Text Analysis with R for Social Science Research

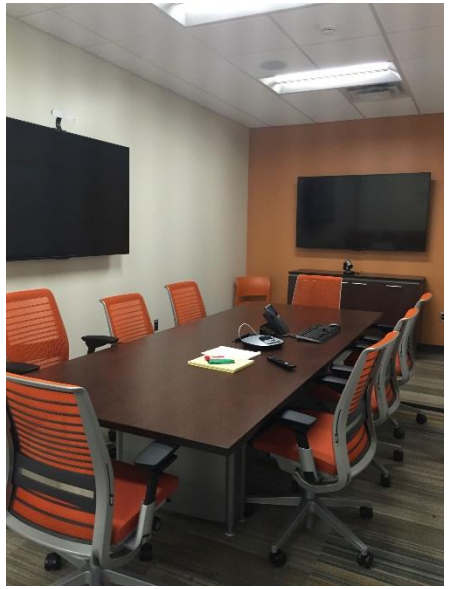
Project Mosaic Workshop

Date 10/6/2016

Ryan Wesslen, Computing & Informatics

rwesslen@uncc.edu

Project Mosaic



- Project Mosaic: What do we do?
  - Build research methods capability in social sciences
  - Facilitate research across social science disciplines
  - Promote social science research
- Project Mosaic Services
  - Social sciences research incubator
    - Facilitate connections
    - Bring people together to exchange ideas and pursue external funding
    - Information sharing on research funding opportunities
  - Consulting
    - Free to UNC Charlotte faculty, staff and graduate students
  - Workshops
    - Open to entire campus community
    - Provides cutting-edge tools for research and a forum for researchers to network within campus

Workshop Agenda

- Overview of Text Analysis

Federalist Papers

- Case 1: Text pre-processing & exploratory analysis
- Case 2: Disputed authorship problem
  - Ridge Regression
- Case 3: Exploring topics in the papers
  - Topic Modeling (LDA)

Learning Objectives

Level	Background	Learning Objective
Beginner	No background in R or text analysis	Learn text terminology and high level of text analysis approaches (e.g. pre-processing, classification, topic modeling)
Intermediate	Familiar with either R or text mining (not both)	Ability to rerun code independently and learn the associated R packages (quanteda, topicmodels) functionality.
Advanced	Proficient in both R and text analysis	Ability to customize code and answer advanced section questions at the end of each case.

Workshop materials

All workshop materials can be found here:

www.github.com/wesslen/federalist-papers-workshop

Introduction to Text Analysis

Why analyze text?

- Growing
- Interesting
- Untapped

Big Data: Internetlivestats.com

www.internetlivestats.com

3,329,578,181
Internet Users in the world
sources more info watch all

1,003,120,764
Total number of Websites

198,705,447,734
Emails sent today

4,121,187,578
Google searches today

3,819,411
Blog posts written today

551,821,139
Tweets sent today

9,186,032,878
Videos viewed today on YouTube

74,921,054
Photos uploaded today on Instagram

115,084,634
Tumblr posts today

Language Technology

mostly solved

Spam detection (Classification)

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

→ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Source: Dan Jurafsky

Why else is text analysis difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad
the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

sarcasm

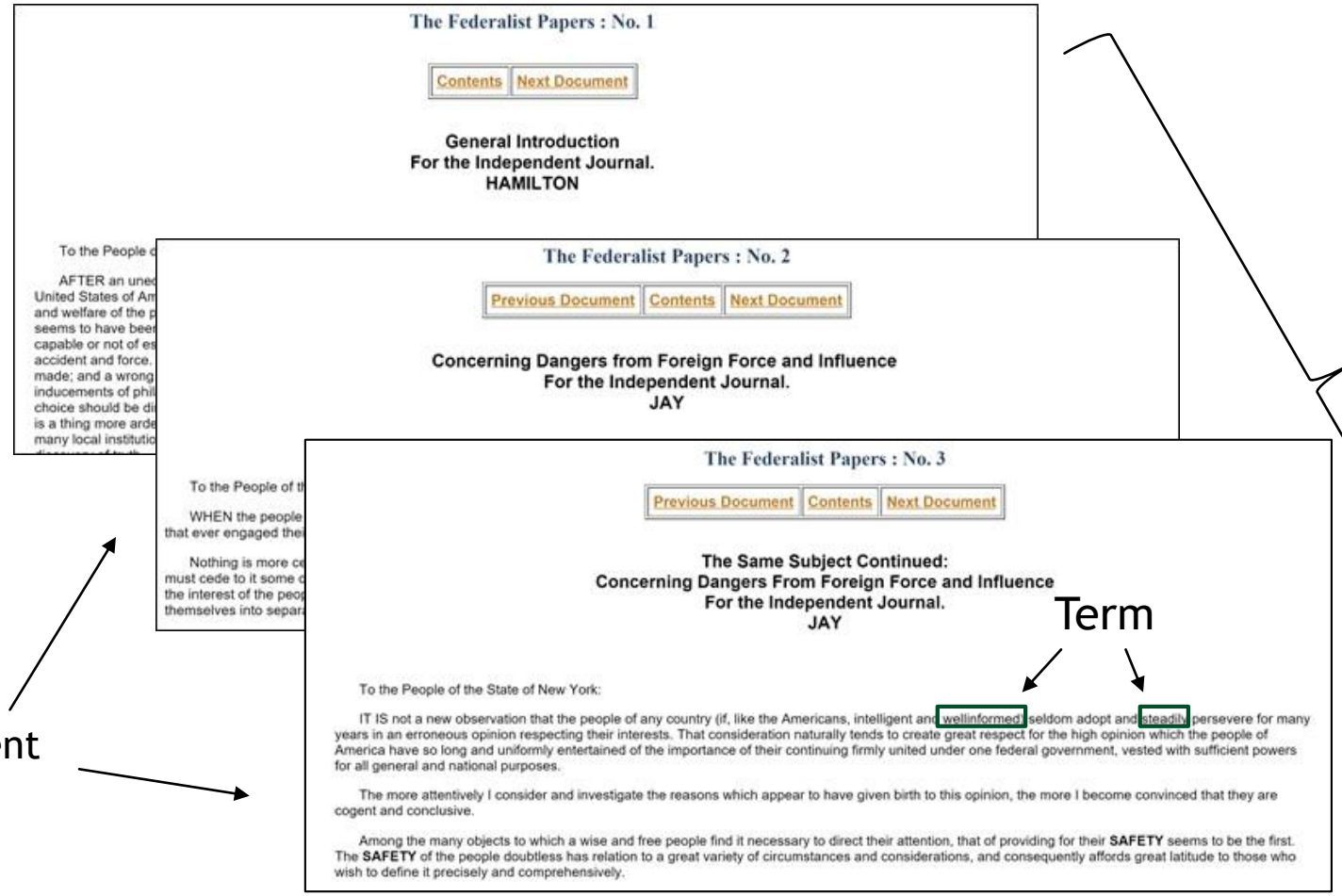
A: I love Justin Bieber. Do you like him to?
B: Yeah. Sure. *I absolutely love him.*

