

1 Basic Information

Name Peter Olson
Email p.olson@wustl.edu
Student ID 441666
Repository GitHub repository link

2 Background and Motivation

During my junior year, I took Dr. Anupam Basu's Introduction to Digital Humanities course, in which we discussed topic models. In our class, we explored a tool called Serendip, which visualizes topics as a matrix of relationships between texts in the corpora and topics. Then, this semester, I took Dr. Neumann's Analysis of Network Data class, and it occurred to me that Serendip's matrix presentation of the topic model is visually similar to an adjacency matrix.

I want to take a distinctly different approach than Serendip, though. While Serendip focuses on drilling down into specific relationships and presenting a large volume of information at once, I want to focus more on exploring the topic-story relationship in a more approachable, intuitive manner.

3 Project Objectives

I would like to build an interface that allows for the creation and exploration of simple topic models, with a focus on drawing out the abstract relationship between a topic and a set of stories. I'd like to emphasize exploring the relationship as a two-way street, not just from the perspective of looking at the topic model of individual stories.

4 Data

Because topic modelling is such a versatile tool, I can create corpora of almost any source of textual data. Hypothetical sources include the lyrics of popular songs, well-known novels, plays, or even technical documents, such as white-papers and specifications.

At least initially, I'm planning on following in the foot steps of Introduction to Digital Humanities and creating a corpus of books from Project Gutenberg.

5 Data Processing

As far as manual cleanup of the corpus itself, I'm going to need to remove any extraneous text, so that the source files are only the raw text documents for the corpus. In the case of

documents from Project Gutenberg, I'll need to remove the header and footer that Project Gutenberg attaches. Other than the header and footer, the text files available there are relatively free of markup and extraneous information.

I'm planning to do the topic model generation in the browser. There are several Javascript libraries (such as jsLDA) that implement topic modelling without relying on running a server. I also have a rudimentary implementation of topic model generation that I could use, but it's definitely inferior to libraries developed by actual machine-learning developers. I'm just a computer engineer.

6 Visualization Design

The center of my visualization is going to be a force-network visualization of the topic-story relationships. Accompanying this representation, I'm going to also have accompanying histogram views of the occurrences of topics in individual stories, as well as a pane that simply displays detailed information about a story or topic if selected in the force-network.

The thesis of my design is that I want to push that the relationship between topics and documents isn't necessarily hierarchical. That is to say that topics aren't exclusively observations on documents, and documents aren't solely representations of topics. By representing both topics and documents as nodes in a force-network, they're given explicitly equal roles in the way that the network is constructed. I'm toying around with the idea of having the ability to switch which class of node generates force within the graph, so that each class of node can be positioned in the graph in relation to the other.

My design very much benefits from related work done in Serendip and SerendipSlim, as those tools act as useful foils for my visualization. I explicitly want to detach from the hyper-analytical approach that Serendip takes. Serendip is mostly interested in analyzing the topic model itself, whereas I am more focused on exploring the relationships created between stories and topics.

7 Must-Have Features

- Topic-Model generation
- Presentation of resulting topics
- Network graph of topic-story relationships
- Alternative content-view for presenting per-story topic breakdown
- A "details" pane that exposes more information about a topic or story when selected.

