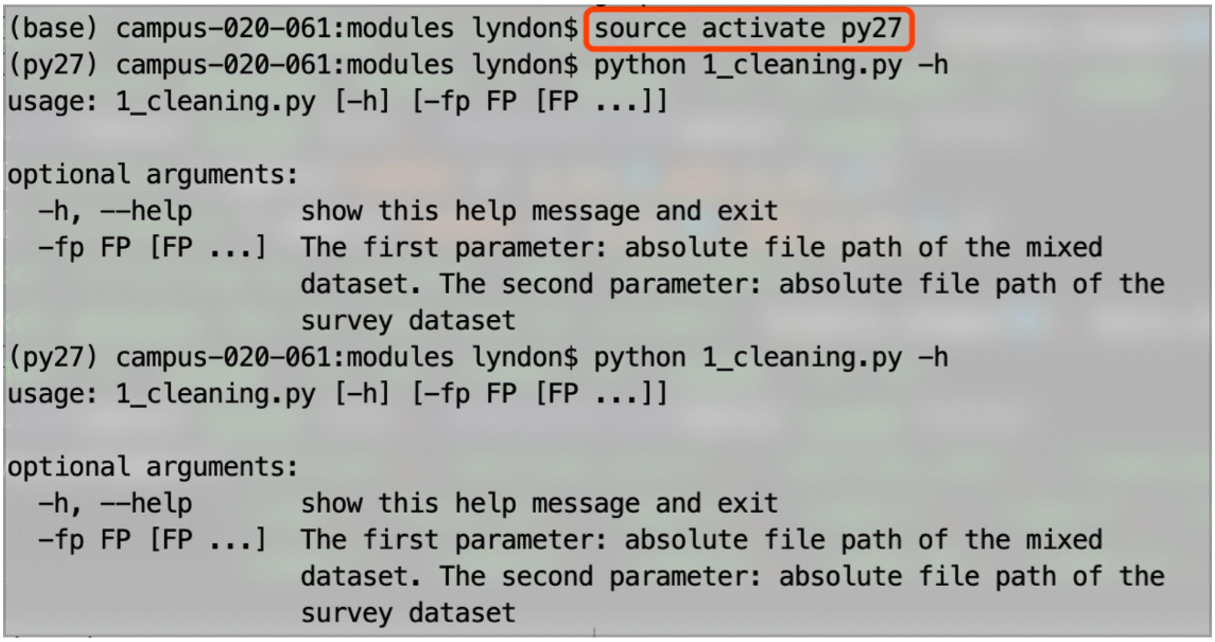
Note for the Gaze Project

> This is a `README` file for predicting the data from Gaze Project
> Written by Lyndon, built on Feb. 8, 2020
> Revised by Lyndon on Feb. 25, 2020