tricky entity names

Where is *A Bug's Life* playing …
Let It Be was recorded …
… a mutation on the *for* gene …

Source: Dan Jurafsky (modified)

Basic Text Terminology

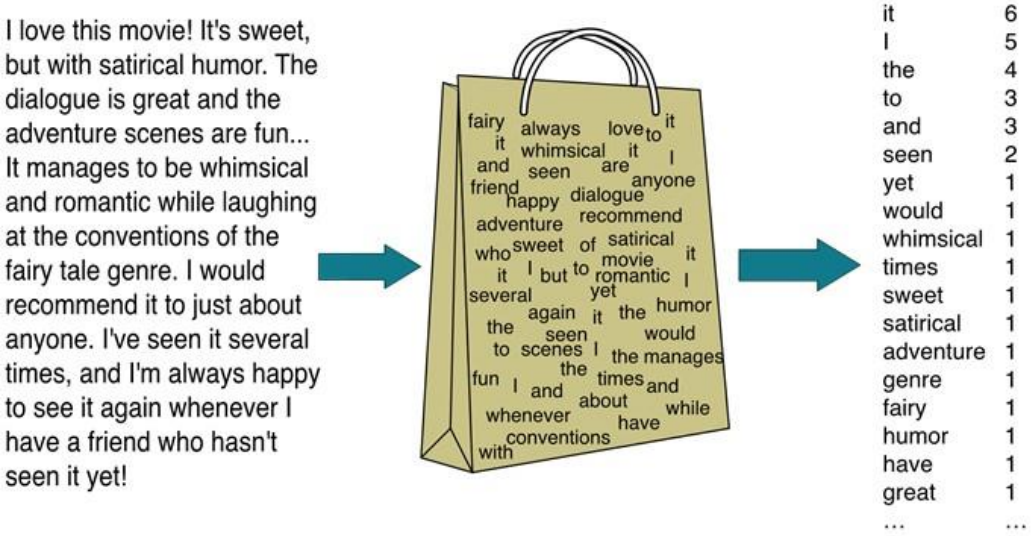


Corpus

Term

Document

“Bag of Words” Approach



Source:
Chris Manning

- Simplest way to quantify text
- Counts the term count per document
  - Document-Term Matrix
- Ignores word order

- N-grams (uni-,bi-,tri-, etc)
- Good at classification
  - Like Spam Filter
- Bad at semantic meaning

Preprocessing

Tokenize → Clean → Stem → Filter

Then	a	hurricane	came,	and	devastation	reigned

then	a	hurricane	came	and	devastation	reigned

then	a	hurricane	came	and	devastation	reigned

~~then~~	~~a~~	hurricane	~~came~~	~~and~~	devastation	reigned

- Tokenization
- Cleaning: Lower case, white space, punctuation
- Stemming and/or Lemmatization
- Filter: remove stop words

Case 1: Preprocessing & Exploratory Analysis

Intro to text analysis in R

Text Analysis with the Quanteda Package

We will use the R quanteda package:

