Topic Modeling in R

Michele Claibourn

February 8, 2016

Approaches for Working with Text

Grimmer and Stewart 2013

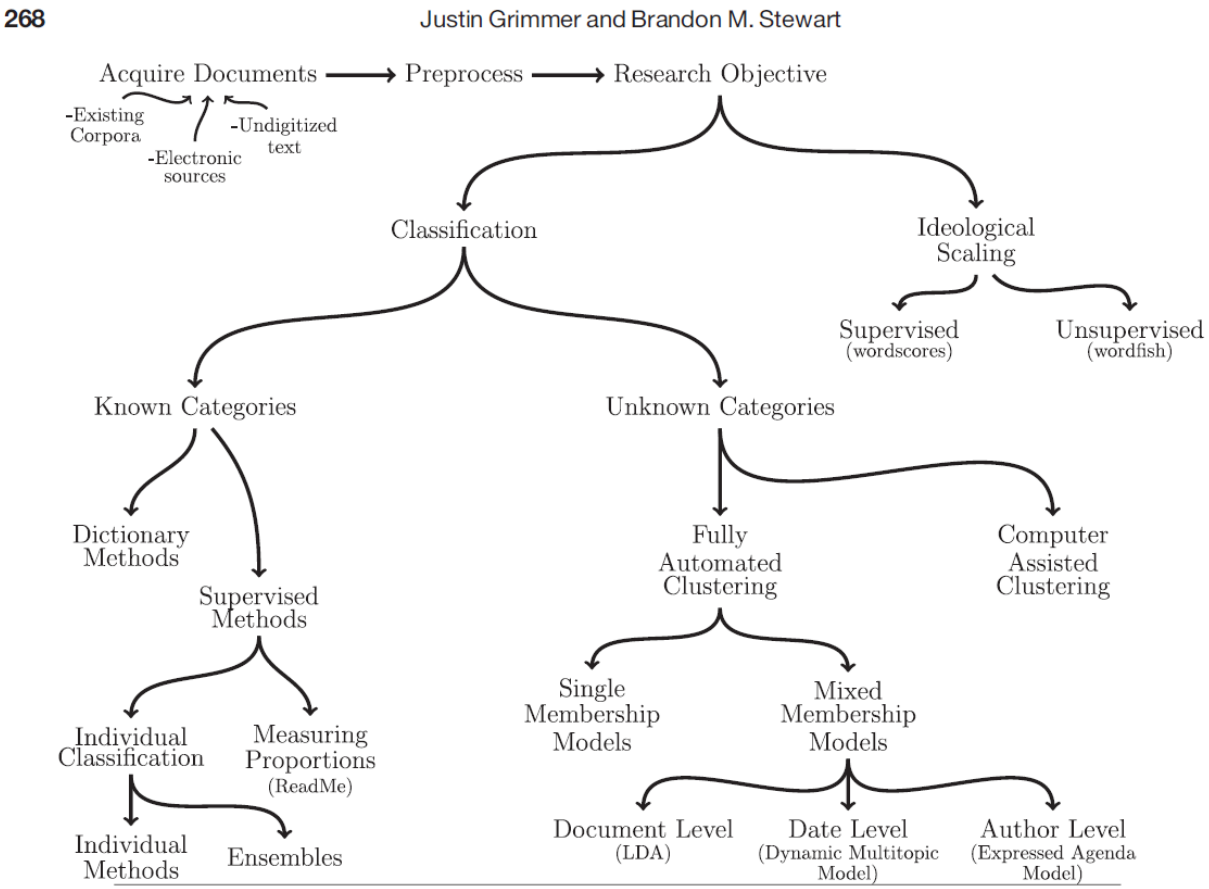


Fig. 1 An overview of text as data methods.

