Workshop: Working with unconventional data using R

Working with text and a little bit of maps

Valeria Rueda

School of Economics, University of Nottingham

[valeria.rueda@nottingham.ac.uk](mailto:valeria.rueda@nottingham.ac.uk)

**This version: 22/03/2022**

# Learning goals

At the end of the session, you should…

* Have jumped in the R pool
* Familiarized yourself with regular expressions and their use
* Created one map with ggplot
* Getting started with text pre-processing for text analysis: stemming, cleaning
* Getting started with TF-IDF Representations

# Structure (subject to change)

1. Regular Expressions
   1. Introduction to regular expressions
   2. Playing with regular expressions
   3. Making a map
2. Pre-processing a corpus of text
   1. Cleaning
   2. Stop words, stemming and tokenizing
3. Representing text as data 1
   1. TF-IDF representations: theory
   2. One application using ECB speeches
4. Regressions with text
   1. A brief ML glossary: supervised vs unsupervised models; labelled vs unlabelled data
   2. Feature selection and classification using a LASSO
5. Further topics:
   1. Dictionary methods
   2. Word embeddings

This is a modular class, and we will likely not cover everything. You will get the Jupiter Notebook for further reference and the data to run all the tutorials.

# Prerequisites

## Materials

**Laptops with a working version of R installed**. To work with R, RStudio or Jupyter are recommended. Rstudio may be the easiest to sort out.

Make sure you have tested your version of R and basic commands, such as read.csv(), are running before the lecture[. This video](https://youtu.be/Eq8Xnueb-50) presents a tutorial on how to do load a dataset in R.

To install a package, run the command: install.package(“package”). For instance: install.package(“tidyverse”).

You should also install the following packages before the class: tidyverse, lubridate, reshape2, ggmap, sf`, tidytext, stopwords, SnowBallC, glmnet, gamlr, topicmodels, textdata, ranger.

In other words, the following script should be running in your machines:

## #General data handling

## library(tidyverse)

## library(lubridate)

## library(reshape2)

## #Maps

## library(ggmap)

## library(sf)

## sf::sf\_use\_s2(FALSE) ## s2 in sf version 1.0 slows down the code too much

## #Text analysis

## library(tidytext)

## library(stopwords)

## library(SnowballC)

## library(topicmodels)

## library(textdata)

## # ML

## library(ranger) #Random Forests

## library(glmnet) #LASSO

## library(gamlr) #LASSO choice lambda AIC

## Prior knowledge

This is a hands-on class: it is structured around activities to be done by students during the session. We will discuss all challenges faced.

NO prior knowledge of R is required. This class can be taken as an opportunity to try R for the first time.

Why R?: Text analysis is more commonly done using Python. However, R has a slightly lower entry cost, and perhaps has more applications in economics. Perhaps there are then more gains to learning R from scratch than Python from the average economists. More and more packages for text analysis are being developed for R. If you can, learn both.

## Useful References

### General references for learning the methods

Bholat, Hansen, Santos, and Schonhardt-Bailey (2015), “Text Mining for Central Banks”, *Centre for Central Banking Studies,* Bank of England Report ISSN: 1756-7270.

Gentzkow, Kelly, and Taddy (2019), “Text as Data”, *Journal of Economic Literature,* 57(3):535-574

Grimmer and Stewart (2013), “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts”, *Political Analysis,* 21:267-297

Hvitfeldt and Silge (2021), “[Supervised Machine Learning for Text Analysis in R](https://smltar.com/)”, Open-Source ebook.

Silge and Robinson (accessed 2022), “[Text Mining with R: A Tidy Approach](https://www.tidytextmining.com/index.html)”, Open-Source ebook

### Advanced or more specific

Loughran and Bill McDonald, 2011, When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks, Journal of Finance, 66:1, 35-65.

[Documentation for Loughran and Bill McDonald, 2011](https://sraf.nd.edu/loughranmcdonald-master-dictionary/): Available at the University of Notre Dame Software Repository for Accounting and finance

Tim Loughran and Bill McDonald, 2016, Textual Analysis in Accounting and Finance: A Survey, *Journal of Accounting Research*, 54:4,1187-1230.

Stanford Graduate School of Business course on Youtube: “[Machine Learning and Causal Inference](https://youtu.be/Z0ZcsxI-HTs)”