Text As Data - Lecture 1

Machine Learning and Big Data

Vincent Bagilet

2025-11-26

Introduction

Text as Data

- A plethora of text data
- Nowadays, we have both the technology and methods to study them massively

- Text data is **unstructured**:
  - Basically not sorted in a sort of “table-like” format
  - Info we want mixed with info we do not want
  - Need to throw away some info ⇒ select what to keep

- Text data is very **high-dimensional**

- We often want to relate text data to **metadata**
  - *eg* who, when, on what topic?

Think About Your Own Question

- What type of text data? What source?
- How would you get the data?
- Which research question?
- How would you go about studying this?

Outline and Resources

Housekeeping

Assignment

- **Exercise** to do for the end of next week
- You can start today, even if we will not have covered everything
- Available on *Portail des études*

Material

- Available on *Portail des études* AND on the course website

Objectives

- **Why** is text data useful **in economics**?
- What is a typical **empirical workflow**?
- How to **concretely implement** these analyses in Python?
- What are **recent developments** in the field?
- Link this section to the rest of the class

Relation to the rest of the class

- More on the data analysis and big data side than on the ML one
- Here, focus on **pre-processing steps**, *ie* prepare the data to use it in:
  - A ML algorithm
  - An econometric analysis
- Can use the tools and algorithms you saw in the rest of the class on text data
  - Examples?

Outline

1. Introduction
2. Applications in economics
3. Workflow for analysis
4. Pre-processing
5. Representation
6. Dictionary based methods
7. Machine learning methods
8. Introduction to deep learning methods
9. Validation
10. Use in econometric analyses

Economics Resources

ANNUAL REVIEWS

Annual Review of Economics

Text Algorithms in Economics

Elliott Ash[1] and Stephen Hansen[2,3]

[1]Center for Law and Economics, ETH Zurich, Zurich, Switzerland

[2]Department of Economics, University College London, London, United Kingdom; email: stephen.hansen@ucl.ac.uk

[3]Centre for Economic Policy Research, London, United Kingdom

Annu. Rev. Econ. 2023. 15:659–88

First published as a Review in Advance on July 5, 2023

The *Annual Review of Economics* is online at economics.annualreviews.org

https://doi.org/10.1146/annurev-economics-082222-074352

JEL codes: C18, C45, C55

Keywords

text as data, topic models, word embeddings, large language models, transformer models

Abstract

This article provides an overview of the methods used for algorithmic text analysis in economics, with a focus on three key contributions. First, we introduce methods for representing documents as high-dimensional count vectors over vocabulary terms, for representing words as vectors, and for representing word sequences as embedding vectors. Second, we define four core empirical tasks that encompass most text-as-data research in economics and enumerate the various approaches that have been taken so far to accomplish these tasks. Finally, we flag limitations in the current literature, with a focus on the challenge of validating algorithmic output.

OPEN ACCESS

Journal of Economic Literature 2019, 57(3), 535–574
https://doi.org/10.1257/jel.20181020

Text as Data†

Matthew Gentzkow, Bryan Kelly, and Matt Taddy*

An ever-increasing share of human interaction, communication, and culture is recorded as digital text. We provide an introduction to the use of text as an input to economic research. We discuss the features that make text different from other forms of data, offer a practical overview of relevant statistical methods, and survey a variety of applications. (*JEL* C38, C55, L82, Z13)

1. *Introduction*

New technologies have made available vast quantities of digital text, recording an ever-increasing share of human interaction, communication, and culture. For social scientists, the information encoded in text is a rich complement to the more structured kinds of data traditionally used in research, and recent years have seen an explosion of empirical economics research using text as data.

To take just a few examples: In finance, text from financial news, social media, and company filings is used to predict asset price movements and study the causal impact of new information. In macroeconomics, text is used to forecast variation in inflation and unemployment, and estimate the effects of policy uncertainty. In media economics, text from news and social media is used to study the drivers and effects of political slant. In industrial organization and marketing, text from advertisements and product reviews is used to study the drivers of consumer decision making. In political economy, text from politicians' speeches is used to study the dynamics of political agendas and debate.