8 Optional Features

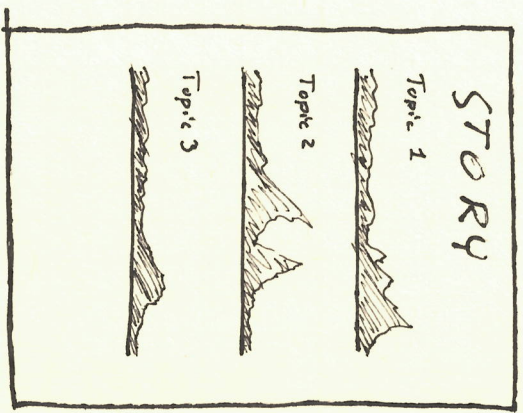
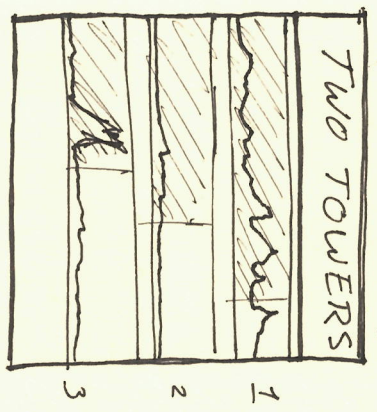
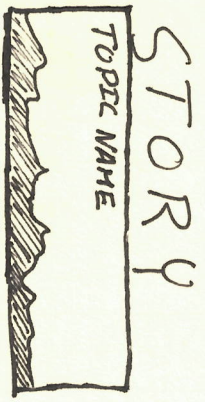
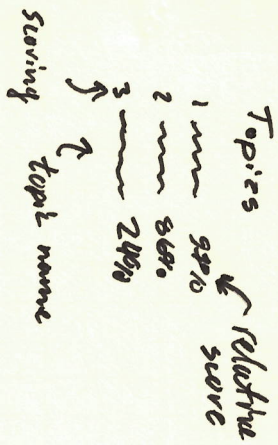
- Customization of topic-model parameters
- Coding the weight of the relationship strength into the link between stories and topics
- Some means of filtering the documents displayed
- Preset configurations for the network
- Lasso selection of nodes in the network display

9 Project Schedule

	Nov. 9th	Topic-model debug output Dummy network with working forces
Milestone I	Nov. 13th	Card for topics on the sidebar
Milestone II	Nov. 27th	Filtering topics Link network with topics
In-Class Studies	Nov. 29th	
	Dec. 7th	Outline presentation Process book
	Dec. 14th	Finish presentation Glitz and color
Submission + Presentation	Dec. 19th	

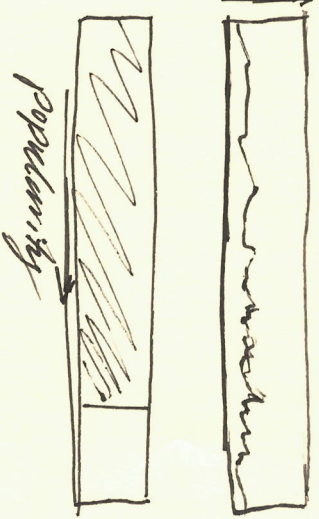
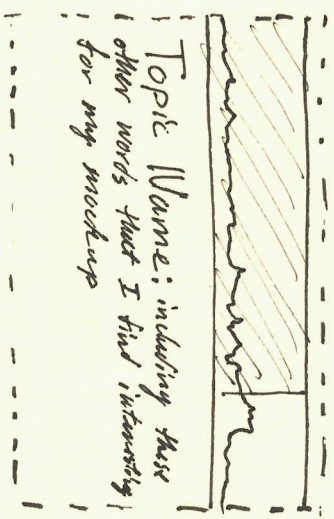
Question: how do you communicate topic relevance to the corpus?

↳ how does one see the topics? → Intrinsically suggests a listing



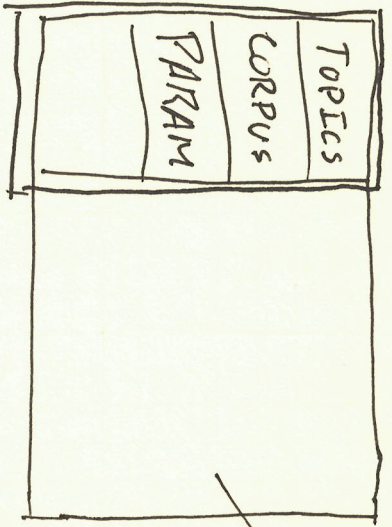
Topics through the story driven by the story

