Syllabus

Department of Economics and Business
CENTRAL EUROPEAN UNIVERSITY

Coding 2: Python Programming, Visualization and Text Analysis (Finance)

Instructor: Arieda Muço

- **Credits:** 2 (4 ECTS)
- **Term:** Fall 2019-2020
- **Teaching Assistant:** Marton Fleck
- **Course level:** MS in Finance

Course description

This course will teach how to perform visualization and extract meaningful data and information from text and other shapes of data such as mp3, png, pdf, excel, csv, etc. It will introduce the student to the basics of Python programming and manipulation and mining of text. It will cover the most commonly used tools for data visualization and information retrieval. We will begin with a general introduction of the Python framework and an understanding of how text is handled by Python. Then cover the use of regular expressions, cleaning and preparing data and text for use. We will then apply basic natural language processing methods and machine learning tools to demonstrate how (text) classification is performed. We will also explore methods for sentiment analysis, topic detection and modelling.

Learning outcomes

By the end of the course students should have a firm grasp of performing analysis on data generated from text. Also, students will know how to perform basic string processing, identify names, dates, and locations, tokenize text, convert text and words to vectors, identify the grammatical parts of a sentence, etc.

Course Requirements

Students should attend class, submit quizzes, read assigned material, and solve problem sets.

Assessment

Final assessment will consist of the following:

1. Problem Sets (40% of final grade).
2. Individual Project (40% of final grade)

   An individual project is required to pass the course.

3. Quizzes (20% of final grade). Students will be asked to solve a quick programming exercise, in class, on topics covered previously.

Late submissions will not be graded.

Topics

We will attempt to cover the following topics if time permits.

1. Variables, Strings and Numbers, Conditionals and Functions, Sequences, Collections and Iterations, Lists, Dictionaries, NumPy and Operations, Pandas, Series, Regular Expressions
2. Data Visualization and Maps
3. Natural Language Toolkit (NLTK), Description of text mining and data processing techniques such as Normalization, Tokenization, Lemmatization, Stemming, Parts of Speech (POS), Named Entity Recognition (NER)
4. Term Weighting and Document Classification, Bag of Words (BOW), Term Frequency, Inverse Document Frequency, Term Frequency - Inverse Frequency (TF-IDF)
5. Supervised Learning Models such as Naive Bayes, Logistic Regression, Support Vector Machines, Sentiment Analysis
6. Unsupervised Learning and Topic Modelling such as Latent Dirichlet allocation (LDA), Distributed Multinomial Regression (DMR), Cluster Analysis

Course material

- A Shaw, Zed. "Learn Python the hard way."
- VanderPlas, Jake. "Python data science handbook: essential tools for working with data. " O'Reilly Media, Inc."
- D. Manning, Christopher, Raghavan, Prabhakar and Schütze Hinrich. "Introduction to Information Retrieval"
- Jurafsky, Dan and H. Martin, James. "Speech and Language Processing"
- Guido, Sarah and Müller, Andreas. "Introduction to Machine Learning with Python: A Guide for Data Scientists"

Technical/laptop requirements (Students are required to bring their own laptop in class.)

Short bio of the instructor Arieda Muço is an Assistant Professor of Economics at the Department of Economics and Business at the CEU. Arieda received her Ph.D. in Economics from the Stockholm School of

Economics in 2017. During her Ph.D. studies, Arieda was a visiting researcher at the Centro de Estudios Monetarios y Financieros (CEMFI) in Madrid. Her research lies in the intersection of Political and Development Economics. In her research she combines natural language processing, machine learning, and causal inference methods. Currently, her research focuses on tools to fight corruption and inefficiencies in public administrations such as motivating whistleblowers, providing information to voters, and the interplay between campaign finance and state capture.