Exploratory Text Mining

School of Information Studies
Syracuse University

Typical Text Mining Tasks

Exploratory text mining

Predictive text mining

Exploratory Text Mining

Corpus statistics

Document clustering (k-Means)

Topic modeling (LDA)

Corpus Statistics

Word frequency

KWIC (keyword in context)

Document Clustering

Cluster documents based on their similarities and differences

Similarity/distance measure

The k-Means algorithm

Applications of Document Clustering

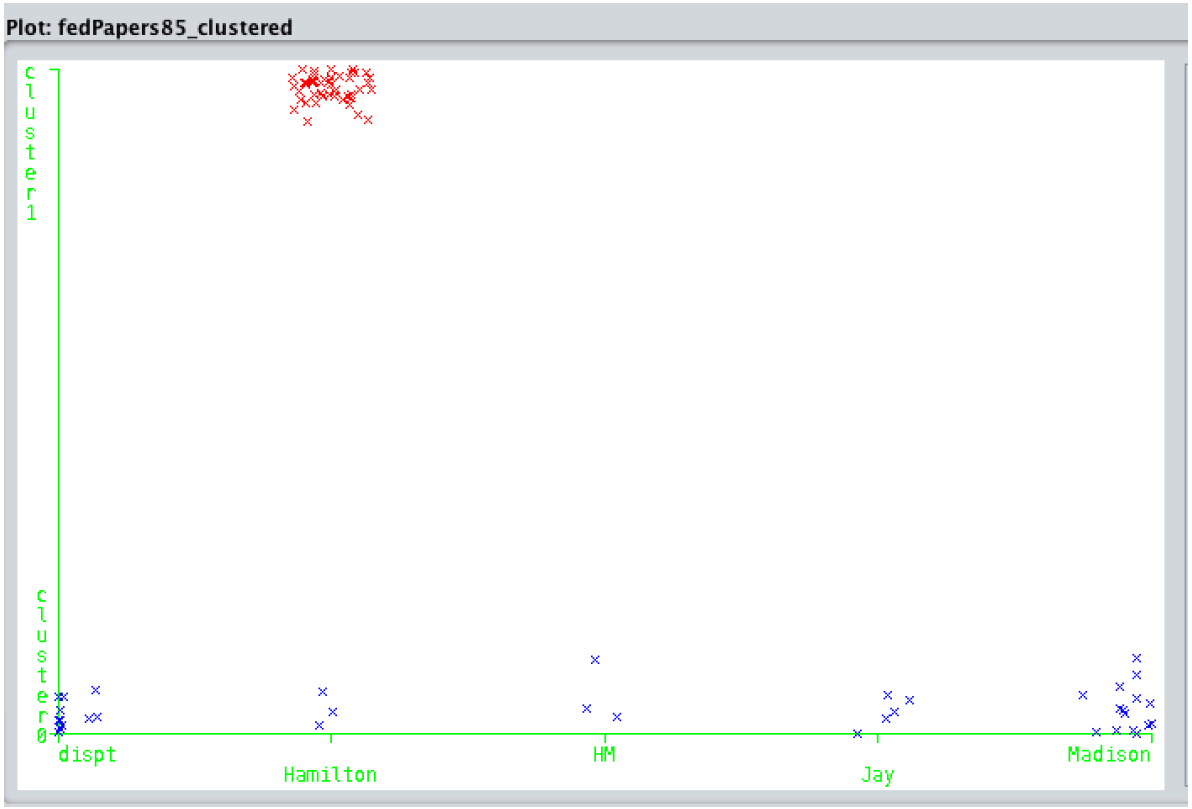
Authorship attribution

- Plagiarism detection

Grouping students, job applicants, etc.

Clustering search results

Document Clustering for Solving Mystery in History



Topic Modeling

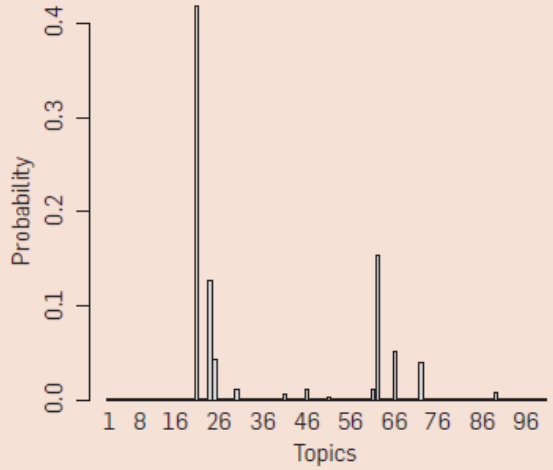
Finding the main topics in a text collection