Approaches for Working with Text

O'Connor, Bamman, and Smith

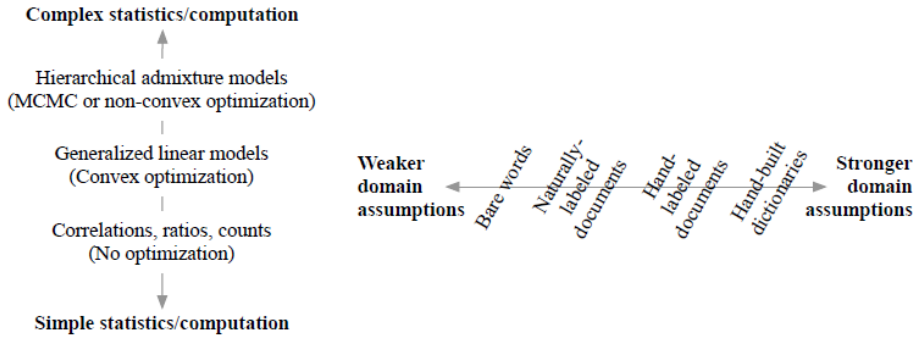


Figure 1: Schematic of model complexity versus domain assumptions for various computational text analysis methods. Statistical models are listed with their respective inference/training algorithms; computational expense increases with model expressiveness.

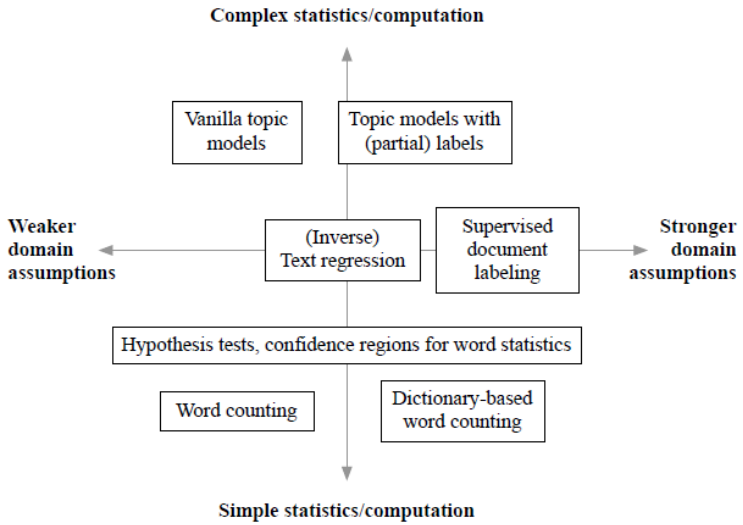


Figure 2: Typical methods used in computational text analysis. Compare to Table 1 in [27].

Processing Text

The goal is to build a corpus, a collection of documents

- Acquire text, metadata
- Preprocess text: retain useful information

The Bag of Words Approach

Treats a text like a set of words with identified frequencies (aka the Vector Space Model), by discarding grammar and word order. The mapping is not one-to-one. So

> *Make peace, not war.*
> *Make war, not peace.*

are equivalent in the bag-of-words model.

The Bag of Words Approach

Treats a text like a set of words with identified frequencies (aka the Vector Space Model), by discarding grammar and word order. The mapping is not one-to-one. So

