Technical Assignment

You should write code to solve all the tasks below. Your code can be submitted either as a singular script or as multiple scripts for each task, this is entirely your personal choice, and either is acceptable.

Ideally this assignment will be completed using either Python or R. If you have a strong desire to use a different coding language, then please contact us to discuss prior to starting the task.

Task 1

You have been sent an SQLite database file that includes three tables (posts, subforum, and users). Write code to open this database and create a data frame from the SQL code included.

Task 2

The `partition` column in your data frame defines which rows are to be used for training a model and which are to be used for testing a model. Using only the training data, build a model that uses the `text` column to predict the `forum_id`.

Your model does not have to be optimised or parameter tuned in anyway. You can build your model how you wish, however we recommend converting the text into vectors and then using the vectors to create a predictive classifier (all using the default parameters). For example using `sklearn.feature_extraction.text.TfidfVectorizer` and `sklearn.naive_bayes.MultinomialNB` if using Python, or `TfIdfVectorizer from library(superml)` and `multinomial_naive_bayes from library(naivebayes)` if using R.

Task 3

Use the test data from your data frame to create a score of how good your model is. NOTE: You will not be scored on the accuracy of your mode.

Task 4

The file `new_posts.json` contains a selection of new data to be applied to your model. Write code that:

- Converts the json file to a data frame where each line of the json file is a new row in the data frame,
- Run each row through your predictive model, adding a column to your data frame that defines the forum id your model says the post should be assigned too,
- Create a final 'reassign' column that contains a null value if your predicted forum id matches with the one that was in the original data, and your predicted id if there is a mismatch.
- Export the dataframe as a csv file.