→ This

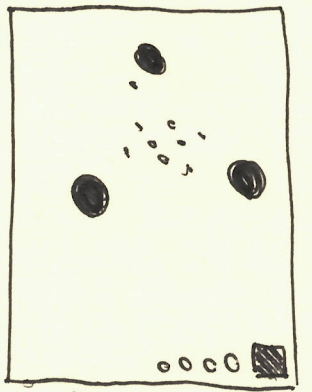


~~100%~~

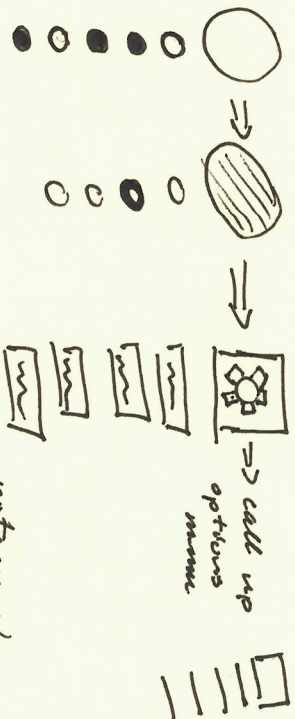
Main window wireframe



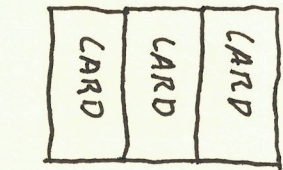
force network...
or not?
general work space



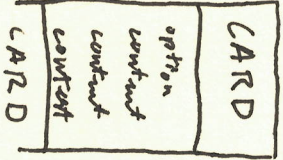
floating buttons
Ⓟ - options & mouse & keyboard



call up options menu

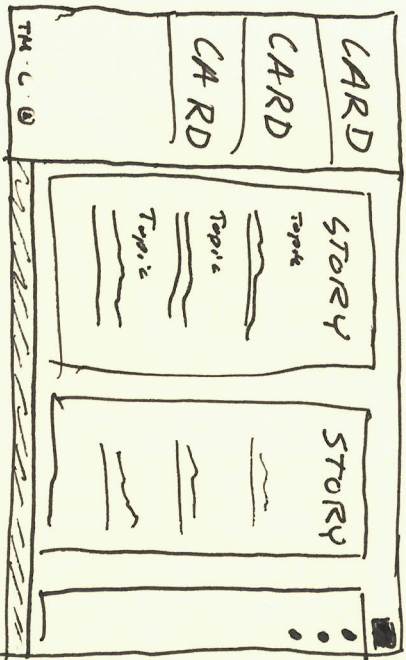


x click



very usual
hardanger-
menus-style
interaction

governed by object-model
generation & manipulation

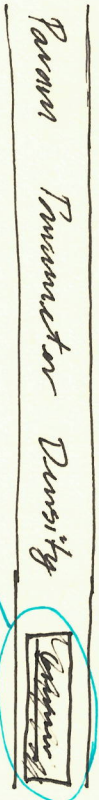
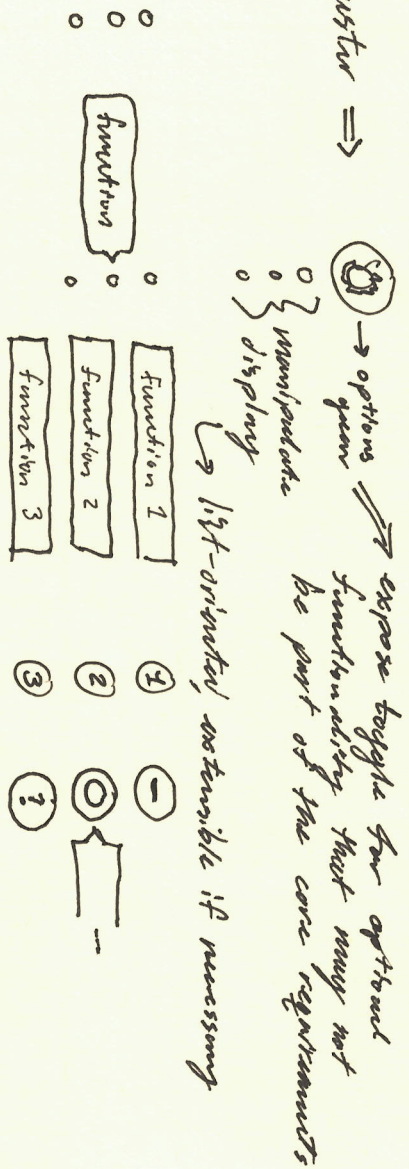
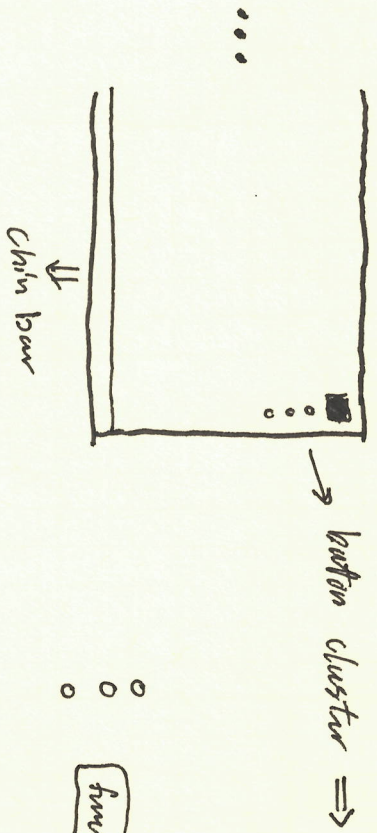


