

CCT College Dublin Continuous Assessment

|  |  |  |  |
| --- | --- | --- | --- |
| Programme Title: | MSc in Data Analytics | |  |
| Cohort: | MSc in Data Analytics SB+ (Feb 23 start) | |  |
| Module Title(s): | Advanced Data Analytics  Big Data Storage and Processing | |  |
| Assignment Type: | Individual | Weighting(s): | Advanced Data Analytics – 40%  Big Data Storage and Processing – 40% |
| Assignment Title: | Integrated CA1 Sem 2 MSc in Data Analytics | |  |
| Issue Date: | 06/09/2023 | |  |
| Lecturer(s): | David McQuaid Muhammad Iqbal | |  |
| 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. | | |
| Method of Submission: | **Moodle** | | |
| Instructions for Submission: | All files to be uploaded separately (Do NOT Zip your submission)  **Expected files : Written report (word document ONLY, No PDF’s),** **Code files (jupyter notebook), Screencast presentation, Data sets, GitHub link** | | |
| Submission Date: | 6th October 2023 | Feedback Date: | Three weeks after final submission [Including late and PMCs] |

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 MLO’s

1. Debate the theory and application of different types of neural networks. (Linked to PLO 1, PLO 2)
2. Analyse a set of requirements to determine the type of Neural Network for a particular problem set. Document and justify the choices made to stakeholders and peers through insight gained from the process. (Linked to PLO 4, PLO 5)

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 |

**Assessment Task**

**Note your chosen topic of research cannot be the same as your topic chosen for your RP&E Module**

**Submissions that are suspected of plagiarism and/or inclusion of AI (CHATGBT, BARD etc…) Generated content will be referred to the college authorities.**

**Note ALL Students are required to use Git for any Assignments that they are working on.**

This means that ALL changes must be committed to Git during your assignment. (Not just a single commit at the end!) This is to allow you to display your incremental progress throughout the assessments, give you practice for your capstone/thesis, allows you to create an online portfolio that can be used to showcase your work and to ensure that there are no problems with final uploads (as all your work will be available on GitHub). It is expected that there will be a minimum of 10 commits (with many of you making very many more).

You may Only use your CCT email for your git account, private/work email-based accounts will not be accepted. You must also include your lecturer's CCT email as a collaborator on your account.

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

In this continuous assessment, you are expected to carry out research and produce a research paper based on your chosen topic **combining Advanced data analytics (specifically Neural Networks) and Big Data Storage & Processing.** Some example topics of interest include but are not limited to.

|  |  |
| --- | --- |
| ‒ | Deep Learning using Big Data |
| ‒ | Big data architectures that incorporate Neural Networks |
| ‒ | Programming paradigms for Data Analytics |
| ‒ | Data streaming / encoding / compression |
| ‒ | Algorithms for parallel and distributed computing incorporating Neural Networks |
| ‒ | Application of distributed computing and Neural Networks for specific domains / problem areas / use cases |
| ‒ | Data Lakes |

The rationale behind this continuous assessment is to practice your critical evaluation skills based on the state of the art in regard to your chosen area of interest. Note that critical evaluation doesn’t necessarily imply a negative evaluation of a work, but it is your structured methodological review of its contributions as well as limitations.

**Technical demonstration:**

Demonstrate a practical example of your chosen area of interest that should include Big Data and Advanced Data Analytics. This will be a coded example and will require you to upload your code in a Jupyter notebook.

Demonstrate your research along with your practical deployment using a video recording (5 – 7 minutes) or Screen cast. All code must in a Jupyter notebook and be uploaded as a zip file on Moodle. [0 - 20] Your paper should include:

* Overview of the chosen topic, including objective statement and Research Question. Presentation of state of the art, including research methodologies and key of the papers you reviewed. [0 - 20]
* Literature review based on the chosen papers for the topic of your choice including proper citations and Harvard style referencing. [0 - 15]
* Critical evaluation of the key findings, specifically their implications and limitations, and highlighting any contradicting viewpoints and research gaps. [0 - 30]
* Conclusions you have drawn based on your research. [0 - 15]

**‒ It is important to remember that a research paper is not just a summary of articles**.

‒

The expected word count should be (4000 – 6000) (excluding references, titles, citations and quotes) in length and should follow the IEEE format1. You should draw on appropriate academic literature for your review (use a MINIMUM of 10 references). Tools like Google Scholar will be of key value to achieve this. Moreover, tools such as Zotero or Mendeley will be very useful for reference management and proper reference formatting. [Ignore private information in the IEEE template except StudentID, Name Course]

SUBMISSION:

* The final report must be submitted to Moodle on/before the deadline. Submissions received after the deadline will be subject to penalties.
* Format of word file name should follow the format, studentID\_Integrated\_CA (NO PDF’s!)
* Code files (jupyter notebook),
* Screencast presentation,
* Data sets,
* GitHub link

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:

* 4000 - 6000 words (excluding references, titles, citations and quotes)
* Word Document for report, Jupyter notebook for code, Screencast for practical demonstration.
* IEEE Format
* 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.

