Text as Data
Probability Models

Robert Stine
Department of Statistics
Wharton School of the University of Pennsylvania

www-stat.wharton.upenn.edu/~stine

Comments from Second Lecture

Sentiment analysis

R file shows an “quick and dirty” example of building a dictionary using supervised data

Latent semantic analysis

Review the underlying latent variable motivation

Interpreting DTM as a covariance matrix is key

Need time for covering the R portion of LSA

Consequently…

Topic models will run over into the next lecture

Meet tomorrow, earlier for a bit less time: **5 – 6:30**

Naive Bayes

Naive Bayes

First example of generative probability model

Too “naive” to generate actual text, but …

First hint of a “language model”

Supervised method

Designed to produce a classifier

Implies need a categorical response

Different example

Introduce Federalist Papers

Refer to article/book by Mosteller and Wallace for history

Example with fewer, but longer documents than wine example

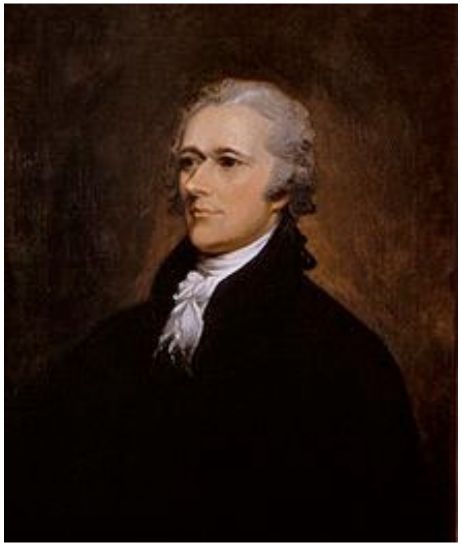
Federalist Papers

Federalist papers

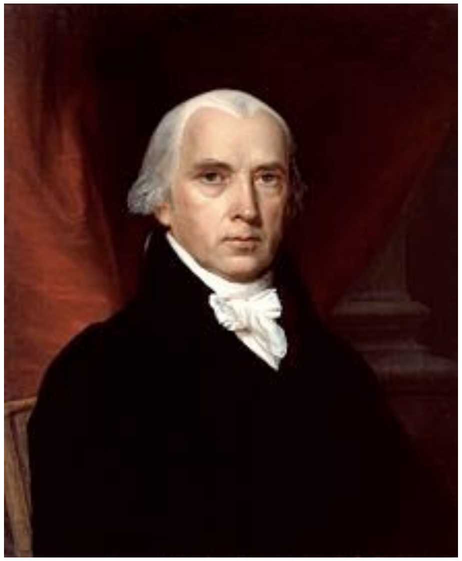
85 essays advocating US Constitution in 1787-1788

Revisit text by Mosteller and Wallace (1964)
Who wrote the 12 disputed Federalist papers?

Supervised classification



Hamilton
51



Madison
14

Jay
5

Hamilton & Madison
3

Wharton
Department of Statistics

Initial Processing

Data

Get data from Project Gutenberg

Build data frame with names of authors

Illustrates some different things you may need to do with your data

Sample

To the People of the State of New York: AFTER an unequivocal experience of the inefficacy of the subsisting federal government, you are called upon to deliberate on a new Constitution for the United States of America…

Preprocessing

Want a document-term matrix for identifying useful words

Downcase, remove stop words and punctuation

Classification

Predict class membership

Assign class label given collection of document characteristics (e.g., word present/absent)

Assign to category $\hat{Y}$ that maximizes conditional probability

$$\max_y P(Y=y | X_1, X_2, \ldots X_k)$$

Complication

Suppose k is large, possibly larger than number of cases

Lack enough examples to build conditional probability from frequencies

Example: Federalist papers

70 written by Hamilton or Madison, but thousand-word vocabulary

Naive Bayes is competitive in cases with few training examples

Provided its assumption is “reasonable”

Note: could use LSA
with logistic regression

Naive Bayes

Employ Bayes rule

$P(Y|X)\ P(X) = P(X|Y)P(Y)\ \ -> P(Y|X) = P(X|Y)P(Y)/P(X)$

$\max_y P(Y{=}y|\ X_1, X_2, \ldots X_k) = \max_y P(X_1, X_2, \ldots X_k|Y)\ P(Y)$

