Text Representation/ Vectorization

School of Information Studies
Syracuse University

What Is Text Mining?

Knowledge discovery from text data.

Computers do not understand human language; they just count.

Three-step process:

1. Convert text documents to numeric vectors
2. Find patterns in the numeric vectors
3. Explain the patterns' semantic meaning

What to Count?
Text Representation

Which six words/phrases did you choose to describe yourself?

If you were a text document, then the six words are a representation of you as a document.

Are your description words similar or different from other students' words?

Text Representation

A document can have many properties

Which property to represent?

- Topic
- Sentiment and opinion
- Genre
- Author
- Writing style
- Confidence
- …

Bag-of-Words (BoW) representation

How to Count? Vectorization

Step 1: Create a dictionary of all unique words.

1. "glasses"
2. "smart"
3. "tired"
4. …

Step 2: Represent every document as a word vector: each word is an attribute/feature.

	"glasses"	"smart"	"tired"	…
Jack	1	1	0	
Jill	0	1	0	
Ben	1	1	1	

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

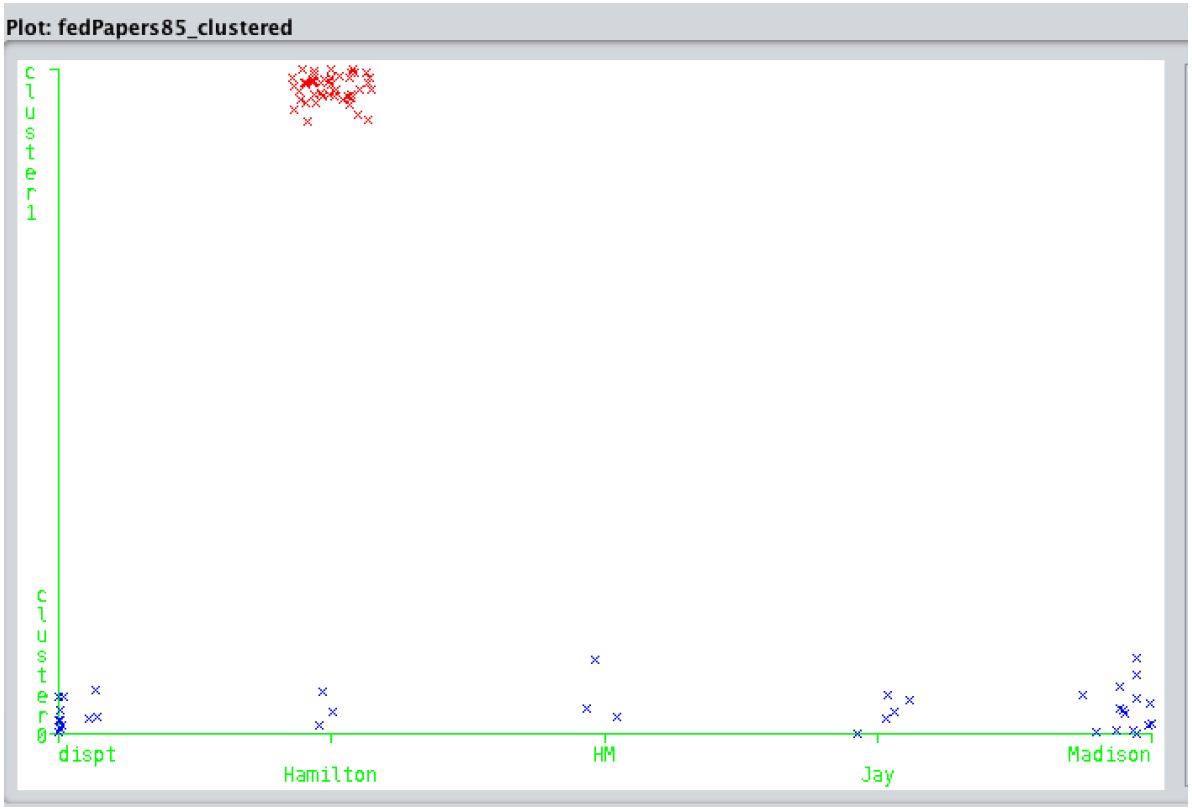
Authorship attribution

- Plagiarism detection

Grouping students, job applicants, etc.

