Big Data – Exam Admission Assignment

June 8, 2020

1 Overview

In order to gain admission to the Big Data exam (summer term 2020), you have to submit a group project. The overall goal of this project is to analyze some real-world data using Apache Spark. You will have to write a program using Apache Spark and explain your approach.

Submissions are to be submitted via OLAT – an option to do so is coming soon. The submission deadline is **June 30th, 2020**. Your project will then be reviewed. If your project does not meet the requirements to pass, you will have another two weeks to improve and to resubmit it.

The following sections provide you with further information regarding the task itself, the associated rules and guidelines, and some examples.

Disclaimer: If you do not pass this assignment, you are not eligible to take the exam. The only exception is if you failed last year's exam and you have to retake it this year. If this is the case, you are allowed to participate anyway.

2 Problem Definition

Overall, your project should include the following steps:

1. Import your dataset to Spark.
2. Preprocess the data
3. Perform some meaningful analysis (e.g., using machine learning methods). It is also part of this task to get acquainted with subjects that were not covered in the lecture or that you are not familiar with.
4. Present/visualize your results.
5. Describe your approach.

2.1 Programming Part

- Submit all of your code.
- Your code must actually run.
- Submit a video (screen cast) which documents that your project works. In the audio track, you are supposed to explain what you did. Upload the video to a provider of your choice (such as NextCloud), and only submit a working link.

2.2 Guidelines

- Group size: up to 5 people
  - Write the name and matriculation number of each group member as a comment at the beginning of your source code.
- Language: English
- The deadline will not be extended.
- Do not copy somebody else's work! Utilizing small snippets of code from the internet is acceptable, but the overall work must be your own.
- We will only accept submissions via OLAT. (Thus, no submissions via email, etc.)
- Again: do not upload your video to OLAT, directly. Upload it to NextCloud, Google Drive, etc. and only submit a link. Make sure that the video is accessible for external viewers.

3 Examples

In this section, we present some examples of tasks you could implement. However, you can also come up with a project yourself. The only requirement here is that your idea must be approved by one of us. So, if you have an idea, just send us an email and describe it.

3.1 Example Projects

3.1.1 Example 1: Change in CO_2 Emission

The aim is to analyze the change in CO_2 emissions from 2004 to 2014. Perform the following tasks:

1. Import the data to Spark.
2. Preprocess the data: handle null values.
3. Compute the change.
4. Apply k-means on the change. Choose a reasonable amount of clusters.
5. Visualize the data, for instance as in Fig. 1. You can use GeoPandas[1] for this.

3.1.2 Example 2: COVID-19 Development in Australia

Analyze the development of confirmed COVID-19 infections[2] from 2/1/20 to 3/21/20 in Australia and predict possible infection rates for 3/22/20 and 3/23/20. Compare your predictions with the actual infection rates.

1. Import the data to Spark.

[1] `https://geopandas.org/gallery/plotting_with_geoplot.html`

[2] `https://github.com/CSSEGISandData/COVID-19/blob/master/archived_data/archived_time_series/time_series_19-covid-Confirmed_archived_0325.csv`

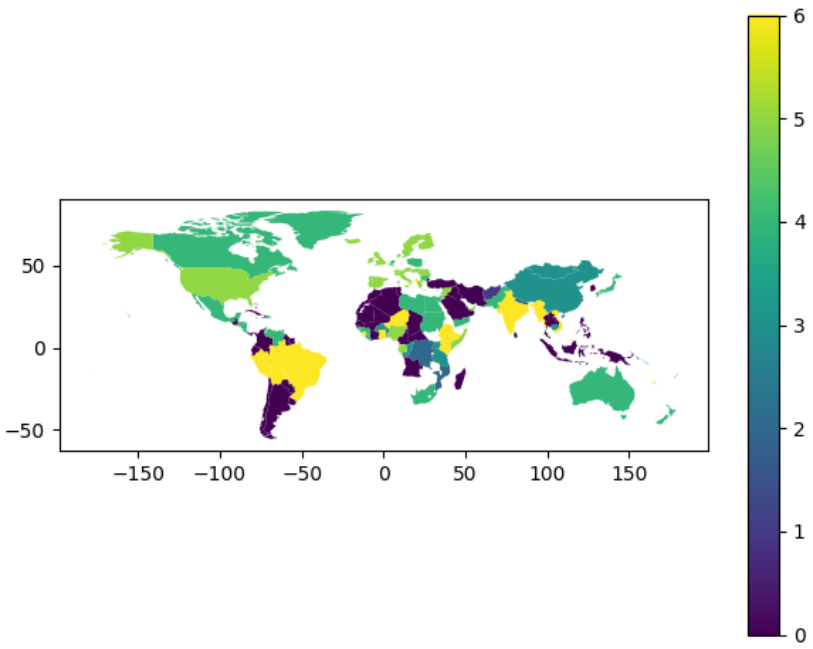

Figure 1: Visualization of the change in CO_2 emission.

2. Preprocess the data:
   (a) Handle null values
   (b) Select only the relevant rows (i.e., those that include data about Australia)
   (c) We want to consider the entire country, so you should merge all states.
3. Take the number of confirmed infections as well as the corresponding point in time (the first day could be encoded as 0, the second as 1, and so forth) and apply a regression method (e.g., linear regression[3]) to predict further values.
4. Compute the difference to the actual values for 3/22/20, and 3/23/20, respectively.

3.2 Datasets

Here are some examples of datasets you could use. Nevertheless, there are many more, so feel free to look for another one yourself.

- CO_2 Emissions[4]
- BBC Datasets[5]
- COVID-19 data[6][7]
- Twitter data

[3] https://spark.apache.org/docs/2.2.0/ml-classification-regression.html#linear-regression
[4] https://data.worldbank.org/indicator/EN.ATM.CO2E.PC
[5] http://mlg.ucd.ie/datasets/bbc.html
[6] https://github.com/CSSEGISandData/COVID-19
[7] https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases