]
So what do these values tell us

• If you repeatedly multiply a vector by a matrix, (and then normalize), then you will eventually get the primary Eigen vector.

• The primary Eigen vector is, sort of, the general direction in which the matrix turns the vector.
How is this all relevant to the class?

• Instead of thinking of 2-dimension or 3-dimension vectors, imagine vectors in T dimensions
  • T = number of different terms.
  • Each doc will be a vector in this space.
  • Similarity between the docs = normalized dot product
  • Store the link structure of the web in a matrix
  • Eigen values / vectors – PageRank
Class exercises
Lessons Learned Today

• Correlation Analysis
  • Correlation matrices
  • Normalized correlation matrices – Association clusters
  • Scalar clusters

• Review of linear algebra
  • Matrix inverse
  • Eigenvalues and eigenvectors