Text / Document Visualization

Slides Adopted from:
Dr. Shixia Liu
Text is Everywhere

✓ We use documents as primary information artifact in our lives
✓ Our access to documents has grown tremendously in recent years due to networking infrastructure
  • WWW
  • Digital libraries
  • ...
  
Is This You?
The Big Question

What can information visualization provide to help users in understanding and gathering information from text and document collections? -- Text Analytics

Analysis vs Analytics:
Analysis is the examination process itself where analytics is the supporting technology and associated tools
Text Analytics

How can I find information buried inside the piles of text?

- **Terracotta Army - Wikipedia, the free encyclopedia**
  [2] Mount Li is also where the material to make the terracotta warriors originated. In addition to the warriors, an entire man-made necropolis for the ...
  en.wikipedia.org/wiki/Terracotta_Army - 58k - Cached - Similar pages

- **Museum of Qin Terra Cotta Warriors and Horses**
The Terra Cotta Warriors and Horses are the most significant archeological excavations of the 20th century. It is a sight not to be missed by any visitor to ...
  www.travelchiguide.com/attraction/shaanxi/xian/terra_cotta_army/ - 17k - Cached - Similar pages

- **Terra Cotta Pit 1**
Museum of Qin Terra Cotta Warriors and Horses - Pit 1 ... There are more than 6000 terracotta warriors and horses in Pit No. 1, marshaled into battle line ...
  www.travelchiguide.com/cityguides/xian/terra_cotta.htm - 14k - Cached - Similar pages

- **Terracotta Warriors - The Museum**
Terracotta Warriors Museum, Dorchester, brings together all the wonder of the discovery of the many treasures of the first Emperor of China.
  www.terracottawarriors.co.uk/ - 14k - Cached - Similar pages

- **Terracotta Warriors - A Fantastic Tourist Attraction in China ...**
Terracotta Warriors Tours: Private tours to Terracotta Warriors, and other Xian ... 1 which contains 6000 life-size terracotta warriors and horses ...
  www.chinavista.com/travel/terracotta/warrior01.html - 6k - Cached - Similar pages

- **Terracotta Warriors - A Fantastic Tourist Attraction in China ...**
Terracotta Warriors Tours: Private tours to Terracotta Warriors, ... Let us go to Xian to have a look at the Museum of Qin Terracotta Warriors. ...
  www.chinavista.com/travel/terracotta/main.html - 6k - Cached - Similar pages
## Text Analytics

### What is in my text?

<table>
<thead>
<tr>
<th>What’s inside the NHTSA Data:</th>
<th>What are the major causes of injuries</th>
<th>What did my customers say about my hotels</th>
</tr>
</thead>
<tbody>
<tr>
<td>450,000+ documents</td>
<td>70,000+ patient emergency room records</td>
<td>3000+ customer-posted reviews</td>
</tr>
</tbody>
</table>

### Information Understanding: Text Summarization
<table>
<thead>
<tr>
<th>What is in my text?</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Which hotel features do my customers like/dislike</strong></td>
<td><strong>How customers’ sentiment have changed toward my hotels</strong></td>
</tr>
<tr>
<td>3000+ customer reviews</td>
<td>3000+ customer-posted reviews</td>
</tr>
</tbody>
</table>

**Insight Discovery: Sentiment Analysis**
## Text Analytics

### What is in my text?

<table>
<thead>
<tr>
<th>What are the correlations of tire problems and highway death in the NHTSA Data:</th>
<th>What are the correlations of patient gender and the cause of injury</th>
<th>Compare the customers’ attitude toward our product with theirs for our competitors</th>
</tr>
</thead>
<tbody>
<tr>
<td>450,000+ documents</td>
<td>70,000+ patient emergency room records</td>
<td>thousands of e-opinion postings</td>
</tr>
</tbody>
</table>

---

**Decision Making and Problem Solving: Text Analysis++**
Document Visualization Pipeline

Text/Document Mining

✓ Preprocessing
  • Data cleaning, extract keywords, etc.

