Name: Chaoran Li
Student ID: 2021489307 (UTD ID) cxl190012 (NET ID)
Course Number: CS 6364.002

Homework 1 Writeup Part

Task 1: Test Python Environment

```
Task 1: Test Python Environment
hello world
```

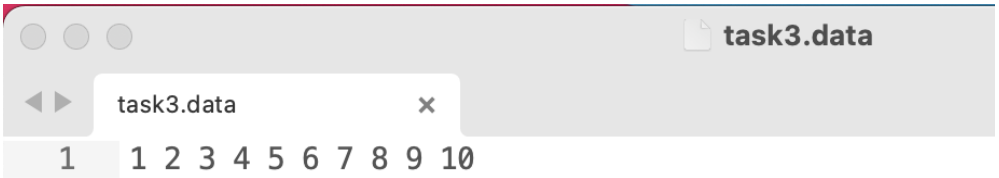
The code is very simple and the above result shewed that our python environment works well.

Task 2: Define Object

```
Task 2: Define Object
items: [1, 2, 3, 4, 5]
```

For this task, I directly assigned a list to object items and returned it to main. Hence, I can use it in below tasks.

Task 3: File Reading




```
Task 3: File Reading
items1: [1, 2, 3, 4, 5]
items2: [6, 7, 8, 9, 10]
```

After creating task3.data, I firstly read whole file as a string and splitted it with space. Then, I directly splitted the list into two half because I could not see other instructions for how to split the elements.

Task 4: Data Structure

First, wrote a program to use all three functions: items(), keys() and values(). These three functions are iterable and can be used to search in dictionary:

Code:

```python
def task4_show_dict_functions(dic):
    print("Show three dictionary functions:")
    print("items(): " + str(type(dic.items())))
    for (k, v) in dic.items():
        print("{0} -> {1}".format(k, v))
    print("keys(): " + str(type(dic.keys())))
    for k in dic.keys():
        print(k)
    print("values(): " + str(type(dic.values())))
    for v in dic.values():
        print(v)
```

Then, I chose to tranverse items() to print required result.

```
Print the required results:
school: UAlbany
address: 1400 Washington Ave, Albany, NY 12222
phone: (518)442-3300
```

Task 5: Data Serialization

```
Task 5: Data Serialization
{
    "school": "UAlbany",
    "address": "1400 Washington Ave, Albany, NY 12222",
    "phone": "(518)442-3300"
}
Save dict to task5.json.
<class 'dict'>, {'school': 'UAlbany', 'address': '1400 Washington Ave, Albany, NY 12222', 'phone': '(518)442-3300'}
Read dict from task5.json.
```

task5.json

```
{
    "school": "UAlbany",
    "address": "1400 Washington Ave, Albany, NY 12222",
    "phone": "(518)442-3300"
}
```

I first saved the dictionary to task5.json. Then, I read it back. As the above result shown, the return result is a dictionary and is just the same as we assigned before.

Task 6: Data Serialization

For this, I still use json module to solve this question. I also learned from my friend that he used pickle module to solve this question. I can give another solution by pickle if necessary.
I firstly use json.dumps() to save objects to file. To meet the requirement that I should save objects to file one by one, I firstly create an empty file. Then, I append the string to file each time. I use "\n" to split objects from each other. Then, when I need to load them back, I read the file by line and find target object from it. Below is the output of my code.

```
Task 6: Data Serialization
[1, 2, 3, 4, 5]
{'school': 'UAlbany', 'address': '1400 Washington Ave, Albany, NY 12222', 'phone': '(518)442-3300'}
Save data structures to task6.data.
items: <class 'list'>, [1, 2, 3, 4, 5]
data: <class 'dict'>, {'school': 'UAlbany', 'address': '1400 Washington Ave, Albany, NY 12222', 'phone': '(518)442-3300'}
Read data structures from task6.data.
```

task6.data

```
[
    1,
    2,
    3,
    4,
    5
]
{
    "school": "UAlbany",
    "address": "1400 Washington Ave, Albany, NY 12222",
    "phone": "(518)442-3300"
}
```

