

A Glimpse into Text Mining

Malte Schierholz

University of Mannheim, MZES
Institute for Employment Research (IAB)

Malte.Schierholz@iab.de

October 20, 2017

Overview




- 1 Introduction and Applications
- 2 Preprocessing
- 3 Supervised Learning for Text Classification
- 4 Topic Models
- 5 Resources

Fields of Research

Three major perspectives:

- Natural Language Processing (Association for Computational Linguistics 2017)
 - Machine translation: English document → German document
 - Part-of-speech tagging: Detect for each word if it is a noun, a verb, an adjective, ...
 - Automatic Question Answering: Select an appropriate answer for a given question (*Chat Bots*)
 - ...
- Information Retrieval: Search for complex content ... (Manning et al. 2009)
 - ... in the Internet (*Google Search*)
 - ... in academic libraries (*OPAC library catalogues*)
- Text Mining, Text as Data or “Distant Reading”
(Feldman & Sanger 2007; Grimmer & Stewart 2013; Gentzkow et al. 2017; Jänicke et al. 2015; Schulz 2011)
 - Information overload: Text is everywhere, but it is too much to read it all
 - How can we still gain insights from it?

Text as Data-Framework

Document	Outcome		
	Estimate	True value	
	\longrightarrow	\hat{V}_1	V_1
	\longrightarrow	\hat{V}_2	V_2
\vdots			\vdots
	\longrightarrow	\hat{V}_n	V_n

- Computers can calculate numbers (estimates) from large documents
- Humans must evaluate if the estimates are useful

Applications

Some examples

- Authorship: Did Philip Wright or his son Sewall write an appendix in which instrumental variables were invented?
- Stock Prices: Can one forecast changing stock prices from companies' annual reports or from newspaper articles?
- Google Flu: Using billions of search queries, can one estimate the flu prevalence for specific regions?
- Media Slant: By comparing political speeches in Congress with newspaper articles, what is the newspaper's position on a left-right scale of political ideology?
- Content analysis: What are dominant topics in some text collection (e.g., articles from a newspaper, Congress speeches) and how does the focus of attention shift over time?

(Gentzkow et al. 2017)

Overview

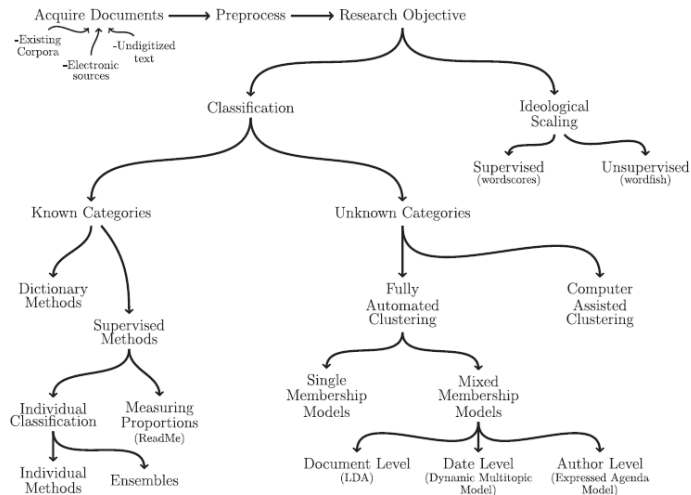





Fig. 1 An overview of text as data methods.

(taken from Grimmer & Stewart 2013)

Preprocessing

Many methods, but preprocessing is common to most

Document		Numeric Vector		Outcome	
				Est.	True
	→	C_1	→	\hat{V}_1	V_1
	→	C_2	→	\hat{V}_2	V_2
⋮		⋮			⋮
	→	C_n	→	\hat{V}_n	V_n

Preprocessing

Document:

Time flies like an arrow. Fruit flies like a banana.

Same document after cleaning and processing:

	arrow	banana	fli	fruit	like	time
$C_i =$	1	1	2	1	2	1

Steps taken:

- ➊ Remove punctuation
- ➋ Lowercase letters
- ➌ Remove stopwords (like “a”, “the”)
- ➍ Stemming (“flies” → “fli”, based on a linguistic algorithm)
- ➎ Count word frequency

Preprocessing

Document:

Time flies like an arrow. Fruit flies like a banana.