- https://cran.r-project.org/web/packages/quanteda/vignettes/quickstart.html

Case 1: Preprocessing

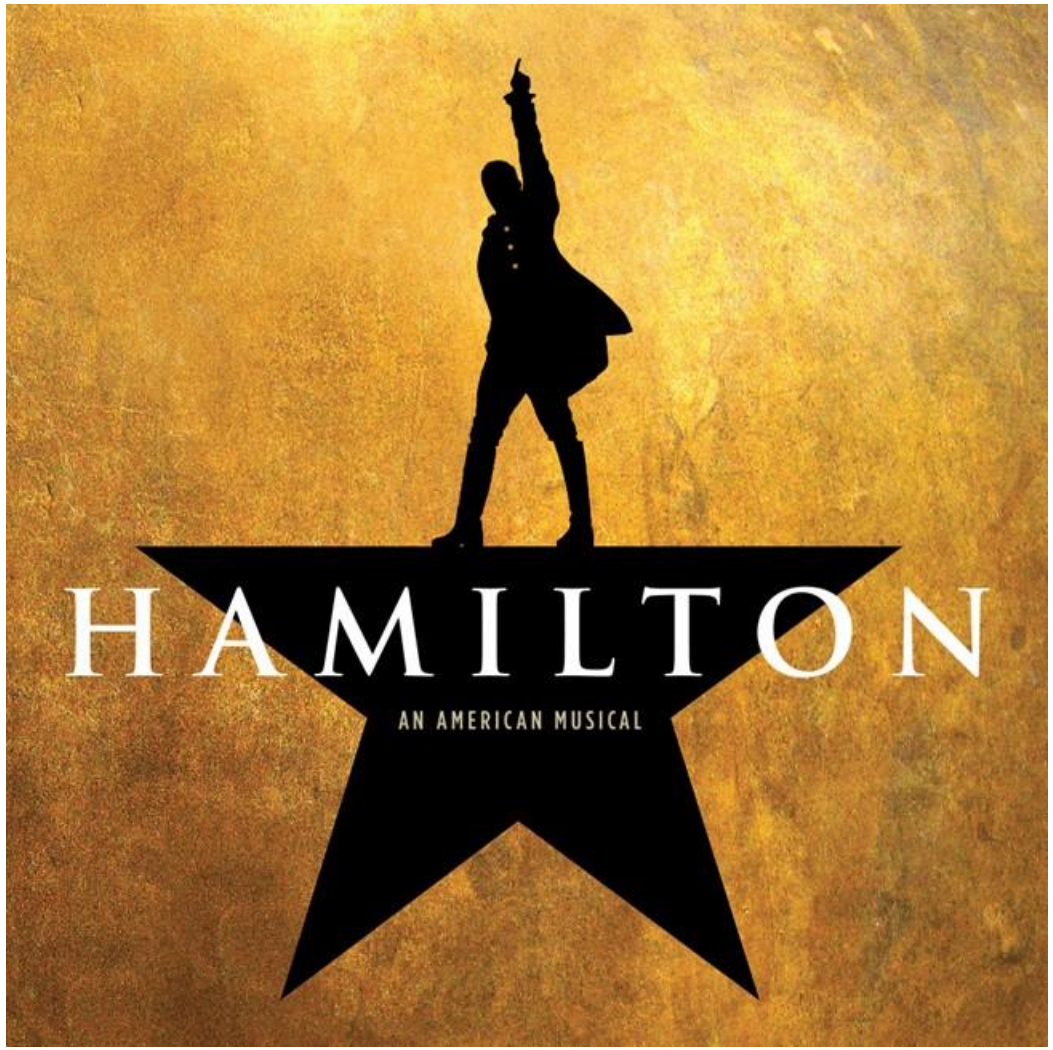
- The exercise can be found on the github site:

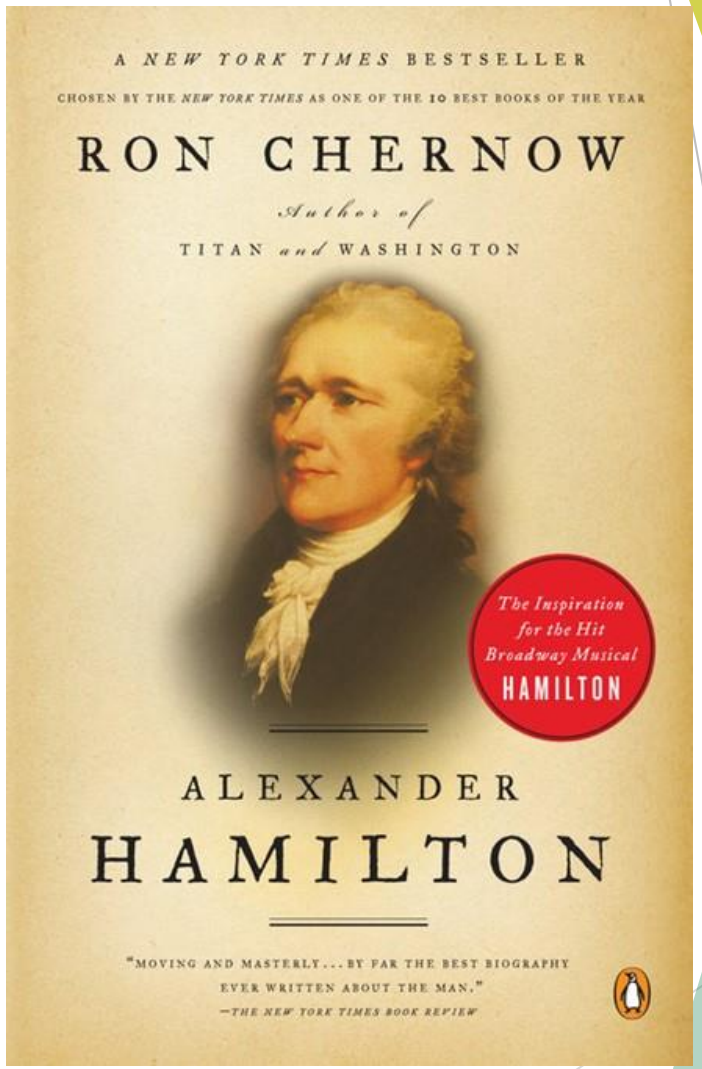
https://github.com/wesslen/Federalist-Papers-Workshop/blob/master/part1/preprocessing-01.Rmd