Task 7: Data Preprocessing

I firstly read some tweets and use tweet.keys() to analyze the keywords of all tweets.

```
dict_keys(['lang', 'favorited', 'truncated', 'text', 'created_at', 'retweeted', 'source', 'place', 'user', 'retweet_count', 'id', 'favorite_count'])
dict_keys(['lang', 'favorited', 'truncated', 'text', 'created_at', 'retweeted', 'source', 'user', 'urls', 'retweet_count', 'id', 'favorite_count'])
dict_keys(['lang', 'favorited', 'truncated', 'text', 'created_at', 'retweeted', 'user_mentions', 'source', 'user', 'urls', 'retweet_count', 'id', 'favorite_count'])
```

Since I found 'id' is the keyword, below work is just reading all tweets and collecting all ids.

```
[429129916446031872, 429117247307923456, 429315798893461505, 429079910091476992, 429030956888899584, 429152645702352896, 429016075921940481, 428898232690044928, 428956

Present with first 20 tweet ids:
429129916446031872, 429117247307923456, 429315798893461505, 429079910091476992, 429030956888899584,
429152645702352896, 429016075921940481, 428898232690044928, 428956892602191873, 429111068368326656,
429167217976958976, 429379822314614785, 429235096919359488, 428986795586383872, 429238790310203392,
429509723516194817, 429509617756811264, 429509135437017089, 429509033045659648, 429508763708817408,
...
```

I chose first 20 tweet ids for representation since the result is quite long.

