INSTRUCTIONS

Machine Learning Models for predicting age, gender, and personality of the user, based on text, image, and relational data ssets.

Text

Import numpy, keras, tensrflow, nltk(optional), and set up python jupyter notebooks with traditional conda or pip install, in order to emulate my environment. You can also utilize Google Colab, it may be more convinient. All of the code performing above the baseline is within the working_model folder, all of the other code is a draft that could be improved on or is not good for the provided form of text data, such as doc2vec, which did not perform well on the given textual dataset. To run in the linux command line with proper arguments, follow the format:

- start_bash [input directory with test dataset] [output directory]

Where the code containing the model, referenced by start_bash, can be modeled after our start.by python file in ensemble and contains proper calls to the machine learning algorithms running the prediction algorithm from the already trained machine learning models.

In order to train a text-based model, download and modify files located in main folder – src.

Working Models for **gender**: Logres over text with preprocessing (76%), BERT with simple NN (69%) , with simple NN (67%), Ensemble (81%), SVM (low accuracy)

Working Models for **age**: Logres over text with preprocessing (61%)

Brackets for gender:

- male = 0
- female = 1

Brackets for age:

- xx-24 = 0
- 25-34 = 1

- 35-49 = 2
- 50-xx = 3

Image

Requires Tensorflow, Keras, Pillow, and Numpy in a Python Environment. The code provided in GenderImageClassifer is a Convolutional Neural Network that is intended to classify gender.

Training a Model: Setup a folder for the CNN. Then copy the contents of GenderImageClassifer into this folder. Create a folder that will hold your training set. This folder must have subfolders that are labeled with the classifications.

Ex.

- image_postchange
  - 0
    - (All Your Male Example Images)
  - 1
    - (All Your Female Example Images)

Change the path to the training data (ex. "image_postchange/") in CatergoricalGenderClassiferCNN.py and optionally edit the settings. Run CatergoricalGenderClassiferCNN.py with Python.

Making Predictions: Put an image in the same folder as ModelTestCategoricalClassifer.py and make sure that the model that was trained is present there as well (Named "image_model" by default). Edit the path to the image in ModelTestCategoricalClassifer.py Run ModelTestCategoricalClassifer.py with Python.

Relations

Relations related code was located at files: "Relations_KNN", "Relations_LogisticRegression", and "Relations_SoftmaxRegression" were used for internal testing and training only and the code would not be runnable unless the correct file paths were added in the code where necessary, pointing to appropriately formatted data. Parts of the code were also commented and uncommented depending on which

attribute was being targeted at the time. If you wish to run the code, comments have been left indicating what file paths should be passed.

"Relations_LogisticRegression" and "Relations_SoftmaxRegression" both require numpy and output their models as a text file.