The LDA algorithm

- Every topic is a probability distribution of all words in the vocabulary

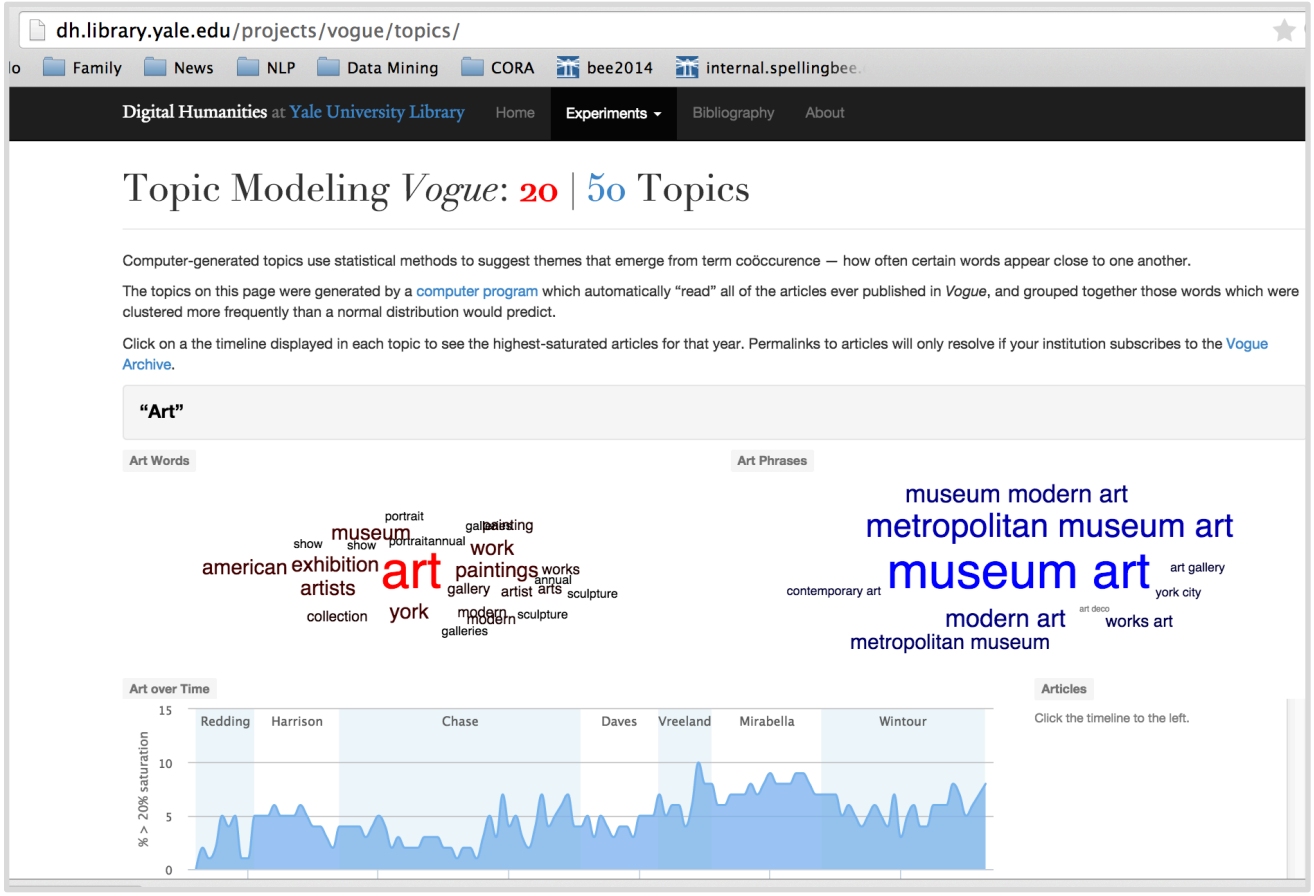
Topics in the Science Journal

Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Trend of Topics in Vogue



http://dh.library.yale.edu/projects/vogue/topics/

School of Information Studies
Syracuse University