DATA8001 Assignment 2

Summary

The DATA8001 Assignment 2 is worth 50% of your overall module score.

Download the zip file from Canvas corresponding to your student id and unzip the contents into your local assignment folder and ensure your files are similar to Figure 1.

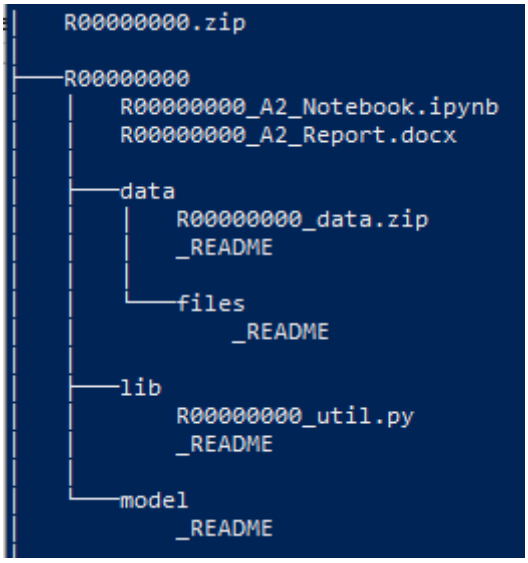

Figure 1 - Example Assignment Folder & Files

Assignment Sections (50%)

Data ETL – 10%

There are 2,000 news articles in the data/R00000000_data.zip file (replacing R00000000 with your CIT student number). Unzip the news articles (data/R00000000_data.zip) into the **data/files** folder.

Each news article can be viewed in Notepad and is in the format:

```
<REPORTER>Student Name</REPORTER>
<DATE>News Article Date</DATE>
<CATEGORY>News Article Category</CATEGORY>
<HEADLINE>News Article Headline</HEADLINE>
<ARTICLE>News article Text</ARTICLE>
```

Create a single dataframe containing the 2,000 news articles with the headings: **[news_category, news_headline, news_article]** and save the dataframe as data/R00000000_processed.csv replacing R00000000 with your CIT student number.

All code required to reproduce the data ETL process should be placed in the Python library file (at the bottom where indicated): lib/R00000000_util.py and able to be called from the Jupyter Notebook: R00000000_A2_Notebook.ipynb.

Data Modelling – 20%

Create 3 multi-class classification models to classify news article categories using the sample data provided to train & test your models. For each model, briefly explain in your report why you selected this model and its accuracy (overall & individual class) on your data. Also provide your recommendations for best models and settings based on your research in the report.

In your report explain your choice of text pre-processing technique (e.g., bag of words, TF-IDF etc.) for each model and also include what text preparation methods you employed (e.g., lowercase, stemming etc.).

For each model, use some form of model parameter optimisation (e.g., grid search, partial grid search etc.) to determine the best model parameters and ensure the models are not overfitted (i.e., they generalise to unseen data).

For each model show the model classification report and confusion matrix in your Jupyter notebook.

Split your dataset into a training set (80%) and a test set (20%) using the seed (random_state=8001).

Using the Python class provided in lib/R00000000_util.py, save the objects to the model folder as: [model/R00000000_model_1.pkl, model/R00000000_model_2.pkl., model/R00000000_model_3.pkl]

All code required to reproduce the modelling process should be placed in the Python library file: lib/R00000000_util.py and able to be called from the Jupyter Notebook: R00000000_A2_Notebook.ipynb.

The pickled model files should be loaded and called from the Jupyter Notebook and available to process unseen test data including any transformations required to ensure the models work. The models can be called from the Jupyter Notebook:

```
R00000000_model, news_category = util.load_run_model(model_id=model_id,
student_id=STUDENT_ID, news_headline=news_headline, news_article=news_article)
```

Report & Questions (15%)

Write a max 2-page report outlining the steps taken to complete the assignment. Identify any areas you feel are worth mentioning during the ETL, visualisation of modelling steps including any insights developed.

Answer 2 exam type questions (max 300 words) each. Note – due to the “open-book” nature of this assignment, a clean, concise and well-thought-out answer of your “own” viewpoint is expected, this is not a “cut and paste” exercise!

Presentation (5%)

Presentations for the assignment will take place on **Tuesday 4th May 2021** between 6pm and 10pm. Each student has 5 minutes to present their work. How you demonstrate your work is entirely up to you (e.g., PowerPoint, Jupyter notebooks, videos, mimes etc.). Only students present on the evening can be scored!

Note: DO NOT submit any PowerPoint files as part of your project submission, they will not be graded.

Submission Details

Assignments are due to be uploaded as a zip file via CIT Canvas no-later than **5pm on Monday 3rd May 2021**.

Students should upload a zip file with the same name as the downloaded zip file (e.g., R00000000.zip) containing their completed work containing **ONLY** the folders & files listed in Figure 2.

Files:

- **R00000000_A2_Notebook.ipynb** – completed notebook to call ETL and modelling processes.
- **R00000000_A2_Report.docx** – 2-page report and 2 answer exam type questions
- **data/R00000000_processed.csv** – clean dataset
- **lib/R00000000_util.py** – all the Python code required to recreate your work.
- **model/R00000000_model_1.pkl** – your pickled 1st model object (ML model and transformations)
- **model/R00000000_model_2.pkl** – your pickled 2nd model object (ML model and transformations)
- **model/R00000000_model_3.pkl** – your pickled 3rd model object (ML model and transformations)

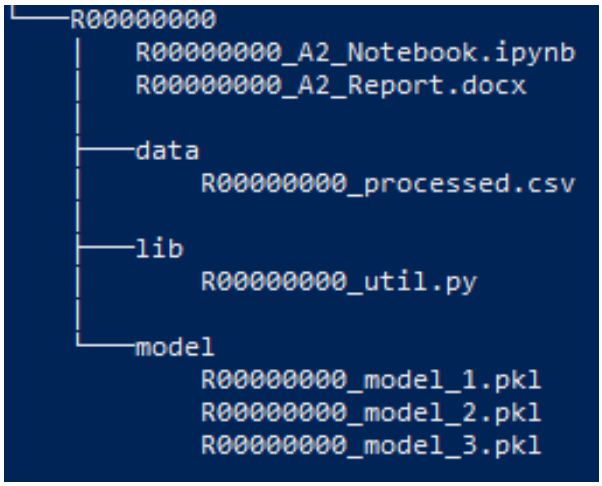

Figure 2 - Example submission Folder & Files