BTS | Barcelona Technology School

Data Science Foundations Curriculum

Master in Big Data Solutions

1 Subject description

Subject: Data Science Foundations
Year: 2020-21
Quarter: 1st
Degree: Master in Big Data Solutions
Number of credits: 7 **/ Hours of class:** 64 **/ Hours of homework:** 64

Teaching Staff:

Name: Victor Pajuelo Madrigal
Email: victor.pajuelo@bts.tech
Linkedin: https://www.linkedin.com/in/victor-pm/

Schedule:

- 4-hour sessions, two days per week, from 9:00 – 13:00
- Weekly home work: 4 - 8 hours, approximately

2 Subject introduction and goals

We live in a data-centric world, with more than 2500 petabytes of data generated every day and growing, as the cost of data collection and storage goes down. However, knowledge requires applying intelligence to that information, and the techniques to process this sheer amount of data are increasingly complex, and require a mix between a disciplined, mathematical approach with an important dose of creativity and multidisciplinarity.

By the end of this course, students will have acquired the following skills:

- Understand the three main blocks of a data science pipeline: ETL, analysis and production
- Manage data using Python:
  - Perform basic operations with data using pandas DataFrames and NumPy arrays.
  - Learn the basics of relational databases and how to write Structured Query Language (SQL) queries to retrieve and manipulate the data
  - Interact with a database using Python
- Perform time series analysis and forecasting in Python
- Learn how to fit data to probability distributions and regression models
- Use Python/pandas to perform general statistical analysis of the data
- Learn the differences between prototype, project and product in Data Science
- Learn how to deploy and put your analysis into production

3 Teaching methodology

The course is divided in three parts: "Practical Data Manipulation in Python", "Basic Data Analysis" and "Data Science in Production". In all the sessions we will use interactive materials (Jupyter notebooks) that blend theory and explanations with code and visualizations. For each session the students will be asked to complete an individual assignment, similar to what was seen in class, so they can practice the concepts at home. Besides, at the end of each part, group assignments will be proposed that will let the students practise their communication and organization skills.

The course is meant to teach and introduce the basics of a Data Science workflow, starting from Data Ingestion & Manipulation, Data Analysis and Model Inference/Production approaches. The teaching mindset is with practicality in mind, being able to have a quick prototype of a basic workflow that can serve as a proof-of-concept for an investment round pitch or the basic start of a bigger endeavour.

Apart from the technical assignments, occasionally students will be invited to participate in local meetups, discuss their impressions of the Python ecosystem, comment new releases and versions of the software in use, and collaborate with some open source projects.

4 Contents

4.1 Part I: Practical Data Manipulation in Python

After a quick review over the structure of a Data Science workflow the course will dive into data manipulation. The practical data manipulation means common actions that should be performed with data prior to further analysis (descriptive, predictive or prescriptive one). Such actions involve cleaning, filtering, joining, etc. Special attention is given to the unstructured data that should be converted into a structured (tabular) data in order to enable the knowledge extraction.

1. Basic data operations on relational data
    a. Recap of input/output and on-disk formats
    b. Cleaning noisy data, normalizing
    c. Filtering rows and columns
    d. Joining data from multiple sources
    e. Split-apply-combine workflows: groupby
2. Loading and processing images and text
    a. Image loading, pre-processing and filtering
    b. Image pre-processing for object detection and segmentation
    c. Text pre-processing, normalization, stemming, stopword removal
    d. Converting text to vectors and computing text similarity
3. Date manipulation
    a. Basic data types, timezones
    b. Resampling, shifting and windowing

4.2 Part II: Basic Data Analysis

After a review of all the fundamental tools, we will study a structured, simplified version of a data analysis process with clearly identified steps, we will apply them to several different data types, and we will finish with an overview of analytical techniques.

4. The process of data analysis
    a. Exploratory data analysis and insights finding
    b. Feature engineering
5. Data types
    a. Image processing
    b. Time series analysis
    c. Text processing
6. Overview of main analytic techniques
    a. Regression on time-series, text and images
    b. Classification on text and images

4.3 Part III: Data Science in production

After a quick review of Data Manipulation and Analysis, the last part of the course will address the production stage of a Data Science workflow. We will learn to build a small prototype that can be used as a proof-of-concept. For this we will explore methods to automated ETL, package our analytics model and deliver it. More often than not, Data Science forgets about this last step, which is the last mile between actionable information and the customer/user.

