BIA660 Team 4 – Final Project
Jungho Park, Devangi Rajput, Jeel Sutaria, Manasa Prakash, Vadhish Parikh

Instructions on how to run scripts

The following scripts are included in our project:

- 01_Scraping.py
    - Scrape full job descriptions from indeed.com.
    - Uses selenium for scraping
    - Must have valid chromedriver in path
    - The data will be saved in 'filedir/data' folder
- 02_Preprocessing.py
    - Contains merge_csv, load_csv, vectorize_job_desc_data, vectorize_job_title_data, save_vectors
    - The script will make 'combined_jobs.csv' file inside 'data' folder
    - Csv file will be transformed into dataframe and divided into 'all_X' and 'all_Y' for vectorization
    - Main function will execute functions and split train and test set
    - Each X and Y will be vectorized by relevant functions
    - Vectors will be saved in 'filedir/data/vecdata' folder
- 03_Classificaiton.py
    - This script performs importing vectors for training, gridsearch, and voting classifier
    - Best parameters will be automatically feed into the classification algorithm
    - KNN, Decision Tree, and Logistic regression models are used
    - In console, best params for each model, best score for each model, voting classifier accuracy score, and confusion matrix will be shown
    - In 'filedir/results' folder, 'results.csv' file will be created that includes the predicted label for each line in the test file.
        - 0 = 'data scientist'
        - 1 = 'software engineer'
    - In 'filedir/results' folder, confusion_matrix.png will be created

**** This script was running and executed in local Anaconda-Spyder**
**** Next page has detailed instructions**

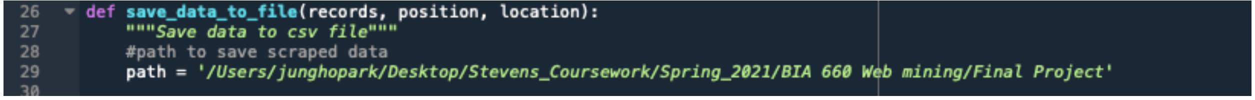
BIA660 Team 4 – Final Project
Jungho Park, Devangi Rajput, Jeel Sutaria, Manasa Prakash, Vadhish Parikh

Step by step instructions for running scripts

1. 01_Scraping.py

To execute scraping, please change the following

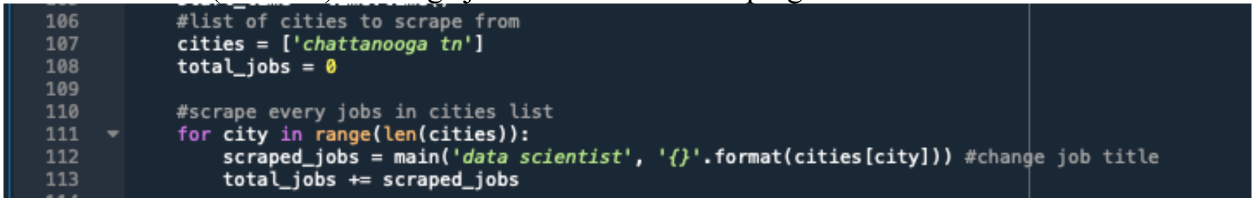
- Filepath (line 29) – where you want to store data in local machine



```
26  def save_data_to_file(records, position, location):
27      """Save data to csv file"""
28      #path to save scraped data
29      path = '/Users/junghopark/Desktop/Stevens_Coursework/Spring_2021/BIA 660 Web mining/Final Project'
30
```

- Cities array (line 107) – list of city names for scraping

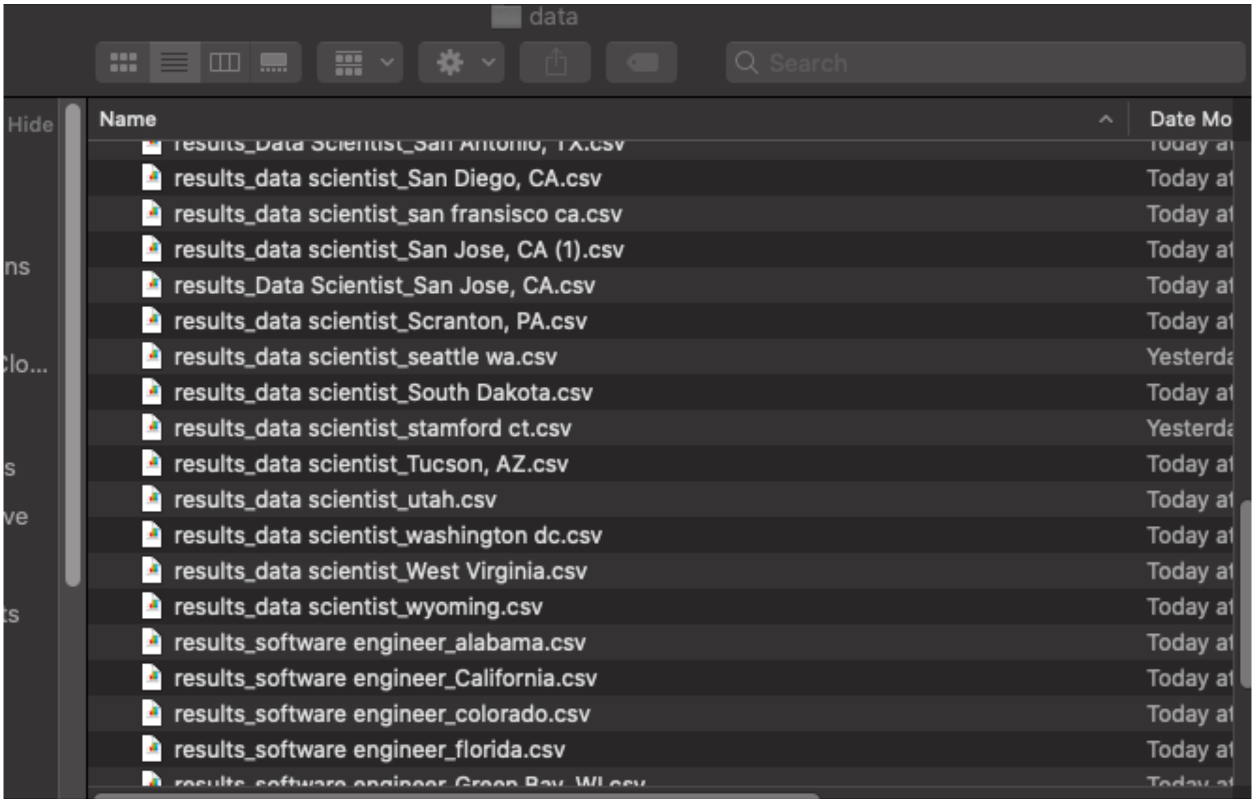
- Job title (line 112) – change job title name for scraping



```
106         #list of cities to scrape from
107         cities = ['chattanooga tn']
108         total_jobs = 0
109
110         #scrape every jobs in cities list
111         for city in range(len(cities)):
112             scraped_jobs = main('data scientist', '{}'.format(cities[city])) #change job title
113             total_jobs += scraped_jobs
```

