MSA Assessment Introduction

This is an individual assessment where you will be ranked against your peers. The MSA team has set a minimum standard that we require all students to *pass but we will be* expecting to see work well above and beyond the minimum specifications (as per industry standards). Your project will need to be drastically different from our example project.

Business case

An international company which specialises in developing English language curriculums for educational institutions worldwide is looking to use NLP to optimise development of their curriculum.
Every year the company has to select a set of readings of different complexity for each level of their curriculum. The company wants a model that can assess the readability (complexity) of the text and assign it to appropriate curriculum level.

The company decides to outsource the development of the model through a competition held on Kaggle website (this scenario is hypothetical, but the competition is real). Once the competition is complete the company will select the best model. Our task is to develop a model for the company and recommend suitable books for each year level for New Zealand schools.

Basic Requirements

1. Participate in the following Kaggle competition https://www.kaggle.com/c/commonlitreadabilityprize/overview
   If you are not able to participate in the competition before the deadline, we would expect a more thorough evaluation of the model in your technical report.
   - The deadline for entering this competition is **July 26, 2021** (Please sign up to Kaggle and accept the competition rules before this date. The competition submission deadline is **August 2, 2021**). We highly encourage you to enter and participate in the competition (however, it is not compulsory).
   - The competition contains training (train.csv) and test sets (test.csv) for your use. Your goal is to train a model that can output a readability score given an excerpt. You use the excerpts from the training set to learn the "target" (readability) values. Once you have trained your model you should make predictions of readability scores of the test set excerpts (test.csv). Your predictions should be in the form of

(sample_submission.csv) available in the “Data” tab on the Kaggle website.

   - The competition only accepts notebook submissions; hence you will have to create a notebook on Kaggle. To do that navigate to the “Code” tab in the competition name and click “New Notebook”. This should start a JupyterLab like interface.
   - Once the notebook is created click “Settings” in the right-side bar and toggle “Internet” button to off. This turns off the internet for the notebook (it's somewhat an anti-cheat method for the competition which doesn’t let you look up answers elsewhere on the web). If you don’t complete this step, you would get a “Cannot submit” error later.
   - You can use this notebook [Link] to see how to read in test and train datasets and how to structure your submission file.
   - Once you created your submission, click “Save Version” button in the top right. Then navigate to the competition screen and click “Submit Predictions”. You should be able to choose the correct notebook and make a submission. Once you submit predictions you should see your score for the submission both in the “My Submissions” tab and in the “Leaderboard” tab.

2. Submit a **technical report** (either a Jupyter notebook or pdf) outlining how you approached the Kaggle competition detailing all steps taken in the **data preparation**, **model selection**, and **evaluation** stages.
3. Scrape paragraphs from the text of at least the top 100 books from https://www.gutenberg.org/ebooks/search/?sort_order=downloads
   - Clean and process this data into a .csv file.
   - Using the machine learning model in the previous part, evaluate the readability of the top 100 books.
4. Based on the provided business case, use your machine learning model and scraped data to prepare a short (2-3 page) report aimed at high school English teachers that teach classes from Year 9 to Year 13, with suggestions of books for each year level. The report should contain, at a minimum, an **executive summary**, **description of the approach** used (in a way that can be understood by the target audience), and a **list of book recommendations** for each year level. This report will be targeted towards teachers, so it should be aimed at a non-technical audience.

Advanced Requirements

We will fail students that do not tangibly attempt the following. Please complete as many as you can, feel free to work on your own advanced features. It will be at the discretion of your marker to determine if this feature meets the minimum requirement, you will also be expected to thoroughly explain your advanced feature during the presentation.

Deploy the model trained in the Basic Requirements section on to Azure and expose it as an API. And then complete at least 3 of:

1. Create a PowerBI dashboard to visualize the readability data. This dashboard should supplement the report created in part 4 of the basic requirements.
2. Create a transformer based model for the Kaggle competition dataset using the Huggingface library and compare its performance with your competition entry (if the model created for the basic requirements is a transformer based model, this requirement would already be met). Submit either a Jupyter notebook or pdf.
3. Host a Microsoft Bot Framework chatbot that can understand the intents of user messages using LUIS. The chatbot should at least expect a string of a passage and output the complexity of the passage / readability score. This can be done by interfacing the bot with the deployed model's API. Test the Bot using Bot Framework Emulator. Expand the bot so that it becomes a useful tool for the high school teachers mentioned in part 4 of the basic requirements.
4. Connect the Bot to different channels (Microsoft Teams / Facebook Messenger / Telegram).
5. Create a Web Application that would be helpful to the business case target audience. Be as creative as you want.

A set of supporting documents including a template for the non-technical report, a guide to deploying a model and workshop recordings are available at: https://github.com/NZMSA/2021-Phase-2-Data-Science. This repository will be available after the first day of the workshop (25/07/2021)

Submission Format

You will be required to complete two specific submission tasks.

1) Upload your source code, documentation and all resources to your public GitHub repo and submit it using the link below. This is due on **11:59PM 09/09/2021**.

2) Present to the MSA team in-person at the assessment centre on **11/09/2021** (Note: Due to fluctuating COVID alert level, there is a possibility that this will be moved online. If this occurs, we will let students know via email and the Facebook Group ahead of time.)

   a. Presentations are 15-minute slots.
   b. 7 minutes of presentation. You should do a structured presentation using slideshows.
   c. 7 minutes of Q&A from MSA team, we will be assessing your technical ability in depth.
   d. Holistically, we will be assessing your intricate understanding of your solution, its application in the real world and your presentation as well as soft skills.

Academic Integrity

The MSA Programme strictly enforces academic integrity and students that do not uphold academic integrity will not be allowed to proceed within the programme. Any work you submit should be your work and differentiated from others if you have collaborated or troubleshooted with other parties. This includes any sample solutions and tutorials we provide. Please clarify with us at msaccelerate@hotmail.com if you have questions.

MSA Certificate

If you pass phase 2, you will be able to request a certificate from the MSA Team verifying your completion of the MSA Program.

Work Placement information

Work placement opportunities are available in the Auckland region across our partners. We ask employers to pay a minimum compensation of $2000. Many employers will pay full wages. If there are more students who have passed Phase 2 than the number of placements, priorities will be given to students who have shown in-depth knowledge and dedication through the submission and the presentation.

Deadline: 11:59PM 9th September 2021

Submission Form:
https://forms.office.com/r/MYLqKtTMkH

You can book your presentation slot at the end of the submission form.

Have a question?
Ask on the Discord Server or Facebook group