Digital Humanities Across Borders

Class 8: Topic Modeling

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA)

1. There are a fixed number of patterns of word use, groups of terms that tend to occur together in documents. Call them *topics*.

Latent Dirichlet Allocation (LDA)

1. There are a fixed number of patterns of word use, groups of terms that tend to occur together in documents. Call them *topics*.
2. Each document in the corpus exhibits the topics to varying degree.

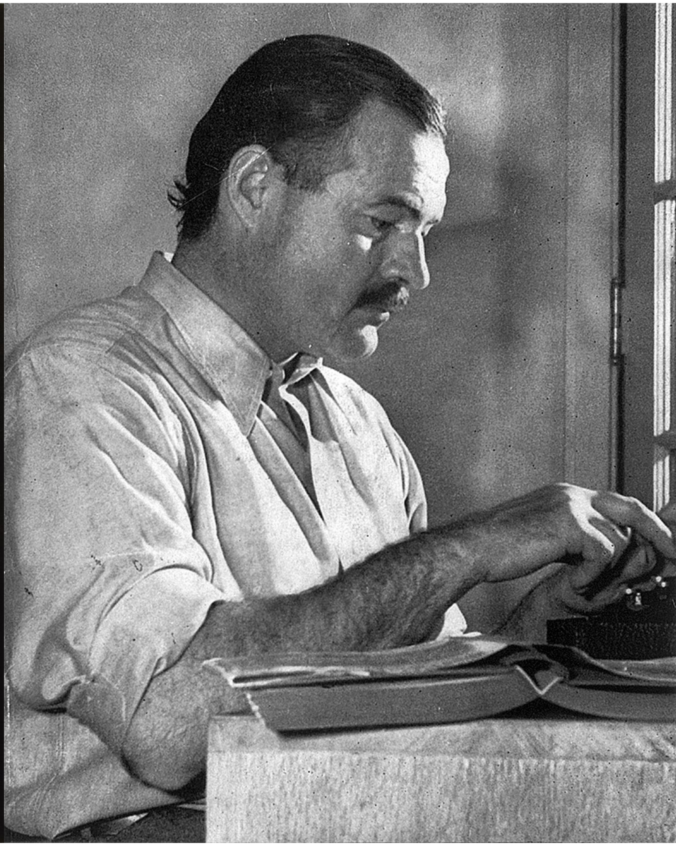
Latent Dirichlet Allocation (LDA)

1. There are a fixed number of patterns of word use, groups of terms that tend to occur together in documents. Call them *topics*.
2. Each document in the corpus exhibits the topics to varying degree.

Topic modeling uncovers the structure of these "topics" by identifying the probability that every word in the text is in a given topic.

The LDA Buffet

A parable by Matt Jockers



How does LDA work?

1. First choose the topics, each one from a distribution over distributions.
2. Then, for each document, choose topic weights to describe which topics that document is about.
3. Finally, for each word in each document, choose a topic assignment — a pointer to one of the topics — from those topic weights and then choose an observed word from the corresponding topic.
4. Each time the model generates a new document it chooses new topic weights, but the topics themselves are chosen once for the whole collection

tf-idf

term frequency - inverse document frequency

"Simply put, a tf-idf score is the frequency of a word multiplied by the total number of documents and divided by the number of documents containing the word"

- David Hoover

"It's all text and no subtext"

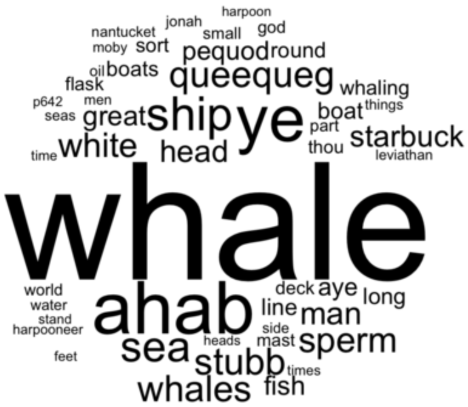
The algorithm is constrained by the words used in the text; if Freudian psychoanalysis is your thing, and you feed the algorithm a transcription of your dream of bear-fights and big caves, the algorithm will tell you nothing about your father and your mother; it'll only tell you things about bears and caves.

