Needs-Driven Workflow Design

Types and levels of analysis determine data, algorithms & parameters, and deployment

Information Visualization MOOC

Unit 4 – “What”: Topical Data

Workflow Design

Relevant Research Disciplines:
Linguistics, Computer Science, Artificial Intelligence

Reference

http://ivmooc.cns.iu.edu
Needs-Driven Workflow Design

Types and levels of analysis determine data, algorithms & parameters, and deployment

Data

READ

ANALYZE

VISUALIZE

DEPLOY

Validation
Interpretation

Visually encode data
Overlay data
Select visualiz. type

Graphic Variable Types
Modify reference system, add records & links
Visualization Types (reference systems)

Needs-Driven Workflow Design

Types and levels of analysis determine data, algorithms & parameters, and deployment

Data

READ

ANALYZE

VISUALIZE

DEPLOY

Validation
Interpretation

Visually encode data
Overlay data
Select visualiz. type

Graphic Variable Types
Modify reference system, add records & links
Visualization Types (reference systems)
Read Data

Data Repositories:
- Google ngrams datasets, text from millions of books scanned by Google
- Scholarly Database, http://sdb.cns.iu.edu

Major Data Formats:
- TXT
- CSV

Preprocessing—Text Normalization

Sample text: Emergence of Scaling in Random Networks

- Lowercase: The example text becomes "emergence of scaling in random networks."
- Tokenize: The text blob is split into a list of individual words. The example text becomes "emergence of scaling in random networks."
- Stem: Common or low-content prefixes and suffixes are removed to identify the core concept. The example text becomes "emerg of scale in random network."
- Stopword: Low-content tokens like "of" and "in" are removed (see the complete stopword list). The example text becomes "emerg scale random network."
- Identification of synonymy and polysemy.
Topical Analysis

• Frequency analysis
• Clustering/Classification
• Sentiment analysis
• Burst analysis, see Unit 1

• Dimensionality reduction, see ARIST chapter.

Using a Dictionary and Thesaurus

Visual Thesaurus http://www.visualthesaurus.com/vocabgrabber/
Sorted by relevance, occurrences, select ‘geography’ words
Visualizing Topical Data

- **Charts:** Wordle Word cloud
- **Tables:** GRIDL
- **Graphs:** MDS plots, circular visualization, Crossmaps, Google n-gram
- **Geospatial maps:** SOM maps
- **Network graphs:** Tree visualizations, word co-occurrence networks, concept maps, science map overlays

Black ones are exemplified on subsequent slides.

---

Chart Example: Word Cloud

Wordle.net of Titles – create your own at [http://wordle.net](http://wordle.net)

Layout: Oval space filling; frequent words are closer to center
Type font size: Word frequency
Font color: No meaning, but different colors help legibility
Wordle.net of Titles – New Layout, non-deterministic

Wordle.net of Titles – Same Layout, different colors
An n-gram is a subsequence of $n$ items from a given sequence. The items in question can be phonemes, syllables, letters, words, or base.

Network Graph-Science Map

Reference system with proportional symbol overlay and legend. See also **Unit 4**—Design and Update of a Classification System: The UCSD Map of Science
Legend
Circle area: Fractional Journal Count
Undisclosed = 95
Minimum = 0
Maximum = 25
Color: Discipline

How To Read This Map
The UCSD map of science depicts a network of 554 subdiscipline nodes that are aggregated to 13 main disciplines of science. Each discipline has a distinct color and is labeled. Overlaid are circles, each representing all records per unique subdiscipline. Circle area is proportional to the number of fractionally assigned records. Minimum and maximum data values are given in the legend.
Co-Occurrence Network of IMDb Movie Title Words


<table>
<thead>
<tr>
<th>Rank</th>
<th>Rating</th>
<th>Title</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.2</td>
<td>The Shawshank Redemption (1994)</td>
<td>857,280</td>
</tr>
<tr>
<td>2</td>
<td>9.2</td>
<td>The Godfather (1972)</td>
<td>625,241</td>
</tr>
<tr>
<td>3</td>
<td>9.0</td>
<td>The Godfather, Part II (1974)</td>
<td>400,229</td>
</tr>
<tr>
<td>4</td>
<td>8.9</td>
<td>Pulp Fiction (1994)</td>
<td>669,105</td>
</tr>
<tr>
<td>5</td>
<td>8.9</td>
<td>The Good, the Bad and the Ugly</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>8.9</td>
<td>12 Angry Men (1957)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>8.9</td>
<td>Schindler’s List (1993)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>8.9</td>
<td>The Dark Knight (2008)</td>
<td></td>
</tr>
</tbody>
</table>

What words occur most frequently? Which words are used together?

Text Preprocessing—Normalize

- Lowercase words
- Stem
- Remove stop words
Text Preprocessing—Extract Network

- Tokenize
- Extract co-occurrence network

Without preprocessing:
Nodes: 470
Isolated nodes: 48
Edges: 905

With preprocessing:
Nodes: 405
Isolated nodes: 75
Edges: 319
Zoom into giant component of uncleaned network

Zoom into giant component of cleaned network
Movie Similarity Network Based on Shared Words

Four movies with 10 unique title words:

<table>
<thead>
<tr>
<th>Title</th>
<th>Unique words</th>
</tr>
</thead>
<tbody>
<tr>
<td>star war episod vi return jedi</td>
<td>episod jedi return star trek war vi vi Total#words</td>
</tr>
<tr>
<td>star war episod v empir strike</td>
<td>episod empri strike</td>
</tr>
<tr>
<td>star trek</td>
<td>0 0 0 0 1 0 0 0 0 2</td>
</tr>
<tr>
<td>star war</td>
<td>0 0 0 0 1 0 0 0 0 2</td>
</tr>
</tbody>
</table>

Semantic network of movies based on shared words:

*Vertices 4
1 "star war episod vi return jedi" 6
2 "star war episod v empir strike" 6
3 "star trek" 2
4 "star war" 2
*Edges 6
1 2 3
1 3 1
1 4 2
2 3 1
2 4 2
3 4 1

Complete matrix has 250 movies and 397 normalized words.

Movie Similarity Network

Semantic network of movies based on shared words:

But, titles have different length!

Solutions:
- Normalize by #words
- Use only important words (tf-idf)
- Apply LSA or topic detection
- Run Poisson-based language models

See also **Unit 4—Comparison of Text- and Linkage-Based Approaches**

Complete matrix has 250 movies and 397 normalized words.
Relevant Tools

- TexTrend (OSGi/CIShell compatible), http://textrend.org
- VOSviewer, http://vosviewer.com

See many more at http://www.kdnuggets.com/software/text.html

Please post your favorite to Twitter, Flickr using tags “ivmooc” and “#topictools.”