**User Guide**

**General Notes:**

* I have attempted 2 approaches for part 1:
  + Wrote a customized code/script that identifies questions and non-questions (***Part1\_script.py***)
  + Implemented modelling approach using Logistic Regression model (***Part1\_Logistic Regression.py***)
* I have attempted 2 approaches for part 2:
  + Unsupervised learning: Implemented k-means clustering (***Part2\_K-Means\_Text.py***)
  + Supervised learning: Implemented multi-label classification model (***Part2\_MultilabelClassification.py***)
* I have provided the csv files for modelling for both the solutions (for reference if you want to use. The name of csv file for part1 is ‘***Part1Method1***.***csv***’ and for part 2 is ‘***Part2Method1.csv***’

**Notes for Part1.py and Part2.py solution approach:**

* In both Part 1 and Part2 question, the python code has 2 methods
  + For training the model
  + For using the trained model
* The 1st method will take 2 parameters:
  + Path for input csv file on which one wants to train the model
  + Path for output tsv file to see the result of predictions on test data
* The 2nd method will take 2 parameters:
  + Path for input csv file for which one has to use the model
  + Path for output tsv file to see the result of actual predictions

**Assumptions/Instructions to run the code:**

* Part 1
  + ***Part1\_Logistic Regression.py***: For 1st method , input Csv file for training the model will have 2 columns: 1st column has text and 2nd column has code(0 for non-questions and 1 for questions)
  + ***Part1\_Logistic Regression.py***: Input Csv file for 2nd method will have only 1 column that contains the text which user wants to identify into questions and non-questions
  + ***Part1\_script.py***: Input csv file will have only 1 column that has texts (questions and non-questions)
* Part 2
  + ***Part2\_MultilabelClassification.py***: I have considered part-2 as a multi-label classification problem. As a labels/categories/subtypes of question I have chosen 3 subtypes for ‘questions’ text: Animals, Countries and Famous Personalities
  + ***Part2\_MultilabelClassification.py***: For 1st method, input csv file will have 4 columns: 1st column-questions, 2nd column for Animals label (have value 0 or 1), 3rd column for Countries label and 4th column for famous personalities label
  + ***Part2\_MultilabelClassification.py***: For 2nd method, the input csv file will have only 1 column that contains the questions which user wants to identify that it belongs to which subtype/s.
  + ***Part2\_K-Means\_Text.py***: The input csv file will have 1 column that has texts (both questions and non-questions)