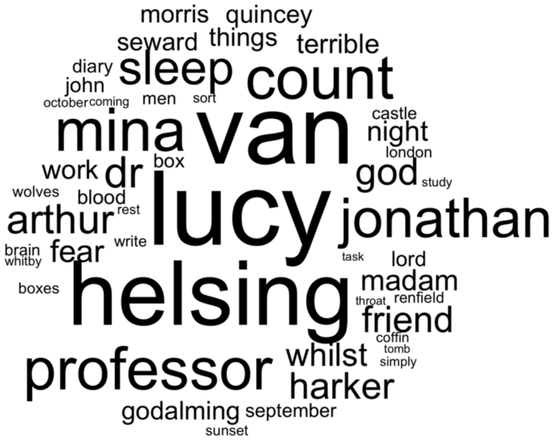
-Scott Weingart, http://www.scottbot.net/HIAL/index.html@p=19113.html

Pre-processing is key

word clouds from Matt Jockers

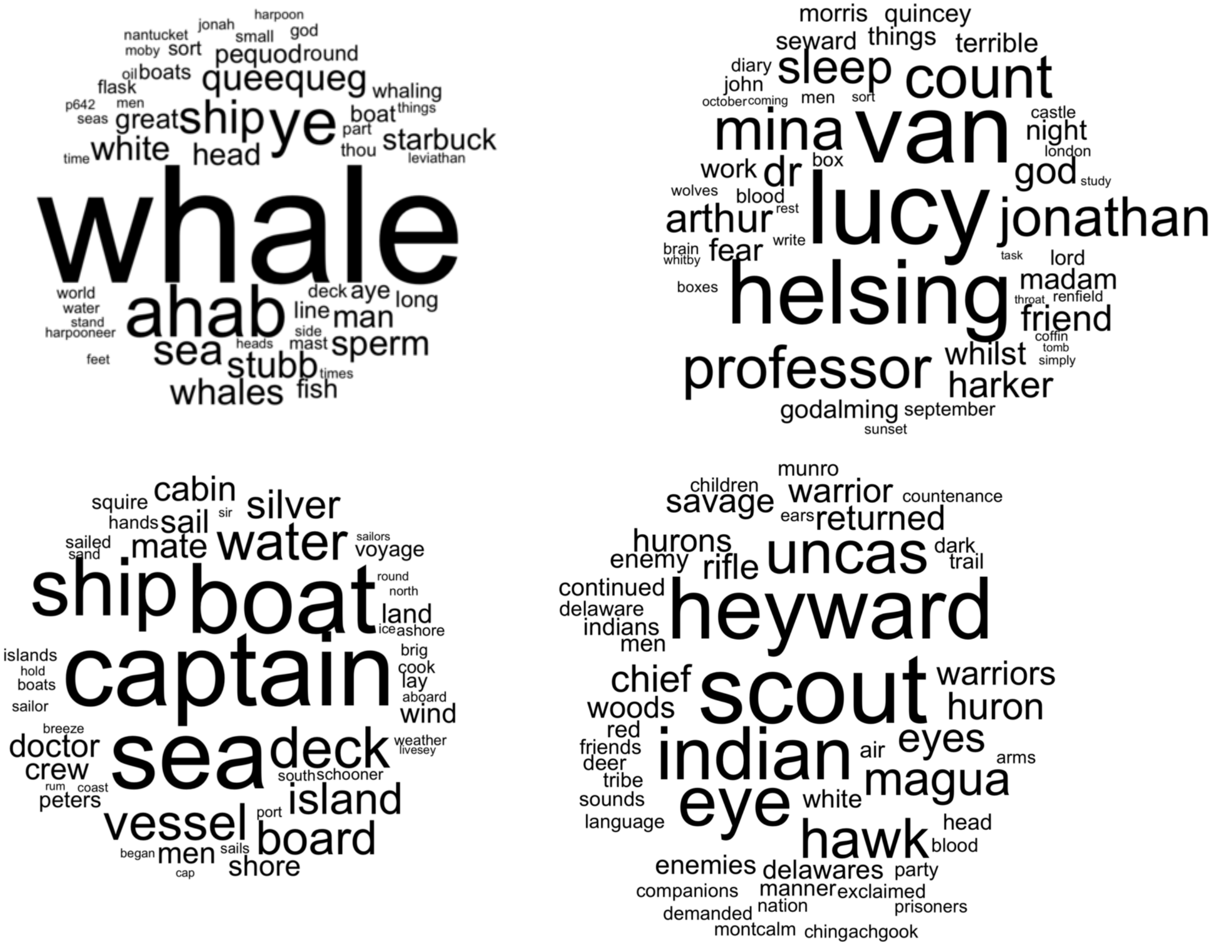
Pre-processing is key





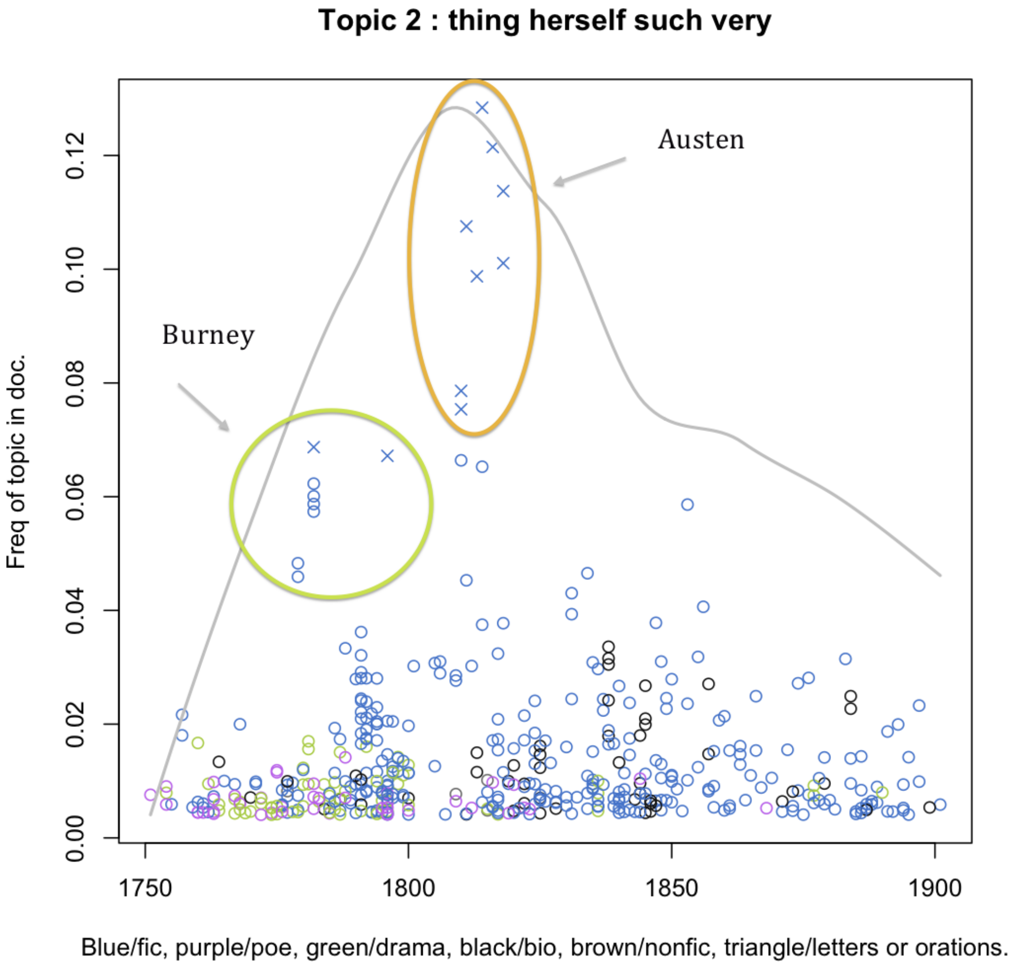
word clouds from Matt Jockers

Pre-processing is key



word clouds from Matt Jockers

Author signal



Graph from Ted Underwood

Assumptions

To make topic models present new raw material for humanists to read, analysts generally assume that an individual topic produced by the algorithm has two properties. First, it is coherent: a topic is a set of words that all tend to appear together, and will therefore have a number of things in common. Second, it is stable: if a topic appears at the same rate in two different types of documents, it means essentially the same thing in both. Together, these let humanists assume that the co-occurrence patterns described by topics are meaningful; topics are useful because they describe things that resemble "concepts," "discourses," or "fields."

- Benjamin Schmidt

Information retrieval vs. literary analysis

A poorly supervised machine learning algorithm is like a bad research assistant. It might produce some unexpected constellations that show flickers of deeper truths; but it will also produce tedious, inexplicable, or misleading results. If it is not doing the desired task, it is time to give it some clearer instructions, or to find a new one.

- Benjamin Schmidt

Trying out the Topic Modeling Tool

Tutorials by Miriam Posner:

https://github.com/miriamposner/tmt_get_started

http://miriamposner.com/blog/very-basic-strategies-for-interpreting-results-from-the-topic-modeling-tool/