Simplifying assumption

Know prior probabilities (such as equal) so P(Y) drops out

$\max_y P(Y{=}y|\ X_1, X_2, \ldots X_k) = \max_y P(X_1, X_2, \ldots X_k|Y)$

X_j are conditionally independent given Y

$\max_y P(Y{=}y|X_1, X_2, \ldots X_k) = \max_y P(X_1|Y)\ P(X_2|Y) \cdots P(X_k|Y)$

Rationale in language

Reduces problem to product of frequencies from 2x2 contingency tables in case of words/text

Naive Bayes

Simple analysis

Identify whether a word appears or not (0/1) rather than count

Component probabilities $P(X_w|Y)$ reduce to relative frequency of a word appearing in the papers written by each author

Complication: unseen word

Apply naive Bayes to new document

New document contains a word not seen previously

Should we set the probability to zero? There's not chance that the author would use that word?

Solution: Good-Turing smoothing

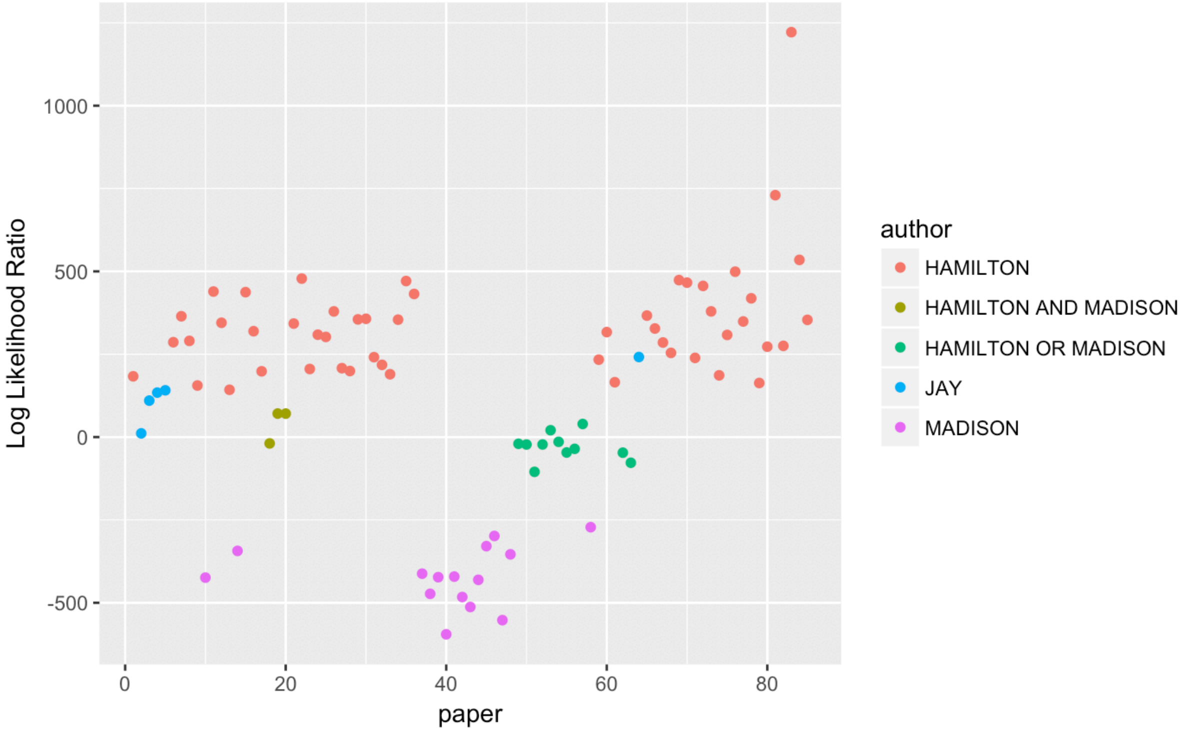
Spread probability over entire vocabulary to cope with OOV

Results

Test case

Apply to Federalist Papers known to have been written by either Hamilton or Madison

Naive Bayes assigns most to Madison.