- Run Script

- Running Scripts will store scraped data in 'filedir/data' folder

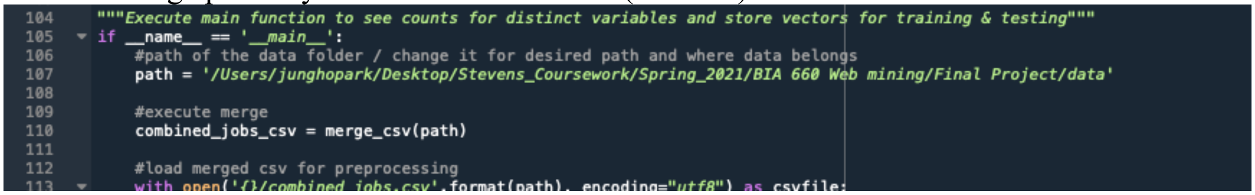


```
results_Data Scientist_San Antonio, TX.csv          Today at
results_data scientist_San Diego, CA.csv            Today at
results_data scientist_san fransisco ca.csv         Today at
results_data scientist_San Jose, CA (1).csv         Today at
results_Data Scientist_San Jose, CA.csv             Today at
results_data scientist_Scranton, PA.csv             Today at
results_data scientist_seattle wa.csv               Yesterda
results_data scientist_South Dakota.csv             Today at
results_data scientist_stamford ct.csv              Yesterda
results_data scientist_Tucson, AZ.csv               Today at
results_data scientist_utah.csv                     Today at
results_data scientist_washington dc.csv            Today at
results_data scientist_West Virginia.csv            Today at
results_data scientist_wyoming.csv                  Today at
results_software engineer_alabama.csv               Today at
results_software engineer_California.csv            Today at
results_software engineer_colorado.csv              Today at
results_software engineer_florida.csv               Today at
results_software engineer_Green Bay, WI.csv         Today at
```

BIA660 Team 4 – Final Project
Jungho Park, Devangi Rajput, Jeel Sutaria, Manasa Prakash, Vadhish Parikh

2. 02_Preprocessing.py

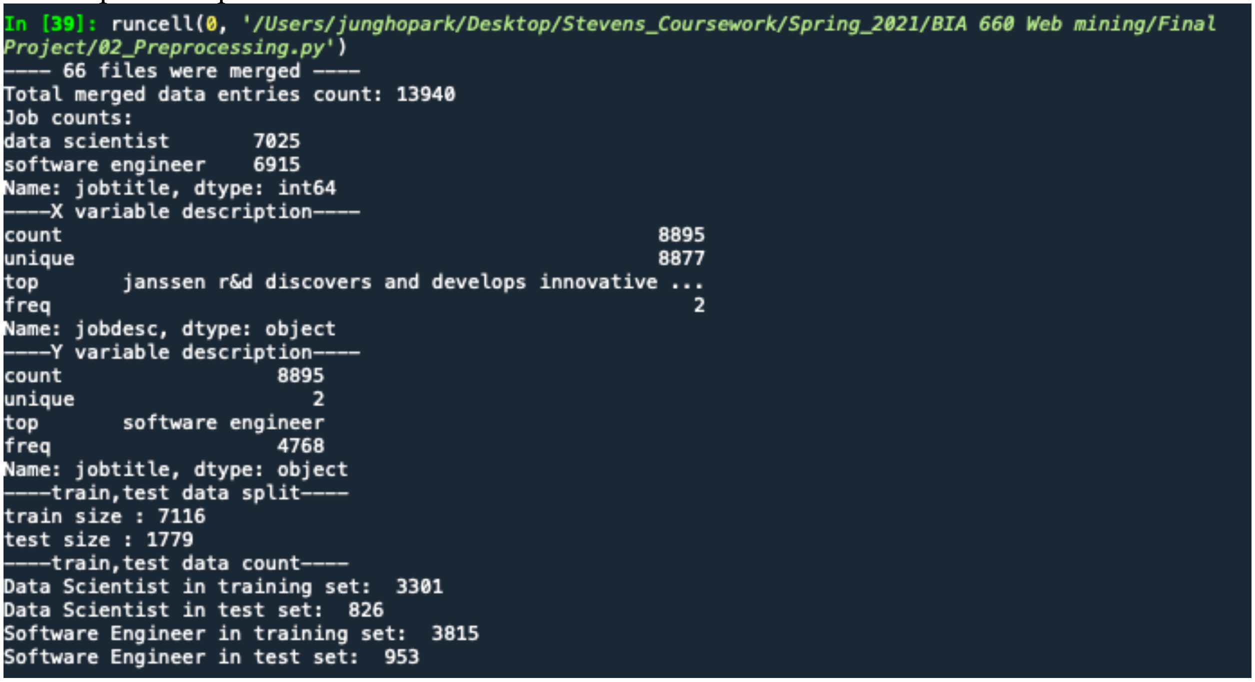
- Change path to your desired local folder (line 107)



```
104  """Execute main function to see counts for distinct variables and store vectors for training & testing"""
105  if __name__ == '__main__':
106      #path of the data folder / change it for desired path and where data belongs
107      path = '/Users/junghopark/Desktop/Stevens_Coursework/Spring_2021/BIA 660 Web mining/Final Project/data'
108  
109      #execute merge
110      combined_jobs_csv = merge_csv(path)
111  
112      #load merged csv for preprocessing
113      with open('{}/combined_jobs.csv'.format(path), encoding="utf8") as csvfile:
```