Case 2: Disputed Authorship Problem

Supervised Learning

Alexander Hamilton





Genius.com's "Non-Stop" Lyrics

genius.com/7873684

If I could grant you peace of mind
Would that be enough?

[BURR]
Alexander joins forces with James Madison and John Jay to write a series of essays defending the new United States Constitution, entitled The Federalist Papers. The plan was to write a total of twenty-five essays, the work divided evenly among the three men. In the end, they wrote eighty-five essays, in the span of six months. John Jay got sick after writing five. James Madison wrote twenty-nine. Hamilton wrote the other fifty-one!

[BURR]
How do you write like you're
Running out of time?
Write day and night like

[ALL WOMEN]
Running out of time?

GENIUS ANNOTATION 1 contributor

True story, bro. The Federalist Papers were written between October 1787 and June 1788, coming it at nearly 175,000 words in total, and spawned an entire political party, fittingly named the Federalists.

> The Federalist essays appeared up to four times a week, the printer often pacing the hall outside while Hamilton finished up his latest salvo. Regardless of who composed them, the essays appeared under the pen name Publius, in homage to the man who had led the people of Rome in establishing a republic after they overthrew their king.

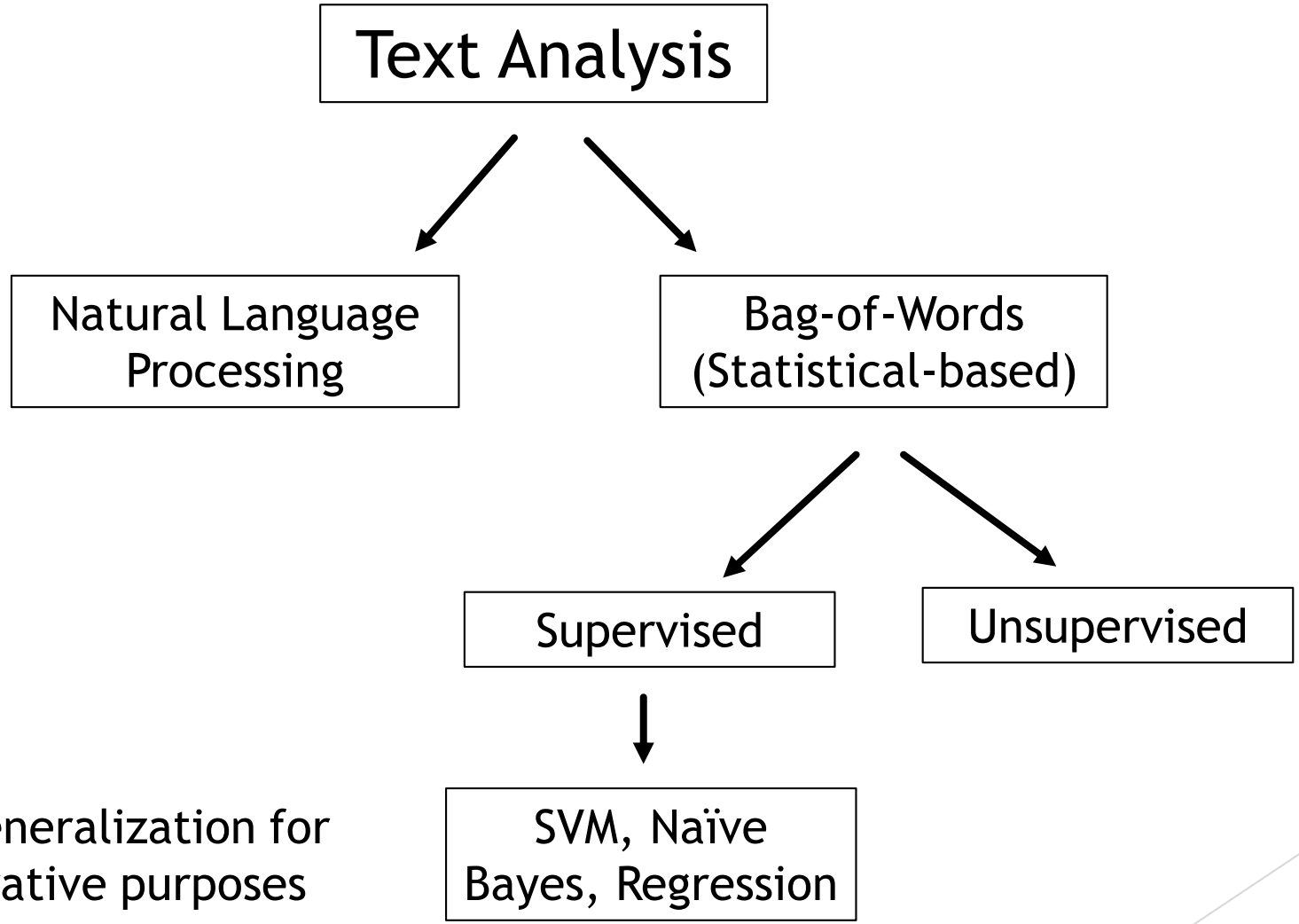
The essays began by enumerating the weaknesses of the Articles of Confederation, and then introduced the strengths of the centralized government outlined in the Constitution. Hamilton, happily in his wheelhouse at that

