Final project instruction

1. Classification script

Train the model and create prediction output

1. Use the given ‘training_set’ file to generate a classifier and then do a prediction
2. Please change the file name for testing if you need

location:

Notice

load the testing data hear, could change the file name

```
In [8]:
1 #### load the testing data, could change the file name here!
2 test_text=loadTest('test.csv')
```

3. Create a new CSV file named ‘pred_result’ including the original job description and a new ‘Title’ column.

2. Scraping script

Scrape the text (job description) and title from the Indeed website

1. Collected 5099 Ads for Data Scientist
2. Collected 5144 Ads for Software Engineer
3. Extract the text from the HTML and create a CSV named ‘training_set’
4. Because of the robot detection of anti spider mechanism on the Indeed website, we need to run scrape codes many times, we change URLs and the number many times to add all data to the final file

```
# need to change urls as needed
sde_url = 'https://www.indeed.com/jobs?q=Software%20Engineer&l=California&vjk=28c2b38ed84e3b75'
ds_url = 'https://www.indeed.com/jobs?q=data%20scientist&l=California&vjk=5ef83d01db120213'
# sde_url = 'https://www.indeed.com/jobs?q=Software+Engineer&l=California&start=10'
# ds_url = 'https://www.indeed.com/jobs?q=data+scientist&l=California&start=10'

# define the number of jobs when scraping, do not set this number over 800
# because it may cause robot detection of anti spider and the program crash
number = 500
#number = 100
```