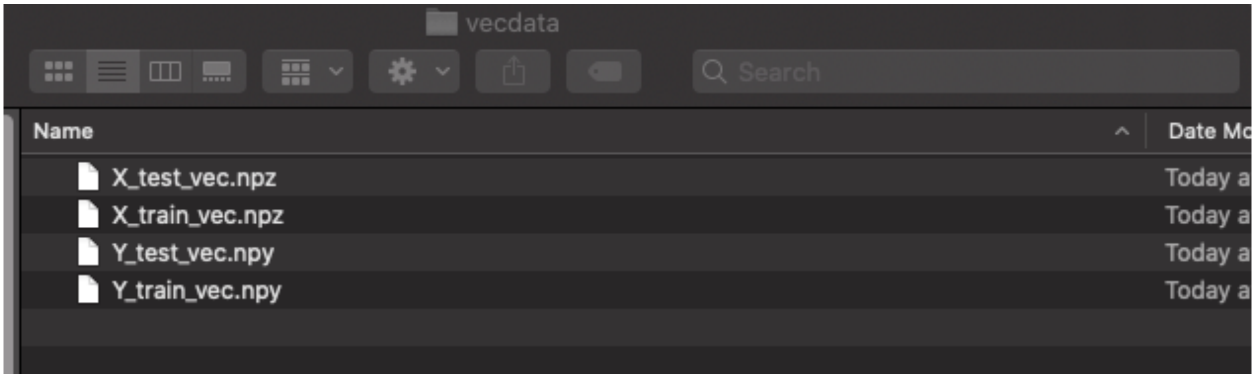
- Run 02_Preprocessing.py
    - Will store 'combined_jobs.csv' in 'filedir/data' folder
    - Will store X and Y vectors in 'filedir/data/vecdata' folder

- Output example



```
In [39]: runcell(0, '/Users/junghopark/Desktop/Stevens_Coursework/Spring_2021/BIA 660 Web mining/Final
Project/02_Preprocessing.py')
---- 66 files were merged ----
Total merged data entries count: 13940
Job counts:
data scientist       7025
software engineer    6915
Name: jobtitle, dtype: int64
----X variable description----
count                                                   8895
unique                                                  8877
top       janssen r&d discovers and develops innovative ...
freq                                                       2
Name: jobdesc, dtype: object
----Y variable description----
count                  8895
unique                    2
top       software engineer
freq                   4768
Name: jobtitle, dtype: object
----train,test data split----
train size : 7116
test size : 1779
----train,test data count----
Data Scientist in training set:  3301
Data Scientist in test set:  826
Software Engineer in training set:  3815
Software Engineer in test set:  953
```

- Vecdata folder

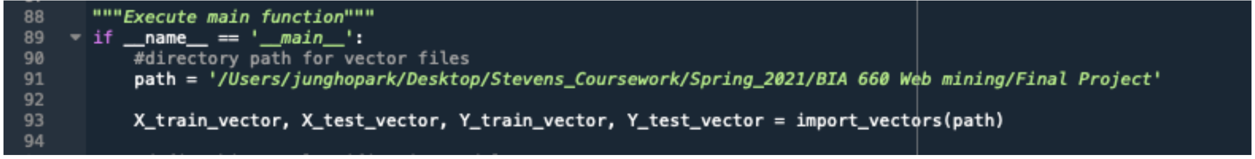


X_test_vec.npz — Today a
X_train_vec.npz — Today a
Y_test_vec.npy — Today a
Y_train_vec.npy — Today a

