Università di Pisa

DESTEC
Dipartimento di Ingegneria
dell'Energia, dei Sistemi, del Territorio e delle Costruzioni

Natural Language Processing with R: Basics and Applications

held by

Vito Giordano, Ph.D.
Assistant Professor in Engineering Management

Course Description

In today's digital age, we are surrounded by vast amounts of data, and a significant portion of this data is unstructured text. From social media posts and customer reviews to scientific papers and legal documents, text data holds immense potential for extracting valuable insights and knowledge. The ability to analyze and understand text data has become increasingly important in various domains, including research, business, and everyday life. Natural Language Processing, a subfield of Artificial Intelligence, focuses on extracting meaningful information from unstructured text. It involves techniques that enable us to process, analyze, and interpret text data, thereby unlocking its potential value. The growth of generative artificial intelligence models, such as ChatGPT, has further emphasized the importance of NLP and our need to comprehend and harness this technology.

This course provides an introduction to Natural Language Processing (NLP) using the R programming language. The course covers the foundational concepts of NLP, including data analysis, text wrangling, and visualization using the tidyverse framework. PhD Students will gain hands-on experience with various NLP techniques such as vectorization, bag of words, tf-idf, and topic modeling using the latent Dirichlet allocation (LDA) algorithm. Additionally, PhD students will explore a case study focused on analyzing scientific papers or patent textual data using NLP techniques.

Course Duration: 10 hours

Course Outline

Topic1: Introduction to Data Analysis with the Tidyverse (3 hours)

- Introduction to the tidyverse framework
- Exploratory data analysis with dplyr and ggplot2

- Data visualization and summarization techniques

Topic 2: Text Wrangling and Preprocessing (1 hours)

- Introduction to text data and its peculiarities
- Text cleaning and preprocessing techniques
- Regular expressions for text manipulation

Topic 3: Text Vectorization (1 hours)

- Introduction to text vectorization
- Bag of words representation
- Term Frequency-Inverse Document Frequency (tf-idf)
- Theory on Non-Contextual and Contexutal Embeddings

Topic 4: Introduction to NLP Algorithms (3 hours)

- Sentiment analysis and text classification
- Named Entity Recognition (NER) and Part-of-Speech (POS) tagging
- Introduction to topic modeling and Latent Dirichlet Allocation (LDA)

Topic 5: Case Study - Analyzing Scientific Papers or Patent Textual Data (2 hours)

- Introduction to the case study
- Data acquisition and preparation
- Applying NLP techniques to extract insights from scientific papers or patent data
- Presenting and visualizing the results of the case study

Note: The course outline is subject to change based on the pace of learning and the needs of the students.

CALENDAR:

Lesson 1 - 31st January 2024, 15:00 – 18:00 (3 hours)
Lesson 2 - 1st February 2024, 16:00 – 18:00 (2 hours)
Lesson 3 - 7tht February 2024, 15:00 – 18:00 (3 hours)
Lesson 4 - 8tht February 2024, 16:00 – 18:00 (2 hours)

Note: The course dates are subject to change based on the specific requirements and scheduling of the teacher.

PLACE: Sala Poggi, DESTeC
Note: The course place is subject to change based on the specific availability of classrooms.

RESOURCES:

- Textbook 1: "R for Data Science" by Garrett Grolemund and Hadley Wickham
- Textbook 2: "Text Mining with R: A Tidy Approach" by Julia Silge and David Robinson
- Online tutorials and documentation
- Research papers and articles on NLP advancements and applications