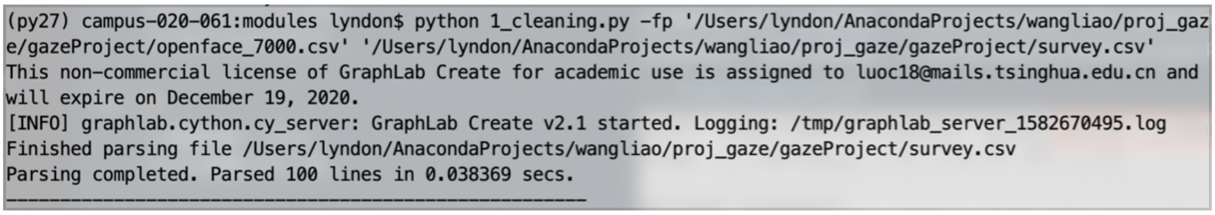
Step one: clean the data

1. Script name: `1_cleaning.py`
2. This script is for data cleaning
3. Please do data cleaning on your PC (local env), using `graphlab` in Python 2.x env, because there is something wrong with `graphlab` in the remote server. Also, please do not try the `pandas` accelerator `modin.pandas`, it will encroach all memory in your PC.
   - Install GraphLab
4. How to run it?
   - For help



```
(base) campus-020-061:modules lyndon$ source activate py27
(py27) campus-020-061:modules lyndon$ python 1_cleaning.py -h
usage: 1_cleaning.py [-h] [-fp FP [FP ...]]

optional arguments:
  -h, --help        show this help message and exit
  -fp FP [FP ...]   The first parameter: absolute file path of the mixed
                    dataset. The second parameter: absolute file path of the
                    survey dataset
(py27) campus-020-061:modules lyndon$ python 1_cleaning.py -h
usage: 1_cleaning.py [-h] [-fp FP [FP ...]]

optional arguments:
  -h, --help        show this help message and exit
  -fp FP [FP ...]   The first parameter: absolute file path of the mixed
                    dataset. The second parameter: absolute file path of the
                    survey dataset
```

   - Execute



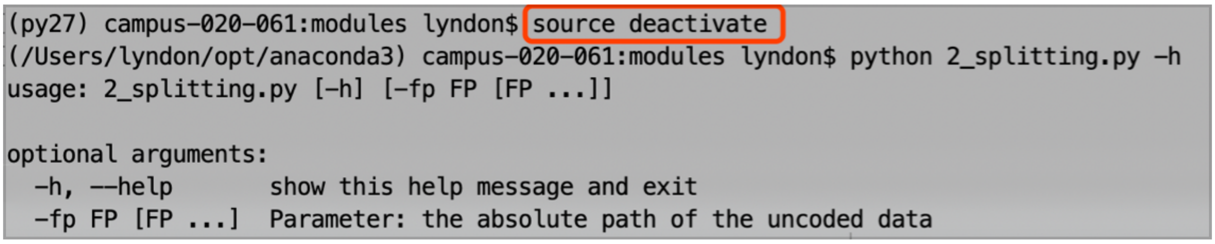
```
(py27) campus-020-061:modules lyndon$ python 1_cleaning.py -fp '/Users/lyndon/AnacondaProjects/wangliao/proj_gaz
e/gazeProject/openface_7000.csv' '/Users/lyndon/AnacondaProjects/wangliao/proj_gaze/gazeProject/survey.csv'
This non-commercial license of GraphLab Create for academic use is assigned to luoc18@mails.tsinghua.edu.cn and
will expire on December 19, 2020.
[INFO] graphlab.cython.cy_server: GraphLab Create v2.1 started. Logging: /tmp/graphlab_server_1582670495.log
Finished parsing file /Users/lyndon/AnacondaProjects/wangliao/proj_gaze/gazeProject/survey.csv
Parsing completed. Parsed 100 lines in 0.038369 secs.
------------------------------------------------
```

5. Output: Two new `csv` files (`coded.csv` , `uncoded.csv`) will be generated under the same file path of the mixed dataset