Task 8: Data Preprocessing: tweets filtering

task8.data


```
1 {"lang": "en", "favorited": false, "truncated": false, "text": "Due to lack of police protection, #Christians in #Pakistan become easy targets for crime
and abuse. http://t.co/6Ld0Zs9h29 ^ws", "created_at": "Thu Jan 30 14:31:21 +0000 2014", "hashtags": ["Christians", "Pakistan"], "retweeted": false,
"source": "<a href=\"http://www.hootsuite.com\" rel=\"nofollow\">HootSuite</a>", "user": {"id": 39288218, "profile_sidebar_fill_color": "2E2D33",
"profile_text_color": "FFFFFF", "followers_count": 26122, "location": "Washington, D.C.", "profile_background_color": "66201A", "listed_count": 968,
"utc_offset": -18000, "statuses_count": 15823, "description": "Your bridge to the persecuted Church.", "friends_count": 13051, "profile_link_color":
"66201A", "profile_image_url": "https://pbs.twimg.com/profile_images/2426077357/fhwptbxflx6nwksesu6j_normal.jpeg", "profile_background_image_url":
"http://a0.twimg.com/profile_background_images/839990601/4d2636234a53eebae33401cb42a5a341.jpeg", "screen_name": "persecutionnews", "lang": "en",
"profile_background_tile": false, "favourites_count": 9, "name": "persecution.org", "url": "http://t.co/4cnV9ebLhE", "created_at": "Mon May 11 16:51:18
+0000 2009", "time_zone": "Eastern Time (US & Canada)", "protected": false}, "urls": {"http://t.co/6Ld0Zs9h29": "http://ow.ly/t6F0U"}, "retweet_count":
27, "id": 428898232690044928, "favorite_count": 6}
2 {"lang": "en", "favorited": false, "retweeted_status": {"lang": "en", "favorited": false, "truncated": false, "text": "So AAP will continue to cry foul
at Delhi police whenever a crime is reported rather than take responsibility like a govt should.", "created_at": "Fri Jan 31 16:47:31 +0000 2014",
"retweeted": false, "source": "web", "user": {"id": 318201824, "profile_sidebar_fill_color": "EAE8F0", "profile_text_color": "EBC42A", "followers_count":
3748, "location": "Planet Earth or up in the air ", "profile_background_color": "0D0F0E", "listed_count": 90, "utc_offset": 14400, "statuses_count":
59369, "description": "I am a leaf in the wind, watch how I soar. The black magic lies in my eyes, read my tweets for more!", "friends_count": 137,
"profile_link_color": "1B00B3", "profile_image_url": "https://pbs.twimg.com/profile_images/425949923486355456/knJaq8FK_normal.jpeg",
"profile_banner_url": "https://pbs.twimg.com/profile_banners/318201824/1389304438", "profile_background_image_url": "http://a0.twimg.com/
profile_background_images/378800000121260163/78be652bc2442b44c92560948a1e901d.jpeg", "screen_name": "blackmagicgal1", "lang": "en",
"profile_background_tile": false, "favourites_count": 28240, "name": "Sue", "created_at": "Thu Jun 16 03:38:35 +0000 2011", "time_zone": "Abu Dhabi",
"protected": false}, "retweet_count": 7, "id": 429294884885762048, "favorite_count": 5}, "truncated": false, "text": "RT @blackmagicgal1: So AAP will
continue to cry foul at Delhi police whenever a crime is reported rather than take responsibility like a go\u2026", "created_at": "Sat Feb 01 07:01:12
+0000 2014", "retweeted": false, "user_mentions": [{"screen_name": "blackmagicgal1", "id": 318201824, "name": "Sue"}], "source": "<a
href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>", "user": {"id": 143528722, "profile_sidebar_fill_color": "DDEEF6",
"profile_text_color": "333333", "followers_count": 9534, "location": "CITY OF DREAMS (BOMBAYY)", "profile_background_color": "C0DEED", "listed_count":
55, "utc_offset": 19800, "statuses_count": 114631, "description": "\u03b5\u03b7gi\u043f\u03b5\u03b5\u044f-\u0441\u044fick\u03b5\u03c4
Fr\u03b5\u03b1\u03ba-\u043c\u03c5f\u0441-SACHINIST-\u0432\u0435\u03b9\u03b7g
\u043d\u03bcma\u03b7-\u03b2\u03b5\u03b2\u03bf\u04a3\u043elic-\u0432l\u03b5\u03b5ds
\u03b2l\u03c5\u03b5-\u043csdi\u0434\u0438-Y\u03c5Vir\u03b1\u03c4,\u03ba\u03c1,\u03c9\u03b1\u03c4\u03c4\u03bf-\u0442\u044f\u0443i\u03b7 \u03c4\u03bf
\u03b2 S\u03c5mo\u03b7\u03b5-keep safe distance u may fall in LOVE", "friends_count": 4782, "profile_link_color": "0084B4", "profile_image_url":
"https://pbs.twimg.com/profile_images/428255065573306368/JTG0s4hv_normal.jpeg", "geo_enabled": true, "profile_banner_url": "https://pbs.twimg.com/
profile_banners/143528722/1384433695", "profile_background_image_url": "http://abs.twimg.com/images/themes/theme1/bg.png", "screen_name": "naushad1002",
"lang": "en", "profile_background_tile": false, "favourites_count": 33818, "name": "\u265a\u03b7\u03bf\u03c4 \u0455\u043e
\u0434\u0448\u03b5\u0455\u043e\u043c\u03be\u265a\u2122 ", "url": "http://t.co/IrPQ3GNjZ4", "created_at": "Thu May 13 18:28:47 +0000 2010", "time_zone":
"Mumbai", "protected": false}, "retweet_count": 7, "id": 429509723516194817}
3 {"lang": "en", "favorited": false, "retweeted_status": {"lang": "en", "favorited": false, "truncated": false, "text": "Crime/Lawlessness has always been
a problem in Delhi. Home minister and Delhi Police are answerable. Media shud stop blaming AAP.", "created_at": "Sat Feb 01 06:51:32 +0000 2014",
"retweeted": false, "source": "web", "user": {"id": 57591654, "profile_sidebar_fill_color": "DDEEF6", "profile_text_color": "333333", "followers_count":
55, "location": "France", "profile_background_color": "C0DEED", "utc_offset": -32400, "statuses_count": 779, "description": "Software Architect and a AAP
supporter", "friends_count": 199, "profile_link_color": "0084B4", "profile_image_url": "https://pbs.twimg.com/profile_images/
426101684515778560/_Te8AkHt_normal.jpeg", "geo_enabled": true, "profile_banner_url": "https://pbs.twimg.com/profile_banners/57591654/1390426885",
"profile_background_image_url": "http://abs.twimg.com/images/themes/theme1/bg.png", "name": "shivendra gaur", "lang": "en", "profile_background_tile":
```


In Task 7, we could find the keyword for this task is ‘created_at’. I chose first 10 recently tweets and found them all created at Feb 01, 2014.