Clustering search results

Document Clustering for Solving Mystery in History



Topic Modeling

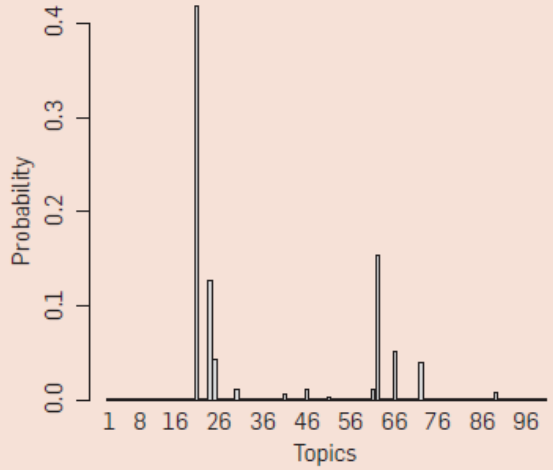
Finding the main topics in a text collection

The LDA algorithm

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

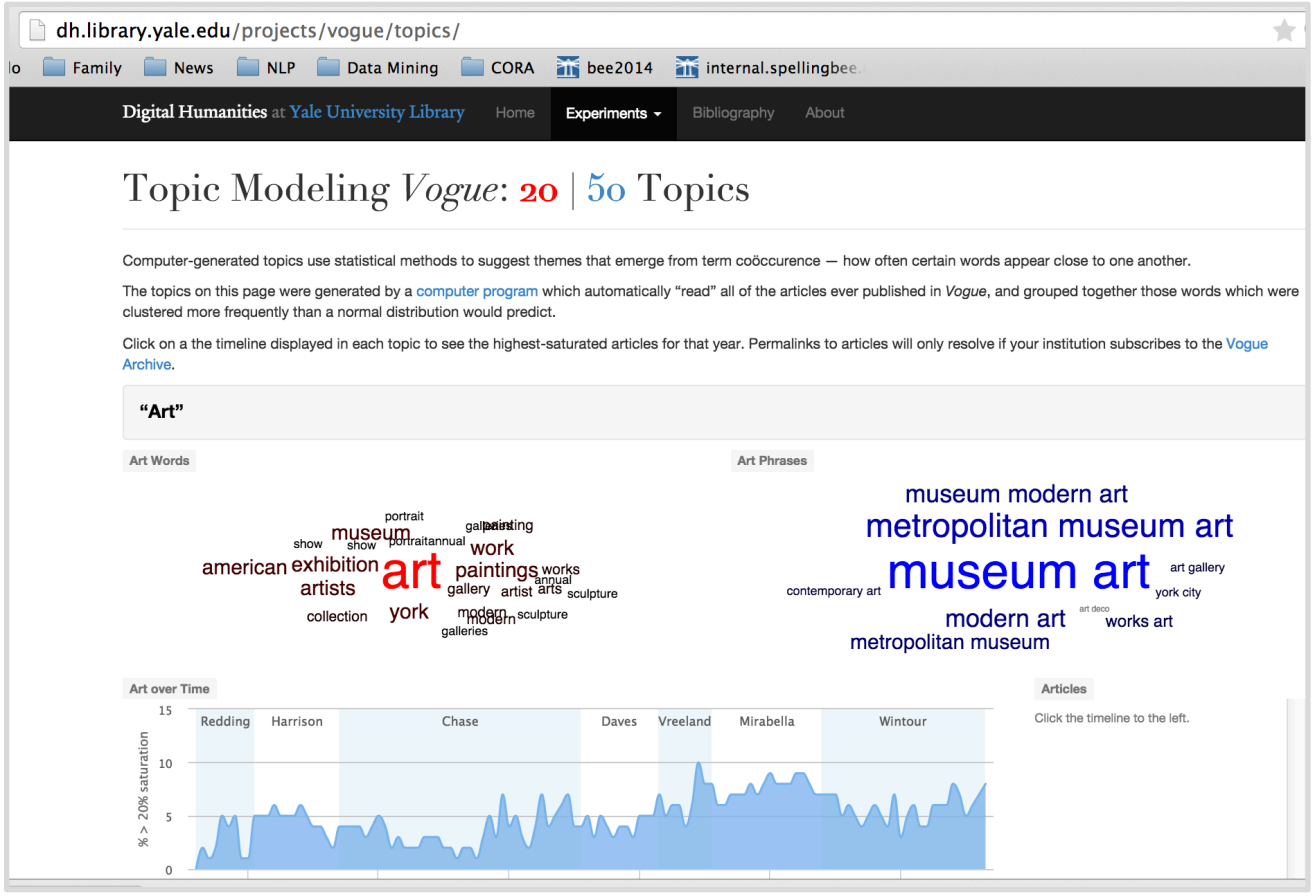
Topics in the Science Journal

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

Predictive Text Mining