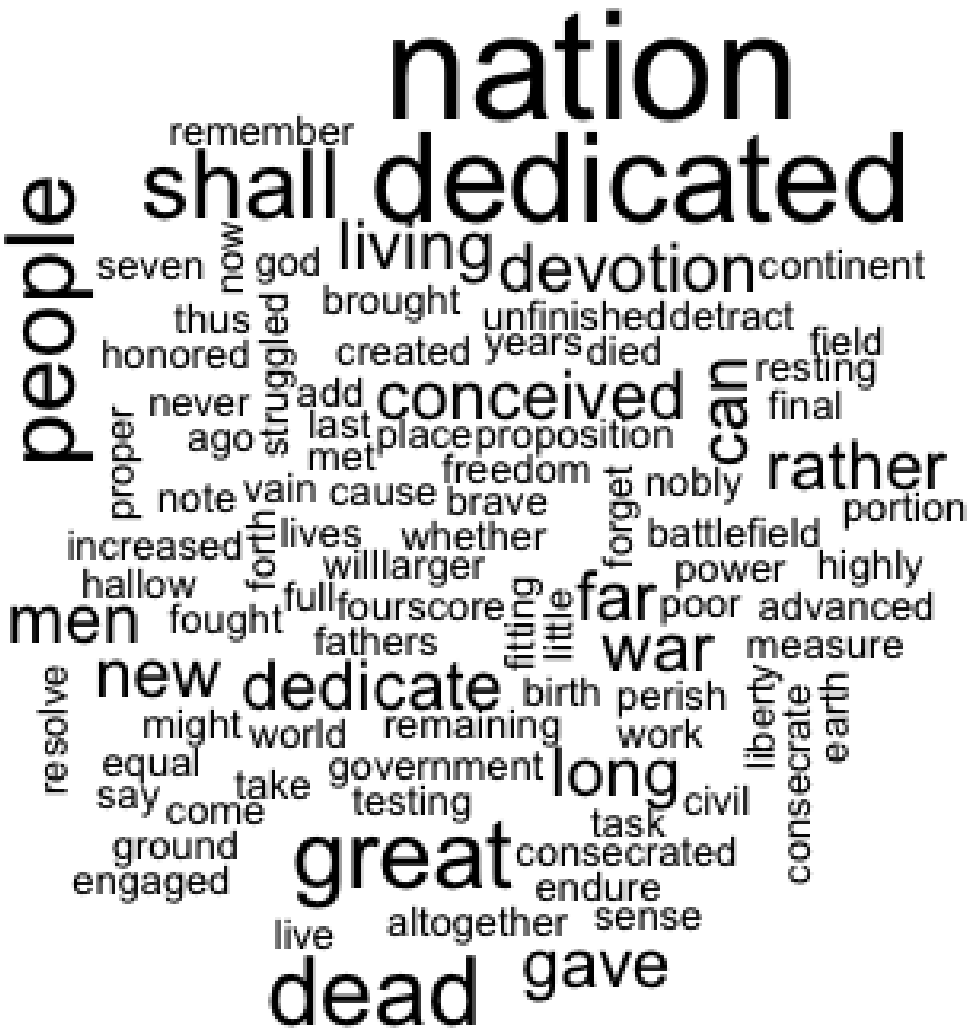
> *Make peace, not war.*
> *Make war, not peace.*

are equivalent in the bag-of-words model. How could this possibly work?

- It might not
- Central tendency in text
- Human supervision

The Bag of Words

Michele Claibourn



The Term Document Matrix/Vector Count

Terms	Freq	Terms	Freq	Terms	Freq	Terms	Freq
add	1	detract	1	gave	2	might	1
advanc	1	devot	2	god	1	nation	5
ago	1	die	1	govern	1	never	1
altogeth	1	earth	1	great	3	new	2
battlefield	1	endur	1	ground	1	nobli	1
birth	1	engag	1	hallow	1	note	1
brave	1	equal	1	high	1	now	1
brought	1	far	2	honor	1	peopl	3
can	2	father	1	increas	1	perish	1
caus	1	field	1	larger	1	place	1
civil	1	final	1	last	1	poor	1
come	1	fit	1	liberti	1	portion	1
conceiv	2	forget	1	littl	1	power	1
consecr	2	forth	1	live	4	proper	1
contin	1	fought	1	long	2	proposit	1
creat	1	fourscor	1	measur	1	rather	2
dead	3	freedom	1	men	2	remain	1
dedic	6	full	1	met	1	rememb	1

The Original

Fourscore and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation or any nation so conceived and so dedicated can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field as a final resting-place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But in a larger sense, we cannot dedicate, we cannot consecrate, we cannot hallow this ground. The brave men, living and dead who struggled here have consecrated it far above our poor power to add or detract. The world will little note nor long remember what we say here, but it can never forget what they did here. It is for us the living rather to be dedicated here to the unfinished

Preprocessing Text

- Acquire text, metadata
- Preprocess text: retain useful information
  For example
  1. Remove capitalization, punctuation
  2. Discard stop words (the, at, a, be, it)
  3. Exclude words occurring in less than/more than X% of documents
  4. Create equivalence class via stemming (lemmatizing)
     - `family, familie, families', familial` $\rightarrow$ `famil`

Term Document Matrix/Document Term Matrix

The result: a document represented as a vector that counts the number of times each of the unique words occur

$$W_i = (W_{i1}, W_{i2}, \ldots, W_{iM}) \tag{1}$$

The collection of count vectors in a matrix is called the term-document matrix.

