LDAExplore
Visualizing Topic Models Generated Using LDA
Ashwinkumar Ganesan, Kiante Brantley,
Shimei Pan & Jian Chen
What is Topic Modeling?

- Find hidden topics
- Process Large Sets of Documents
- Group words & information together
- Understand topics & documents
It should not be thought from above that Soviet party line is necessarily disingenuous and insincere on part of those who put it forward many of them are too ignorant of outside world and mentally too dependent to question (*self-hypnosis,* and who have no difficulty making themselves believe what they find comforting and convenient to believe. Finally we have the unsolved mystery as to who, if anyone, in this great land actually receives accurate and unbiased information about outside world. In atmosphere of oriental secretiveness and conspiracy which pervades this government, possibilities for distorting or poisoning sources and currents of information are infinite. The very disrespect of Russians for objective truth—indeed, their disbelief in its existence—leads them to view all stated facts as instruments for furtherance of one ulterior purpose or another. There is good reason to suspect that this governemt is field of international law. This publication has been written with the expectation that the military attorneys making use of it will be provided with a basic understanding of the legal system governing the international community. International law is an area of jurisprudence which challenges. It quite often fails to provide concise "textbook answers" to problems which reach a degree of complexity far greater than that found in any other legal system. Entrained with the task of regulating the conduct of international sovereign entities, it is a legal framework which develops on a daily basis. Its successes go largely unnoticed, while its failures gain almost instantaneous notoriety and condemnation. It is a jurisprudential system particularly unused for complainant personalities and regimented minds. Hopefully, military attorneys will not view the often evident imprecision of international law as a fatal weakness but as an opportunity afforded its practitioner to develop an efficient and viable legal system. Constructive criticism and the ability to apply concepts and rules to practical international legal problems must be based on a working knowledge of the subject matter. The achievement of this end underlies the purpose of this publication.
• Probabilistic method to find hidden topics
• Word distributions
• Topic distributions
• Link topic, words & documents
Problems with LDA

- Hidden Parameters
- Number of Topics
- Number of Iterations
- Naïve or Knowledgeable Users
- Understanding Documents
Need For Visualization
Example User Work Flow

- **Topics**
  - Check Word Distributions
  - Isolate useful topics

- **Documents**
  - Remove unnecessary topics
  - Filter the documents based on Rank or Probability

- **Cluster**
  - Combine Similar Documents
  - Refine topics and provide feedback for a new model

Annotate Topics
Related Visualizations

SERENDIP ³

TIARA ⁴
Problems With Visualizations

- Document Exploration without modeling User Feedback
- Hyper parameters that need configuration
- Topics Generated are difficult to understand
- Large Document Corpus require Multiple Views / Scrolling
- Real Time computation of model is difficult
Design Goals

• Visualizing Topics in the document corpus
• Topic Document Relations
• Filtering Documents
• Performing Set Operations
• Clustering Topics & Documents
• Topic Annotations
Visualization Pipeline

<table>
<thead>
<tr>
<th>BACKEND</th>
<th>FRONTEND</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extract Data</td>
<td>Render Design Using D3</td>
</tr>
<tr>
<td>Transform to JSON</td>
<td>Extract JSON</td>
</tr>
<tr>
<td>Perform LDA</td>
<td></td>
</tr>
</tbody>
</table>

JSON FORMAT

11
Overall Design

Topic – Document Relations

Topic Information

Individual Filter

Word to document search
Topic Word Distributions

- Shows the Top 10 Words
- Rectangular sizes show which word is more probable
Topic – Document Relations
Keyword Searches

