CCT College Dublin Continuous Assessment

Programme Title:	*MSc in Data Analytics*		
Cohort:	*MSc in Data Analytics FT*		
Module Title(s):	*Advanced Data Analytics* *Big Data Storage and Processing*		
Assignment Type:	*Individual*	Weighting(s):	*Advanced Data Analytics – 60%* *Big Data Storage and Processing – 60%*
Assignment Title:	*MSC_DA_BD_ADAv4 Repeat*		
Issue Date:	*Jun/July 2024*		
Due Date:	*28th July 2024*		
Late Submission Penalty:	Late submissions will be accepted up to 5 calendar days after the deadline. All late submissions are subject to a penalty of 10% of the mark awarded. Submissions received more than 5 calendar days after the deadline above will not be accepted and a mark of 0% will be awarded.		
Method of Submission:	Moodle Use the submission link on the Advanced Data Analytics Module page		
Instructions for Submission:	*Please do not ZIP your files. ALL files must be uploaded individually (to a maximum of 20 files)* *Expected files : Written report (word document only, NO PDF's) ,Code files (Jupyter notebook (.ipynb) ONLY, NO PYTHON FILES), Data Files, Screencast for practical demonstration. Note that the maximum number of Jupyter Notebooks is 4*		
Feedback Method:	Results posted in Moodle gradebook		
Feedback Date:	*After exam board August 2024*		

Learning Outcomes:
Please note this is not the assessment task. The task to be completed is detailed on the next page.
This CA will assess student attainment of the following minimum intended learning outcomes:
Big Data Storage and Processing
MLOs

1. Critically assess the data storage and management requirements of a given data project from a modern perspective and evaluate limitations of legacy approaches to Big Data. (Linked to PLO 3)
2. Assess the design concepts and architectural patterns of distributed Big Data systems and analyse the components that form their technology stack. (Linked to PLO 1, PLO 2)
3. Critically evaluate and select a Big data environment suitable for retrieving and processing a given Big Data set, perform data management and select appropriate analytic algorithms for the required scale and speed. (Linked to PLO 2, PLO 3)

Advanced Data Analytics

3. Analyse a set of requirements to determine the type of Advanced Data Analysis for a particular problem set. Document and justify choices made to stakeholders and peers through insight gained from the process.(linked to PLO 4, PLO 5)
4. Develop a solution, reliant on temporal data (e.g., social media feed, sensor data) to solve a given problem set.(linked to PLO 1, PLO 2)

5. Critically assess the existing state of the art in Natural Language Processing and propose a strategy toward optimisation.(linked to PLO 1, PLO 2, PLO 4)

Attainment of the learning outcomes is the minimum requirement to achieve a Pass mark (40%). Higher marks are awarded where there is evidence of achievement beyond this, in accordance with QQI *Assessment and Standards, Revised 2013*, and summarised in the following table:

Percentage Range	QQI Description of Attainment
	Level 9 awards
70% +	Achievement includes that required for a Pass and in most respects is significantly and consistently beyond this
60 – 69%	Achievement includes that required for a Pass and in many respects is significantly beyond this
40 – 59%	Attains all the minimum intended programme learning outcomes
35 – 39%	Nearly (but not quite) attains the relevant minimum intended learning outcomes
0 – 34%	Does not attain some or all of the minimum intended learning outcomes

The CCT Grade Descriptor describes the standard of work for grade boundaries summarised below. The full descriptor is available on Moodle.

Grade	90-100%	80-89%	70-79%	60-69%	50-59%	40-49%	35-39%	<35%
Performance	Exceptional	Outstanding	Excellent	Very Good	Good	Acceptable	Fail	Fail

Acceptable and Unacceptable Use of AI

Acceptable and Unacceptable Use of AI	● The use of generative AI tools (e.g. ChatGPT, Dall-e, etc.) is permitted in this assignment for the following activities: o Brainstorming and refining your ideas; o Fine tuning your research questions; o Finding information on your topic; o Drafting an outline to organise your thoughts; and o Checking grammar and style. ● The use of generative AI tools is not permitted in this course for the following activities: o Impersonating you in classroom context o Completing group work that your group has assigned to you o Writing a draft of a writing assignment o Writing entire sentences, paragraphs, papers, code fragments, functions, scripts to complete class assignments. ● You are responsible for the information you submit based on an AI query. Your use of AI tools must be properly documented and cited. ● Any assignment that is found to have used generative AI tools in an unauthorised way will be subject to college disciplinary procedures as outlined in the QA Manual. ● When in doubt about permitted usage, please ask for clarification.

