CloudSEK

ML Intern - Assignment

Question:

Headline Creation and Categorizing a given text are important tasks that require human intelligence and manual effort. Scaling up the task cannot be done without AI.

In this assignment you have two tasks :

1. **Headline Generation :** Given a news description, generate a headline for the news.
2. **Category Prediction :** Given a news description, predict the category of the news article.

Heads-up: You can use any approach/model but don't use a completely pretrained model. You can take a pretrained model and fine tune it for this task.

Dataset:

https://drive.google.com/drive/folders/1zl9upS84ap60Mw9QnJSXP8dh9lr-wnE3?usp=sharing

The dataset has around ~**210k** news articles split into **train dataset (170k)** and **test dataset (~40k)**. The two csv files have three columns :

1. **Headline :** The headline of the news article
2. **Category :** Category of the news article
3. **Summary :** Abstract of the news article

You need to predict the headline and category based on the news summary.

Instructions (to be followed strictly else your submission won't be considered)

1. Maintain below folder structure:

a. train_title_generation.py →Final code used for training the models shared with us.

b. train_category_prediction.py →Final code used for training the models shared with us.

c. test_title_generation.py →should load the model and emit title for given list of texts

d. test_category_prediction.py → should load the model and predict category for given list of texts

e. requirements.txt → libraries needed to execute your scripts

f. model.pkl → Two model files in pickle format

g. report.pdf → A small report briefly explaining your approach and the metrics chosen for measuring the performance and accuracy of the models

2. Share your zip file over the google drive with access to **lasya.ippagunta@cloudsek.com** and **bofin.babu@cloudsek.com**

3. The test script should log the accuracy of your model on the evaluation set kept with us.

4. Share your model and code in the format shared

5. Share over gdrive and name the folder as **headline-accuracy_category-accuracy_your-name.zip**, the **headline-accuracy** and **category-accuracy** here are your accuracy scores on your test data and name should be your name. For example : **89_95_lasya.zip**

Judgment:

1. Both your ML and instruction following skills are judged in this assignment.
2. How neatly can you write and structure your code.
3. The basic structure is already shared
4. Model Performance

Constraints:

You are free to use your favorite libraries/frameworks. Kindly avoid using Matlab or any paid tools.

Submission:

Within **4 days** from when the assignment was received.