**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:** | Storage and Processing of Project Tweets using advanced data analytics techniques (LSTM and Time series autoregressive model) |
| **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. |

**TIME SERIES FORECASTING OF TWEETER SENTIMENTS, PROCESSED USING SPARK AND STORED IN MONGODB**

**Introduction**

-Big data sources, social media)

- processing and storage of big data

- sentiment analysis of social media

Time series forecasting methods

**METHODS**

**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]**

**Data Description**

**MongoDB**

**Pyspark**

**Data Population, Processing and Storage in MongoDB using Pyspark**

1. **Data Processing using SPARK** (Utilisation of a distributed data processing environment (e.g., Hadoop Map-reduce or Spark), for some part of the analysis)
2. Project tweets was stored in MongoDB database(s) prior to processing by Spark.
3. The data can be populated into MongoDB database using Spark etc.
4. 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.).

**COMPARATIVE ANALYSIS OF MYSQL AND MONGODB**

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.

-

**Advanced data analytics**

**Text Processing**

**Sentiment Analysis**

* Provide evidence and justification of your choice of sentiment extraction techniques.

**Text Classification**

**TIME SERIES FORECASTING**

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.

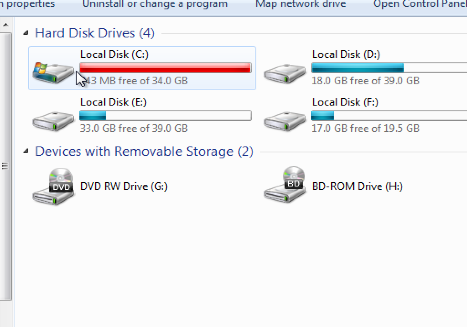
**Deliverables:**

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**

**Data Storage and Processing**

Data storage and processing were conducted using pyspark and MongoDB. The activities involved are as shown below: -



[This Photo](http://superuser.com/questions/484525/extending-c-drive-not-possible) by Unknown Author is licensed under [CC BY-SA](https://creativecommons.org/licenses/by-sa/3.0/)

Spark Data Frame

MongoDB

Database

Spark

1.

**MongoDB**

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. **E**valuation and justification of the hyperparameter tuning techniques that you have used **[0-20]**
3. Your analysis of any change sentiment that occurs and your forecast of the sentiment 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](http://40.115.124.2/sp/subjects/guide.php?subject=harvardref) 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](http://learningspace.cct.ie/subjects/index.php).
* For additional support with subject matter content students are advised to contact the [CCT Student Mentoring Academy](https://moodle.cct.ie/mod/forum/view.php?id=55148)
* For additional support with IT subject content, students are advised to access the [CCT Support Hub](https://moodle.cct.ie/course/view.php?id=1861).