Coursework for COMP4107 Big Data 22-23

Overview

The coursework will be in the form of a research-like project in the field of Big Data and Machine learning with Apache Spark. The aim of it is to deal with Big Data analysis applications by research and then implement and optimise data analysis problems according to real scenarios.

The coursework consists of three key parts: 1) main submission: coursework report and data repository, 2) Individual report, 3) presentation.

Please read this document carefully to see project titles/explanations and further details.

Important dates

- Project allocation: 10th April 2023
- Final submission deadline: 9th May 2023
- Presentations: 10th May 2023

Copying Code and Plagiarism

You may freely copy and adapt any of the code samples provided in the lab exercises or lectures. You may freely copy code samples from the Spark documentation, which have many examples explaining how to do specific tasks. This coursework assumes that you will do so and doing so is a part of the coursework. You are therefore not passing someone else's code off as your own, thus doing so does not count as plagiarism.

You can, and you should look at other code/papers online, but you need to reference any source/material that you have used as inspiration, and highlight what's your contribution. Turnitin/JPlag will detect any use of external sources automatically. Successful completion means that you are able to explain your solution during the presentations. The university takes plagiarism extremely seriously and this can result in getting 0 for the group project, the entire module, or potentially much worse.

Getting Help

You MAY ask the module convenor for help in understanding the **group project requirements** if they are not clear (i.e. what you need to achieve).

Talk to us during the labs, after the lecture, or post your questions on Moodle. Any necessary clarifications will then be added to the Moodle page or posted on the discussion forum so that everyone can see them. You may **NOT get help from anybody else (other than your group mates) to actually do the project (i.e. how to do it)**, including Saeid or the lab helpers.

We might however help each other with technical issues using Apache Spark, but not to solve your particular parallelisation task. Please use the Discussion Forum to ask any questions about the group project description or issues when using Spark.

Project Allocation Instructions

Each project should be completed in groups of at most **four students**. The group contributions on the project will be assessed, but also each individual effort. Each group should select one person as the Team leader. The team leader will be in charge of organising meetings, team coordination, group submissions and communications.

During the lab session in Week 7, you will form your project group and choose your project preferences. You can talk to the convenor for better understanding of the projects and your justification for your preferences. However, you may not get what you preferred if more groups prefer the same project. The group leader should send the module convenor of their group members name and student ids wither their top 3 project preferences by email at the end of week.

Coursework Report (main submission)

The report must be clearly presented in English, **4-6 pages (and must not exceed 8 pages**), including tables, figures, references, and appendixes, in IEEE Computer Society proceedings format as a PDF file. The report templates (in Latex is available on Moodle).

This report should be submitted (with Data repository- see Section C) via Moodle by the due date. The folder should be named by the group id (to be assigned) and project code, e.g., BD01.

The submission should cover the following:

1. Title: A representative name that describes what you have done.
2. Author names: To be included ONLY in the final version.
3. Abstract (maximum 250 words) to briefly outline the objective or problem statement of the project and includes information on the method, research results, and conclusions of the research.
4. Introduction (maximum one a double-column page) presents the background to your study, introduces your topic and aims, gives an overview of the paper the problem you solved and outlines the project contributions.
5. Literature review (maximum two double-column pages) is a survey of scholarly sources (such as books, journal articles, and theses) related to your specific topic and/or research question. This aims to link your submission to existing knowledge.
6. Methodology (maximum one and half double-column page) focuses on the reasoning for the certain techniques and methods that you used in the context of the study. This describes and explains your chosen methods, it is very important to correlate them to your research questions and/or hypotheses.
7. Result and discussion (maximum 2 and half double-column pages) contains a description of the main findings of the research, interprets the results, and provides the significance of the findings.
8. Conclusion (maximum 500 words) allows you to have the final say on the issues you have raised in your paper, synthesize your thoughts, demonstrate the importance of your ideas, and outline a new view of the subject.
9. Reference (include at least 30 high-quality references, in-text cited) according to IEEE referencing format.

Data Repository

All the code, data, and documentation files (latex package of your submission) should be submitted with the main submission via Moodle. You do not need to upload an entire big dataset, but smaller datasets should have been used to test the software.

All your code should be written in Python 3 and using the Apache Spark library (version 3.0.1 or above).

Note: Your code should be compiled and run successfully on CSlinux server-Spark.

The data repository should contain a **README** file explaining all the included materials, and references to other codes/papers you have used for inspiration.