Assessment Task
Students are advised to review and adhere to the submission requirements documented after the assessment task.

It is Required that you use GitHub Classroom as your version control repository etc with regular commits of code and report versions. You may be called to a Viva to defend your work.

Please find the GitHub Classroom link below:
https://classroom.github.com/a/4oouDYVk

You may not upload a PDF document for your report, It MUST be a word document. This is Not a Scientific Paper and must be formatted in standard report format including index, appendix, etc.

In this continuous assessment, You are required to identify and carry out an analysis of a large dataset gleaned from the twitter API and Yahoo Finance and is available on Moodle as "stock-tweet-and-price.zip" This data should be stored as requested below, and you are then required to analyse the data and make a CLOSE price forecast of 1 day, 3 days, and 7 days for at least 5 of the companies using the tweets AND the financial price data.
Context

Data in file "stocktweet.csv"

Data collection period : Jan 2020 - Dec 2020
Number of Tweets : 10,000

It contains the following 4 fields:

- ids: The id of the tweet (eg. 100001)
- date: the date of the tweet (eg. 01/01/2020)
- ticker: The ticker value for the company (eg AMZN)
- tweet: the text of the tweet (eg. $AMZN Dow futures up by 100 points already ðŸ¥³)

Data in folder "stockprice"

Data collection period : Jan 2020 - Dec 2020
38 companies historical price data in csv format.
- Tickers:
'AAPL', 'ABNB', 'AMT', 'AMZN', 'BA', 'BABA', 'BAC', 'BKNG', 'BRK.A', 'BRK.B', 'CCL', 'CVX', 'DIS', 'FB', 'GOOG', 'GOOGL', 'HD', 'JNJ', 'JPM', 'KO', 'LOW', 'MA', 'MCD', 'MSFT', 'NFLX', 'NKE', 'NVDA', 'PFE', 'PG', 'PYPL', 'SBUX', 'TM', 'TSLA', 'TSM', 'UNH', 'UPS', 'V', 'WMT', 'XOM'

Each CSV file contains the following 7 fields:

- Date : the date of the stock price (eg. 01/01/2020)
- Open : the opening value of the stock price that day (eg. 123.33)
- High : the highest value of the stock price that day (eg. 125.45)
- Low : the lowest value of the stock price that day (eg. 121.54)
- Close : the closing value of the stock price that day (eg. 122.49)
- Adj Close : the adjusted closing value of the stock price that day (eg. 122.49)

- Volume : the number of stocks traded that day (eg. 100805600)

Following your analysis, you are then required to make a time series forecast of the CLOSE price for at least 5 of the companies using the tweets AND the financial price data at 1 day, 3 days and 7 days going forward. This forecast must be displayed as a dynamic dashboard.

Your project must incorporate the following elements:

- Utilisation of a distributed data processing environment (e.g., Hadoop Map-reduce or Spark), for some part of the analysis.
- Source dataset(s) can be stored into an appropriate SQL/ NoSQL database(s) prior to processing by MapReduce / Spark (HBase / HIVE / Spark SQL /Cassandra / MongoDB / etc.) The data can be populated into the NoSQL database using an appropriate tool (Hadoop/ Spark etc.)
- Post Map-reduce processing dataset(s) can be stored into an appropriate NoSQL database(s) (Follow a similar choice as in the previous step)
- Store the data and then follow-up analysis on the output data. It can be extracted from the NoSQL database into another format, using an appropriate tool, if necessary (e.g. extract to CSV to import into R/ Python etc.).
- Devise and implement a test strategy in order to perform a comparative analysis of the capabilities of any two databases (MySQL, MongoDB, Cassandra, HBase and CouchDB) in terms of the performance. You should record a set of appropriate metrics and perform a quantitative analysis for comparison purposes between the two chosen database systems.
- Provide evidence and justification of your choice of sentiment extraction techniques.
- Explore at least 2 methods of time-series forecasting including at least 1 Neural Network and 1 autoregressive model (ARIMA, SARIMA etc...) . (Hint: that this is a Short time series, How are you going to handle this?)
- Evidence and justify your choices for your final analysis and include your forecasts at 1 day, 3 days and 7 days going forward.
- Your dashboard must be dynamic and interactive. Include your design rationale expressing Tufts principles.