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 3 \\ 0 & 2 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 5 \end{pmatrix}$$

$\boldsymbol{X} = W \times D$ matrix

- D = Number of documents
- W = Number of features

Topic Models: the goal

- To aid in automatic discovery of thematic content
- Bag of words assumption
- Documents can belong to multiple categories
- Requires no training set, no annotation of text
- Requires a decision about k, number of topics

We observe documents and words; topics are part of latent (hidden) structure. The goal is to infer this latent structure given the words and documents.

Topic Models: the goal

- To aid in automatic discovery of thematic content
- Bag of words assumption
- Documents can belong to multiple categories
- Requires no training set, no annotation of text
- Requires a decision about k, number of topics

We observe documents and words; topics are part of latent (hidden) structure. The goal is to infer this latent structure given the words and documents.

Latent Dirichlet Allocation (LDA) model!

Topic Models: The Generative Model

The model assumes a data-generation process; given a number of topics:

- Across a fixed vocabulary, every topic is a distribution over the vocabulary
- Each document is composed of a random mixture of topics
- To generate a document, d, randomly choose a distribution over topics. For each word in the document
  1. Probabilistically draw one of the k topics from the distribution over topics
  2. Given the topic, draw a word from the distribution of terms
  3. Rinse and repeat

The topic model aims to "reverse" this stylized stochastic process to infer estimates of the originating topic and term distributions.

The Model, with added Greekiness

Slightly more mathematically: Topics, and characteristics of topics (i.e., terms), are drawn from distributions.

- For each topic k, draw a distribution over words, $\phi_k \sim Dir(\beta)$
- For each document d,
  - Draw a vector of topic proportions, $\theta_d \sim Dir(\alpha)$
  - For each word i,
    - Draw topic assignment, $z_{d,i} \sim Mult(\theta_d)$ where $z_{d,i} \in \{1, \ldots, K\}$
    - Draw a word, $w_{d,i} \sim Mult(\phi_k, z_{d,i})$ where $w_{d,i} \in \{1, \ldots, V\}]$

The Model, with added Greekiness

V	number of words in vocabulary
K	number of topics
α	postive vector of length V
β	positive vector of length K
θ_d	document topic proportions
$z_{d,i}$	word topic assignment
k	index over topics
d	index over documents
i	index over words in document

The Model, with added Greekiness

The joint distribution of the observed and hidden random variables.
Topic proportions for documents come from a Dirichlet distirubtion

$$p(\theta_d|\alpha)$$

The Model, with added Greekiness

The joint distribution of the observed and hidden random variables.
Topic proportions for documents come from a Dirichlet distirubtion

$$p(\theta_d|\alpha)$$

Draw a topic assignment from θ_d from a multinomial

$$p(\theta_d|\alpha)p(z_{d,i}|\theta_d)$$

The Model, with added Greekiness

The joint distribution of the observed and hidden random variables.
Topic proportions for documents come from a Dirichlet distirubtion

$$p(\theta_d|\alpha)$$

Draw a topic assignment from θ_d from a multinomial

$$p(\theta_d|\alpha)p(z_{d,i}|\theta_d)$$

Each topic comes from some distribution that's appropriate over topics, a Dirichlet

$$p(\theta_d|\alpha)p(z_{d,i}|\theta_d)p(\phi|\beta)$$

The Model, with added Greekiness

The joint distribution of the observed and hidden random variables.
Topic proportions for documents come from a Dirichlet distirubtion

$$p(\theta_d|\alpha)$$

Draw a topic assignment from θ_d from a multinomial

$$p(\theta_d|\alpha)p(z_{d,i}|\theta_d)$$

Each topic comes from some distribution that's appropriate over topics, a Dirichlet

$$p(\theta_d|\alpha)p(z_{d,i}|\theta_d)p(\phi|\beta)$$

Draw the word conditional on z and all ϕs from a multinomial

$$p(\theta_d|\alpha)p(z_{d,i}|\theta_d)p(\phi_k|\beta)p(w_{d,i}|z_{d,i},\phi_k)$$

Bayesian, conjugacy, Dirichlet, oh my!