School of Information Studies
Syracuse University

Predictive Text Mining
Text Categorization

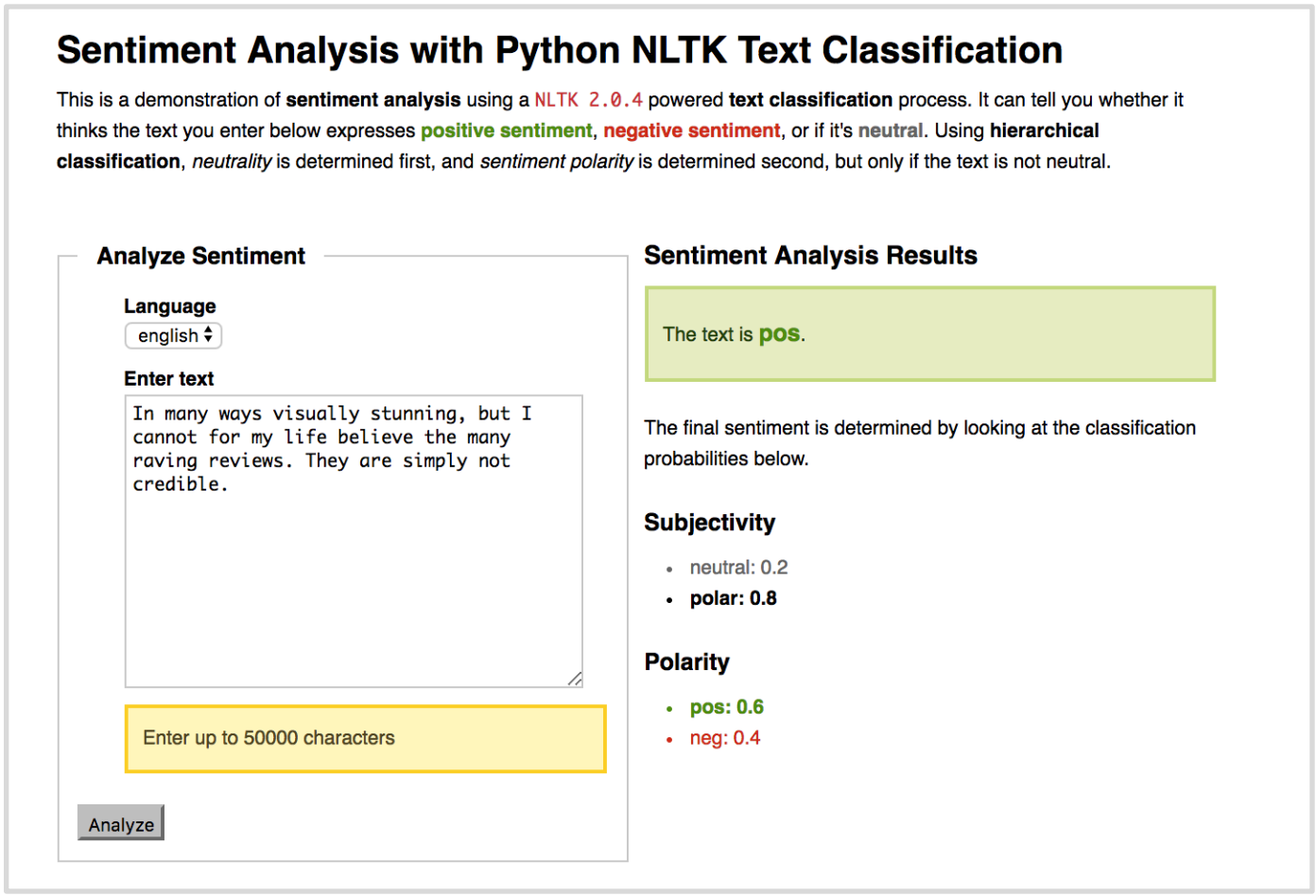
Given pre-defined categories and some training examples, automatically assign documents to categories

Applications

- Sentiment classification
- News topic classification
- Genre classification

Naïve Bayes and SVMs algorithms

Sentiment Analysis



What to Categorize?