Wharton
Department of Statistics

Topic Models

Bayesian Methods

Simple model

Naive Bayes, a “set the baseline” method

Introduces common independence assumption used in other models

Hierarchical model

Latent variables

More realistic generative model, but still bag-of-words

Unsupervised, like LSA

- Supervised version also available but not seen implemented in R

Linked to vector space models

Topic Models

Conceptual model for the generation of text

Language is an expression of an idea or “topic”

Presidential address might move from domestic economics to foreign policy to health care.

Current topic determines the chances for various word choices

The words “inflation” or “interest rate” are more likely to appear when discussing economic policies rather than foreign policy

Hierarchical model

Identify the number of topics

Define a probability distribution for each

Each document mixes words drawn from topics

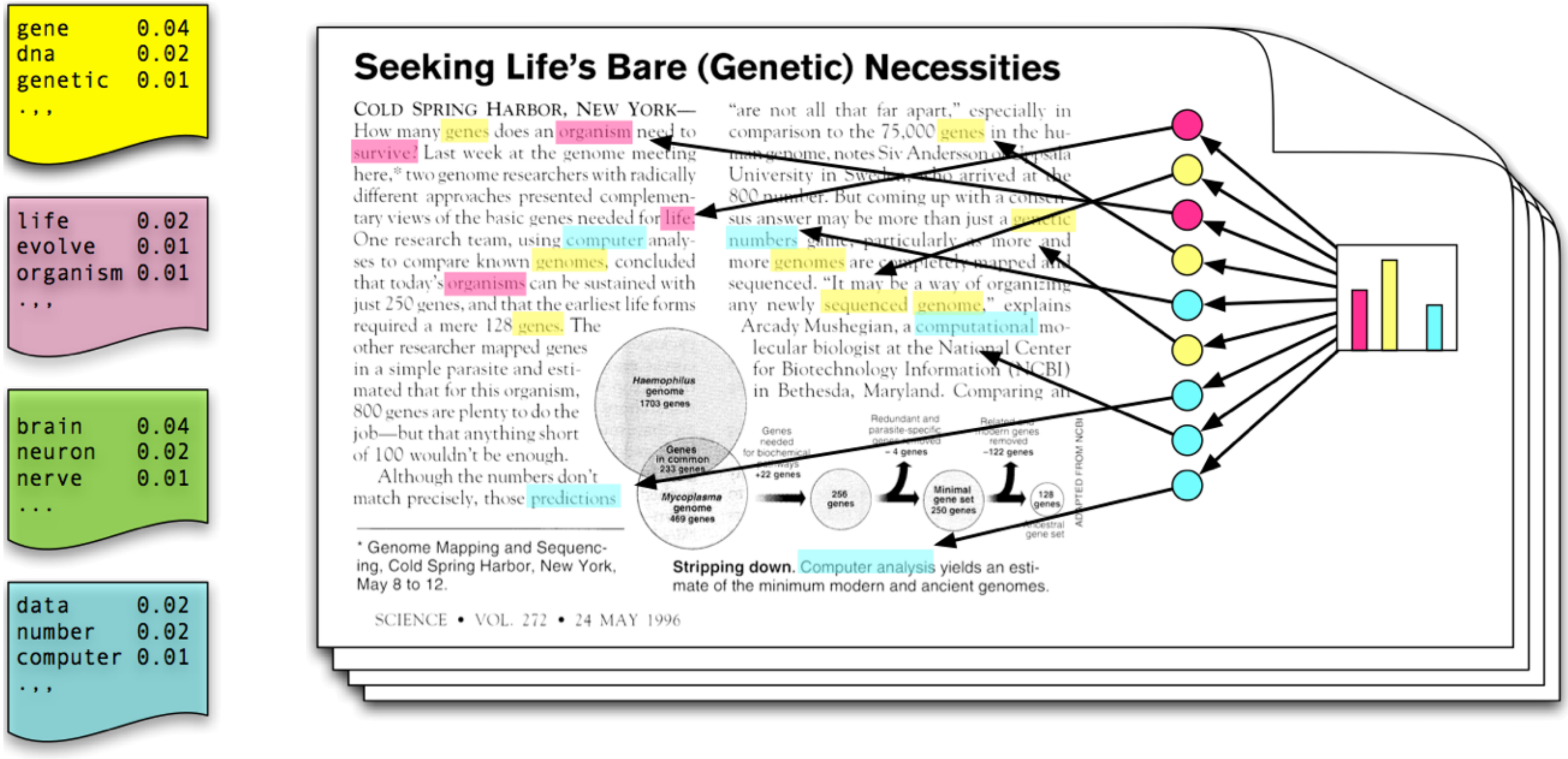
Conditional independence, given topic (like naive Bayes)