3. 03_Classificaiton.py

BIA660 Team 4 – Final Project
Jungho Park, Devangi Rajput, Jeel Sutaria, Manasa Prakash, Vadhish Parikh

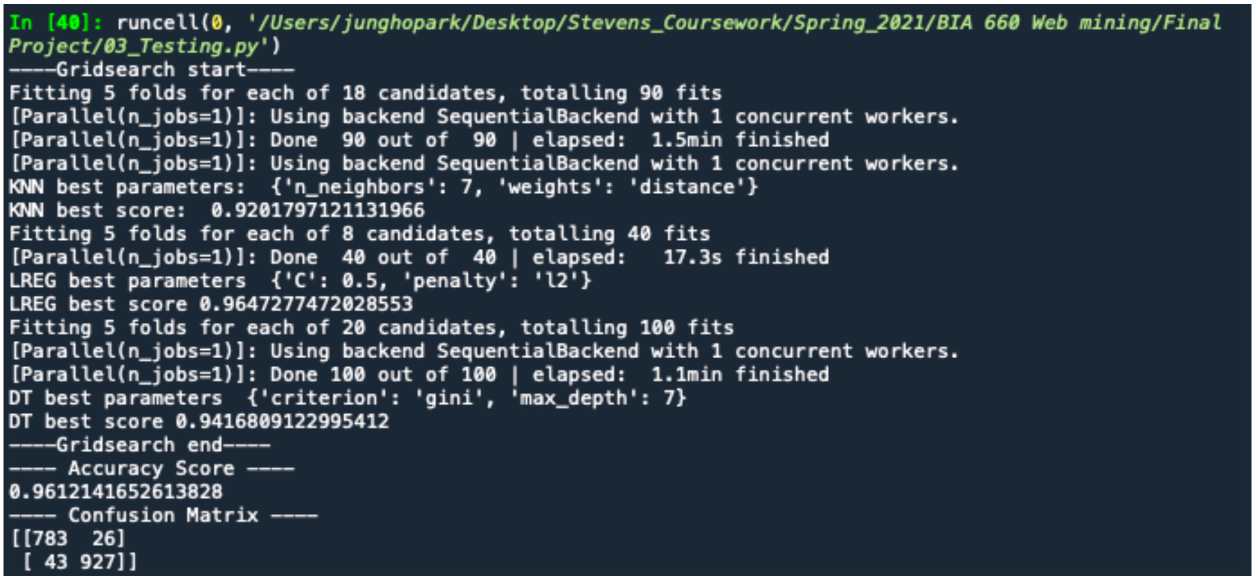
- Change path to desired local folder (line 91)



```
88  """Execute main function"""
89  if __name__ == '__main__':
90      #directory path for vector files
91      path = '/Users/junghopark/Desktop/Stevens_Coursework/Spring_2021/BIA 660 Web mining/Final Project'
92  
93      X_train_vector, X_test_vector, Y_train_vector, Y_test_vector = import_vectors(path)
94  
```

- Run 03_Classification.py

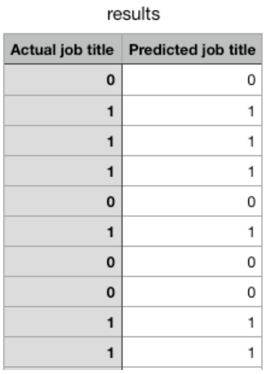
- Console output: show gridsearch result, VT accuracy and confusion matrix



```
In [40]: runcell(0, '/Users/junghopark/Desktop/Stevens_Coursework/Spring_2021/BIA 660 Web mining/Final
Project/03_Testing.py')
----Gridsearch start----
Fitting 5 folds for each of 18 candidates, totalling 90 fits
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done  90 out of  90 | elapsed:  1.5min finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
KNN best parameters:  {'n_neighbors': 7, 'weights': 'distance'}
KNN best score:  0.9201797121131966
Fitting 5 folds for each of 8 candidates, totalling 40 fits
[Parallel(n_jobs=1)]: Done  40 out of  40 | elapsed:   17.3s finished
LREG best parameters  {'C': 0.5, 'penalty': 'l2'}
LREG best score 0.9647277472028553
Fitting 5 folds for each of 20 candidates, totalling 100 fits
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:  1.1min finished
DT best parameters  {'criterion': 'gini', 'max_depth': 7}
DT best score 0.9416809122995412
----Gridsearch end----
---- Accuracy Score ----
0.9612141652613828
---- Confusion Matrix ----
[[783  26]
 [ 43 927]]
```