We don't observe everything; we infer from posteriors.

Hello, Bayesian mixture model!

Bayesian probability, generically, says that

$$\begin{aligned} \textit{posterior} &\propto \textit{likelihood} \times \textit{prior} \\ p(\theta|D) &\propto L(D|\theta) \times p(\theta) \end{aligned}$$

Bayesian, conjugacy, Dirichlet, oh my!

We don't observe everything; we infer from posteriors.

Hello, Bayesian mixture model!

Bayesian probability, generically, says that

$$\begin{aligned} \textit{posterior} &\propto \textit{likelihood} \times \textit{prior} \\ p(\theta|D) &\propto L(D|\theta) \times p(\theta) \end{aligned}$$

The joint distrbution for the LDA fits in this framework

$$p(\theta|\alpha)p(z|\theta)p(\phi|\beta)p(w|z,\phi)$$

Use the joint distribution to compute the conditional distribution of the hidden variables given the observed variables (aka, the posteriors).

Bayesian, conjugacy, Dirichlet, oh my!

We don't observe everything; we infer from posteriors.

Hello, Bayesian mixture model!

Bayesian probability, generically, says that

$$\begin{aligned} \textit{posterior} &\propto \textit{likelihood} \times \textit{prior} \\ p(\theta|D) &\propto L(D|\theta) \times p(\theta) \end{aligned}$$

The joint distrbution for the LDA fits in this framework

$$p(\theta|\alpha)p(z|\theta)p(\phi|\beta)p(w|z,\phi)$$

Use the joint distribution to compute the conditional distribution of the hidden variables given the observed variables (aka, the posteriors).

- The Dirichlet is a convenient prior for the multinomial – a conjugate prior
- The α of the Dirichlet for topic proportions controls the mean shape and sparsity of θ

A Note on Estimation

The challenge is computing the conditional distribution of the topic structure given the observed documents. Two (primary) approaches:

- Sampling-based algorithms: collect samples from the posterior to approximate it with an empirical distribution (e.g., Gibbs sampling).
- Variational methods: deterministic alternative, posit a parameterized family of distributions over the hidden structure and then find the member of that family that is closest to the posterior.

More on choice of K

- Compare likelihoods across a range of k
- Compare a function of the likelihood (perplexity) across a range of k on held-out data, e.g.,
  - Split data into training and testing sets
  - Apply model estimated on trained data to test data and measure perplexity

$$perplexity(w) = exp\left\{-\frac{\sum_{d=1}^{D} log(p(w_d|\Phi, \alpha))}{\sum_{d=1}^{D} N_d}\right\}$$

    How "perplexed" is the model using the test data? Lower perplexity is better.
  - Only useful for comparisons; values have no inherent meaning!

Cannot replace more manual evaluation!

More on topic quality

Manual approaches

- Consider words most frequently assigned to a topic
- Look at document topic assignments
- Read documents dominated by given topics

More on topic quality

Manual approaches

- Consider words most frequently assigned to a topic
- Look at document topic assignments
- Read documents dominated by given topics

Experimental approaches (Chang et al 2009)

- Word intrusion: can subjects identify spurious word inserted into list
- Topic intrusion: can subjects identify spurious topic assigned to document

More on topic quality

Manual approaches

- Consider words most frequently assigned to a topic
- Look at document topic assignments
- Read documents dominated by given topics

Experimental approaches (Chang et al 2009)

- Word intrusion: can subjects identify spurious word inserted into list
- Topic intrusion: can subjects identify spurious topic assigned to document

Experimental measures were non- or negatively correlated with likelihood measures!