Task 9: File operation

For this task, assume we do not need to sort tweets in each group. We found 18 different labels which also means 18 files in task9-output folder. Some of them only contains one tweet.

```
Task 9: File operations
01-31-2014-05, [{'lang': 'en', 'favorited': False, 'truncated': False, 'te
01-31-2014-18, [{'lang': 'en', 'favorited': False, 'truncated': False, 'te
01-31-2014-02, [{'lang': 'en', 'favorited': False, 'truncated': False, 'te
01-30-2014-23, [{'lang': 'en', 'favorited': False, 'truncated': False, 'te
01-31-2014-07, [{'lang': 'en', 'favorited': False, 'truncated': False, 'te
01-30-2014-22, [{'lang': 'en', 'favorited': False, 'truncated': False, 'te
01-30-2014-14, [{'lang': 'en', 'favorited': False, 'truncated': False, 'te
01-30-2014-18, [{'lang': 'en', 'favorited': False, 'truncated': False, 'te
01-31-2014-04, [{'lang': 'en', 'favorited': False, 'truncated': False, 'te
01-31-2014-08, [{'lang': 'en', 'favorited': False, 'truncated': False, 'te
01-31-2014-22, [{'lang': 'en', 'favorited': False, 'truncated': False, 'te
01-31-2014-12, [{'lang': 'en', 'favorited': False, 'truncated': False, 'te
01-30-2014-20, [{'lang': 'en', 'favorited': False, 'truncated': False, 'te
01-31-2014-13, [{'lang': 'en', 'favorited': False, 'truncated': False, 'te
02-01-2014-07, [{'lang': 'en', 'favorited': False, 'retweeted_status': {'
02-01-2014-06, [{'lang': 'en', 'favorited': False, 'retweeted_status': {'
02-01-2014-05, [{'lang': 'en', 'favorited': False, 'retweeted_status': {'
```

Task 10: NLP and Sentiment Analysis

Actually, I spent a lot of time on this. Because I use python 3.8 in my environment, but pattern only support python 3.6 for pip install.
You can use this link to help you install pattern for your environment:
https://github.com/clips/pattern
Besides, my IDE required me to install ntlk even I never use it in my code. My guess is that module pattern relies on ntlk module. And in python 3, you should use:
from pattern.text.en import positive, sentiment
instead of
from pattern.en import positive, sentiment

Show the sentiment score of each tweet:

```
Task 10: NLP and Sentiment Analysis
(-0.07777777777777778, 0.4666666666666666)
(-0.4, 0.7)
(0.0, 0.06666666666666667)
(0.13636363636363635, 0.45454545454545453)
(-0.15000000000000002, 0.15000000000000002)
(0.0, 0.0)
(0.0, 0.0)
(0.15416666666666667, 0.6041666666666667)
(0.13636363636363635, 0.45454545454545453)
(-0.18333333333333338, 0.8666666666666667)
(0.0, 0.0)
(0.0, 0.0)
(0.35714285714285715, 0.5714285714285714)
(0.0, 0.0)
(-0.1953125, 0.07083333333333333)
```