The most important way that text differs from the kinds of data often used in economics is that text is inherently high dimensional. Suppose that we have a sample of documents, each of which is w words long, and suppose that each word is drawn from a vocabulary of p possible words. Then the unique representation of these documents has dimension p^w. A sample of thirty-word Twitter messages that use only the one thousand most common words in the English language, for example, has roughly as many dimensions as there are atoms in the universe.

A consequence is that the statistical methods used to analyze text are closely related to those used to analyze high-dimensional data in other domains, such as machine learning and computational biology. Some methods, such as lasso and other penalized regressions, are applied to text more or less exactly as they are in other settings. Other methods, such as topic models and multinomial inverse regression, are close cousins of more general

* Gentzkow: Stanford University. Kelly: Yale University and AQR Capital Management. Taddy: University of Chicago Booth School of Business.

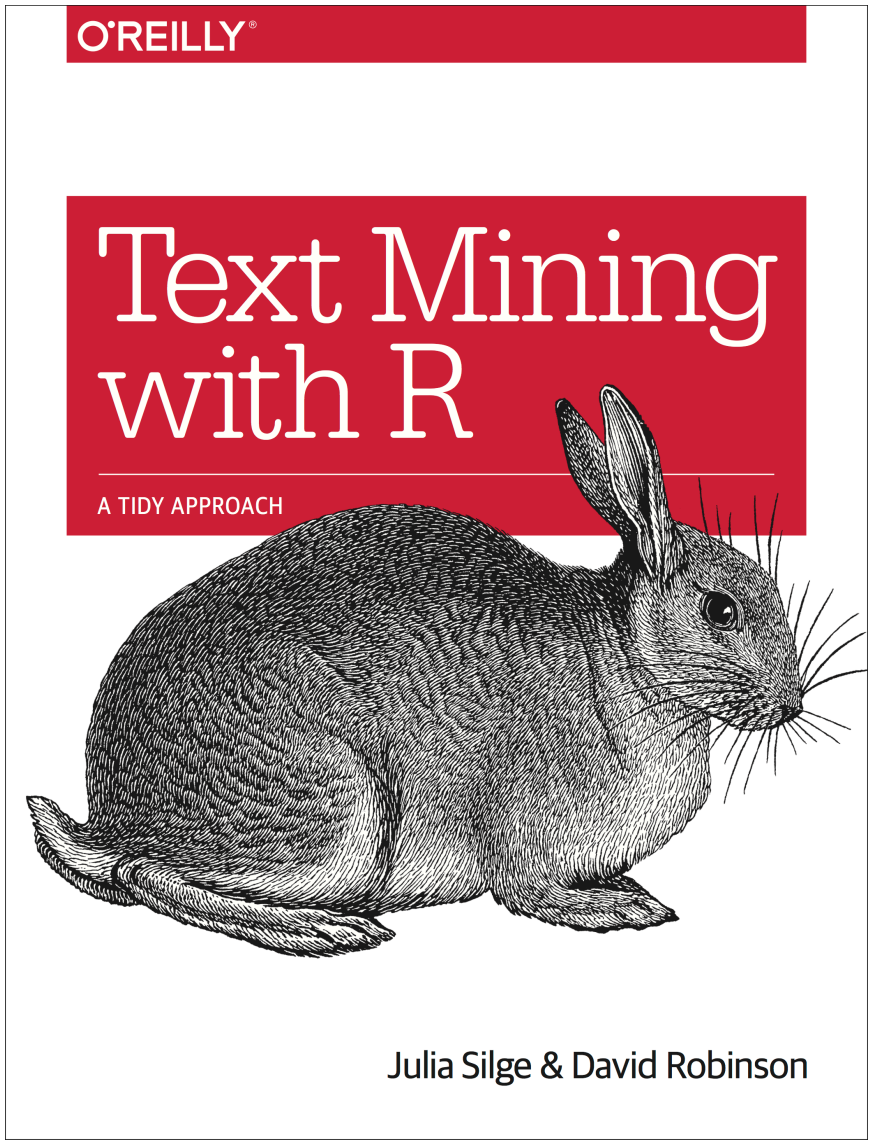
† Go to https://doi.org/10.1257/jel.20181020 to visit the article page and view author disclosure statement(s).

