18P033
Text Mining and Web Scraping for Economics in R

Winter Term - 0 ECTS

Prof. Bruno Conte

Prerequisites to Enroll

Possess a laptop computer with the required software installed (R + RStudio).

Overview and Objectives

The objective of this course is to give students the computational tools to retrieve, manage and reshape data originally structured in text format. Its focus will be on how to deal with this type of data, obtained either from text files or the web, and how to manipulate it so to be used in practical applications in economic research.

Course Outline

Part I. Introduction to R and Text data

Objectives: Make students familiarized with R language. Covering: creation of vectors, lists, datasets; string variables, pasting, cutting, searching text patterns in strings. Texts as factors for memory efficiency. Appending/merging datasets with text data. Exporting datasets to .csv/.dta formats.

Part II. Text Mining

Objectives: Put students in contact with most common text mining tools. Covering: importing text from text and pdf files, Optical Character Recognition in R; text cleaning, search of words/patterns, splitting texts by patterns, reshaping text data as time series, creating word clouds; efficient ways of storing data with text.

Part III. Web Scraping

Objectives: Allow students to retrieve text data from the web instead of directly from text/pdf files. Rvest package from R for more complex/refined routines. Definition of html sessions, targeting web objects with x-patches, filling formularies out and retrieving outcome, creating datasets from web data for applications in Economics.

18P033
Text Mining and Web Scraping for Economics in R

Winter Term - 0 ECTS

Prof. Bruno Conte

Required Activities

Sessions are going to be practical in a follow-up framework. Students are going to be required to bring their own computer, follow the illustrative examples and do the “hands-on” exercises in class.

Competences

☒To (be able to) communicate with determination and in the English Language, the results and implications of the required analytical study using a language that the receiver can relate to.

☒To work within a heterogeneous team of researchers as economic analyst using specific group techniques.

☒That students know how to apply the acquired knowledge and their ability to solve problems in new or unfamiliar environments within broader (or multidisciplinary) contexts related to their field of study.

☒That the students be able to communicate their conclusions and the knowledge and the ultimate reasons that sustain them to both, specialized and non-specialized publics in a clear and unambiguous way.

☒That students possess the learning skills that allow them to continue studying in a way that will be largely self-directed or autonomous.

Learning Outcomes

☒ Applies the empirical tools of economic analysis to evaluate public policies.

Useful References

Peng, Roger D. “R programming for data science.” Lulu. com (2015).

Paradis, Emmanuel. "R for Beginners." (2002).

Sanchez, Gaston. “Handling and processing strings in R”. Berkeley: Trowchez Editions (2013).