Federalist Paper setup

- Not so true story, bro (about how many papers each wrote)
- Reality: the authorship of twelve papers is disputed
  - Hamilton claimed authorship before he was killed; Madison disputed those claims eight years later.
  - Adair (1944), Moesteller & Wallace (1963), Fung (2003), Collins et al (2004)

- In this case, our goal is to build a binary supervised algorithm on the known authored papers.
  - The y variable is authorship (1 = Hamilton, 0 = Madison) and the x variables are the word counts.
- We will then apply the model to the disputed papers to predict their authorship.

Text Analysis Methodologies



Overgeneralization for illustrative purposes

Predictive Models: Classification

- Classification models predict ***class labels***.
- ***Class labels*** = categories
  - For example, binary (yes or no), ordinal (high, medium, low) or nominal (dog, cat, kangaroo)
- Classification models use **supervised algorithms** as the class labels ("y variables") are known (observed).
- Determining the disputed Federalist papers is a binary classification problem as the author of the disputed papers is one of two authors: Hamilton or Madison.

Types of Classification Models

- There are many different models (algorithms) that can be used for classification problems.
  - Examples: Naïve Bayes, Decision Tree, Support Vector Machine, Neural Networks
- We are going to use Ridge Regression.

Ridge Regression

Suppose we have N documents, with each document i having label $y_i \in \{-1, 1\} \rightsquigarrow \{\text{liberal, conservative}\}$
We represent each document i is $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$.

$$
\begin{aligned}
f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) &= \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}' \boldsymbol{x}_i\right)^2 \\
\widehat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} \left(y_i - \boldsymbol{\beta}' \boldsymbol{x}_i\right)^2 \right\} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{Y}
\end{aligned}
$$

Problem:

- J will likely be large (perhaps $J > N$)
- There many correlated variables

Source: Grimmer, 2014, "Text as Data" course week 15

Ridge Regression

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \underbrace{\lambda \sum_{j=1}^{J} \beta_j^2}_{\text{Penalty}}$$

where:

- $\beta_0 \rightsquigarrow$ intercept
- $\lambda \rightsquigarrow$ penalty parameter

Source: Grimmer, 2014, "Text as Data" course week 15

Ridge Regression

- Ridge regression is based on regression framework but with regularization parameters.
- Regularization = add in penalty for the number of X variables
- Regularization is performed to reduce overfitting given the large number of X variables (words).

- StackExchange on why Ridge Regression works well for text classification
- StackExchange on Interpretation of Ridge Regression

Ridge Regression

- Pro:
  - Prevents overfitting with many features (x variables)
  - Interpretable coefficients (compared to Naïve Bayes or SVM)
- Con:
  - Decision of the lambda value (need of cross-validation)
  - Abandons unbiased estimator

A more thorough (statistical) overview of Ridge Regression

Cross-Validation

Intuition:

- Create K training and test sets ("folds") within training set.
- For each k in K, run classifier and estimate performance in test set within fold.
- Pick value of λ (regularization parameter) that gives better performance
- Train classifier with full dataset and compute performance on test set
- Why? Avoids overfitting