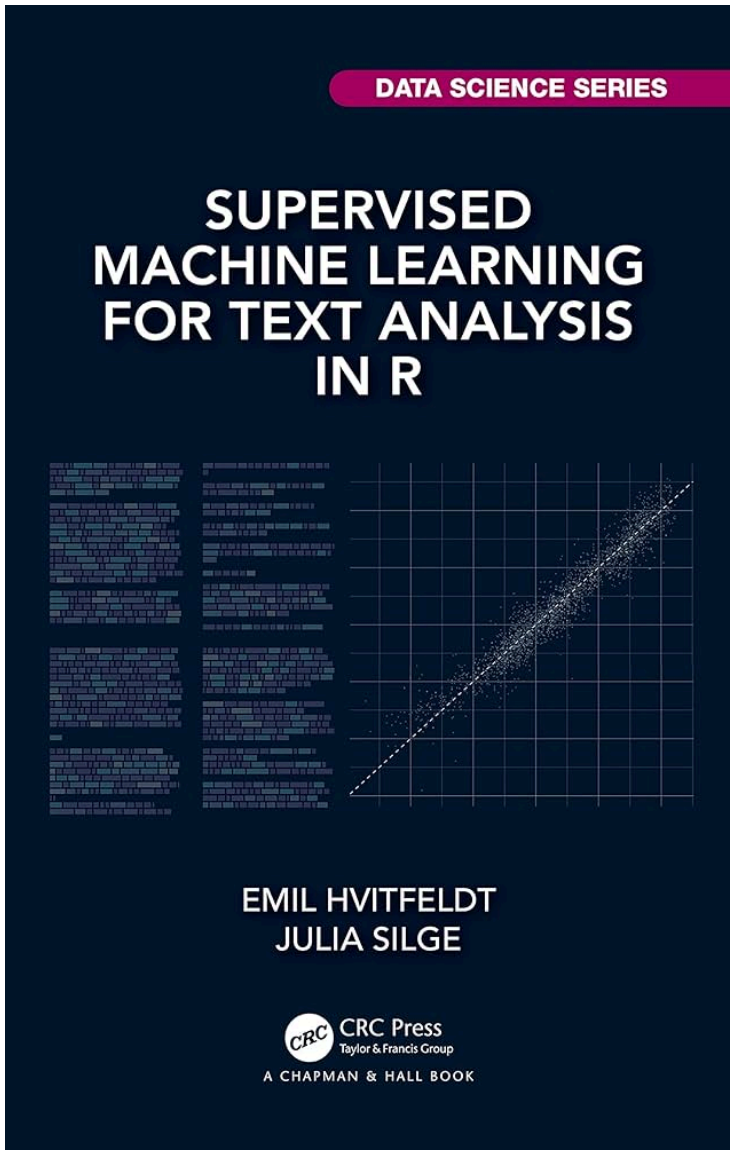
Elliott Ash's NLP class

Python Resources

- spaCy 101: tutorial for `spaCy`, a Python library for NLP
- Natural Language Processing with Python: general Python book on NLP (with the `nltk` library)
- Companion Python notebooks to Ash and Hansen (2023)

R Resources


O'REILLY
Text Mining with R
A TIDY APPROACH
Julia Silge & David Robinson


DATA SCIENCE SERIES
SUPERVISED MACHINE LEARNING FOR TEXT ANALYSIS IN R
EMIL HVITFELDT
JULIA SILGE
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Applications in Economics

Measuring Document Similarity

- **What?**
  - Make pairwise document comparisons
  - Forming clusters of related documents
- **Examples**
  - Cagé, Hervé, and Viaud (2020) use the distance between online news articles and social media posts to group items into common stories
  - Biasi and Ma (2022) build a measure of **syllabi**'s distance from **frontier knowledge** (academic articles) and then relate this metric to socio-economic variables
- **How?**
  - Represent documents in some sort of vector form and then compute cosine similarity
  - There are different ways of defining vectors