✓ Feature Extraction
  • Keywords, word frequency distribution, topics

✓ Feature Evaluation
  • Similarity evaluation, clustering, etc.
Typical Text Mining Techniques

✓ Term frequency
  • TF-IDF
✓ Text clustering
  • K-means, etc.
✓ Topic model
  • LDA, HDP, PLSA
TF-IDF: term frequency-inverse document frequency

\[ W_{ik} = T_{ik} \times \log \left( \frac{N}{n_k} \right) \]

- \( W_{ik} \) = Weight of term \( k \) in document \( i \)
- \( T_{ik} \) = Frequency of term \( k \) in document \( i \)
- \( N \) = Total number of documents
- \( n_k \) = Number of documents having term \( k \)
Text/Document Visualization Tasks

Visualization for IR
Helping search

Visualizing text
Showing words, phrases, and sentences

Visualizing document sets
Words, entities & sentences
Analysis metrics
Concepts & themes
Grokker

A web search engine (based on Yahoo Search Engine) that clusters its results and presents them in a unique circular map.
Result Maps

Treemap-style vis for showing query results in a digital library
Visualization for IR
Helping search

Visualizing text
Showing words, phrases, and sentences

Visualizing document sets
Words, entities & sentences
Analysis metrics
Concepts & themes
Tag/Word Clouds

✓ Visual representation of text to depict keyword metadata (tags)
✓ To show the importance of word/concept through visual means (color, size, etc.)
  • Tags: User-specified metadata (descriptors) about something
  • Sometimes generalized to just reflect word frequencies
Flickr Tag Cloud

All time most popular tags

Africa animals architecture art australia autumn baby band barcelona beach berlin bird birthday black blackandwhite blue bw california cameraphone Canada canon car cat chicago china christmas church city clouds color concert coca cola day de dea dog england europe fall family fashion festival film florida flower flowers food football france friends fun garden geotagged germany girl girls graffiti green halloween hawaii hiking holiday home house india ireland island italy japan july kids lake landscape light live london macro me mexico mountain mountaineering museum music nature new newyork newyorkcity night nikon nyc ocean old paris park party people photo photography photos portrait red river rock rome san sanfrancisco scotland sea seattle show sky snow spain spring street summer sun sunset taiwan texas thailand tokyo toronto tour travel tree trees trip uk urban usa vacation vancouver washington water wedding white winter york zoo

What are tags?

You can give your photos and videos a "tag", which is like a keyword or category label. Tags help you find photos and videos which have something in common. You can assign up to 75 tags to each photo or video.
Wikipedia Word Cloud
Wordle: http://www.wordle.net

“I Have a dream” Wordle
Wordles
ManiWordle
Word Tree

Genesis (KJV)

god said

unto

let there be

behold

abraham,

h

serpent, because thou hast done this, thou art cursed above all cattle, and above every beast that is brie

noah,

the end of all flesh is come before me; for it is that thou hast done this, thou art cursed above all cattle, and above every beast that is brute.

this is the token of the covenant, which I make between me and you and every

sarah thy wife shall bear thee a son indeed; and thou shalt call his name is

thou shalt keep my covenant therewith, and my testimonies that I shall command thee this day.

as for sarai thy wife, thou shalt not let it not be grievous in thy sight be

thy name is jacob: thy name shall not

i am god almighty: be fruitful and multiply, and replenish the earth, and sute

jacob, arise, go up to bethel, and dwell there: and make there

under the heaven be the waters bring forth abundantly the grass, the herb the living crea
Web Seer

✓ Sensemaking with text data
✓ Web Based
Study behavioral patterns across the social web
SparkClouds

Tag Cloud + Sparkline to represent trend of Tag use.
Document Cards

Representing document’s as a mixture of images (by color histogram classification) and key terms (by Text Mining)
Visualization for IR
Helping search

Visualizing text
Showing words, phrases, and sentences

Visualizing document sets
Words, entities & sentences
Analysis metrics
Concepts & themes
Theme River

- Nationalization of property begins
- Cuba and Soviet relations resume
- Castro confiscates American refineries
- Eisenhower breaks relations
- Bay of Pigs

- Yankee (63)
- Reform (48)
- Soviet (49)
- Oil (44)
- Imperialists (29)
- Cooperatives (16)
- Cane (7)

Months: May to July 1961

Dates: May 1961 to May 1962
Y axis encodes topic significance

(1) A layer represents a topic

(2) Topic keywords

(3) Doc #

~10,000 emails in 2008
Tiara Demo

Interactive, Time-based Visual Email Summarization

Shixia Liu, Michelle X Zhou, Shimei Pan, Weihong Qian, Weijia Cai, Xiaoxiao Lian

IBM Research
TIARA:
A visual text summarization system

Text collection

Text Preprocessing

Text content + meta data

Text Summarization

Summarization results

Data Transformation

Transformed results

Visualization

User Interaction
LDA Data Transformation

A set of topics
\{T_1, \ldots, T_i, \ldots, T_N\}

A set of topic probabilities
\{\ldots, P(T_i \mid D_k), \ldots\}

A set of keywords
\{W_1, \ldots, W_j, \ldots, W_M\}

A set of word probabilities
\{\ldots, P(W_j \mid T_i), \ldots\}

Rank the topics to present most interesting ones first

Select keyword sub-sets for time-based summary

{\ldots}_{t-1}, {\ldots, W_j, \ldots}_t, {\ldots}_{t+1}
Topic Ranking

✓ Rank topics by “strength”

\[ \text{rank}(T_k) = f(\mu(T_k), \sigma(T_k), \alpha(T_k)) \]

\[ \mu(T_k) = \frac{\sum_{m=1}^{M} N_m \hat{\theta}_{m,k}}{\sum_{m=1}^{M} N_m} \]