not even story view

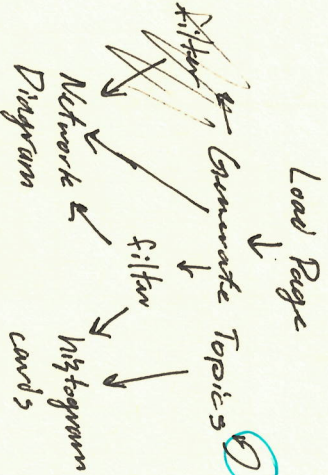
↳ To chinbar or not to chinbar?

*

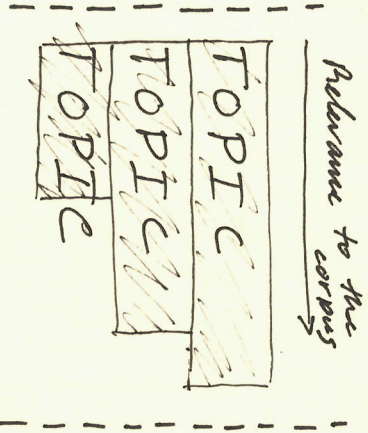
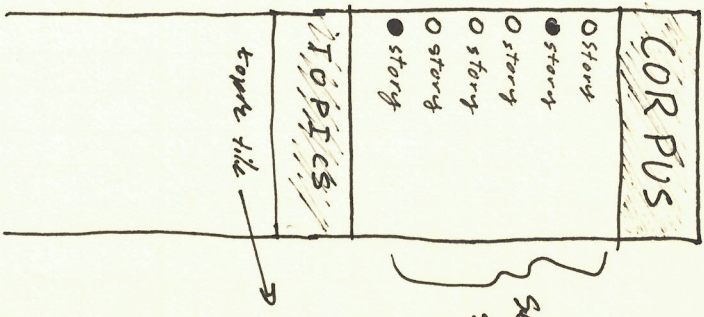
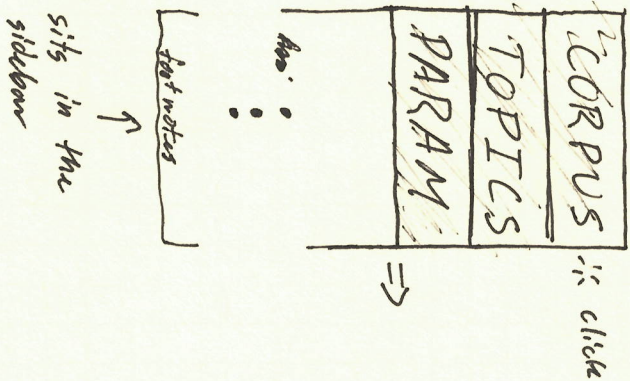
Network Interface



Keep pages
 points of
 user interaction
 visible at all times
 User flow?