Source: Pablo Barberá

Case 2: Disputed Authorship Problem

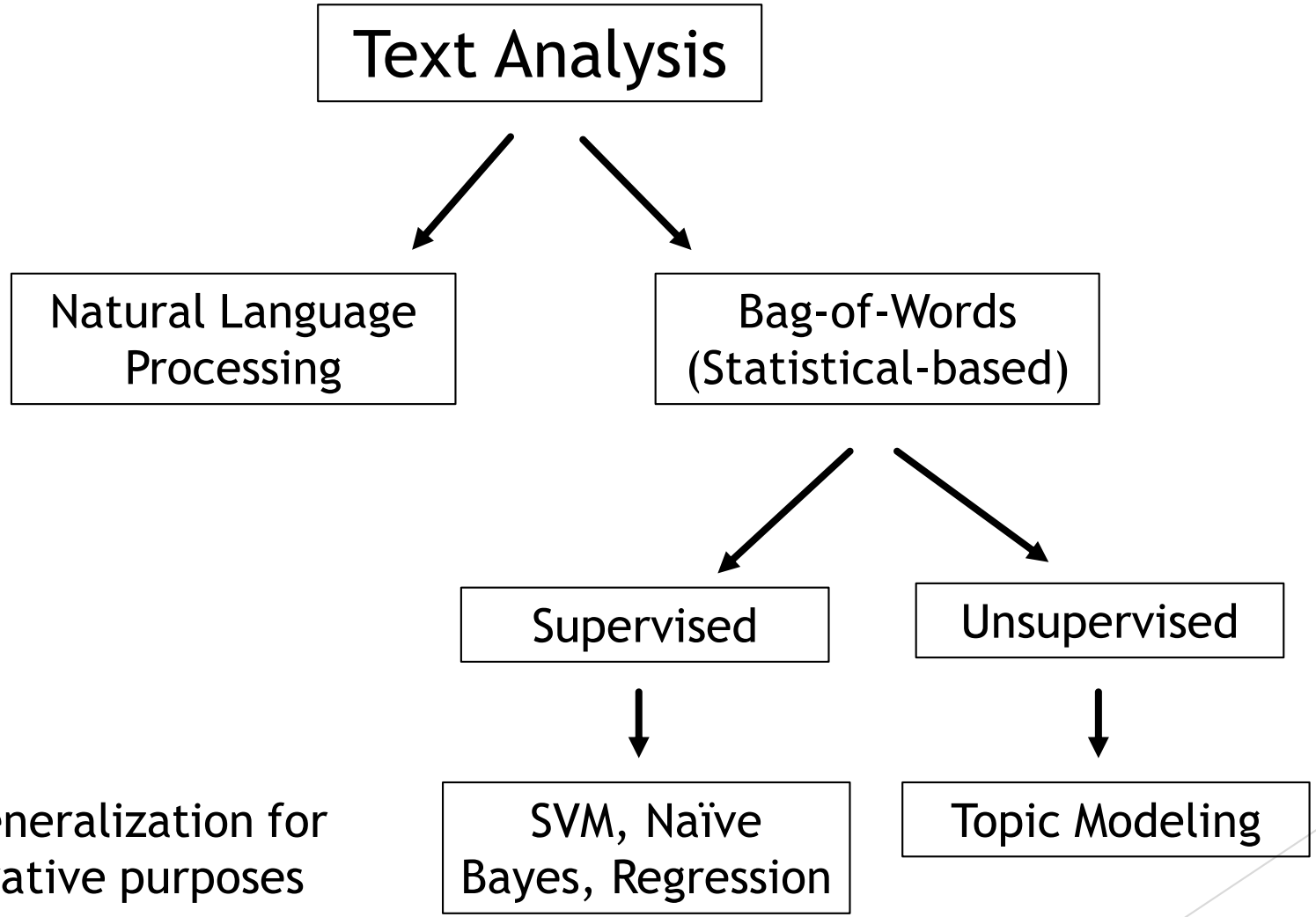
- The exercise can be found on the github site:

https://github.com/wesslen/Federalist-Papers-Workshop/blob/master/part2/supervised-02.Rmd

Case 3: Topic Modeling (LDA)

Unsupervised Learning

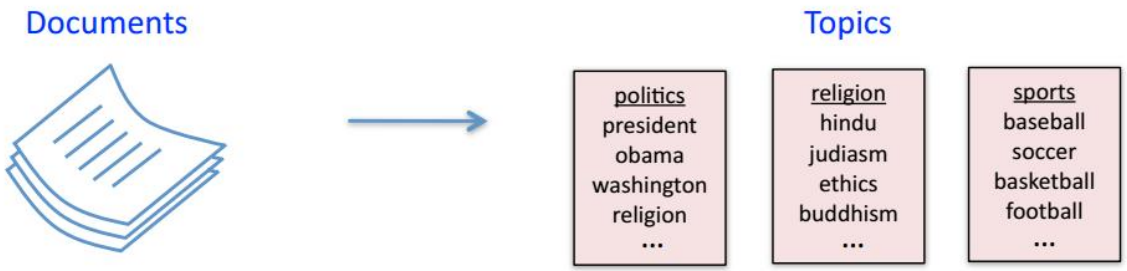
Text Analysis Methodologies



Overgeneralization for illustrative purposes

Topic Models

- **Topic models** are powerful tools for exploring large data sets and for making inferences about the content of documents

