Text Mining and Sentiment Analysis

Prof. Annamaria Bianchi
A.Y. 2024/2025

Course Introduction

UNIVERSITÀ DEGLI STUDI DI BERGAMO | Dipartimento di Scienze Economiche

Syllabus

Instructor: Annamaria Bianchi

Instructor email: annamaria.bianchi@unibg.it

Office Hours: during the course: every Wednesday 14.30-15.30 (office 202)

Tutor: Jurgena Myftiu

Lectures: 4 hours a week:

Monday	10.30-12.30
Tuesday	14.30-16.30

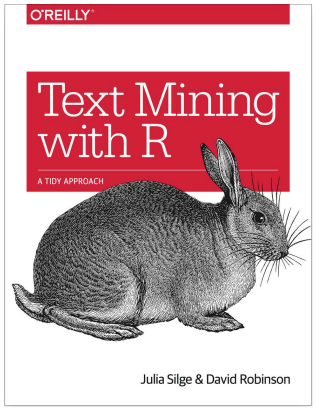
Textbook

Title: Text Mining with R

Authors: Silge, J. & Robinson, D.

Editor: O'Reilly

Link: https://www.tidytextmining.com/index.html

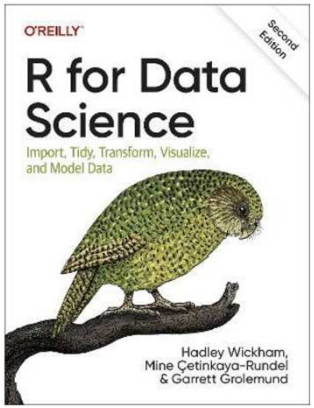


Title: R for Data Science, Second Edition

Authors: Wickham, H., Cetinkaya-Rundel, M. & Grolemund, G.

Editor: O'Reilly

Link: https://r4ds.hadley.nz/



References

+ Lecture notes
+ E-learning course with lecture notes and other relevant course material

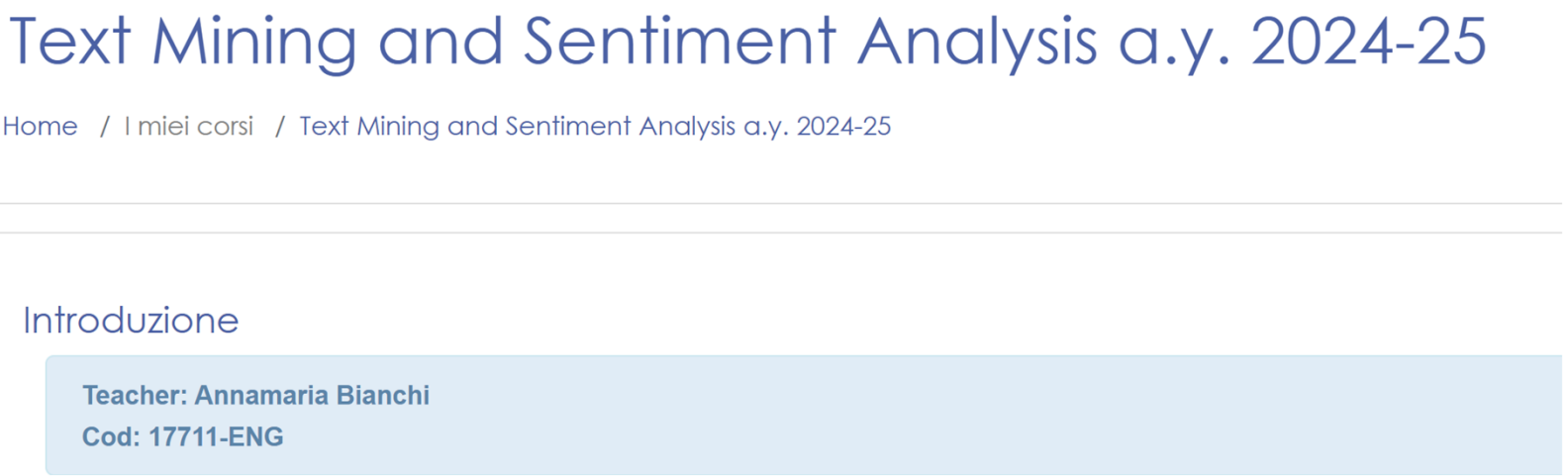
+ Supporting material on a lecture-by-lecture basis

  Links & online materials, pdf files, books references, …

Class web-page

On the E-learning platform you can access lecture notes and other relevant course material as well as important announcements

https://elearning15.unibg.it/enrol/index.php?id=6314



On the first access only, you need to insert the following keyword: **Bianchi5611**

What is Text Mining?