Simple Image of Model

Each document mixes words from collection of topics

topic = probability distribution over words

Original details: Blei, Ng, and Jordan 2003


gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed − 4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Figure from Blei, Intro to Topic Models

Probability Model

Beta:Binomial
as
Dirichlet:Multinomial

conjugate prior

Latent Dirichlet allocation (LDA)

Assume K topics

Discrete dist over vocabulary $P_k \sim \text{Dirichlet}(\alpha), k = 1,\ldots, K$

β in literature

Parameter α controls sparsity of the distribution

Each document mixes topics

Distribution over topics in doc_i $\theta_i \sim \text{Dirichlet}, i = 1,\ldots, n$

θ_i are probabilities

Word probability $P(w \text{ in doc } i) = P_z(w)$ $z \sim \text{Multi}(\theta_i)$

Number of words within doc allowed to be random/fixed

Wharton
Department of Statistics

Expected Word Counts

Matrix product determines counts

Let K x m matrix P denote the matrix with probability distribution P_k in the k^{th} row.

Let the nxK matrix T denote the mix of topics in the documents, with the mix for document i in row i.

Then the expected number of word tokens of type j in document i is $(T\ P)_{ij}$.

Factorization connects to LSA

Topics models imply a factorization of the expected count matrix, the document term matrix C

$$E(C) = n_i\ T\ P$$

and the SVD is one way of factoring C!