Moreover, some brief documentation(s) should be provided for each code file to explain the code structure and describe how to use the code and data.

Individual Report

Each member of a team is expected to submit a two-page report including the following:

1. The student information including full name, email, and ID.
2. A table of participation marks: this should provide marks to show how group members contributed/collaborated to the project. This table should have three columns, 1) student full name, 2) a mark out of 10 and 3) one (or maximum two) sentence(s) for making justification.

Student name	Mark (out of 10)	Marking justification

3. A brief explanation of the individual role in the project outlining the offered contributions (Maximum one page).
4. A discussion of individual understandings, findings, and reflections on the project and team-working (Maximum one page).

This report should be submitted by each student via Moodle, a separate submission forum. The format should be:

"Student Number. Pdf" (e.g., 20029784.pdf)

This will be used to assess the role of each individual in the project. This report also might be used to ask relevant questions during the presentation.

Presentation

All the projects will be presented orally 10th May. For this, all group members will be required to be present to deliver a presentation and answer questions from attendees and the Chairs (module convenors).

The presentations are open for all Big data module students.

All oral sessions will follow a similar structure to how they are held in a typical physical conference format.

- The Session Chair will introduce each group.
- The authors will deliver the presentation (10 minutes maximum) for the audience. Please make sure you practice this before the session.
- Once the presentation has concluded, the Session Chair will facilitate a live Q&A period (5 minutes approx.) with the audience and the session chair.
- Process repeats for each subsequent paper in the session.

Presentation tips

- Please use a template to make your presentation slides, it will be available on Moodle.
- Suggest having maximum 10 slides, 1 slide a minute.
- You need to upload your presentation file on Moodle, (one day) before the presentation.
- Start off with a brief introduction of yourselves and the key focus of your coursework.
    - This should outline the contributions of each student in the project.
- The outline of the presentation should be similar to the structure of the coursework (e.g. introduction and motivation, methodology, experimental set-up, results, and conclusions).
- It is a 10-min presentation, do not aim to show every single aspect of your project, focus on the most important things/findings. Key points: Make sure you state clearly your motivation and how novel and good your solution is.
- Be ready for any question and discuss your contribution. You can prepare additional slides/files for any potential questions you expect. You may be asked to explain your solution, and even show your code, so please make sure that one appointed member of the group can share the screen and show the coursework and the code.
- All members should contribute during the presentation.

Coursework Marking Criteria

- **Individual Report (10 marks)**: Each member of the team is expected to submit an individual report.
    - **Active role**
        - Does the student actively participate in the project?
        - What are the student's contributions to the project? How they are relevant and valuable?
    - **Understanding**
        - Does the report show a fair student's reflection and understanding of the projects studied?
        - Does the report clearly highlight the key findings of the student in the project?
- **Main Submission (70 marks)**: Each group should submit a paper-based description of their big data solution together with the code produced.
    - **Motivation and context**
        - Do the team understand the need for a big data solution for their project?
        - Do they provide a compelling argument as to why is needed?
    - **Design/Methodology**
        - Explanation of the methodology.
        - Is the design efficient? (e.g. does it reduce as much as possible data movement across workers?)

        - Innovation of the proposed methodology. Does it provide a standard divide and conquer strategy, or has the team thought of an elaborated strategy to alleviate the bottleneck of the methods to tackle Big Data.
    - **Experiments and results**
        - Are the experiments well designed to test the proposed solutions?
        - Do the results support the original motivation?
        - Is the analysis coherent?
    - **Writing**
        - Clear description, reproducibility
        - Quality of visual elements, illustrations, tables.
        - Quality of References
    - **Code and data – software quality**
        - Efficiency, suitability of Spark operations to solve the problem
        - Documentation
- **Group Presentation (20 marks)**: Each group is asked to deliver a 10-min presentation summarising their contribution + 5 mins for Q&A.
    - Quality and clarity of the presentation
    - Response to questions from the panel and public
    - Understanding of their solution
    - Individual participation in the presentation. All members are expected to participate equally.

The Project Titles

The aim of this coursework is to offer you an opportunity to put your hands on designing/developing a Big Data solution. All the project ideas presented in this section are general descriptions and will allow you to have a preliminary idea of what the project will be, but we are expecting you to do research and determine the final shape of the project.

