MMA 865

Big Data Analytics

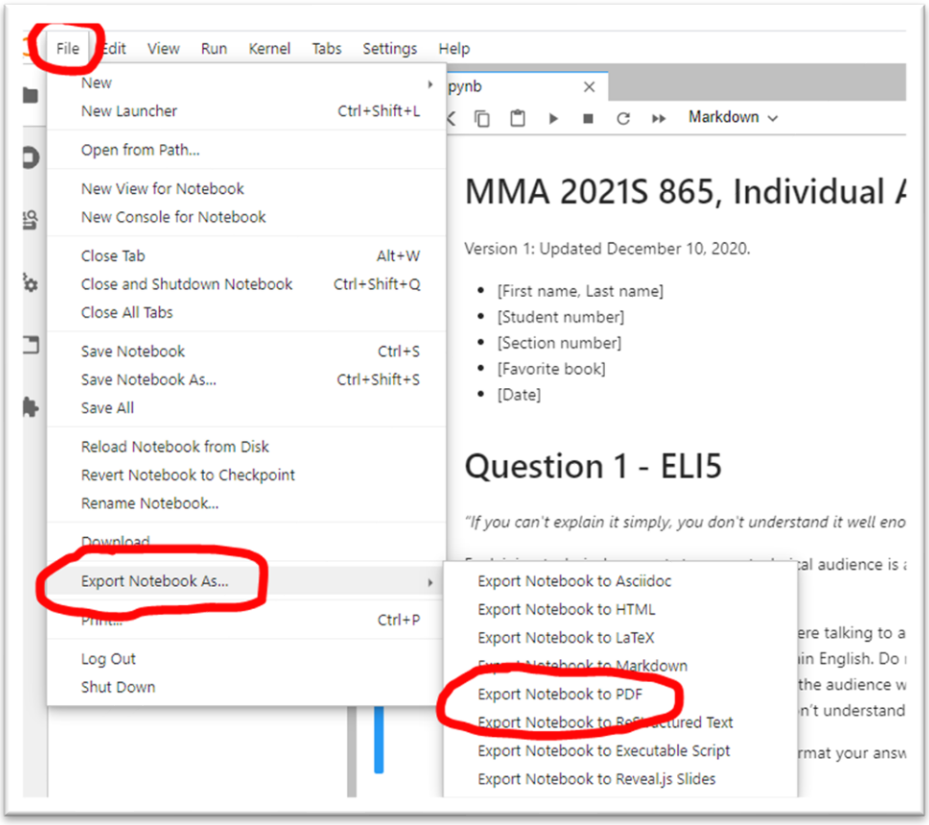
Moez Ali

Individual Assignment

Last Updated August 7, 2024

OVERVIEW

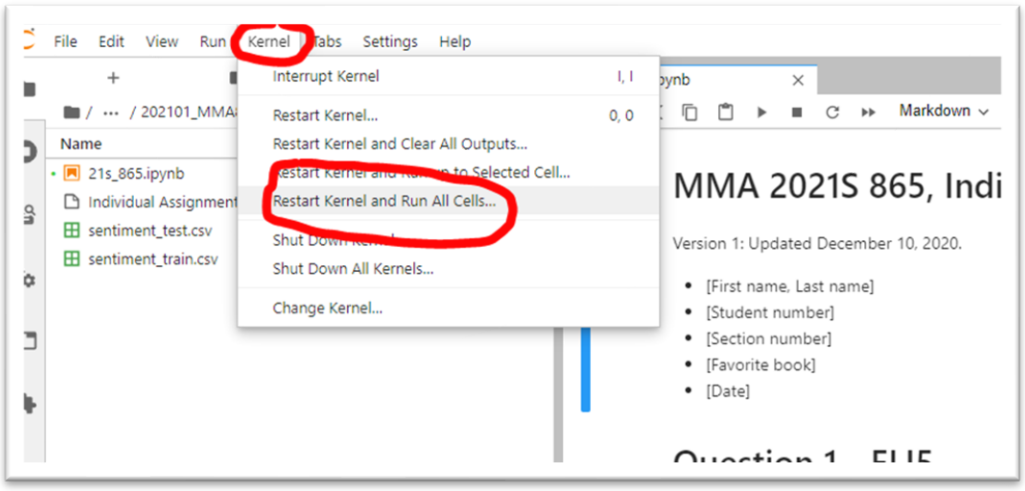
This assignment has two parts. You must complete **the first part** of the assignment in the accompanying Python Jupyter Notebook file named *MMA_865.ipynb*. Once completed, export the Notebook as a PDF or HTML file.