Simulated Example

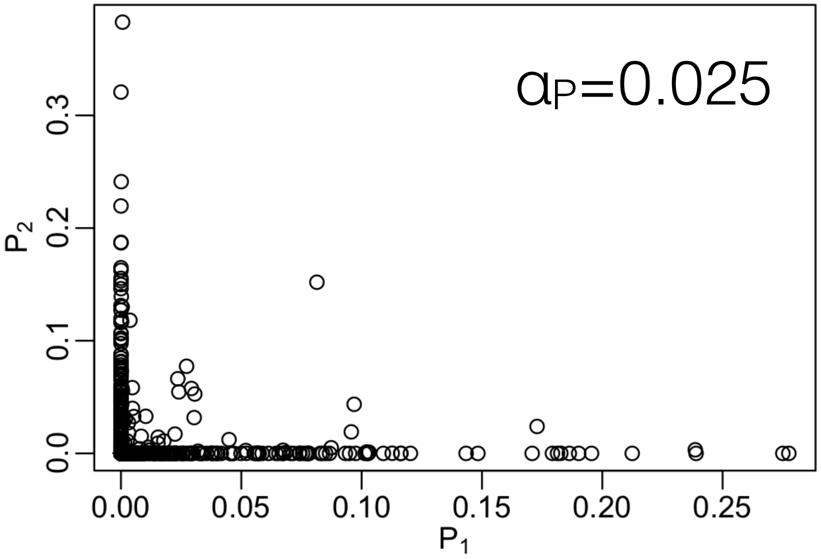
Simulate data from a topic model

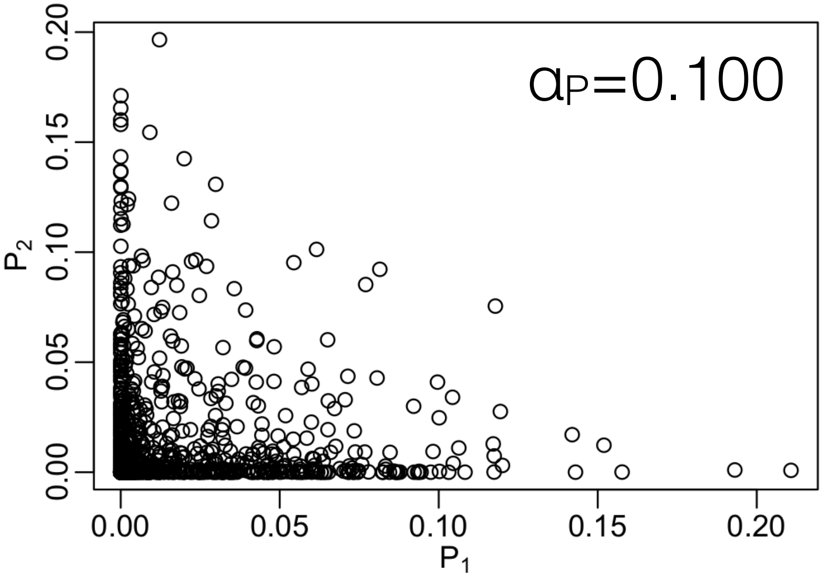
Pick the number K of topics

Pick size m of the vocabulary and the number of documents n

Choose α_P that controls “sparsity” of topic distributions

Small α_P produces nearly singular distributions with little overlap.