Important to bear in mind: We won't have enough resources (i.e. computing nodes) to perform real big data. You are not expected to run your code on extremely big datasets which might take a very long time. However, that doesn't prevent us from developing and designing Big Data solutions. You simply need to use smaller sets. As a thumb rule, I wouldn't be expecting you to run anything taking longer than 1 day of execution unless this is just of a 'sequential' program that you are using as a baseline. This means that you are expected to use subsets of the original datasets if they are very big. Note that creating those subsets might not be trivial in all cases, as you still want them to be representative (i.e. preserve the class distribution!).

Deep Learning: Whilst it is ok to use Deep Learning techniques, this is usually best parallelised using GPUs which is not the goal of this module (and we don't have clusters of GPUs or multi-GPUs). You can certainly parallelise Deep Learning on Spark (i.e. using TensorFlow On Spark), but this is not our goal. You are however able to use deep learning, if necessary, after processing Big Data with Apache Spark, but this shouldn't be the focus of your project.

General Tips:

- For each project a dataset is available, however, **you are responsible for data sufficiency and the accuracy of your solution (the proposed model)**. This means, your solution should work well if we extend the dataset (increased volume only). For this, you should measure the scalability of your solution (size-up, speedup). Hence, you may need to find some other dataset(s) to add and test your model, if you think your solution is not accurate enough yet. To this end, you must justify your reasons for adding/choosing the complementary dataset.
- You are allowed for suggesting a new dataset. **Your suggested dataset needs to be approved by the module convener**.
- Some of the datasets might need pre-processing and data cleaning. You are responsible to show that the dataset is cleaned and ready to be used by your project.
- An experimental design is super important when talking about Big Data. Estimating the potential length of all the experiments you want to carry out is key to manage the time. E.g. I would never try to test my code or any minor changes on a 5 million instance dataset.
- You need to justify that your approach is optimised.

BD01: A Predictive Data Analysis for Traffic Accidents

Predictive analysis has the capacity to forecast upcoming events using time-series variables. This project should propose a predictive model for traffic accidents by analysing the available dataset below and according to the traffic patterns and data features. For this, your predictive analysis should study the correlation between the accident and traffic parameters such as weather condition, region, road type, speed, and time and forecast the accidents. This model should be optimised to achieve accurate results as much as possible.

Datasets and the suggested readings:

- https://arxiv.org/abs/1909.09638
- US Accidents (updated) | Kaggle

BD02: Predicting and Understanding the Crypto Market

Predicting the stock market or the crypto market is probably one of the most appealing time series prediction problems. Many have attempted various time-series prediction algorithms, and algorithms such as Facebook Prophet or LSTMs seem to be the best performing methods. As such, this is not necessarily Big Data. For this to be Big Data, we need to add more variety and maybe even deal with the velocity. The main challenge with the stock market, in general, is the BIG influence of social media (producing sudden changes) and news going around which make people panic sell (decreasing the price) or buy like crazy (increasing the price). One funny thing about bitcoin is how the entire crypto market tends to fall apart like a house of cards, but sometimes some coins remain stronger (which ones? And why?). In this project, you should investigate the use of Big Data solutions to either predict or understand the crypto market. There are many different things you could do, e.g. classifying time series, clustering them, and using that to later make predictions. But again, the focus shouldn't only be on predicting more accurately, but on the innovative use of Big Data solutions.

Datasets and the suggested readings:

- https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory
- https://www.kaggle.com/mczielinski/bitcoin-historical-data

- Predicting Cryptocurrency Prices With Deep Learning - dashee87.github.io
- Time-Series Prediction of Cryptocurrency Market using Machine Learning Techniques - EUDL

BD03: Adaptive Feature Selection for Big Data Applications

Feature selection techniques usually focus on a similarity matrix that assigns a fixed value to pairs of objects. However, they may lead to unreliable or inaccurate results especially if the dataset is huge and includes noisy, unlabelled and/or missing samples. This project refers to a multi-target predictive analysis for Big data applications. For this, you need to use/compare machine learning techniques (e.g., random forest and/or SVM) to simultaneously select the features and learn the similarities for either a single or multiple target prediction (i.e., Google's App Rating). You need to demonstrate that the adaptive feature selection gives optimised results as compared to at least two or three well-known traditional feature selection techniques.