7. The process of ETL (Extract Transform Load)
    a. Data gathering and scraping
8. Packaging your model
    a. Model evaluation and deployment
    b. Containerize the model
9. The last mile to user: visualization and delivery
    a. Visualization and reporting
    b. Common delivery platforms

5 Schedule of contents and activities

Session	Activity at class (4 hours)	Activity at home (4 hours)
PART I. Practical Data Manipulation in Python		
Session 1 6-Oct	**Content** Quick introduction to Data Science workflows. Recap of common data formats: CSV, JSON and parquet. **Activity** Load data in CSV, JSON and parquet formats into pandas DataFrames. • Basic manipulation of data, statistical summary Commit scripts to git.	**Individual Assignment** Select an open source data source in CSV, JSON or parquet format and: • Load data into a pandas DataFrame • Answer a series of questions related to the data • Commit scripts to git
Session 2 8-oct	**Content** How to perform cleaning, filtering and joining of data using Pandas DataFrames. **Activity** Download city bike stations or weather data (JSON format): • Clean and store data • Analyse an hour of city bike data Commit scripts to Git.	**Individual Assignment** Given city bike stations or weather data, answer a series of questions and solve challenges. Commit scripts to Git.
Session 3 20-oct	**Content** Recap of SQL. How to work with a SQLite database. **Activity** Create SQLite database and: • Import data from CSV or parquet • Load data from database to pandas DataFrame and perform basic manipulations Commit scripts to git.	**Individual Assignment** Select an open source data source in CSV, JSON or parquet format and: • Create a SQLite database • Import data into SQLite database • Load data from database to pandas DataFrame • Answer a series of questions related to the data • Commit scripts to git

Session	Activity at class (4 hours)	Activity at home (4 hours)
Session 4 22-oct	**Content** Introduction to image processing in Python and to the scikit-image library. **Activity** Loading image from file, Perform some transformations and image descriptors over it. Commit scripts to Git.	**Individual Assignment** Get data that from public image libraries, perform image loading, filtering and easy classification with the scikit-image library. Commit scripts to Git.
Session 5 27-oct	**Content** Image pipelines for object detection and segmentation using scikit-image **Activity** Detect objects in an image using different approaches seen in class. Commit scripts to Git.	**Individual Assignment** Perform object detection and segmentation using scikit-image through an image captured with your own phone device. Commit scripts to Git.
Session 6 29-oct	**Content** Introduction to text processing in Python and to the spaCy library. **Activity** Loading text from file, sentence tokenize and word tokenize, normalizations, and part-of-speech tagging. Commit scripts to Git.	**Individual Assignment** Get data that is not raw text (JSON or CSV formats), perform text loading, filtering and pre-processing with the spaCy library. Commit scripts to Git.
Session 7 3-Nov	**Content** Text stopping, stemming and lemmatization. TF-IDF and Basic metrics for text similarity (Euclidean distance, cosine similarity). **Activity** Load text from a file, and perform text stopping, stemming and lemmatization with the spaCy library. TF-IDF with the scikit-learn library. Load text from a file, and perform text similarity with spaCy. Commit scripts to Git.	**Individual Assignment** Take a different text file, perform text stopping, stemming and lemmatization with the spaCy library and produce some basic statistics of how the text changed after these operations (e.g., number of final tokens w.r.t. the initial ones). Perform TF-IDF on the text. Take two text files and detect plagiarism (i.e., how similar the two documents are). Commit scripts to Git

Session	Activity at class (4 hours)	Activity at home (4 hours)
Session 8 5-Nov	**Content** Basic time datatypes (date, datetime, timedelta), scalar and vector dates, offsets, timezones. Shifting, windowing, resampling. **Activity** Comparison of Python, numpy and pandas time datatypes, date manipulation on a CSV or parquet dataset. Advanced time data manipulation on a structured dataset using pandas and timezone helpers. Commit scripts to Git.	**Group Assignment** Application of studied techniques to a Kaggle or any other public dataset. Commit scripts to Git.
Part II. Basic Data Analysis		
Session 9 12-Nov	**Content** Exploratory data analysis and insights finding **Activity** Exploratory analysis of a common dataset, visualization of interesting features, insights finding. Commit scripts to Git.	**Individual Assignment** Exercises on different data sources. Commit scripts to Git.
Session 10 17-Nov	**Content** Feature engineering, feature extraction. **Activity** Application of manipulation techniques using creative and brute force strategies to augment a dataset. Commit scripts to Git.	**Individual Assignment** Exercises on different data sources. Commit scripts to Git.