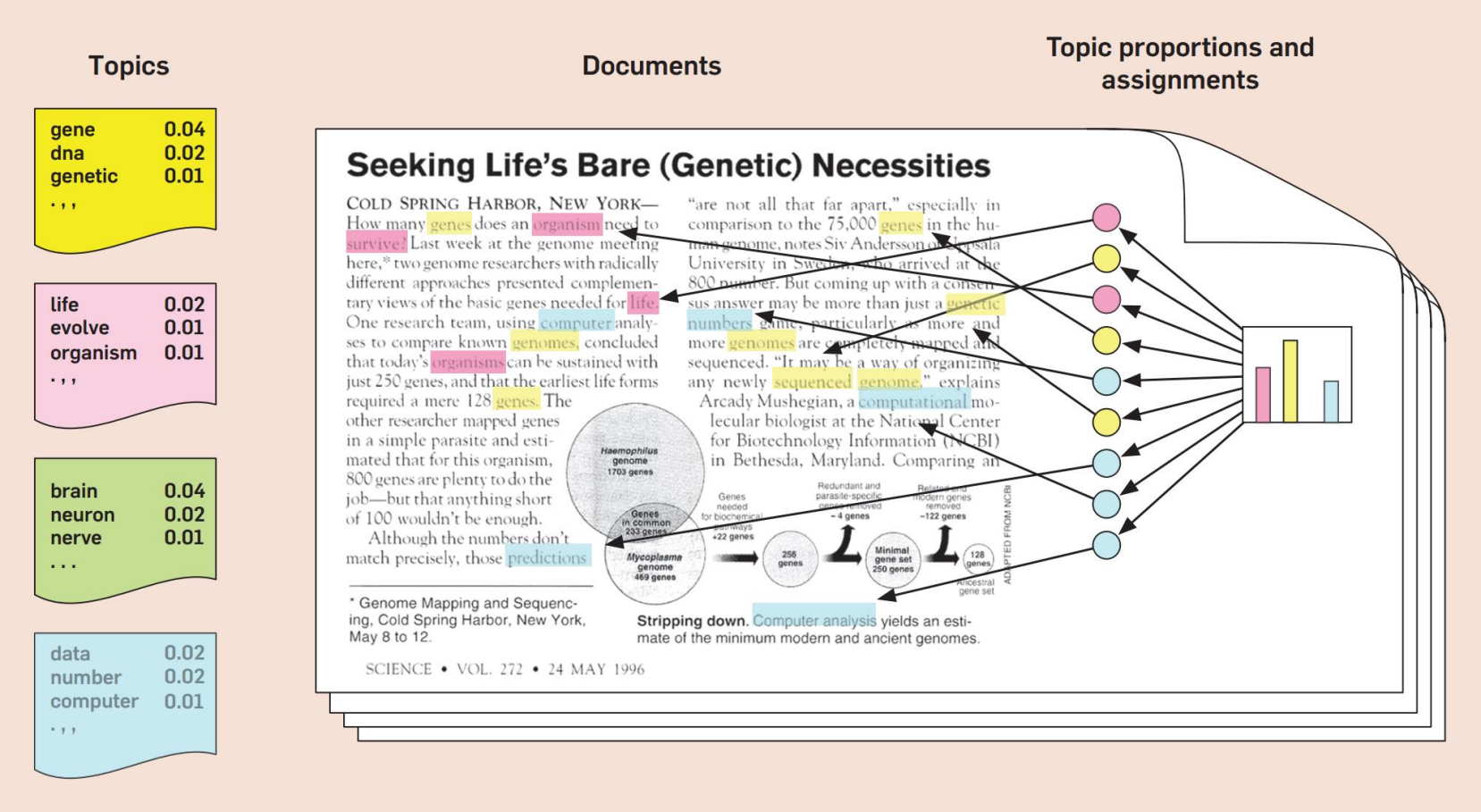


- Many applications in information retrieval, document summarization, and classification

- LDA is one of the simplest and most widely used topic models

Source: Pablo Barberá

Latent Dirichlet Allocation (LDA)



David Blei's "Probabilistic Topic Models" (2012)

topicmodels package in R

- We'll focus on a hands on introduction.
- For more detailed documentation, see: https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf
- Also, we'll try to run LDAVis package:

Sample: https://gallery.shinyapps.io/LDAelife/

Case 3: Topic Modeling

- The exercise can be found on the github site:

https://github.com/wesslen/Federalist-Papers-Workshop/blob/master/part3/topicmodeling-03.Rmd

Learning Objectives

Level	Background	Learning Objective
Beginner	No background in R or text analysis	Learn text terminology and high level of text analysis approaches (e.g. pre-processing, classification, topic modeling)
Intermediate	Familiar with either R or text mining (not both)	Ability to rerun code independently and learn the quanteda package basic functions.
Advanced	Proficient in both R and text analysis	Ability to customize code and answer advanced section questions at the end of each case.