Concept Detection

- **What?**
  - Detect the presence of a concept
- **Example**
  - Djourelova, Durante, and Martin (2024) interpretable news topics using CorEx
- **How?**
  - Dictionary methods / pattern matching
  - Topic models to identify latent concepts
  - Distance between documents and dictionaries
  - Supervised (human classification) learning problem
  - BERT/GPT type of models

Relation Between Concepts

- **What?**
  - How are concepts related?
- **How?**
  - Co-occurrence of dictionaries
  - Word Embedding Association Test (WEAT):
    - Picks two terms at each end of the spectrum (eg rich and poor) and compute the cosine similarity of each term of interest with these “extreme” terms
- **Examples**
  - Ash, Chen, and Ornaghi (2024) on gender attitudes of individual US judges
  - Kozlowski, Taddy, and Evans (2019) locate words on meaningful dimension, *eg*, rich/poor (this has a feel of Bourdieu’s *social space*)

Associating Text with Metadata

- **Impute** an outcome of interest to other documents
- Supervised learning
- *eg* have some data on party label but not for everyone
- **Example**
  - Gentzkow and Shapiro (2010) build a model for party label using US Congressional speeches. Then use it to predict political bias of newspapers

Workflow for Analysis

How are text data structured?

- Text data is a sequence of characters called **documents**
- A **corpus** is a collection of documents
- The unit of analysis (the “document”) depends on the question:
  - Fine enough to fit relevant metadata variation
  - Not unnecessarily fine to reduce dimensionality

From text to a usable format

- Text data is very highly dimensional
- Need to transforming data into a **useful representation** for modeling
- Encode text data into **numeric arrays**:
  - Tokenization and word counts
  - Document-Term Matrix
  - Latent Semantic Allocation
  - Word embeddings
- Both for computational reasons and to have objects we can manipulate and do computations on

Steps of Text Analysis

- Gentzkow, Kelly, and Taddy (2019) summarize a text analysis in three steps:
  1. Represent raw text D as a numerical array C
  2. Map C to predicted values $\hat{V}$ of unknown outcomes V
  3. Use $\hat{V}$ in subsequent descriptive or causal analysis

Workflow, rephrased

1. Collecting data
2. Pre-processing (stemming, lemming, etc)
3. Data transformation (from pre-processed text to numeric arrays)
4. Analysis and modelisation
5. Validation
6. Use in econometric analysis

Type of Analyses

- **Dictionary Methods**: rely on pre-defined lexicons and info associated with specific keywords
- **Rule-Based Methods**: use predefined rules and patterns (*eg* RegEx)
- **Machine Learning Methods**: methods leveraging lightweight ML models
  - Supervised ML
  - Unsupervised ML
- **Deep Learning Methods**: methods leveraging neural networks

There is an important **interpretability-flexibility trade-off**

Gathering data

Common Text Data Sources

- **Political economy**: news articles, parliamentary debates, speeches, social network posts, party manifestos, press releases, etc
- **Economic history**: correspondence, institutional documents, books, etc
- **Labor and Industrial Organisation (IO)**: job adds, product descriptions, etc
- **Finance and macro**: earnings conference calls, central bank speeches, etc

Where to find data

- Digital archives of **news** (Factiva, ProQuest, Europress, etc)
  - Expensive. Some universities have subscriptions.
- Companies **APIs** (*eg* Twitter[1], New York Times)
  - In general there are **wrappers** to directly access the data from Python or R
- Data sets of occurrence of keywords (Google Trends, Google Ngram, Gallicagram)
- **Online** ⇒ scrapping
- **Printed texts**: digitalize and OCR
- Curated lists of text data: here, here, here, and there

Overview of web scraping

- The overall idea is to **write algorithms** that:
  1. **Browse** a website and **download** relevant `html` pages
     - Main challenge: identify the relevant pages
  2. Transform the `html` pages into a **data frame** containing text data
     - Typical structure: date column, title column, author column, text column, etc
     - Main challenge: identify the right tags
- Not limited to text data: you can retrieve anything that is online

HTML pages

- HTML: *HyperText Markup Language*
- Structure of an example html page

```
<html>
<head>
  <title>Page title</title>
</head>
<body>
  <h1 id='first'>A heading</h1>
  <p>Some text &amp;amp; <b>some bold text.</b></p>
  <img src='myimg.png' width='100' height='100'>
</body>
```