---

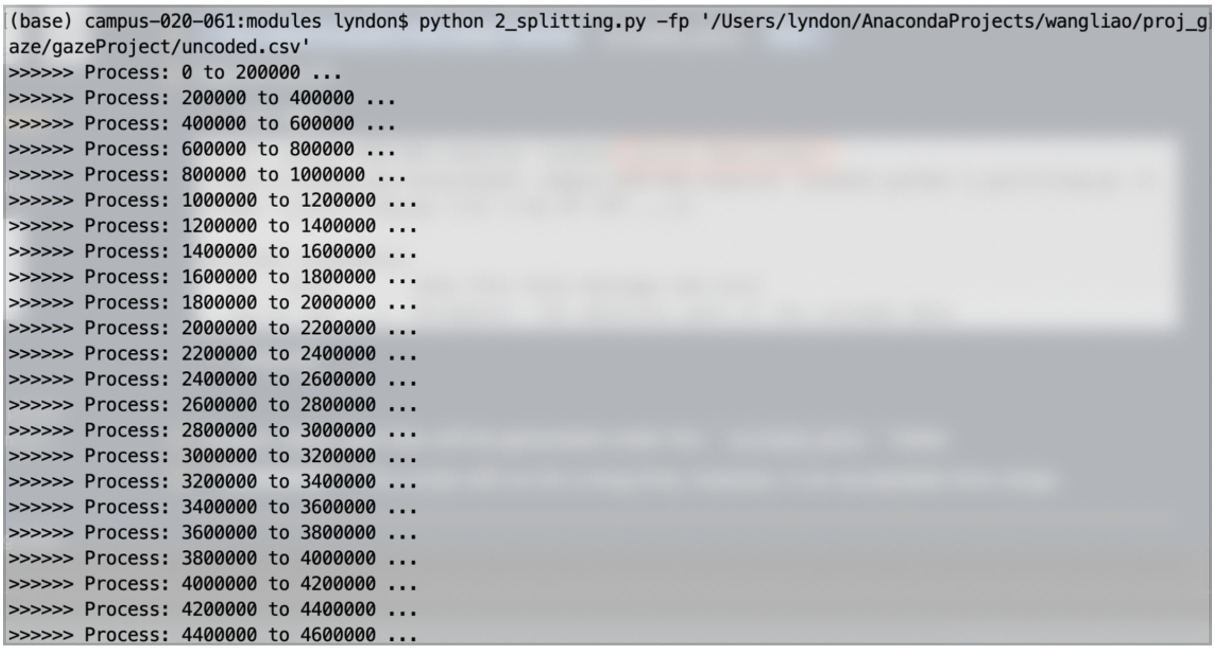
Step two: split the uncoded data

1. Script name: `2_splitting.py`
2. This script is for splitting the huge amounts of uncoded data
3. You need to build a file folder named `uncoded_data` first
4. How to run it?
   - For help



```
(py27) campus-020-061:modules lyndon$ source deactivate
(/Users/lyndon/opt/anaconda3) campus-020-061:modules lyndon$ python 2_splitting.py -h
usage: 2_splitting.py [-h] [-fp FP [FP ...]]

optional arguments:
  -h, --help        show this help message and exit
  -fp FP [FP ...]   Parameter: the absolute path of the uncoded data
```

   - Execute

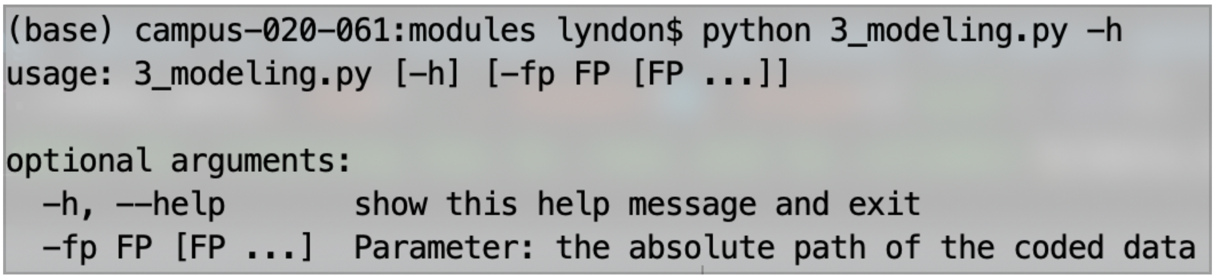


```
(base) campus-020-061:modules lyndon$ python 2_splitting.py -fp '/Users/lyndon/AnacondaProjects/wangliao/proj_g
aze/gazeProject/uncoded.csv'
>>>>>> Process: 0 to 200000 ...
>>>>>> Process: 200000 to 400000 ...
>>>>>> Process: 400000 to 600000 ...
>>>>>> Process: 600000 to 800000 ...
>>>>>> Process: 800000 to 1000000 ...
>>>>>> Process: 1000000 to 1200000 ...
>>>>>> Process: 1200000 to 1400000 ...
>>>>>> Process: 1400000 to 1600000 ...
>>>>>> Process: 1600000 to 1800000 ...
>>>>>> Process: 1800000 to 2000000 ...
>>>>>> Process: 2000000 to 2200000 ...
>>>>>> Process: 2200000 to 2400000 ...
>>>>>> Process: 2400000 to 2600000 ...
>>>>>> Process: 2600000 to 2800000 ...
>>>>>> Process: 2800000 to 3000000 ...
>>>>>> Process: 3000000 to 3200000 ...
>>>>>> Process: 3200000 to 3400000 ...
>>>>>> Process: 3400000 to 3600000 ...
>>>>>> Process: 3600000 to 3800000 ...
>>>>>> Process: 3800000 to 4000000 ...
>>>>>> Process: 4000000 to 4200000 ...
>>>>>> Process: 4200000 to 4400000 ...
>>>>>> Process: 4400000 to 4600000 ...
```