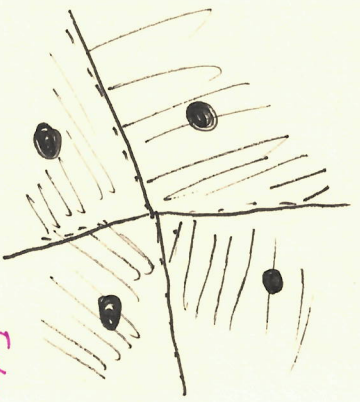
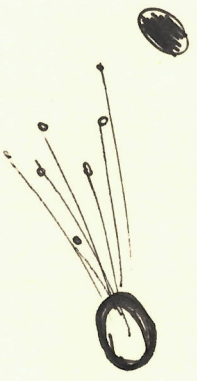
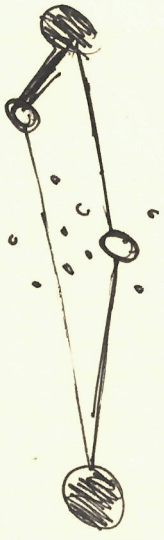
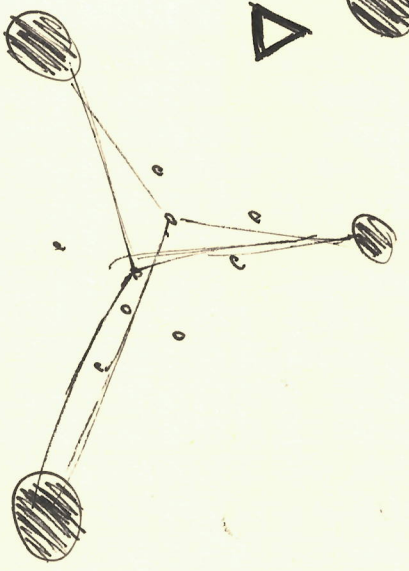
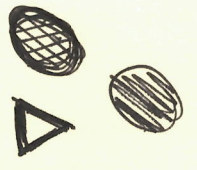
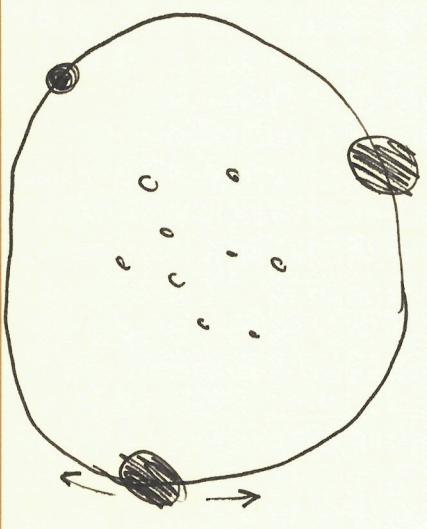
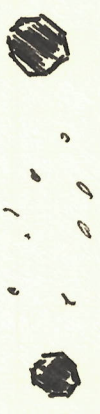
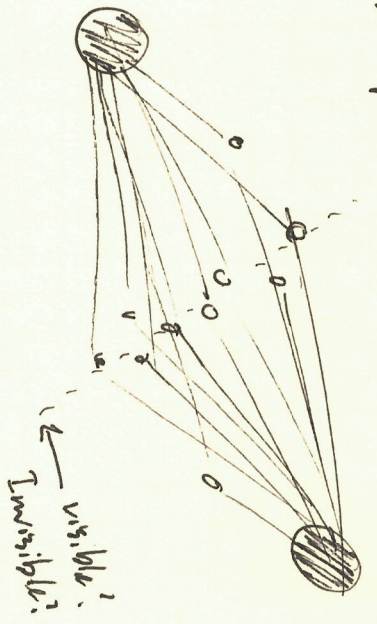
Present topics and stories in a visually intuitive way => how to really drill into detail?
 ↳ Back to the CARD motif



sit in the sidebar or overlay with content window?

"Parameters" pane? **Object model**

Diagram the force network



Voronoi diagram? Overlay? Extra?