CS 6364-002 Homework 1

August 24, 2022

Deadline for submission: **Sep-5-2022**.

REQUIREMENT: You have 9 tasks to finish, please write all codes in one Jupyter notebook. In this Jupyterlab notebook, you can write the python codes for each specific task in a single cell. Please run each cell to generate the output of your codes in the notebook, before you save and submit the notebook.

(`https://www.dataquest.io/blog/jupyter-notebook-tutorial/`)

Check the following example Jupyter notebook for the suggested format: `https://github.com/chenf79/test/blob/master/CS6364_HW1_Example_format.ipynb`

- **Task 1: Test Python Environment**

  Try to install python3 environment, copy and paste following codes into in one cell in the Jupyter notebook (note that python is indentation sensitive), and run it in the terminal with ”python hw1.py” or in your preferred IDE, such as Pycharm or Anaconda.

```
def task1():
    print ”hello world”

if __name__ == '__main__':
    task1()
```

  Note, when you copy and paste the code, please be careful with the proper quotation marks.

- **Task 2: Define Object**

  Define a function called ’task2’ to assign items= [1, 2, 3, 4, 5] as a list object, and print the list.

- **Task 3: File Reading**

  Create a file called ’task3.data.’ and type string ’1 2 3 4 5 6 7 8 9 10’ in it without quotation marks, and then write a function to read the file and load the string as two list-objects items1 = [1, 2, 3, 4, 5], items2 = [6, 7, 8, 9, 10]. Print the lists finally.

- **Task 4: Data Structure**

  It is important to be familiar with the functions of dictionary: items(), keys(), values(), write a program to use all these functions.

  Notes: Dictionary has two ways to initialize
  data = dict()
  data[’school’] = ’UAlbany’
  data[’address’] = ’1400 Washington Ave, Albany, NY 12222’
  data[’phone’] = ’(518) 442-3300’
  data = {’school’: ’UAlbany’, ’address’: ’1400 Washington Ave, Albany, NY 12222’, ’phone’: ’(518) 442-3300’}
  Print the results as follows by accessing the defined dictionary.
  school: UAlbany
  address: 1400 Washington Ave, Albany, NY 12222
  phone: (518) 442-3300

- **Task 5: Data Serialization**

  Use json to store above dictionary in task-5 and then print item, key and values.
  Notes: This task tells how to store a dictionary object to a file and then load the dictionary object from the file. In python, object of any type can be mostly saved in json format to a txt file.

  You need to be familiar with the following two json functions : json.dumps(object), which dumps an object to a json format (string), json.loads(a json format string), which loads a json format string back to the original object. We do not care about if the original object is list, dictionary or others. json.dumps() can automatically recognize.

  Note, json.load(object) only can only load an object, not a file.

- **Task 6: Data Serialization**

  Store a number of different types objects (e.g., list, dictionary, array) to a file and then load the objects from the file.

  Write a function to dump list object items = [1,2,3,4,5] and above dictionary in task-5 to a file called 'task6.data', and then load them from the same file and print.

- **Task 7: Data Preprocessing**

  Read the tweets from the file "CrimeReport.txt" and print the id for each tweet.

  Here are some functions that you will use in the task: open().readlines(), tweet = json.loads(), print tweet.keys(), you will know the keys of tweet dictionary object, then you can find which key relates to tweet id, and you can then retrieve the id of this specific tweet.

- **Task 8: Data Preprocessing: tweets filtering**

  INPUT: "CrimeReport.txt"
  OUTPUT: a file "task8.data" that stores the 10 most recent tweets
  ****
  Suggestions:
  tweet[$'created - at'$] gives the created time of this tweet. Rank tweets based on the time from the earliest to the most recent. Then we can identify the 10 most recent tweets. Some example lines that are **not** directly runnable

  ```
  import datetime
  tweets = []
  for line in open().readlines():
      tweet = json.loads(line)
      tweets.append(tweet)
      #datetime.datetime.strptime(item['created-at'], '%a %b %d %H:%M:%S +0000 %Y')
      #converts the string format of a date time to the datetime object
  sorted_tweets = sorted(tweets, key = lambda item:
      datetime.datetime.strptime(item['created-at'], '%a %b %d %H:%M:%S +0000 %Y'))
  # sorted tweets based on time.
  f = open('output.txt', 'w')
  for tweet in sorted_tweets[-5:]:
      f.write(json.dumps(tweet) + '\')
  f.close()
  ```

  Note, when you copy and paste the code above, please be careful with the proper indentation and quotation mark.

- **Task 9: File operations**

  INPUT :CrimeReport.txt: in this file, each line is a raw tweet json format.
  output-folder: where new results will be stored
  REQUIREMENT: read tweets and separate these tweets in to groups based on the specific hours (Mon-Day-Year-Hour). The tweets related to a specific hour will be stored in a separate file in the folder "**task9-output**"

with the file name "**Mon-Day-Year-Hour.txt**"
OUTPUT: new files generated and stored in the folder "**task9-output**", in which each file stores the tweets corresponding to a specific hour.