5. Output: Many split files will be generated under the `uncoded_data` folder
6. **WARNING**: This script will run for a long time, however, in an acceptable time range.

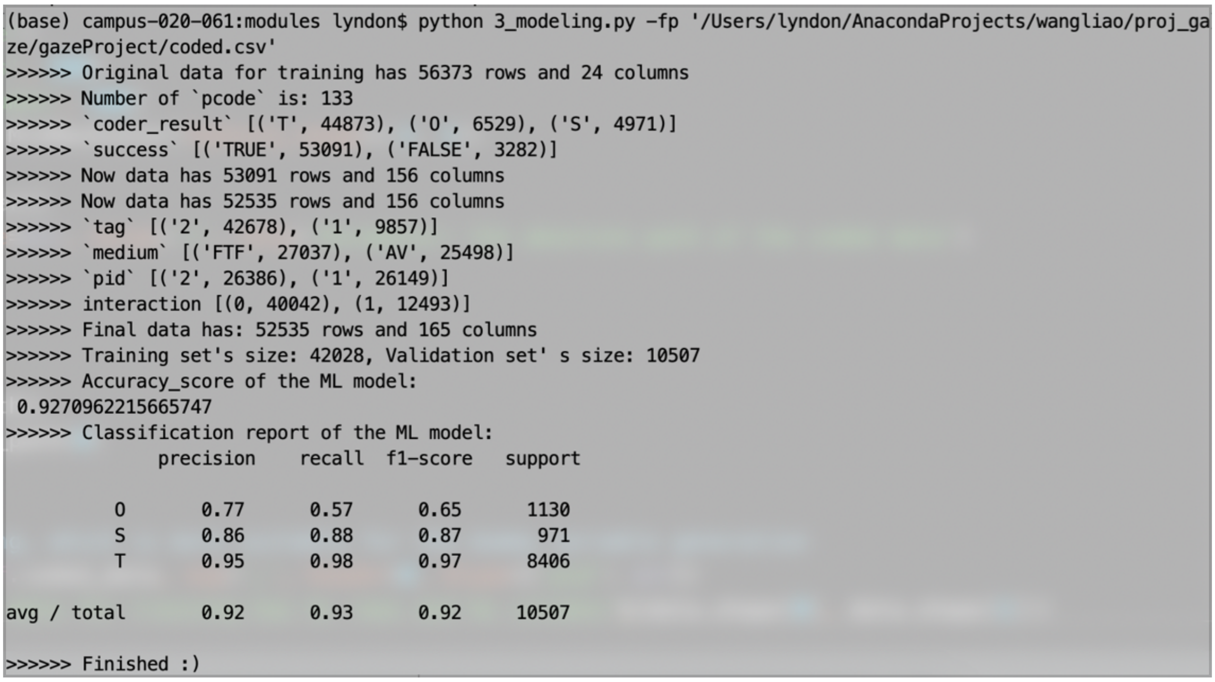
Step three: build the ML model

1. Script name: `3_modeling.py`
2. This script is for training the model using coded data
3. How to run it?
   - For help



```
(base) campus-020-061:modules lyndon$ python 3_modeling.py -h
usage: 3_modeling.py [-h] [-fp FP [FP ...]]

optional arguments:
  -h, --help        show this help message and exit
  -fp FP [FP ...]   Parameter: the absolute path of the coded data
```