The growth of social media networks has brought a large amount of data, which is very valuable for companies in a way to measure how the population is behaving and a tool to understand better consumers. That is the moment that Data Science has developed tools to create methods and extract any kind of information and insights from this huge amount of Data, one of these tools is sentiment analysis.

Sentiment Analysis is part of Natural Language Processing (NLP) that develops tools for locating and extracting sentiments from texts. These systems often extract characteristics from sentences recognizing the opinion, subject, and polarity.

This study presents a comprehensive sentiment analysis from two social midias platforms, Youtube and Tweeter. Using scrapping tec

This study presents a comprehensive analysis of sentiment in user-generated content from two prominent social media platforms, YouTube and Twitter. Leveraging web scraping techniques, a diverse dataset of comments and posts was collected, encompassing a wide array of topics and user demographics. The objective of this research is to assess the efficacy of sentiment analysis models in capturing nuanced emotional expressions within these platforms and to discern potential variations in sentiment expression across them.

The preprocessing pipeline involved text normalization, tokenization, and the application of state-of-the-art sentiment analysis models. A comparative evaluation of popular sentiment analysis techniques, including lexicon-based approaches, machine learning algorithms, and deep learning models, was conducted to ascertain their performance in accurately classifying sentiment in the respective datasets.

The findings of this study reveal intriguing insights into the nature of sentiment expression on YouTube and Twitter. While both platforms exhibit similarities in sentiment distribution, significant disparities emerged in the way sentiment is expressed, possibly reflecting the differing purposes and user demographics of these platforms. Additionally, the study uncovers the influence of multimedia content (videos on YouTube and images on Twitter) on sentiment expression, providing valuable context for future sentiment analysis endeavors.

Furthermore, this research investigates the impact of platform-specific features such as video length, comment length, and user engagement metrics on sentiment classification accuracy. The results shed light on the role of these factors in refining sentiment analysis models for social media data.

In conclusion, this study contributes to the burgeoning field of sentiment analysis by offering a comprehensive comparison of sentiment expression on YouTube and Twitter. The insights garnered from this research have implications for a wide range of applications, including brand sentiment monitoring, content recommendation systems, and social media analytics. The methodologies and findings presented herein provide a valuable foundation for further research in sentiment analysis within the dynamic landscape of online user-generated content.

The growth of social media networks has brought a large amount of data, which is very valuable for companies in a way to measure how the population is behaving and a tool to understand better consumers. That is the moment that Data Science has developed tools to create methods and extract any kind of information and insights from this huge amount of Data, one of these tools is sentiment analysis.

Sentiment Analysis is part of Natural Language Processing (NLP) that develops tools for locating and extracting sentiments from texts. These systems often extract characteristics from sentences recognizing the opinion, subject, and polarity.

# **Ease of Use**

## ***First Stage - Scrapping comments from Youtube***

The aim of this project is to analyse climate change based on twitter dataset and youtube comments. As a initial part of the process. I began by extracting and organizing comments from two videos on Youtube platform. Both videos show the changes the world is facing due to climate change.

The first video used was from the channel ClickView video What is Climate Change? Explore the Causes of Climate Change and the second video was from the channel DW Documentary, Climate Change - Averting catastrophe. A dataset was created using concatenate function to make only one dataset with comments of both videos.

In order to get the comments and be able to scrape from Youtube, firstly it is needed 3 pins of authentication provided by Google. These pins are for API service, API version and Developer key, the last one contains 39 digits. The pins are essential for the data extraction and manipulation.

The objective of using scrapping on Youtube is to collect data and apply sentiment analysis models capturing expressions within the platforms and measure potential variations of sentiments.

## ***Importing Datasets***

The main libraries were imported to use on datasets., such as Pandas, numpy, some o others very important to read sentiment analysis, textblob, nltk.

The first function, and possibly more important is clean the text. Everything that can interfere the modeling process will be dealt with by the function “def cleanTxt”. This function is very useful when the text, specially from social medias contain a lot @, Url’s, HTML punctuation, numerals, lower case characters.

I am using the python Textblob library, Other than the processing of textual data, the primary focus of NLP research is on other aspects of language. In addition, it aims to simplify some essential procedures by using the NLTK and Pattern libraries in the background. An important fact about Textblob is that this library just have support in English.

The TextBlob’s ‘sentiment’ is attribute was used. This attribute returns a sentiment object with 2 values. The first refers to polarity of the sentence, which ranges from the most negative (-1) to the most positive (1). The second value refers to subjectivity of the sentence, that can vary from 0 to 1, and the closer to 1, the more subjective the sentence.

This ind of sentiment analysis is unsupervised, we do not need to pass a base of positive or negative phrases to carry out the classification.

Colocar um grafico

# 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.