Session	Activity at class (4 hours)	Activity at home (4 hours)
Session 11 19-nov	**Content** Time Series introduction and time series forecasting **Activity** Loading and handling time series in Pandas. Check stationarity of a time series. How to make a time series stationary. Use of pandas, statsmodels, pyramid and prophet for time series forecasting. Commit scripts to Git.	**Individual Assignment** Exercises on data series handling and stationarity with different sources of data. Time series forecasting with a different source of data. Commit scripts to Git.
Session 12 24-nov	**Content** Main analytics techniques on text, images and time series. Regression vs. Classification. Text classification and Image classification. **Activity** Using Stanford Text Analysis Tools in Python. Preliminary study on text classification. Using scikit-learn for preliminary image classification. Commit scripts to Git.	**Individual Assignment** Text and image regression and classification exercises. Commit scripts to Git.

Session	Activity at class (4 hours)	Activity at home (4 hours)
PART III. Data Science in Production ***This part will combine a final group assignment and smaller individual assignments.***		
Session 13 26-nov	**Content** Data gathering and scraping, APIs, REST, authentication. **Activity** Use of requests, lxml and beautifulsoup to read data from external APIs, and xlwings and others to read from Excel files. Commit scripts to Git.	**Group Assignment (delivery on last session)** Choose a problem from the ones presented by the instructor and create a prototype workflow to pitch an idea. Use of several analytical techniques on text, image or time series data. Write a report describing all the steps in the data science process, from data gathering to deployment, visualization and delivery. Commit scripts to Git. **Individual Assignment** Data retrieval from public API from Spanish Open Data portals. Commit scripts to Git.
Session 14 1-dec	**Content** Model evaluation, model deployment, principles of software operations and automation. Container and container orchestration principles. **Activity** Transformation of a notebook containing a data analysis pipeline into a series of files, unit testing, metric reporting automation. Commit scripts to Git.	**Individual Assignment** Unit tests and code refactoring on past assignments. Commit scripts to Git.
Session 15 3-dec	**Content** Advanced visualization and reporting, interactive visualization, principles of color and design. Delivery platforms. **Activity** Interactive visualization using matplotlib and plotly, use of Jupyter interactive widgets, Hans Rosling "Gapminder" example. Commit scripts to Git.	**Individual Assignment** Interactive visualization on different data sources. Explore different delivery platforms. Commit scripts to Git.

Session	Activity at class (4 hours)	Activity at home (4 hours)
Session 16 10-dec	**Content** Recap and presentation of advanced tools for the next sessions of Data Science. **Activity** Group presentations of ideas for startup following data acquisition, preparation, analysis and delivery.	**N/A**

The Schedule of activities can be modified according to the program needs.

6 Qualification system

Participation: 30%
Active participation at class is expected during the subject.

Individual assignments: 40%
Continuous assessment by delivery of individual exercises.

Group assignments: 30%
Continuous assessment by delivery of group exercises and in-class presentations.

The 85% of attendance to each subject is required to pass the Master.
There is no final exam.

Conditions to recover the subject
If the final mark is below 5, students will be able to deliver a small project with an real application of the presented techniques.

7 Bibliography

Basic:

The Python 3 official documentation	https://docs.python.org/3.6/
Think Stats. Probability and Statistics for Programmers **(book)**	https://greenteapress.com/thinkstats/
Python Data Science Handbook **(book)**	https://jakevdp.github.io/PythonDataScienceHandbook/
Data Science for Business **(book)**	http://www.data-science-for-biz.com/DSB/Home.html

Complementary:

Kaggle	https://www.kaggle.com
Spanish Open Data Portal	https://datos.gob.es/
Open Data Barcelona	http://opendata-ajuntament.barcelona.cat/en/

Other non-academic resources:

Barcelona Python Meetup	https://www.meetup.com/python-185/

Bibliography and other academic resources will be detailed and updated in the Campus.