   - Execute



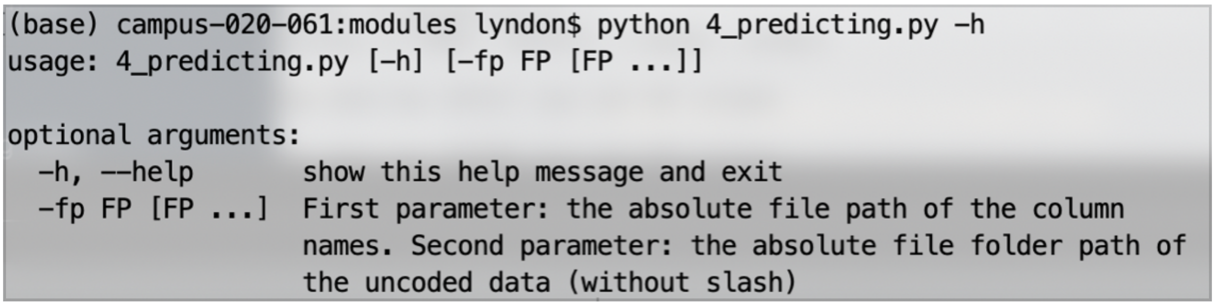
```
(base) campus-020-061:modules lyndon$ python 3_modeling.py -fp '/Users/lyndon/AnacondaProjects/wangliao/proj_ga
ze/gazeProject/coded.csv'
>>>>>> Original data for training has 56373 rows and 24 columns
>>>>>> Number of `pcode` is: 133
>>>>>> `coder_result` [('T', 44873), ('O', 6529), ('S', 4971)]
>>>>>> `success` [('TRUE', 53091), ('FALSE', 3282)]
>>>>>> Now data has 53091 rows and 156 columns
>>>>>> Now data has 52535 rows and 156 columns
>>>>>> `tag` [('2', 42678), ('1', 9857)]
>>>>>> `medium` [('FTF', 27037), ('AV', 25498)]
>>>>>> `pid` [('2', 26386), ('1', 26149)]
>>>>>> interaction [(0, 40042), (1, 12493)]
>>>>>> Final data has: 52535 rows and 165 columns
>>>>>> Training set's size: 42028, Validation set' s size: 10507
>>>>>> Accuracy_score of the ML model:
 0.9270962215665747
>>>>>> Classification report of the ML model:
              precision    recall  f1-score   support

          O       0.77      0.57      0.65      1130
          S       0.86      0.88      0.87       971
          T       0.95      0.98      0.97      8406

avg / total       0.92      0.93      0.92     10507

>>>>>> Finished :)
```

4. Output
   - A text file named `column_name.txt` contains the column names, which is necessary for building the predicting data, it will be saved under the folder contains the coded data
   - Two kinds of serialized model, under the folder contains the coded data