Automated Text Analysis Methods

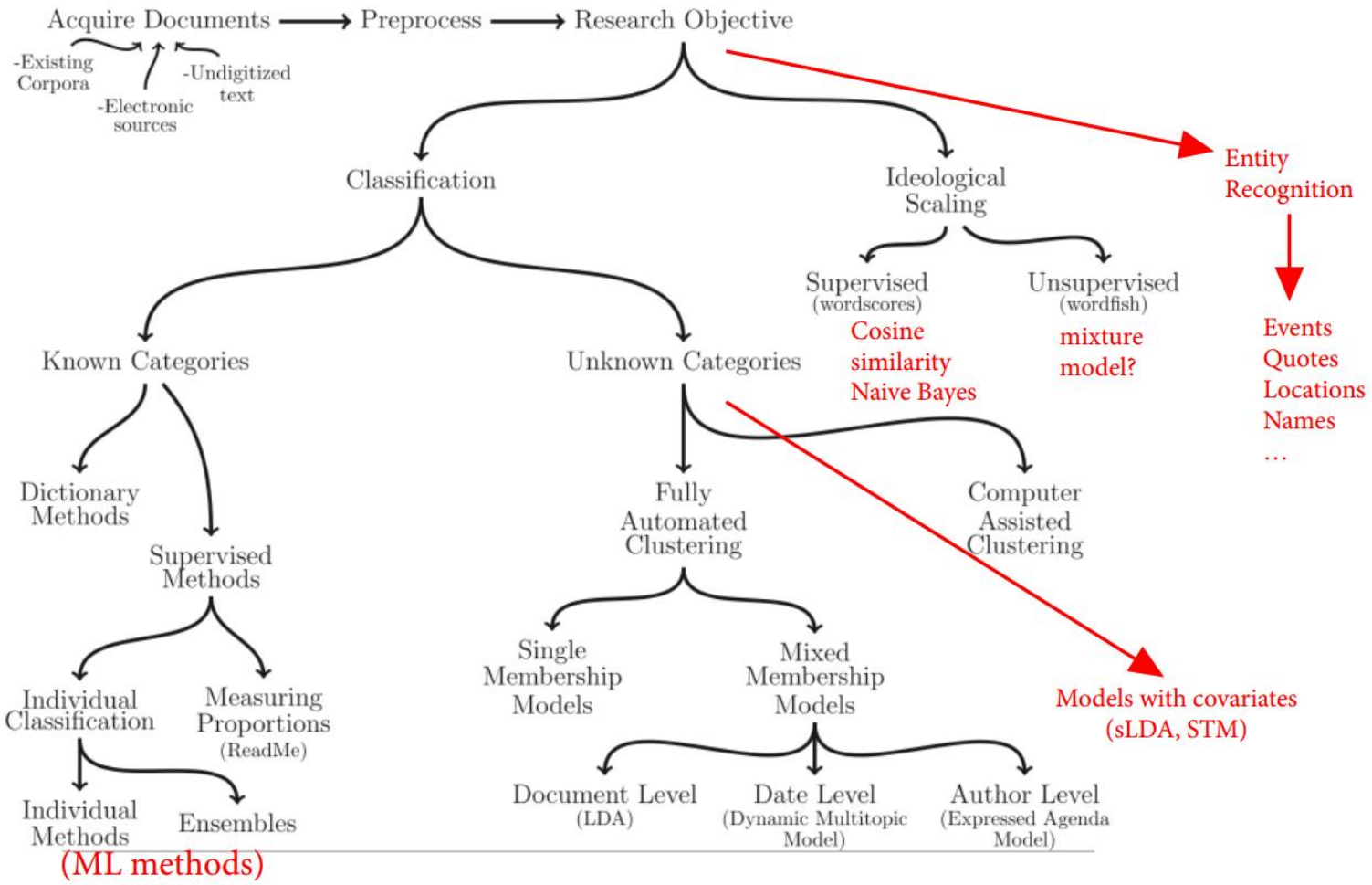


Fig. 1 in Grimmer and Stewart (2013)

Source: Pablo Barberá

Resources for Social Science Text Analysis

- Quanteda package vignette
  - https://cran.r-project.org/web/packages/quanteda/vignettes/quickstart.html
- Justin Grimmer's Text as Data Course
  - http://stanford.edu/~jgrimmer/Text14/
- Topicmodels package vignette
  - https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf
- Stm package vignette for Structural Topic Modeling
  - https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf

More About Services

- Research Incubator
  - Affiliates Program
    - Faculty Affiliates are hand picked for their research expertise
    - Our affiliates leverage the core functionality and expertise of Project Mosaic
  - Seed Grants Program
    - Geared towards the formation of new teams of researchers in the social, behavior and economic sciences
    - Aim is to pursue external funding
- Consulting
  - Project Mosaic offers three types of consulting:
    - Software-centric
    - Dissertation/thesis assistance
    - Research collaboration

Make an appointment on our website!

- Workshops
  - Our workshops fulfill a commitment to enhance data literacy and analytical capabilities of UNC Charlotte researchers

Find workshops online on our Events List.

Contact Project Mosaic

- Jean-Claude Thill is the director of Project Mosaic. A broadly trained geographer, he is a 'Knight' Distinguished Professor of Public Policy at UNC Charlotte.
- Contact Jean-Claude:
  - Email: Jean-Claude.Thill@uncc.edu
  - Phone: 704-687-5931 ext. 75909

- Leonora is the Administrative Support for Project Mosaic. She manages our not-so-massive paperwork, coordinates meetings and assists with administrative functions.
- Contact Leonora:
  - Email: projectmosaic@uncc.edu
  - Phone: 704-687-5931

Visit our website!
Projectmosaic.uncc.edu

Project Mosaic: A Social Sciences Research Initiative
College of Liberal Arts & Sciences

Additional Resources: Consultants

- Shaoyu Li is the head consultant in the Center of Statistics and Applied Mathematics Consulting Center (CSAMC) and works with Project Mosaic to coordinate consulting requests for statistical and mathematical expertise.
- Contact Shaoyu:
  - Email: shaoyu.li@uncc.edu

- Kailas Venkitasubramanian is a research methodologist and manages the consulting service and the workshop program of Project Mosaic. Kailas is experienced in a variety of applied statistical techniques and works fluently on multiple software platforms.
- Contact Kailas:
  - Email: kvenkita@uncc.edu

Questions?