Documents:
- A Data Driven Approach to Hue-Preserving Color-Blending
- A Deeper Understanding of Sequence in Narrative Visualization
- A Design Space of Visualization Tasks
- A Lightweight Tangible 3D Interface for Interactive Visualization of Thin Fiber Structures
- A Model for Structure-Based Comparison of Many Categories in Small-Multiple Displays
- A Multi-Criteria Approach to Camera Motion Design for Volume Data Animation
- A Multi-Level Typology of Abstract Visualization Tasks
- A Novel Approach to Visualizing Dark Matter Simulations
- A Partition-Based Framework for Building and Validating Regression Models
- A Perceptual-Statistics Shading Model
- A Scale Space-Based Persistence Measure for Critical Points in 2D Scalar Fields
- A Study on Dual-Scale Data Charts
- A Systematic Review on the Practice of Evaluating Visualization
- A User Study on Curved Edges in Graph Visualization
- A Visual Analytics Concept for the Validation of Geoscientific Simulation Models
- A Visual Analytics Approach to Multiscale Exploration of Environmental Time Series
- A case study: Tracking and visualizing the evolution of dark matter halos and groups of satellite halos in cosmology simulations.
- A correlational analysis process in a visual analytics environment
- A generic model for the integration of interactive visualization and statistical computing using R
- A visual analytics approach to understanding cycling behaviour
- About the Influence of Illumination Models on Image Comprehension in Direct Volume Rendering
- Acuity-Driven Gaze-based Visualization
- Adaptive Composite Map Projections
- Adaptive Extraction and Quantification of Geophysical Vertices

Search Documents on Top Words: [Search options: different, analysis]
- data, visualization, interactive, users, use, visual, visualizations, study, different, analysis
- data, visualization, visual, different, analysis, user, study, information, approach, design
- data, visualization, visual, visualizations, different, analysis, information, approach, design, user

Key word Searching
• Using TF-IDF as a measure to clean the data
• There is a preset threshold for eliminating the words

• Word Probabilities:

\[
P(w_x, dy) = \sum_{t=1}^{n} \sum_{i=1}^{n} P(w_x, t_i) \times P(t_i, dy)
\]
• There are 5 participants in the survey

• The participants did not have any formal training on the tool

• There is a brief introduction given to the participants at the start of the survey

• Some participants understand topic modeling
Pilot Study Details (2)

Survey Questionnaire

• Overview
• Topics
• Filtering
• Keyword Search
Types Of Questions

• **Tasks Execution questions**
  How many documents are represented in this visual?

• **Understanding questions**
  Does the visual have many things on it?

• **Reasoning questions**
  Which is the least important topic? (Important means highly ranked topic)

• **Usability questions**
  Are the rankings on the axis visible / readable?
• Eliminate Documents
• Connect topics and parallel coordinates
• Combine Searching
• Keep & Exclude can be connected to searching
• Creating a reset button
• Create probability mode & Ranking mode
• Topic – Rank group
• Filtering & Selection memory
Thank You
Questions?
References


Survey Results

**How many documents are represented in this visual?**

<table>
<thead>
<tr>
<th>Number</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>128</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>322</td>
<td>2</td>
<td>100%</td>
</tr>
<tr>
<td>256</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>I can't find the information</td>
<td>0</td>
<td>0%</td>
</tr>
</tbody>
</table>

**How many topics are there in this visual?**

<table>
<thead>
<tr>
<th>Number</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>100%</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>I can't figure it out</td>
<td>0</td>
<td>0%</td>
</tr>
</tbody>
</table>
Survey Results

Is it easy to filter using the above method?

- Yes [1] 1 50%
- Somewhat [1] 1 50%
- No [0] 0 0%

Apply a filter for topic T7 across the range of ranks 1 to 2. What is the relation between topic T7 and T19?

- T7 is more important than T19 [2] 2 100%
- T19 is more important than T7 [0] 0 0%
- Can't Say [0] 0 0%
Survey Results

Only by looking at the given topics grid, which topics do you feel are important?

- T5, T7, T1: 0, 0%
- T6, T8: 0, 0%
- All are the same: 2, 100%

Move the filter on T19 axis from top to bottom. What do you notice?

Keywords change. Relevant documents in the filter range are highlighted in black while others are grayed out. This is not obvious at first though.

They all are about data, visualization and there are documents seem to be evenly distributed throughout the axis.
Survey Issues

- Ambiguity in Questions
- Imprecise human movements
- Filtering issues
- Remote tests
- Adding Ease of Use questions
- Adding a purpose section
- Better Example can be provided
- Data Selection