Datasets and the suggested readings:

- Local Adaptive Projection Framework for Feature Selection of Labeled and Unlabeled Data | IEEE Journals & Magazine | IEEE Xplore
- Google Play Store Apps | Kaggle

BD04: Time-series Classfication of Freight Transport Data in Belgium

This project should focus on Freight Transport classification to understand how it is influenced by various data features such as road network, traffic count and traffic speed. For this, you need to clean the dataset aiming to propose a time-series classification model. There are many classification techniques which can be used for this project. However, you should explain that your designed

approach is well-fitted and optimised. You need to show that the model is able to predict Freight Transportation with maximum accuracy.

Datasets and the suggested readings:

- Freight Transport Data | Kaggle
- https://link.springer.com/article/10.1007/s10618-020-00727-3

BD05: Climate Change Analysis in Brazil

This project should study weather data to predict climate change. For this, you need to select meaningful features (using LDA/PCA, for example) and train the time-series classification model to classify the climate weather dataset. It should demonstrate how the regional climate was changed during the past 2 decades and is influenced by the COVID-19. You need to show that your machine learning approach is the best-fitted one for this Big data application.

Datasets and the suggested readings:

- Climate Weather Surface of Brazil - Hourly | Kaggle
- https://link.springer.com/article/10.1007/s10113-020-01677-8

BD06: Pre-processing for Imbalanced learning in Big Data

One of the funniest things in the context of Big Data is that in many cases we still suffer from data scarcity. This usually happens in real problems where we have very skewed datasets. Machine learning and data

mining technique usually struggle to cope with imbalanced datasets. One way to deal with this is by using pre-processing techniques, which aim to find a better balance of the input data prior to any learning. We have worked a lot on under-sampling and over-sampling approaches for imbalance datasets in big data, but we usually used traditional methods such as: Random under-sampling vs evolutionary under-sampling, or Random oversampling vs SMOTE. In this project, you should investigate alternative approaches to pre-process imbalanced classification datasets (e.g., SMOTE-ENN, RUSBoost, Repeated Edited Nearest Neighbours) or you could attempt to deal with multi-class imbalance problems. You need to demonstrate how your imbalance learning enhances the performance of drought prediction based on the given dataset.

Datasets and the suggested readings:

- https://github.com/triguero/EUS-BigData
- https://www.sciencedirect.com/science/article/abs/pii/S0950705120307279?via%3Dihub
- https://github.com/scikit-learn-contrib/imbalanced-learn
- Predict Droughts using Weather & Soil Data | Kaggle

BD07: An Optimised Classification for Flight Status

This project focuses on flight status classification (e.g., delayed, cancelled, on time, …) according to various data features such as destination, origin, date, weather, and so forth. For this, you need to propose a data pre-processing approach to clean-up the dataset, remove noises and select the

meaningful data features. The prepared dataset should be used to train a classification model which shows how various data features impact on flight status. There are many machine learning techniques to address it, however, you need to show that your approach is the best-fitted one.

Datasets and the suggested readings:

- Airline Delay Analysis | Kaggle
- https://www.sciencedirect.com/science/article/abs/pii/S0168169916308833

BD08: A Multi-label Prediction in Big Data

In multi-target prediction, an instance has to be classified along with multiple target variables at the same time, where each target represents categories or numerical values. There are several strategies to tackle multi-target prediction problems: the local strategy learns a separate model for each target variable independently, while the global strategy learns a single model for all target variables together. Under this project, you have to first define the machine learning strategy (local or global) and then build a machine learning model for mortgage prediction via a hierarchical multi-label (or multi-target) classification approach. You need to support the model optimisation.

Datasets and the suggested readings:

- Single Family Loan-Level Dataset - Freddie Mac
- https://www.sciencedirect.com/science/article/pii/S003132031200430X

BD09: Analysis the Impact of Green Infrastructure on Carbon Monoxide Reduction

This project needs to verify correlations between urban tree plantation and carbon storage by trees and green infrastructures at the city scale. In other words, it should propose a time-series machine learning model to explore the impact of the tree plantation on carbon monoxide reduction. For this, you need to

pre-process two separated datasets and remove meaningless and noisy features, and select/combine the meaningful ones. The prepared dataset should contain tree plantation features (e.g., location, for the tree location, health status) and the carbon level data in New York during past years (e.g., 1995-2015). Your time series clustering approach should be optimised to forecast the impact of trees on carbon monoxide with maximum accuracy.

Datasets and the suggested readings:

- Tree Census in New York City | Kaggle
- Carbon Monoxide Daily Summary | Kaggle
- https://www.sciencedirect.com/science/article/pii/S2590252020300246