Serendip
Topic Model-Driven Visual Exploration of Text Corpora

Eric Alexander, Department of Computer Science, University of Wisconsin-Madison
Joe Kohlmann, Department of Computer Science, University of Wisconsin-Madison
Robin Valenza, Department of English, University of Wisconsin-Madison
Michael Witmore, Folger Shakespeare Library in Washington, D.C.
Michael Gleicher, Member, IEEE, Department of Computer Science, University of Wisconsin-Madison
Serendipity

• „The Three Princes of Serendip”

• Happy accidents

• Fate or extreme cleverness

• Research: How to make coincidence more likely?

• „The bohemian bookshelf“ by Thudt et al
Promoting Serendipity

A. Thudt, U. Hinrichs, and S. Carpendale
„The bohemian bookshelf: supporting serendipitous book discoveries through information visualization“
In Proc. ACM Human Factors in Computing Systems

• Providing multiple access points
• Highlighting adjacencies
• Offering flexible pathways for exploration
• Enticing curiosity and playfulness
Three Views

- **CorpusViewer**: Re-orderable matrix
- **TextViewer**: Examination within one document
- **RankViewer**: Examine specific words
CorpusViewer
Features combatting scale

- Ordering
- Aggregation
- Annotation
- Assigning colors
- Details on demand
TextViewer
Serendip: Topic Model-Driven Visual Exploration of Text Corpora

RankViewer
How are the factors for Serendipity implemented?

- Multiple access points: 3 Views, Ordered to user’s liking
- Highlighting adjacencies: Through ordering and vis
- Flexible pathways: Jumping between views
- Curiosity and playfulness: Interaction & Discovery
2 use cases

• Vis Abstracts
  1127 abstracts
  From SciVis, InfoVis, VAST, BioVis, and PacificVis papers from 2007-2013
  each 30 to 389 words

• Early Modern Literature
  1080 digitized texts
  From English literature published between 1530 to 1799
  each few hundred words to few hundred pages

Formal evaluation still due

• To confirm serendipitous discoveries across multiple scales of data and abstraction
• Problem: How to evaluate serendipity?
• Long-term user studies needed
Discussion

+ strengths

• Suitable for any document and corpus size
• Three layers (whole corpus, single document, single words)
• Simple but effective visualizations
• Easily accessible (Online tool; though topic modelling part still due)

- weaknesses

• Not for quick exploration
• Sceptical about serendipitous discoveries
<table>
<thead>
<tr>
<th><strong>System</strong></th>
<th>Serendip</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What</strong></td>
<td><strong>Data</strong></td>
</tr>
<tr>
<td></td>
<td>Document collection (Text, Metadata, Topics)</td>
</tr>
<tr>
<td><strong>Why</strong></td>
<td><strong>Tasks</strong></td>
</tr>
<tr>
<td></td>
<td>Explore document collections</td>
</tr>
<tr>
<td></td>
<td>Facilitate serendipitous discoveries</td>
</tr>
<tr>
<td><strong>How</strong></td>
<td><strong>Encode</strong></td>
</tr>
<tr>
<td></td>
<td>Reorderable matrix, line graph, bar graph</td>
</tr>
<tr>
<td></td>
<td><strong>Reduce</strong></td>
</tr>
<tr>
<td></td>
<td>Aggregation, Ordering</td>
</tr>
<tr>
<td></td>
<td><strong>Manipulate</strong></td>
</tr>
<tr>
<td></td>
<td>Order, Color, Annotate, Details-on-Demand</td>
</tr>
<tr>
<td><strong>Scale</strong></td>
<td>Up to ~1,000 Documents (various length)</td>
</tr>
</tbody>
</table>