α_p = 0.025 in following

Wharton
Department of Statistics

Simulate the Documents

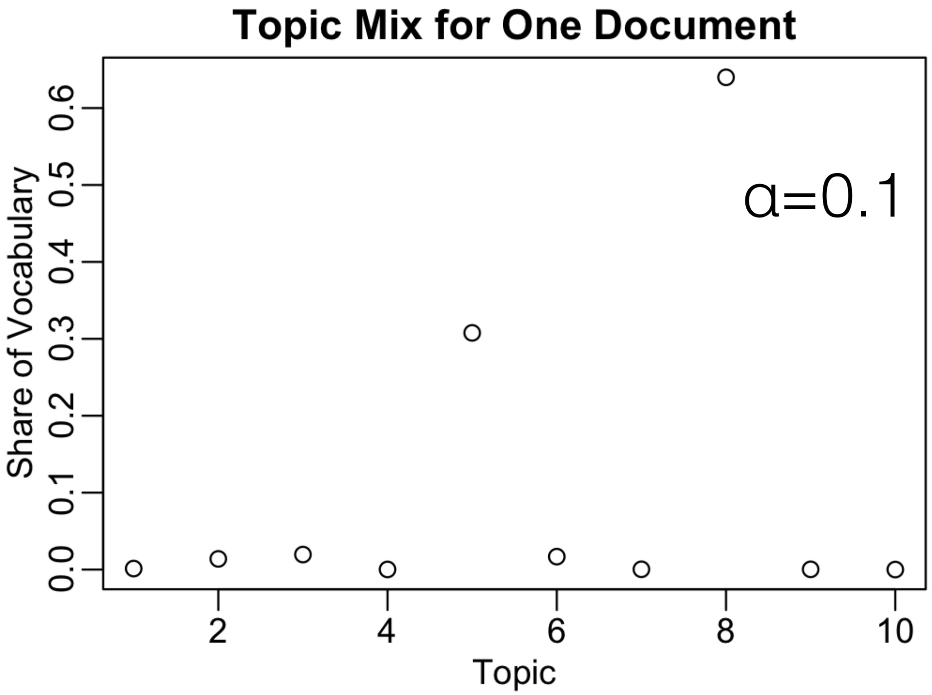
n=3000
m=1000
$n_i \approx 100$

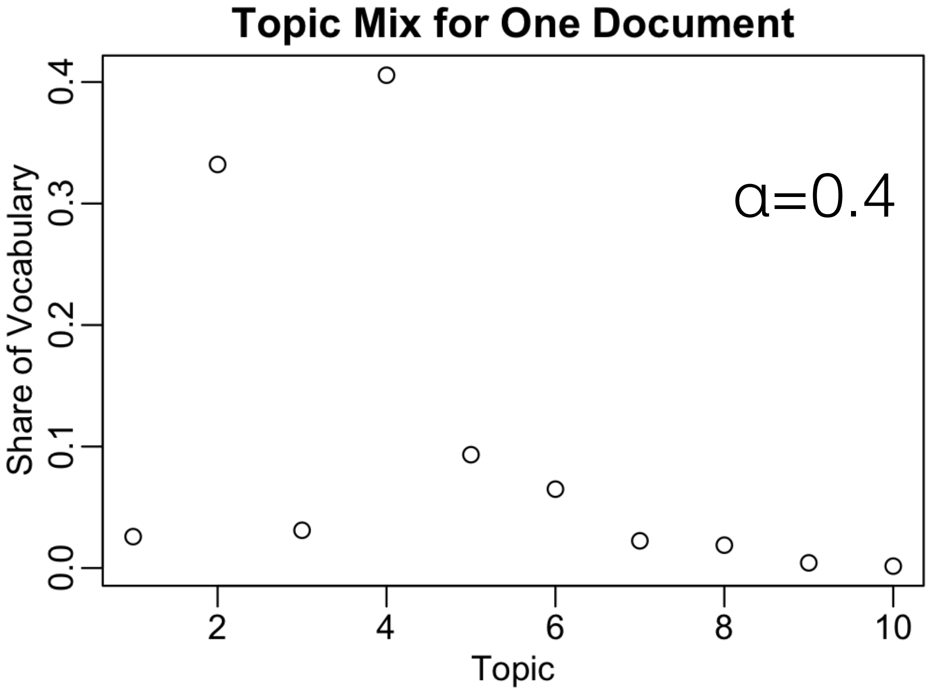
Generate documents

Choose average length of documents (Poisson distribution)

Pick α that controls the mix of topics within documents

Small α produces documents predominantly of one topic.