Same document after cleaning and processing:

	arrow	banana	fli	fruit	like	time
$C_i =$	1	1	2	1	2	1

Preprocessing aims to simplify the document without losing important information, but

- Meaning of words is ignored (e.g. “flies”)
- Word order is ignored (so-called “bag-of-words” representation)

Many more ways exist for processing (e.g. N-grams, letterwise, tf-idf)

→ Optimal approach depends on the research question

Document-Term Matrix

Preprocessing converts a *corpus* (= a set of documents) into a *Document-Term Matrix*

$$C = \begin{pmatrix} C_1 \\ \vdots \\ C_i \\ \vdots \\ C_n \end{pmatrix} = \begin{pmatrix} \text{arrow} & \text{banana} & \text{fli} & \text{fruit} & \text{like} & \text{time} & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 2 & 1 & 2 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & 1 & \dots \end{pmatrix} \quad (1)$$

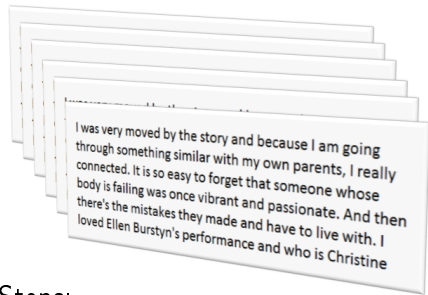
Matrix is ...

- sparse (= many zeros) → Do fast algorithms exist?
- high-dimensional (= several thousand columns) → Avoid overfitting?

→ Specialized statistical methods needed

Example: Supervised Learning for Text Classification I

Data: 5000 movie reviews from the Internet Movie Database (IMDb), specially selected for sentiment analysis



Task: Find the positive reviews
(Information retrieval)

Steps:

- ① Transform raw text to Document-Term Matrix
- ② Learn model from labeled training data
 - Here: Logistic regression with ridge penalty
- ③ Predict sentiment of new data
 - Here: Evaluated on 1000 test observations

Example: Supervised Learning for Text Classification II

How much labor is needed to label the training data? What size should N_{train} have?

$N_{train} = \mathbf{100 \text{ documents}}$

$N_{test} = 1000 \text{ documents}$

Precision 60%

Recall 88%

Accuracy 63%

Metrics for Evaluation

- Precision = $\frac{\text{No. of documents that are retrieved (=1) and positive (=1)}}{\text{No. of documents that are retrieved (=1)}}$
- Recall = $\frac{\text{No. of documents that are retrieved (=1) and positive (=1)}}{\text{No. of documents that are positive (=1)}}$
- Accuracy = $\frac{\text{No. of documents with (Retrieved == positive)}}{N}$

Example: Supervised Learning for Text Classification II

How much labor is needed to label the training data? What size should N_{train} have?

$N_{train} = 100$ documents

$N_{test} = 1000$ documents

Precision 60%

Recall 88%

Accuracy 63%

4000 documents

1000 documents

84%

87%

85%

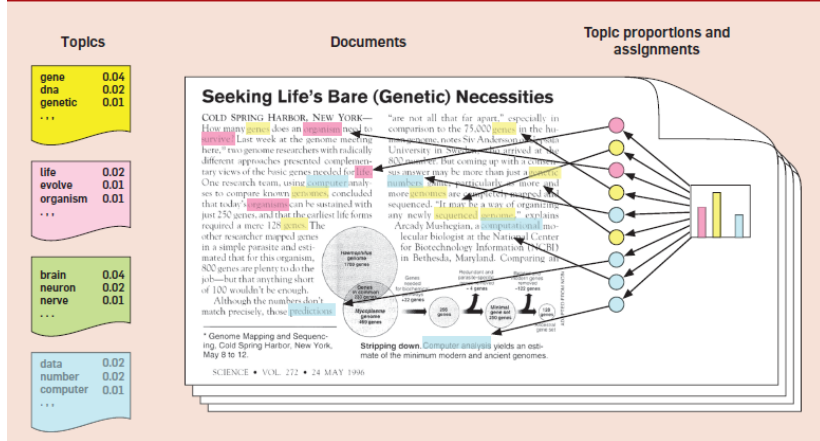
Metrics for Evaluation

- Precision = $\frac{\text{No. of documents that are retrieved (=1) and positive (=1)}}{\text{No. of documents that are retrieved (=1)}}$
- Recall = $\frac{\text{No. of documents that are retrieved (=1) and positive (=1)}}{\text{No. of documents that are positive (=1)}}$
- Accuracy = $\frac{\text{No. of documents with (Retrieved == positive)}}{N}$

Example: Latent Dirichlet Allocation for Topic Modeling

Goal: Discover document topics automatically

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



(taken from Blei 2012)