---

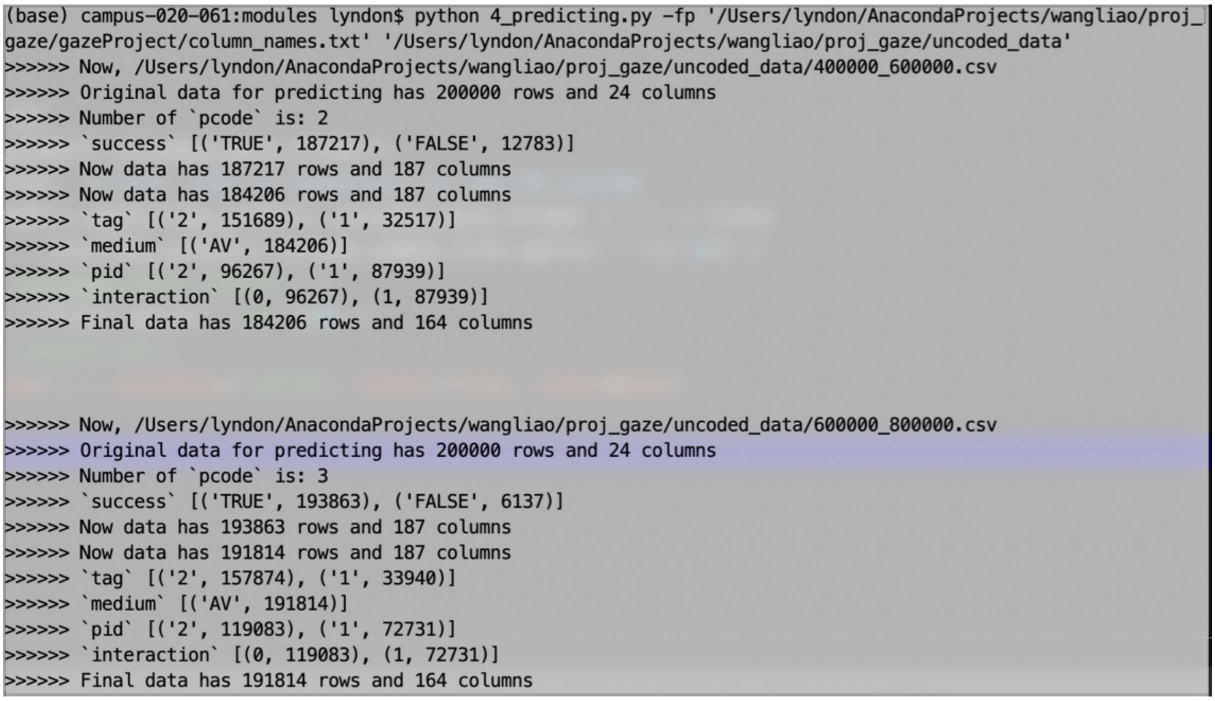
Step four: predict the uncoded data

1. Script name: `4_predicting.py`
2. This script is for predicting the uncoded data (from the `uncoded_data` folder)
3. You need to build a file folder named `uncoded_data_result` first
4. How to run it?
   - For help



```
(base) campus-020-061:modules lyndon$ python 4_predicting.py -h
usage: 4_predicting.py [-h] [-fp FP [FP ...]]

optional arguments:
  -h, --help        show this help message and exit
  -fp FP [FP ...]   First parameter: the absolute file path of the column
                    names. Second parameter: the absolute file folder path of
                    the uncoded data (without slash)
```

   - Execute



```
(base) campus-020-061:modules lyndon$ python 4_predicting.py -fp '/Users/lyndon/AnacondaProjects/wangliao/proj_
gaze/gazeProject/column_names.txt' '/Users/lyndon/AnacondaProjects/wangliao/proj_gaze/uncoded_data'
>>>>>> Now, /Users/lyndon/AnacondaProjects/wangliao/proj_gaze/uncoded_data/400000_600000.csv
>>>>>> Original data for predicting has 200000 rows and 24 columns
>>>>>> Number of `pcode` is: 2
>>>>>> `success` [('TRUE', 187217), ('FALSE', 12783)]
>>>>>> Now data has 187217 rows and 187 columns
>>>>>> Now data has 184206 rows and 187 columns
>>>>>> `tag` [('2', 151689), ('1', 32517)]
>>>>>> `medium` [('AV', 184206)]
>>>>>> `pid` [('2', 96267), ('1', 87939)]
>>>>>> `interaction` [(0, 96267), (1, 87939)]
>>>>>> Final data has 184206 rows and 164 columns


>>>>>> Now, /Users/lyndon/AnacondaProjects/wangliao/proj_gaze/uncoded_data/600000_800000.csv
>>>>>> Original data for predicting has 200000 rows and 24 columns
>>>>>> Number of `pcode` is: 3
>>>>>> `success` [('TRUE', 193863), ('FALSE', 6137)]
>>>>>> Now data has 193863 rows and 187 columns
>>>>>> Now data has 191814 rows and 187 columns
>>>>>> `tag` [('2', 157874), ('1', 33940)]
>>>>>> `medium` [('AV', 191814)]
>>>>>> `pid` [('2', 119083), ('1', 72731)]
>>>>>> `interaction` [(0, 119083), (1, 72731)]
>>>>>> Final data has 191814 rows and 164 columns
```

5. Output: Predictions will be saved in the `uncoded_data_result` folder