Team1 Project
Raj, Sajad, Alan
Code's Documentation

README

IMAGE

The Python Notebook in this folder is commented but here is a gist of it :

- Please install Tensorflow in prior.
- Data is loaded from Google Drive. Edit code to match your dataset path accordingly.
- Images are read and stored as NumPy arrays.
- Images are normalized and the labels are converted to categorical type.
- Data is split and then fed into an instance of our model.
- Accuracy and loss curves are generated by the model's training history.
- The model instance is saved.

Instructions for deployment :

Run the Python notebook after making the necessary changes. Load the saved model in your script and feed it image data in the form of NumPy arrays in the exact dimensions as in the notebook (100*100*3) to generate outputs.

TEXT

Description:

The accompanied file, named '`Models_Based_On_Text.zip`', contains .py files to train machine learning models, namely Logistic and Linear Regression models, in order to predict: Gender and Big Five Personality Traits based on Text dataset. The dataset it is using is provided in two forms: raw .txt files and LIWC.csv file that captures analyses of these .txt files.

To Compile:

The accompanied file, named '`Models_Based_On_Text.zip`', contains three .py files as follows: ['`myTest_text_v3.py`', '`text_model.py`', '`text_model_v2.py`']. The files that have the word 'model' in its name perform two phases: training and testing. The file that contains the word 'myTest' in its name performs only the testing phase. This 'myTest' file is dependent on the other two files. 'text_model' is used to predict gender and 'text_model_v2' is used to predict personality traits. The expected sequence of execution is that you train the models first by running the files that has the word ‘model’ in its name, and then you test them by running the file that contains the word ‘myTest’ in its name if you want to test the models in a different dataset that has the same structure. ‘myTest’ file needs a specific command line args, please refer to the To Run section.

Before Training:

1. Make sure to have the following libs installed:

‘pandas’ , ‘sklearn’, ‘nltk’, ‘joblib’, and ‘xml.dom’.

2. Make sure to have the dataset, namely a ‘text’ folder containing all .txt files and LIWC.csv file, placed in a folder named ‘training’.

To Run:

Training & Testing: is running two .py files [‘`text_model.py`’, ‘`text_model_v2.py`’].

1. Assuming you put the dataset at the same folder where the python scripts are:

cd to directory and run the following commands from terminal:

1st

```
python3 text_model_v2.py
```

This would results in 10 .pkl files with the following names:

```
[‘support_e.pkl','text_model_e.pkl','support_n.pkl','text_model_n
.pkl', ‘support_a.pkl','text_model_a.pkl',
‘support_o.pkl','text_model_o.pkl',
‘support_c.pkl’,’text_model_c.pkl']
```

2nd

```
python3 text_model.py
```

This would results in 2 .pkl files with following names:

```
[‘tfid_vect.pkl’,'text_model.pkl']
```

Each command would also print scores in the terminal. Those scores represent testing on the same dataset after splitting it into training and testing.

Testing only:

This testing means that you have another dataset that has the same structure and you would like to test the generated models on it.

1. Assuming you put the dataset at the same folder where the python scripts are AND you have generated the models by running the previous commands:

   cd to directory and run the following command from terminal:

```
   python3 myTest_text_v3.py -i ~/<your-dataset-dir> -o
~/<output-dir>
```

This would result in generating xml files for each testing example you have provided in your-dataset.

LIKES

File Description

Here is the python file description for likes data analysis.

`userData.py`: Data structure for storing user's data

`inputTools.py`: Functions to input data

`outputTools.py`: Functions to output data

`learningTools.py`: Functions to Learn Age, Gender, LIWC data

`functionTest.py`: Contains test about gender and age prediction with likes data. In this file, likes data is dealt as a matrix in which columns are selected likes ids, gender and age. Five methods are implemented to train the data and do the prediction, which include MNB, logistic regression, CART, adaBoost and linear regression.

For running the project, you could run the start script as suggested in the manual.

```
tcss555 -i /data/public-test-data/ -o ~/output/
```

Or run the start script directly.

```
python fcuntionTest.py -i path/to/test/my-test-data/ -o
path/to/output/directory/
```