α = 0.4 in following

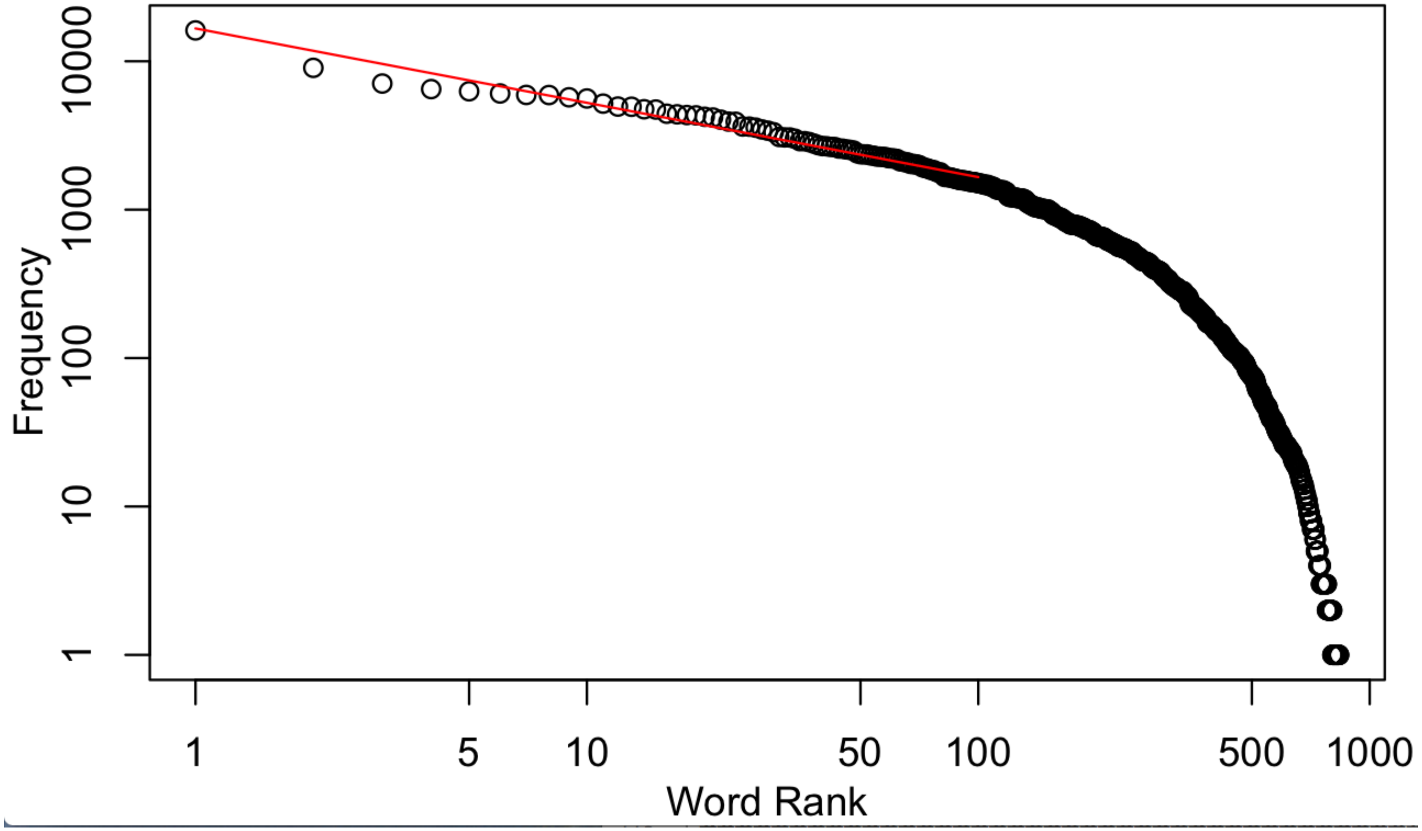
K=10 topics

Word Frequencies

Close to Zipf for early types

More severe concavity that observed in most text

Problem with the probability distributions rather than topic model?