Topics

- Categorize news articles to pre-defined topic categories: politics, finance, sports, science

Genres, styles, authors

Or just anything you can define with training examples

- Quality of writing
- Readability
- Ideology

Annotating Training Examples

Example: sentiment annotation

- Positive, negative, neutral
- Or more granularity

Quality of annotations

Difference Between Text Mining and NLP

School of Information Studies
Syracuse University

Text Mining and NLP

NLP: deep linguistic analysis

- A typical NLP package includes:
  - Part-of-speech tagger
  - Parser
  - Named-entity recognizer
  - Co-reference resolution system

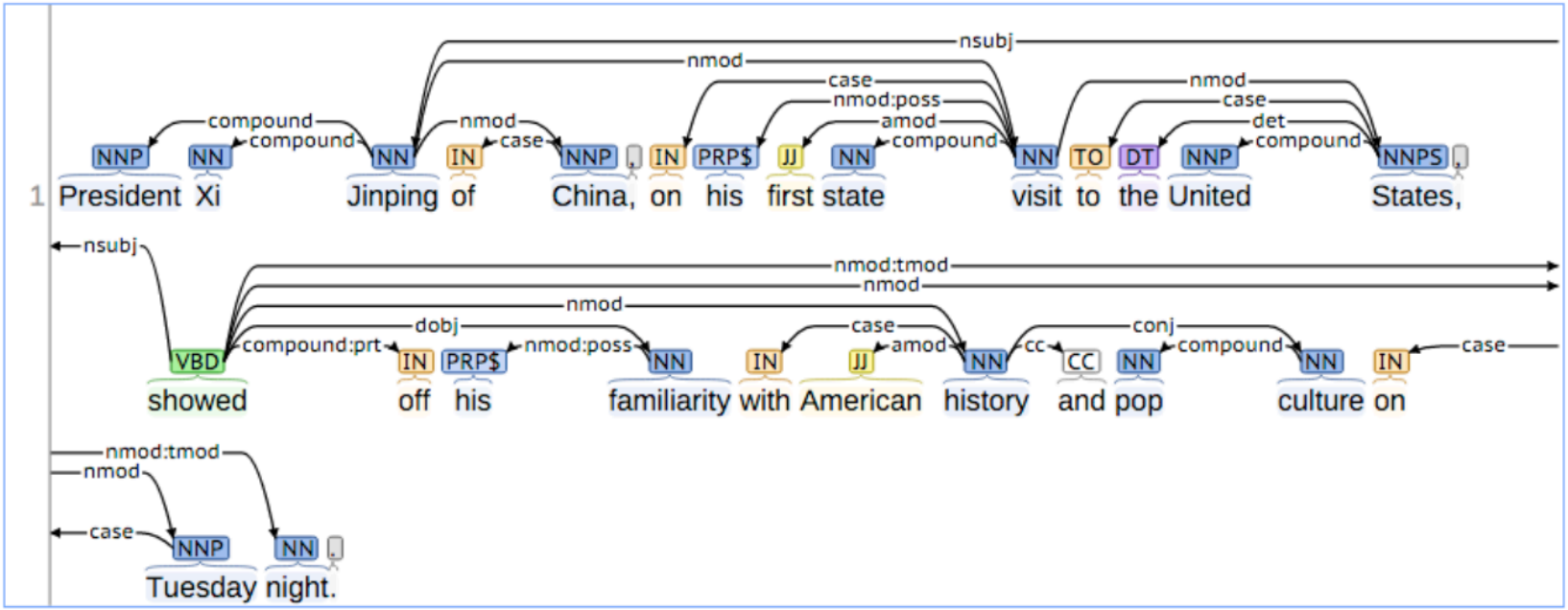
Named Entity Recognition:

1 President Xi Jinping (Person) of China (Loc), on his first (ORDINAL) state visit to the United States (Location), showed off his familiarity with American (Misc) history and pop culture on Tuesday (Date) night (Time).

Coreference:

1 President Xi Jinping of China, on his first state visit to the United States (Mention), showed off his (M) familiarity with American history and pop culture on Tuesday night.

Basic Dependencies:



Text Mining and NLP

NLP

- Deep linguistic analysis
- May take a long time to analyze large text collections

Text mining

- Use shallow features, usually N-grams, to analyze large text collections quickly
- Sometimes use deep NLP features like PoS tags or dependencies for feature engineering

Text Mining and Information Retrieval

Information retrieval (IR): given a query, find relevant documents

Text mining can help information retrieval

- Clustering retrieval results for focused navigation
- Categorize documents for more precise results (e.g., find documents with certain readability levels)

Techniques from IR are used in text mining

- Example: TFIDF weighting

Class Policies

School of Information Studies
Syracuse University

Focus of This Class

Modeling text mining problems

Data collection and preparation

Algorithms

Result evaluation and interpretation

Documentation and presentation

Tips for Success

Curiosity to language and meaning

Critical thinking

Algorithmic thinking

Story telling