Text Mining is an Artificial Intelligence (AI) technique that uses natural language processing (NLP) to transform the free, unstructured text of documents / databases such as web pages, newspaper articles, e-mails, press, post / comments on social media etc. in structured and normalized data. The final aim is to obtain valuable insights from texts.

The main purposes of Text Mining are:

- identify the main thematic groups
- find similarities among documents
- classify documents into predefined categories
- discover hidden associations (links between topics, or between authors, temporal trends, ...)
- extract specific information (ex: names of certain peeple, names of companies, ...)
- train search engines
- ...

What is Sentiment Analysis?

Sentiment analysis (SA) is the process of extracting an author's emotional intent from text

Final aim: classify a sort of polarity that subjects demonstrate towards a topic by creating an index that associates numbers from "completely positive" to "completely negative", passing through intermediate values which reflect the neutrality that a subject assumes with respect to the theme.

Born in the **marketing field** to evaluate the reputation of a certain brand in a shorter time than the classic market research, it is **now applied in various fields** ranging from political vision to the emotional state that generates a social phenomenon up to being supportive in forecasts of monetary economic policy.

Applications. SA supportive in the purchase of goods (the comments / reviews on sites such as Amazon and e-bay), of tourist services (travel suggestions derived from the reviews of Tripadvisor, Booking, etc.), of cultural services (evaluations of universities and / or other types of schools), of personal services (choice of a doctor or an entire medical facility based on feedback obtained from specific blogs).

Recently used by Istat and other National Statistical Institutes to produce 'official' experimental statistics and by the Bank of Italy and other European (and non-European) Central Banks to "predict" the indicators of monetary policy stability so far studied through complex econometric models.

Why analyse text?

We should care about textual information for a variety of reasons:

- Online content from organizations, their competitors and outside sources, such as blogs, continuous to grow
- Social media continues to evolve and affect the public and organizations' public efforts
- The digitilization of formerly paper records is occurring in many industries
- New technologies like automatic audio transcriptions are helping to capture customer touchpoints
- As textual sources grow in quantity, complexity and number, the concurrent advance in processing power and storage has translated to vast amounts of text being stored.

It is illogical for organizations to study only structured information while all these precious resources are available in the form of unstructured natural language. Text represents an untapped input that can further increase knowledge and competitive advantage.

Computation provides access to information in texts that we simply cannot gather using our traditionally qualitative methods of close reading and human synthesis.

Why analyse text?

The **alternative** to text mining may mean ignoring text sources or merely sampling and manually reviewing text.

Consequences of ignoring text:

- Rigorous scientific and analytical exploration requires investigating all sources of information that can explain phenomena
- Not performing text mining may lead an analysis to a false outcome
- Some problems are almost entirely text-based, so not using these methods would mean significant reduction in effectiveness or even not being able to perform the analysis.