- An IMDb example

Webscraping concretely

- There are libraries for web scraping with useful **documentation**:
  - Beautiful Soup (Python)
  - rvest (R)
- Basic pages are **static** and easy to scrape
- More complex to scrape **dynamic** pages
  - Use Selenium
  - It opens a browser and can be used to automate tasks

A few webscraping tips

Tips

- Before scraping, look at whether someone did not already scraped what you want to scrape (*eg* for IMDb)
- To identify relevant pages use the site **sitemap** (website.com/sitemap.xml)
    - Example
- Use **ChatGPT** or Claude for scraping
- Use SelectorGadget or tools from the developer panel of your browser
- Save html pages before processing them
- Read the `robots.txt` page to learn what you can and cannot do

Optical Character Recognition (OCR)

- Some data is either not digitized or digitized but in text/PDF format
- Use software to transform it to text data
- Tesseract and its Python and R wrappers
- When PDFs are digitally made, you can use a text extraction software to directly retrieve the text from the metadata

Pre-processing

Goal

- Going from raw text to something usable in analysis
- Often do run analysis on the raw dataset
- Some information is useful, other less:
  - **What to keep?**
  - **What format?**

Tokenization

- One of the **first steps** of turning raw data into numbers
- Turn it into **tokens**: a meaningful unit of text, such as a word, a character
- Example from the French parliament: *"Ne fumez pas la moquette"*
- Words: *{Ne, fumez, pas, la, moquette}*
- n-grams (here 2-grams): *{Ne fumez, fumez pas, pas la, la moquette}*
- Sentences

```python
import nltk

# Run only once
nltk.download('punkt')
nltk.download('punkt_tab')

sentence = "Ne fumez pas la moquette"
tokens = nltk.word_tokenize(sentence)

tokens
```

```
['Ne', 'fumez', 'pas', 'la', 'moquette']
```

Pre-processing Choices

- Pre-processing choices often affects the results:

What To Keep?

- Not all data is useful
- Sometimes, may remove useful information: *eg* “happy” vs “not happy”
- Remove capitalization?
- Punctuation?
- Numbers?
- Stopwords?
    - Words that carry little information
    - How do you define this set?

Capitalization and punctuation

- Often uninformative
- But sometimes important:
  - Sentence splitting
  - Part-of-speech tagging
  - Named entity recognition
  - Text generation
- Option: keep caps not at the beginning of a sentence

```
1 tokens_lower = [word.lower() for word in tokens]
2 print(tokens_lower)
```

```
['ne', 'fumez', 'pas', 'la', 'moquette']
```

Stop words

- We have seen that might create noise but may also carry meaning
- How to define stopwords?
- Can use existing lists

```
1  stopwords = set(nltk.corpus.stopwords.words('french'))
2
3  print(stopwords)
```

```
{'mais', 'mes', 'eus', 'ayez', 'vos', 'nos', 'fussent', 'étants', 'était', 'eusses', 'étaient', 'aurait', 'étiez', 'étions',
'seras', 'eue', 'sera', 'seriez', 'aura', 'étais', 'vous', 'eussent', 'serai', 'étante', 'nous', 'ait', 'notre', 'des',
'que', 'on', 'étées', 'auriez', 'se', 'l', 'aie', 'ta', 'sommes', 'avaient', 'serions', 'fusse', 'ou', 'de', 'ont', 'étés',
'sois', 'du', 'j', 'ce', 'ayante', 'êtes', 'd', 'ses', 'sont', 'lui', 'avons', 'ces', 'le', 'serait', 'par', 'eut', 'fus',
'tes', 'sa', 'avions', 'aux', 'en', 'son', 'suis', 'aurions', 'eûtes', 'serais', 'ils', 'moi', 'mon', 'fussions', 'auraient',
'avec', 'eux', 'dans', 'elle', 'y', 'un', 'seraient', 'ai', 'avez', 'aies', 'eussions', 'ayant', 'eurent', 'ton', 'étée',
'ma', 'soyez', 'qu', 'eussiez', 'au', 'furent', 'eues', 'il', 'c', 'fûtes', 'ayants', 'ayantes', 'as', 'auront', 'est',
'leur', 'serons', 'pour', 'n', 'les', 'eûmes', 'me', 'fussiez', 'serez', 'même', 'te', 'avais', 'étant', 't', 'toi', 'et',
'aviez', 'soit', 'aurons', 'aurai', 'ayons', 'la', 'sur', 'avait', 'eusse', 'soient', 'votre', 'soyons', 'étantes', 's',
'été', 'm', 'qui', 'pas', 'aurais', 'aient', 'tu', 'fut', 'auras', 'aurez', 'eu', 'eût', 'fût', 'à', 'ne', 'es', 'fusses',
'fûmes', 'je', 'une', 'seront'}
```

```
1  tokens_filtered = [word for word in tokens_lower if word not in stopwords]
2  print(tokens_filtered)
```

```
['fumez', 'moquette']
```