✓ Rank topics by “distinctiveness”

\[ \text{rank}(T_k) = l(T_k) = \frac{\tilde{\nu}_k^T L \tilde{\nu}_k}{\tilde{\nu}_k^T D \tilde{\nu}_k} \]

\[ \sigma(T_k) = \sqrt{\frac{\sum_{m=1}^{M} N_m (\hat{\theta}_{m,k} - \mu(T_k))^2}{\sum_{m=1}^{M} N_m}} \]

Domain-dependent activeness metric

graph degree matrix

doc-topic distribution

for each \( T_k, \nu_k = (\hat{\theta}_{1,k}, \hat{\theta}_{2,k}, \ldots, \hat{\theta}_{M,k})^T \)

\( \tilde{\nu}_k \) is normalized \( \nu_k \)
Visual Text Summary Metaphor

Data to be visualized:

1. Topics: \( \{T_1, \ldots, T_i, \ldots, T_N\} \) and their probabilities
2. Topic keywords by time: \( \ldots \left\{ \ldots, w^i_k, \ldots \right\}_t, \ldots \)
3. For each \( T_i \), Topic strength: \( \{\ldots, S^i(t), \ldots\} \) over time

Visual encoding: Augmented stacked graph
Enhanced Stacked Graph: Key Steps

- Computing geometry of layers
- Layer coloring
- Layer ordering
- Layer labeling
Layer Ordering

✓ Goals
  • Minimize distortion
  • Maximize usable space
  • Ensure semantic coherence

unordered

ordered
Layer Ordering - Comparison (before)
Layer Ordering – Comparison (after)
Interacting with Visual Summary

“people” involved and their relationships
Application Example: Healthcare

✓ Visualize text to facilitate analysis
  • Cause of injury
  • Reason for visit
  • Diagnosis

✓ Multiple fields of text data and their correlation

✓ Leverage structured data to help better illustrate text information
  • Gender + Cause of injury
Correlation between Structured and Text Fields
Correlation between Text Fields

Correlation between two fields, *diagnosis* and *reason for visit*.
TextFlow

Understanding topic evolution and correlations in large text collections

• Find related works in a publication set
• Examine a large collection of emails and instant messages
• Examine online posts to identify the key public opinion and concern
Topic strength evolving over time

Doc #

Topic splitting

Topic merging
Two keywords co-occurring in the topic for a time span

Keyword splitting
TextFlow: Towards Better Understanding of Evolving Topics in Text
Example: IEEE Visualization Publications
Example: Vis Publications

- "structure/layout"
- "exploration.analytics"
- "document/temporal"
Tracking the Evolution of Stories
Each line represents one character
StoryFlow

- Level-of-detail rendering
System

(a) Layout Pipeline

Hierarchy Generation → Ordering → Alignment → Compaction

(b) User Interactions

Adding & Deleting → Dragging → Straightening → Bundling
Jurassic Park
Inception
Inception
Event Tracking: Presidential Election

89,174,308 tweets in total during May 8th to Nov. 13th, 2012
Overall Patterns (1/2)

✓ Three user groups focus mainly on *Election*
  • Grassroots focus on Economy and switched frequently
  • Political figures are more focused
  • Media occasionally switched
Overall Patterns (2/2)

✓ Five significant peaks on *Election*
  • Third debate on foreign affairs
Significant transitions (1/3)

✓ Transition from *Election* to *Defense*

Timeline

- **Defense**
- **Election**
- **Economy**
- **Welfare**
- **Horse Race**

Timeline:

- First debate
- VP debate
- Second debate
- Third debate
- Voting

- benghazi
- Forward
- doesn’t obama owe us answers (on #benghazi)?

- americafoward
Significant transitions (2/3)

✓ Transition from Election to Economy

think Romney is tough on China? Ask the workers of #sensata about that as they train their Chinese replacements
Significant transitions (3/3)

✓ Transition from Election to Welfare

Issue-attention cycle