Station 4: Clustering

Jamie Conklin

Health Sciences Library
Clinical, Academic, & Research Engagement (CARE)
Visualizing Clustering

Compile search results

Cluster based on text similarities in titles and abstracts

Review high priority clusters based on keyword summary
Apply Older Algorithms in a New Way
Clustering Example: Adolescent Perceptions of CF

- Consulted with student
  - used traditional search methods
- Performed broader search across multiple databases
  - psychology and cystic fibrosis
- Applied K-Means clustering
- Identified 2 useful clusters
## Clustering Example: Adolescent Perceptions of CF

<table>
<thead>
<tr>
<th>Cluster</th>
<th># of Studies</th>
<th>Topic Key Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>126</td>
<td>children</td>
</tr>
<tr>
<td>2</td>
<td>73</td>
<td>carrier</td>
</tr>
<tr>
<td>3</td>
<td>61</td>
<td>sleep</td>
</tr>
<tr>
<td>4</td>
<td>97</td>
<td>adolescents</td>
</tr>
</tbody>
</table>
Clustering Example: Adolescent Perceptions of CF

- Once we identified clusters, we used EndNote to collect the articles within those clusters
### Research Question: What factors affect breast cancer risk?

<table>
<thead>
<tr>
<th>Cluster</th>
<th># of Studies</th>
<th>Priority</th>
<th>Keywords/Topic Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>853</td>
<td>M</td>
<td>risk</td>
</tr>
<tr>
<td>1</td>
<td>1377</td>
<td></td>
<td>drug</td>
</tr>
<tr>
<td>2</td>
<td>2099</td>
<td>M</td>
<td>patients</td>
</tr>
<tr>
<td>3</td>
<td>1177</td>
<td>L</td>
<td>activity</td>
</tr>
<tr>
<td>4</td>
<td>1210</td>
<td></td>
<td>Cluster 4: risk</td>
</tr>
<tr>
<td>5</td>
<td>1238</td>
<td></td>
<td>Cluster 5: exposure</td>
</tr>
<tr>
<td>6</td>
<td>1224</td>
<td>L</td>
<td>Cluster 1: drug</td>
</tr>
<tr>
<td>7</td>
<td>1447</td>
<td>L</td>
<td>Cluster 1: drug</td>
</tr>
<tr>
<td>8</td>
<td>2941</td>
<td>L</td>
<td>Cluster 1: drug</td>
</tr>
<tr>
<td>9</td>
<td>3424</td>
<td>M</td>
<td>Cluster 1: drug</td>
</tr>
</tbody>
</table>

**Your Turn:**

Prioritize clusters 1, 4, and 5 in terms of likelihood of containing relevant references.

- H = High priority
- M = Moderate priority
- L = Low priority

When you’re finished turn over to check your answers against Supervised Clustering predictions.
Research Question: What factors affect breast cancer risk?

<table>
<thead>
<tr>
<th>Cluster</th>
<th># of Studies</th>
<th>Priority</th>
<th>Keywords/Topic Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>853</td>
<td>M</td>
<td>risk</td>
</tr>
<tr>
<td>1</td>
<td>1377</td>
<td>L</td>
<td>drug</td>
</tr>
<tr>
<td>2</td>
<td>2099</td>
<td>M</td>
<td>mammalian</td>
</tr>
<tr>
<td>3</td>
<td>1177</td>
<td>L</td>
<td>confidence</td>
</tr>
<tr>
<td>4</td>
<td>1210</td>
<td>H</td>
<td>activity</td>
</tr>
<tr>
<td>5</td>
<td>1238</td>
<td>M</td>
<td>compounds</td>
</tr>
<tr>
<td>6</td>
<td>1224</td>
<td>L</td>
<td>patients</td>
</tr>
<tr>
<td>7</td>
<td>1447</td>
<td>L</td>
<td>patients</td>
</tr>
<tr>
<td>8</td>
<td>2941</td>
<td>L</td>
<td>patients</td>
</tr>
<tr>
<td>9</td>
<td>3424</td>
<td>M</td>
<td>patients</td>
</tr>
</tbody>
</table>
Next Steps…

• K-Means clustering and other clustering algorithms available in:
  • R / SPSS / SAS / etc.
• Keep an eye out for future CE courses
• Coursera courses on clustering
• Consider K-Means clustering as a preprocessing step for a supervised clustering approach with seeds
  • Visit Station 5