The second part of the assignment can be completed in a Word file or any other equivalent word processor of your choice. **<u>You must upload two separate files on the portal.</u>**

PART 1

- In this part you will train an NLP machine learning model for sentiment analysis using supervised machine learning. The accompanying train and test dataset is provided on the portal. There are multiple parts of this question that require analysis. All the questions are in the accompanying Jupyter Notebook MMA865.ipynb file.
- **Please do not submit the Jupyter Notebook (.ipynb) file to the course portal. You must submit only a PDF export of the notebook.**
    - Please name your file MMAB865_FirstnameLastname.pdf
        - E.g., MMAB865_MoezAli.pdf
    - Please make sure you have run all the cells so we can see the output!
    - Best practice: Before exporting to PDF or HTML, make sure to restart the Kernel (to clear any local state) and Re-run all the cells:



Instructions:

- Use Python to complete the task.
- You may use standard Python libraries, including scikit-learn, pandas, and numpy.
- Tips:
    - Submit code that runs without errors.
    - Submit code that is reproducible. E.g., set random number seeds as appropriate. You should be able to run you code again and again and again, from the top of the file to the bottom of the file and get the exact same results each time. I should be able to run your code, from scratch, on my machine, again and again, and get the exact same results that you get.
    - Submit code that is organized. Make your code and output readable. Provide comments to describe what the code is doing and why. Don't leave "old" code laying around. Overall, if your code is clear and easy to read, then we will be happy. When we are happy, we give better marks.

PART 2

- Identify one use-case of Artificial Intelligence at your work. It can be a conventional NLP use-case like sentiment analysis, document classification, information extraction, etc. or it could be use-case that leverage Generative AI and Large Language Models (LLM) like ChatGPT, Bard, Midjourney etc.
- The identified use-case **must** use text data or any other unstructured data like images, audio, video, etc.
- Supervised or unsupervised learning use-cases on structured data (numeric and categorical features only) are **NOT** acceptable for this assignment.
- The use-case may involve ad hoc analysis or a proposed AI functionality for your company's product.

Instructions

- Write a document in 1000 words or less. **You must clearly organize your document in following Sections:**

  - **Section 1:**
    - Define the problem statement.
    - What will happen if you solve this problem and what will happen if you maintain status quo.
  - **Section 2:**
    - Explain how you are proposing to solve this problem with AI techniques.
    - Go deep into the solution workflow without using too much technical jargon.
  - **Section 3:**
    - Quantify the impact of the solution.
    - Document your assumptions for quantification.
  - **Section 4:**
    - Create a high-level plan to implement this use-case.
    - Give details about how you will collect the data, what kind of data is required, what kind of preprocessing you think is required.
    - What type of modeling is required.
    - How you will measure and select the model.

If you do not work and have no possible way to imagine a use-case (even from past jobs) then assume the following scenario:

- You are hired as an AI consultant for the MMA865 course to propose a use-case that you think will improve the student experience or course delivery.
- Write a document in 1000 words or less. **You must clearly organize your document in the sections as described above.**

USE OF GENERATIVE AI / LLM IN THIS ASSIGNMENT

- You are **NOT** allowed to use ChatGPT or any other LLM for any part of this assignment.
- Any type of use of Generative AI / LLM tools constitutes a Departure from Academic Integrity.

GRADING CRITERIA

Part 1:

- Part 1: 15 marks
- Part 2: 5 marks
- Part 3: 20 marks
  Subtotal: 40 marks

Part 2:

- Section 1: 20 marks
- Section 2: 20 marks
- Section 3: 10 marks
- Section 4: 10 marks
  Subtotal: 60 marks

Total: 100 marks