- New csv file that includes predicted label for test file and confusion_matrix.png will be stored in ‘filedir/results’

Csv file

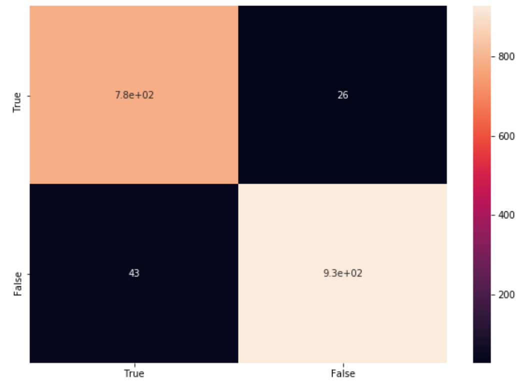


results

Actual job title	Predicted job title
0	0
1	1
1	1
1	1
0	0
1	1
0	0
0	0
1	1
1	1

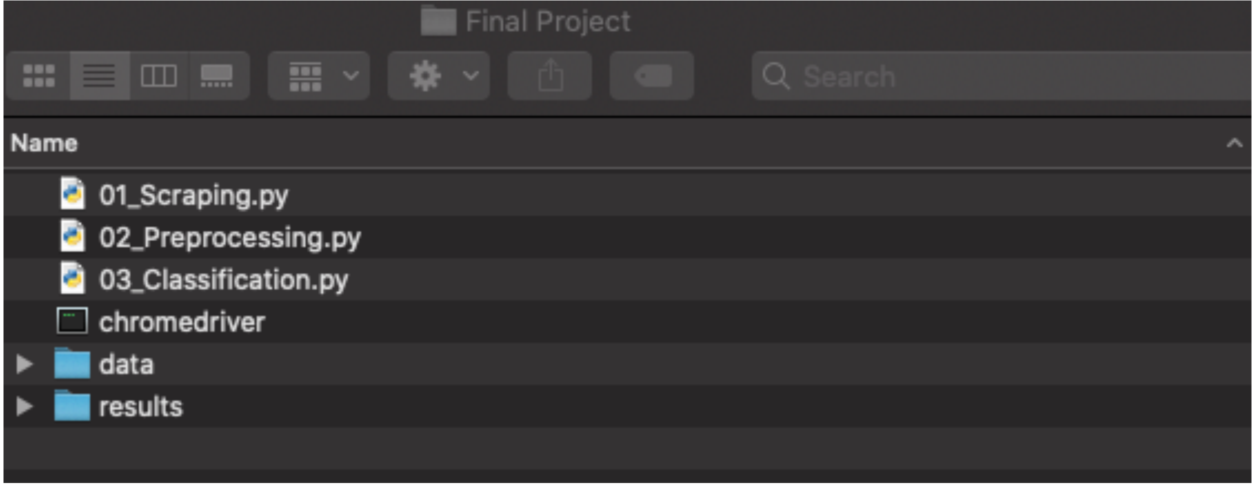
**** 0 = ‘data scientist’ , 1 = ‘software engineer’**

BIA660 Team 4 – Final Project
Jungho Park, Devangi Rajput, Jeel Sutaria, Manasa Prakash, Vadhish Parikh



Confusion_matrix.png

‘Filedir’ Folder outlook with outputs and scripts



01_Scraping.py
02_Preprocessing.py
03_Classification.py
chromedriver
data
results