Deliverables:

The results of the analysis must be presented in the form of a project report. This report should discuss the storage and processing of big data using advanced data analytics techniques. The report should be 3000 ± 10% words in length (excluding references, titles, and code) and must follow the Harvard styles format in addition to employing appropriate referencing methods and academic writing style. The report should include the following:

Big Data

1. Details of the data storage and processing activities carried out, including preparation of the data and processing the data in a MapReduce/ Spark environment;[0-30]
2. Comparative analysis for at least two databases (one SQL and at least one NOSQL) using YCSB.[0-30]
3. A discussion of the rationale and justification for the choices you have made in terms of data processing and storage, programming language choice, that you have implemented.[0-20]
4. Design the architecture for the processing of big data using all the necessary technologies (HADOOP/SPARK,NOSQL/SQL databases and programming). Present your Design in the form of a diagram and discussion in your report .[0-20]

   Note that MapReduce-style processing in this instance is considered to include platforms such as Apache Spark.

Advanced Data Analytics

1. A discussion of the rationale, evaluation, and justification for the choices you have made in terms of EDA, data wrangling, machine learning models and algorithms that you have implemented.[0-40]
2. Evaluation and justification of the hyperparameter tuning techniques that you have used [0-20]
3. Your analysis of the data and your forecast of the CLOSE price for at least 5 of the companies using the tweets AND the financial price data at 1 day, 3 days and 7 days going forward[0-20]
4. Presentation of results by making appropriate use of figures along with caption, tables, etc and your dashboard for your forecast, Discuss Tufts Principles in relation to your Dashboard .[0-20]

SUBMISSION:

Submission Requirements All assessment submissions must meet the minimum requirements listed below. Failure to do so may have implications for the mark awarded.

All assessment submissions must:

- 3000 words +- 10% (excluding references, titles, citations and quotes)
- Word Document for report (No PDF's), Jupyter notebook for code, Screencast for practical demonstration.
- Be submitted by the deadline date specified or be subject to late submission penalties
- Be submitted via Moodle upload
- Use Harvard Referencing when citing third party material
- Be the student's own work.
- Include the CCT assessment cover page.

Additional Information

- Lecturers are not required to review draft assessment submissions.
- In accordance with CCT policy, feedback to learners may be provided in written, audio or video format and can be provided as individual learner feedback, small group feedback or whole class feedback.
- Results and feedback will only be issued when assessments have been marked and moderated / reviewed by a second examiner.
- Additional feedback may be requested by contacting your lecturer AFTER the publication of results, Additional feedback may be provided as individual, small group or whole class feedback. Lecturers are not obliged to respond to email requests for additional feedback where this is not the specified process or to respond to further requests for feedback following the additional feedback.
- Following receipt of feedback, where a student believes there has been an error in the marks or feedback received, they should avail of the recheck and review process and should not attempt to get a revised mark / feedback by directly approaching the lecturer. Lecturers are not authorised to amend published marks outside of the recheck and review process or the Board of Examiners process.
- Students are advised that disagreement with an academic judgement is not grounds for review.
- For additional support with academic writing and referencing students are advised to contact the CCT Library Service or access the CCT Learning Space.
- For additional support with subject matter content students are advised to contact the CCT Student Mentoring Academy
- For additional support with IT subject content, students are advised to access the CCT Support Hub.

CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission

Module Title:	
Assessment Title:	
Lecturer Name:	
Student Full Name:	
Student Number:	
Assessment Due Date:	
Date of Submission:	

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.