IEMS5780 Building and Deploying Scalable Machine Learning Services (Spring 2022)
Assignment 2

Instructions:

1. Do your own work. You are welcome to discuss the problems with your fellow classmates. Sharing ideas is great, and do write your own explanations.
2. All work should be submitted onto the blackboard before the due date.
3. You are advised to submit a single zip/rar file containing the following items.
   a. A .pdf file containing the answers of the written part.
   b. A .ipynb file storing your programs for data preparations and training the model.
   c. A .py file storing your telegram services.
   d. Two .pkl files storing your trained models.
4. Do type/write your work neatly. If we cannot read your work, we cannot grade your work.
5. If you do not put down your name, student ID in your submission, you will receive a 10% mark penalty out of the assignment 2.
6. Due date: ~~28th February, 2022~~ 7th March, 2022 (Monday) 23:59

Short questions (20%)
Answer all questions.

1. You are working for a e-commerce Website. Customers of the Website can purchase various kinds of products online, and they can leave comments on the products, which will be publicly available to any user of the Website. Your supervisor would like you to develop a text classification model to automatically category the comments of the users. Examples of categories include “product quality”, “usability of the Website”, “quality of delivery service” and “product price”.

   a. You are given a small sample of user comments collected on the Website. Suggest three pieces of information you would like to extract from this dataset (by performing some analysis), which will be useful for implementing the machine learning model later. (3%)

   b. Suggest some preprocessing steps that you will apply to the data before using them in your model training. (3%)

   c. Describe how you can convert a comment (represented as a string of characters) into a feature vector to be fed into the machine learning model. (4%)

   d. Name two different machine learning models that can be used to perform supervised learning in this task. Briefly describe how inference is done for these two models. (4%)

   e. You trained a model and it achieved 90% accuracy on the test dataset. (6%)
      i. What other metric(s) would you check to ensure that your model’s performance is good enough?
      ii. Give an example to explain why the model is bad even when it achieved an overall accuracy of 90%.

Practical problems (80%)
Work on the following problems.
If your Python programs cannot be run, you will receive no scores for the problem.
Put down your student ID as a comment in your first line of every program.

You are going to detect whether the title and message is real or fake. Data source has a shape of 7796×4. The first column identifies the news, the second and third are the title and text, and the fourth column has labels denoting whether the news is “REAL” or “FAKE”.
Link: https://drive.google.com/file/d/1er9NJTLUA3qnRuyhfzuN0XUsoIC4a-_q/view

2. Data Preprocessing (20%)

    - Download the data source, news.csv.
    - Split the dataset to 80% training set and 20% testing set.
    - Check and report the ratio of real-to-fake news are roughly the same in both training and testing sets.

3. Training Logistic Regression Models with Adding Bi-Grams to the Model (30%)

    - Prepare pipeline building up using sklearn’s CounterVectorizer and TfidfVectorizer.
    - Add bigram in both vectorizers.
    - Train logistic regression classifiers using the training set.
    - Compute (i) accuracy, (ii) precision and (iii) recall based on the testing set.
    - Save your models in a .pkl file using joblib.

4. Deploying the Model as a Telegram Bot (30%)

    - Ask botfather to create your own bot using the format IEMS5780_<studentID>_bot (e.g. IEMS5780_1155012345_bot).
    - Load the trained models in advance.
    - Ask the user to input title and message, and return both (i) predicted value and (ii) results.
        - If the predicted value < 0.4, then it is a fake message.
        - If the predicted value > 0.6, then it is a real message.
        - Otherwise, we cannot determine if the message is fake or real.

Remarks:
- You may take a look to the online assignment specification in some previous year: https://github.com/billzhonggz/IEMS5780/blob/master/Assignment%201/assignment_1.ipynb
- If your bot ID is used by others, please contact Haozheng for special arrangement on your bot ID.
- You need to submit your API key to your bot. However, you do not need to launch your service after submitting your work. Haozheng will launch and test your services privately.
- You will reuse this bot ID for future assignment submissions.
- You can see the demonstration on @IEMS5780_2022Spring_Demo_bot.