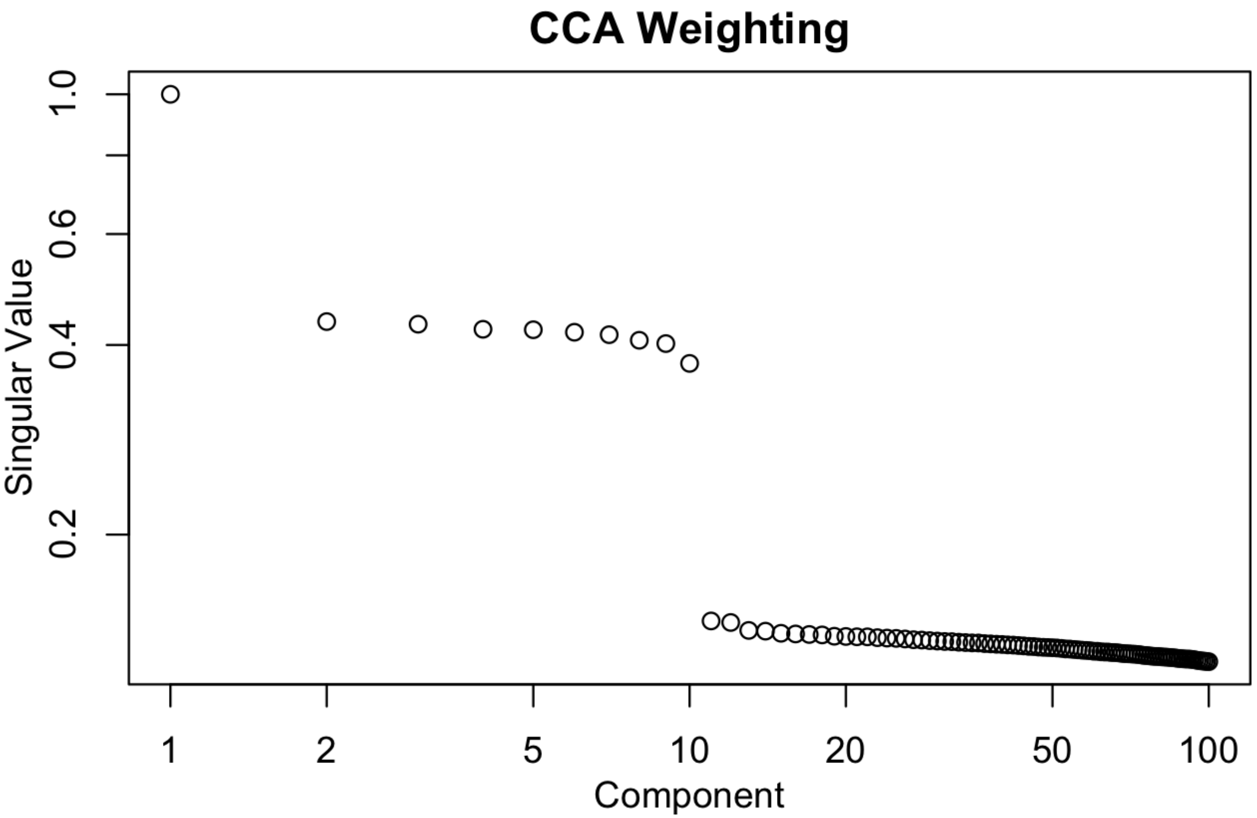
Wharton
Department of Statistics

LSA Analysis

Compute the SVD of the counts

Raw counts and using CCA weights

Number of topics stands out clearly, particularly in CCA

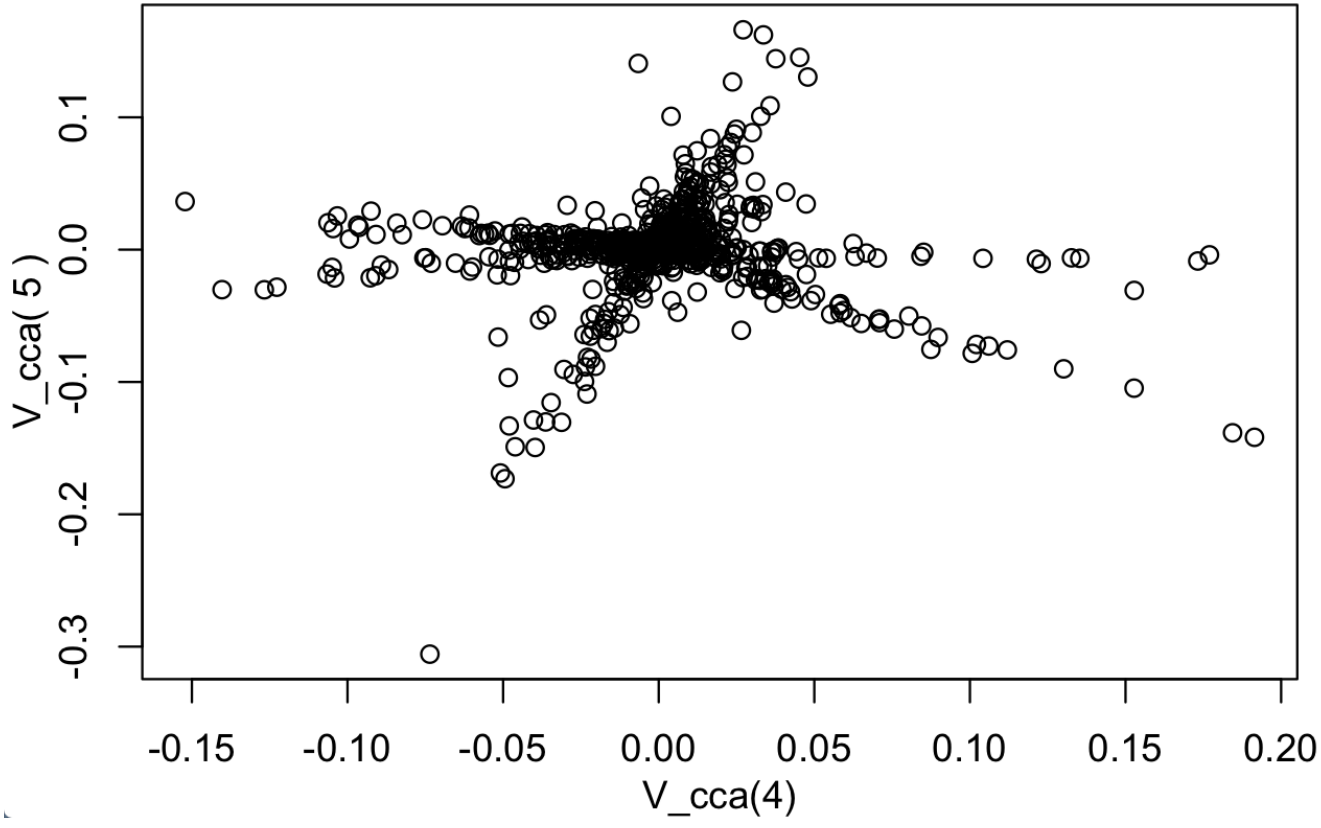


LSA Analysis

Loadings have the "ray-like" behavior

Similar to those in LSA analysis of wine tasting notes

More clearly defined



Topic Model Analysis

Same as used for simulated data

Pick number of topics (e.g., know there are 10)

Input the associated DTM

Results

Indicates which topics most prevalent in documents

Associates word types with the discovered topics

Goodness-of-fit

Obtain overall log-likelihood of fitted model

Vary the number of topics to see how fit changes

Topic Models: Wine

Further notes are in the
Rmd file