- Which list? How to define it?

- Build one
  - By hand, starting from existing ones
  - Remove words that appear the most frequently, *eg* by Inverse Document Frequency (idf):

$$idf(term) = \ln\left(\frac{n_{\text{documents}}}{n_{\text{document containing term}}}\right)$$

- *idf* decreases the weight of commonly used words and increases that of words that are rarely used in the corpus
- *tf-idf*: multiplying the term frequency (*tf*) and its *idf*

Stemming/Lemmatization

- Find the stem of the word:
  - Ruled based
  - Produces grams that are not actual words

```
stemmer = nltk.stem.snowball.SnowballStemmer(language="french")
tokens_stemmed = [stemmer.stem(word) for word in tokens]
print(tokens_stemmed)
```

```
['ne', 'fum', 'pas', 'la', 'moquet']
```

- Lemmatization:
  - Similar but with semantic rules
  - Produces actual words but sentences that do not mean anything

```
import spacy
nlp = spacy.load("fr_core_news_sm")
doc = nlp(sentence)
lemmas = [token.lemma_ for token in doc]

print(lemmas)
```

```
['ne', 'fumer', 'pas', 'le', 'moquette']
```

Summary

NLP in Economics

- Measuring document similarity
- Concept detection
- Relation between concepts
- Associating text with metadata

Overal Approach

- **Get** text data (ready-made, scrap, OCR, etc)
- **Pre-process** the data
- **Transform** data into useful format (a numeric array)
- **Run** analysis. Several methods:
  - Dictionary-based
  - Rule base
  - Machine Learning
  - Deep Learning
- Use output in an **econometric** analysis

References

Ash, Elliott, Daniel L. Chen, and Arianna Ornaghi. 2024. "Gender Attitudes in the Judiciary: Evidence from US Circuit Courts." *American Economic Journal: Applied Economics* 16 (1): 314–50. https://doi.org/10.1257/app.20210435.

Ash, Elliott, and Stephen Hansen. 2023. "Text Algorithms in Economics." *Annual Review of Economics* 15 (1): 659–88. https://doi.org/10.1146/annurev-economics-082222-074352.

Biasi, Barbara, and Song Ma. 2022. "The Education-Innovation Gap." Working {{Paper}}. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w29853.

Cagé, Julia, Nicolas Hervé, and Marie-Luce Viaud. 2020. "The Production of Information in an Online World." *The Review of Economic Studies* 87 (5): 2126–64. https://doi.org/10.1093/restud/rdz061.

Djourelova, Milena, Ruben Durante, and Gregory J Martin. 2024. "The Impact of Online Competition on Local Newspapers: Evidence from the Introduction of Craigslist." *The Review of Economic Studies*, May, rdae049. https://doi.org/10.1093/restud/rdae049.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57 (3): 535–74. https://doi.org/10.1257/jel.20181020.

Gentzkow, Matthew, and Jesse M Shapiro. 2010. "What Drives Media Slant? Evidence From U.S. Daily Newspapers." *Econometrica* 78 (1): 35–71. https://doi.org/10.3982/ECTA7195.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class Through Word Embeddings." *American Sociological Review*, September. https://doi.org/10.1177/0003122419877135.