Example: Latent Dirichlet Allocation for Topic Modeling

Data: 112,000 newspaper pieces (24 million words) published between 1860 and 1865 in the *Richmond Daily Dispatch*

Research Question: Explore changes and continuities in Civil War Richmond

Illustration

- Categorize newspaper articles
 - see <http://dsl.richmond.edu/dispatch/pages/intro>
- Number of fugitive slave ads
 - see <http://dsl.richmond.edu/dispatch/Topics/view/15>
- The numbers are highly correlated with manual counts
 - see <http://dsl.richmond.edu/dispatch/pages/intro>

“Topic modeling and other distant reading methods are most valuable [...] when they reveal patterns that we can’t [easily explain], patterns that surprise us and that prompt interesting and useful research questions.”
(Nelson)

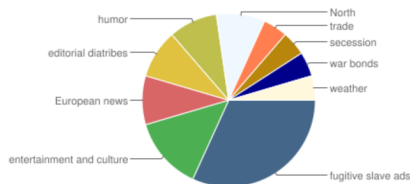
Example: Latent Dirichlet Allocation for Topic Modeling II

Another example. Four years later in December 1865 the following appeared in the *Dispatch*:

Some of our tradesmen advertise the Fenian hat. We should think that the style just at present must be a shocking bad hat.—*New York Tribune*.

The above is not very witty, but very remarkable as coming from Horace Greeley, who is everywhere known as "the philosopher of the old white coat and shocking bad hat."

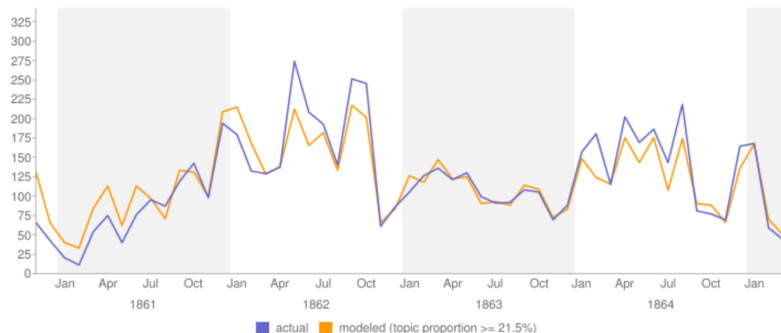
How to classify this?



(taken from <http://dsl.richmond.edu/dispatch/pages/intro>)

Example: Latent Dirichlet Allocation for Topic Modeling III

When overlayed with the actual count of fugitive slave ads in the paper, the accuracy of the model in detecting these ads is apparent, even remarkable.⁸



(taken from <http://dsl.richmond.edu/dispatch/pages/intro>)

Example: Latent Dirichlet Allocation for Topic Modeling

Many more visualizations are possible

- Interactive visualization of fugitive slave ads
 - see <http://dsl.richmond.edu/dispatch/Topics/view/15>

“Topic modeling and other distant reading methods are most valuable [...] when they reveal patterns that we can’t [easily explain], patterns that surprise us and that prompt interesting and useful research questions.”
(Nelson)

Software Resources

Resources for R

- General overview: <https://cran.r-project.org/web/views/NaturalLanguageProcessing.html>
- Easy-to-use text classification: Package `RTextTools`
- General framework: Package `tm`
- Alternative framework: Package `text2vec`

Resources for Python

- Natural Language Toolkit `NLTK`

More open-source software

- Stanford CoreNLP, for syntactic analysis
- MALLET, for statistical text models

and far more exists ...

References



Association for Computational Linguistics (2017)

Wiki of the Association for Computational Linguistics. Online at https://aclweb.org/aclwiki/Main_Page



Blei, David M. (2012)

Probabilistic Topic Models. *Comm. of the ACM* **55**(4). 77–84



Feldman, Ronen, Sanger, James (2007)

The Text Mining Handbook. Cambridge University Press.



Gentzkow, Matthew, Bryan T. Kelly & Matt Taddy (2017)

Text as data. *NBER Working Paper No. 23276*. 1–53



Grimmer, Justin & Brandon M. Stewart (2013)

Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* **21**(3). 267–297

References



Jänicke, Stefan, Franzini, Greta, Cheema, Muhammad & Scheuermann Gerik (2013)

On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges *Eurographics Conference on Visualization*.



Manning, Christopher D., Prabhakar Raghavan & Hinrich Schütze (2008)
Introduction to Information Retrieval. Cambridge University Press



Schulz, Kathryn (2011)

What is distant reading? *New York Times*.

Online at http://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html?_r=0&pagewanted=all
(25.09.2017)