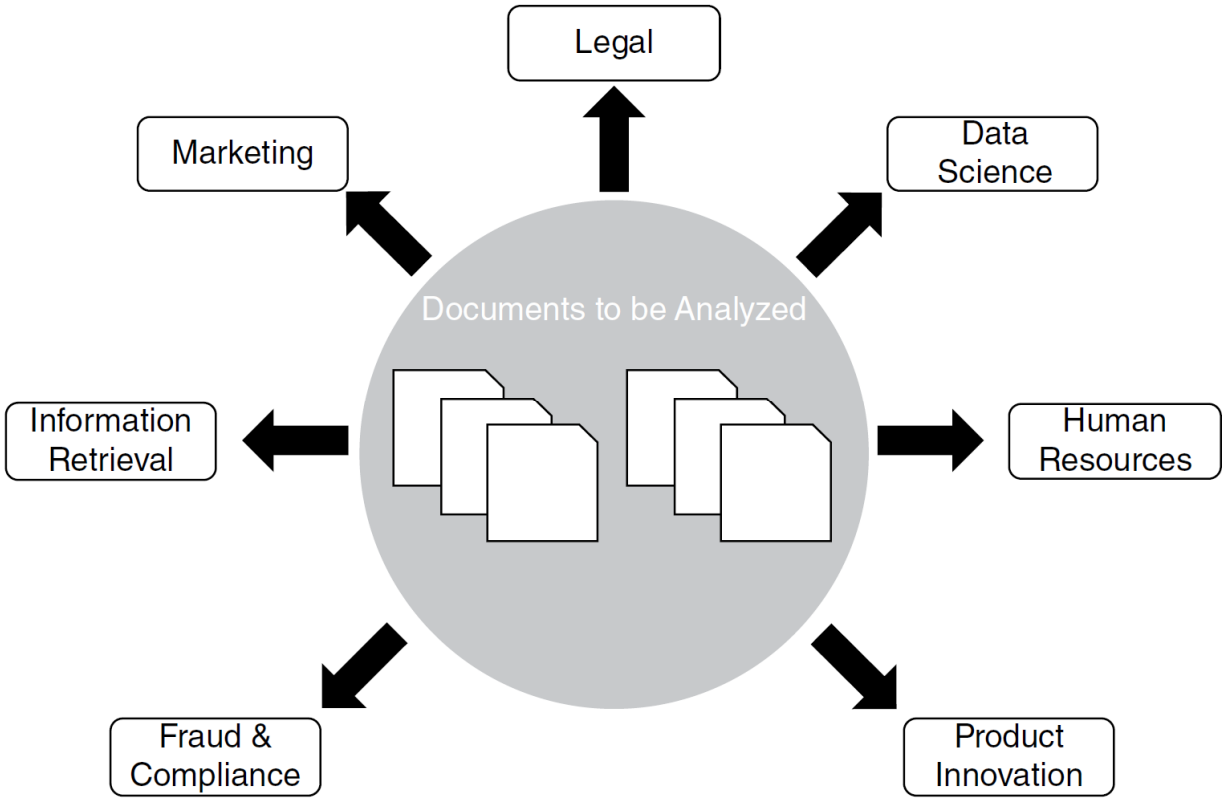
It is not always appropriate to use text for analytics, but if the problem being investigated has a text component, and resource constraints do not forbit it, the ignoring text is not suitable.

Why analyse text?

Potential applications are, automatically:

- Sentiment analysis of messages
- Detailed product identification from descriptions on web sites
- Identify activity code from descriptions on web sites
- Code description of jobs/educations/products
- Classify answers to open questions
- Classify cause of death from medical reports
- …

Why analyse text?



Some possible enterprise uses of text mining

Are there connections with Machine Learning?

Machine learning is an artificial intelligence (AI) technology which provides systems with the ability to automatically learn from experience without the need for explicit programming, and can help solve complex problems with accuracy that can rival or even sometimes surpass humans.

In text mining context, one **can apply machine learning algorithms for textual analysis**, making the text mining process entirely data driven.

Objectives

- Understand what text analytics is.
- Acquire a solid background on text analytics techniques from the theoretical and practical perspective.
- Learn how to convert unstructured text-based character data into structured numeric data.
- Learn how to gain insight from the text, once it has been given a structure.
- Categorize and cluster text to provide economic statistic information.
- Become aware of the pros and cons in the use of unstructured data and devote attention to the quality of the data sources, in a total quality perspective.

Contents

Working with strings

Basics of Text Mining

Pre-processing / cleaning of text data

Common Text Mining Visualization Techniques

Sentiment Analysis

Topic Modeling

String distances

Quality of the data sources

Data extraction from the web (possibly)

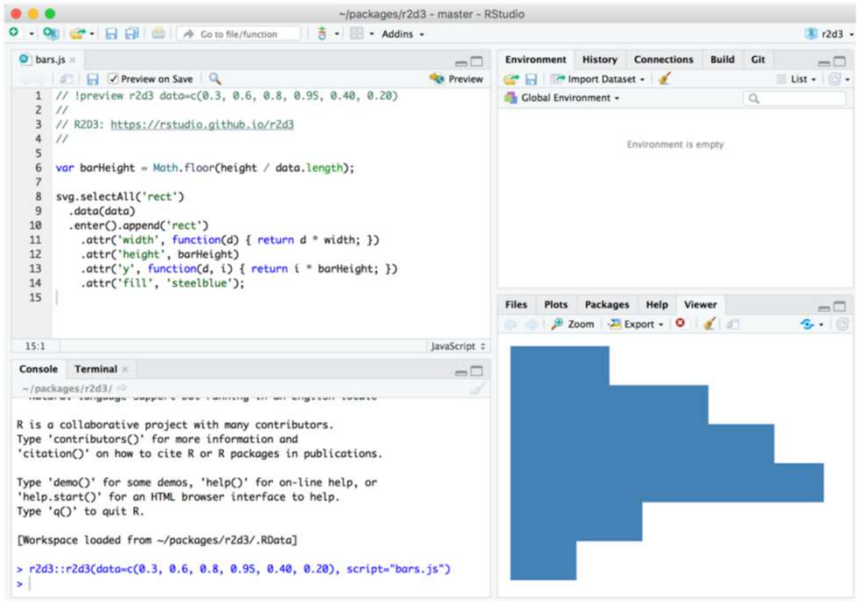
Laboratories using R with real applications