More on topic quality

Quantitative approaches

- Semantic coherence/cohesiveness: high-probability words for the topic tend to co-occur within documents (e.g., Mimno et al 2011)

$$cohesive_k = \sum_{l=2}^{L} \sum_{m=1}^{l-1} log \left(\frac{D(w_{lk}, w_{mk}) + 1}{D(w_{mk})} \right)$$

where D counts the number of time the argument occurs.

More on topic quality

Quantitative approaches

- Semantic coherence/cohesiveness: high-probability words for the topic tend to co-occur within documents (e.g., Mimno et al 2011)

$$cohesive_k = \sum_{l=2}^{L} \sum_{m=1}^{l-1} log \left(\frac{D(w_{lk}, w_{mk}) + 1}{D(w_{mk})} \right)$$

  where D counts the number of time the argument occurs.

- Term Exclusivity: high-probability words for the topic are unlikely to appear within top words of other topics (e.g., Bischof and Airoldi 2012)

$$exclusivity_{w,k} = \frac{\mu_{w,k}}{\sum_{l=1}^{K} \mu_{w,l}}$$

  where $\mu_{w,k}$ is a word's topic-specific frequency (or assignment).

A topic that is both cohesive and exclusive is more likely to be semantically useful.

Extensions, Assumptions

The "basic" LDA assumes documents are exchangeable (independent). Several extensions are intended to relax this:

- Correlated Topic Model (Blei and Lafferty 2007): Dirichlet prior is replaced with a logistic Normal distribution
- Dynamic Topic Model (Blei and Lafferty 2006): parameters change using an evolution model
- Expressed Agenda Model (Grimmer 2010): allows for differences in topic probabilities across authors
- Structural Topic Model (Roberts et al 2013): for topic modeling with document-level covariate information.

Implementations (in R)

- `topicmodels`: provides interface to the C and C++ code by Blei et al (using Variational EM) and by Phan et al (using Gibbs Sampling); works well with `tm`. `https://cran.r-project.org/web/packages/topicmodels/index.html`
- `lda`: uses a collapsed Gibbs Sampler, a different form of input, but is reportedly faster; contains some text processing capability. `https://cran.r-project.org/web/packages/lda/index.html`
- `mallet`: a wrapper for the LDA functionality in Mallet (Java implentation), to load results of a model into R. `https://cran.r-project.org/web/packages/mallet/index.html`
- `stm`: allows researchers to estimate topic models with document-level covariates, also includes tools for model selection, visualization, and estimation of topic-covariate regressions. `https://cran.r-project.org/web/packages/stm/index.html`

Some Principles

1. All quantitative models of language are wrong, but some are useful.
2. Validate, validate, validate
3. Quantitative methods agument humans, not replace them.

References

- David M. Blei and John D. Lafferty. 2006. "Dynamic Topic Models." *International Conference on Machine Learning (ICML)*.
- David M. Blei and John D. Lafferty. 2007. "A Correlated Topic Model of Science." *The Annals of Applied Statistics* 1:17-35.
- Jonathan M. Bischof and Edoardo M. Airoldi. 2012. "Summarizing topical content with word frequency and exclusivity." *International Conference on Machine Learning (ICML)*.
- Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." *Neural Information Processing Systems (NIPS)*
- Justin Grimmer and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21: 267-297.
- Justin Grimmer. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18: 1-35.
- David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. "Optimizing Semantic Coherence in Topic Models." *Empirical Methods in Natural Language Processing (EMNLP)*.
- Brendan O'Connor, David Bamman, and Noah A. Smith. 2011. "Computational Text Analysis for Social Science: Model Assumptions and Complexity." *NIPS Workshop on Computational Social Science and Wisdom of the Crowds.*
- Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. 2013. "The Structural Topic Model and Applied Social Science." *NIPS Workshop on Topic Models: Computation, Application, and Evaluation.*