Statistical software

- R (https://www.r-project.org/) is a *free* software environment (actually it is a programming language) for statistical computing, data analysis and visualisation (graphics).
- R is an *open source project*, which means that it depends on a worldwide community of active developers to grow and evolve.
- R compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.
- The **Comprehensive R Archive Network** (CRAN) is a network of web servers around the world that store identical, up-to-date, versions of code and documentation. Currently, the CRAN package repository features about 22052 packages which extend the base release of R (to load, manipulate data, visualise, model data, to report results, for spatial, financial, time series data and much much much more).
- The R community is very active and supportive (see e.g. https://www.r-bloggers.com/). If you have a problem with R, or you want to code something specific, try with a Google search. In 90% of the cases someone has already implemented what you need (and the code is available in the web).

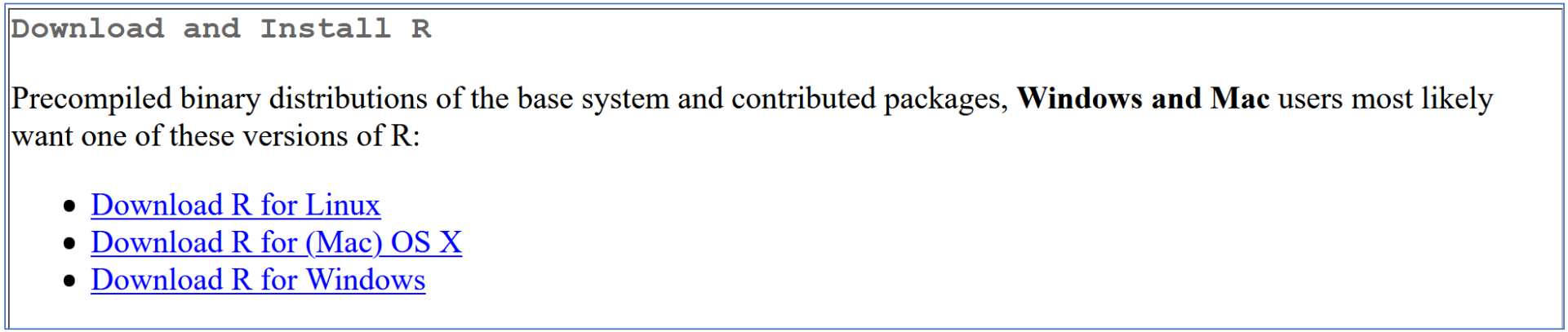
Statistical software

- **RStudio** (https://rstudio.com/products/rstudio/) is a free and open-source *integrated development environment* (IDE) where to run R code.
- To run RStudio you need to install R first. R will do all the calculations for you, while RStudio is just a convenient environment where to work, it's the interface between you and R. So, you can just launch RStudio to use R.



Install the software

1) Install R, go to https://cloud.r-project.org/ and follow the instructions



At the end of the installation you will find an icon (named R) on your desktop and also in the Start Menu.

2) Install RStudio for your own operating system from this link: https://rstudio.com/products/rstudio/download/#download.

Proceed as usual and find the icon on your desktop/dock.

Requirements

No compulsory prerequisite.

Recommended prerequisites: basic knowledge of Statistics (probability, statistical inference). Basic knowledge of the R programming language.

Lecture structure

At the end of each lecture, I will give some exercises/questions that I will correct during the next lecture.

I **strongly suggest** that you try to do these exercises by yourself before the next lecture

Evaluation

Written test consisting of theoretical questions and exercises to be solved using R.

Attendee students, who have regularly (and on time) submitted ALL the assignments, can get a maximum of 3 extra points on the grade of the final exam based on the quality of their work.

Questions?