Then, I use positive function to separate tweets into positive-sentiment-tweets.txt and negative-sentiment-tweets.txt

positive-sentiment-tweets.txt UNREGISTERED


```
1 {"lang": "en", "favorited": false, "truncated": false, "text": "New NRA
  Advice: Don't Cooperate With Police If Your Gun Shows Up At A Crime
  Scene;  http://t.co/f96jCNY3wV #patriots #liberty", "created_at": "Fri
  Jan 31 02:33:17 +0000 2014", "hashtags": ["patriots", "liberty"],
  "retweeted": false, "source": "web", "user": {"id": 34643610,
  "profile_sidebar_fill_color": "DBE1F2", "profile_text_color": "000000",
  "followers_count": 35649, "profile_background_color": "174A82",
  "listed_count": 1989, "utc_offset": -18000, "statuses_count": 38235,
  "description": "Media Matters guy. Author. Bowler. Utica Club aficionado.
  Opinions are my own. Especially ones that offend people.",
  "friends_count": 602, "profile_link_color": "336699",
  "profile_image_url": "https://pbs.twimg.com/profile_images/1093194645/
  Screen_shot_2010-08-01_at_3.43.56_PM_normal.png",
  "profile_background_image_url": "http://a0.twimg.com/
  profile_background_images/9860779/book.jpg", "screen_name":
  "EricBoehlert", "lang": "en", "profile_background_tile": false,
  "favourites_count": 10, "name": "Eric Boehlert", "url": "http://t.co/
  QsWd6rjSG8", "created_at": "Thu Apr 23 15:39:39 +0000 2009", "time_zone":
  "Eastern Time (US & Canada)", "protected": false}, "urls": {"http://t.co/
  f96jCNY3wV": "http://bit.ly/1iRFr4e"}, "retweet_count": 19, "id":
  429079910091476992, "favorite_count": 3}
2 {"lang": "en", "favorited": false, "truncated": false, "text": "Due to
  lack of police protection, #Christians in #Pakistan become easy targets
  for crime and abuse. http://t.co/6Ld0Zs9h29 ^ws", "created_at": "Thu Jan
  30 14:31:21 +0000 2014", "hashtags": ["Christians", "Pakistan"],
  "retweeted": false, "source": "<a href=\"http://www.hootsuite.com\"
  rel=\"nofollow\">HootSuite</a>", "user": {"id": 39288218,
```


negative-sentiment-tweets.txt UNREGISTERED


```
1 {"lang": "en", "favorited": false, "truncated": false, "text": "Active
  crime scene on I-59/20 near Jeff/Tusc Co line. One dead, one injured;
  shooting involved. Police search in the area; traffic stopped",
  "created_at": "Fri Jan 31 05:51:59 +0000 2014", "retweeted": false,
  "source": "<a href=\"http://tapbots.com/software/tweetbot/mac\"
  rel=\"nofollow\">Tweetbot for Mac</a>", "place": {"country_code": "US",
  "url": "https://api.twitter.com/1.1/geo/id/cf44347a08102884.json",
  "country": "United States", "place_type": "city", "bounding_box":
  {"type": "Polygon", "coordinates": [[[-86.926154, 33.267324],
  [-86.598948, 33.267324], [-86.598948, 33.471006], [-86.926154,
  33.471006]]]}, "contained_within": [], "full_name": "Hoover, AL",
  "attributes": {}, "id": "cf44347a08102884", "name": "Hoover"}, "user":
  {"id": 15220806, "profile_sidebar_fill_color": "DDEEF6",
  "profile_text_color": "333333", "followers_count": 118021, "location":
  "Alabama", "profile_background_color": "C0DEED", "listed_count": 1705,
  "utc_offset": -21600, "statuses_count": 76381, "description": "Media
  meteorologist. WeatherBrains host. Weather geek.", "friends_count":
  52014, "profile_link_color": "0084B4", "profile_image_url":
  "https://pbs.twimg.com/profile_images/1890149584/
  spannwantsyou_normal.jpg", "geo_enabled": true, "profile_banner_url":
  "https://pbs.twimg.com/profile_banners/15220806/1381811159",
  "profile_background_image_url": "http://abs.twimg.com/images/themes/theme1
  /bg.png", "screen_name": "spann", "lang": "en",
  "profile_background_tile": false, "favourites_count": 27, "name": "James
  Spann", "url": "http://t.co/7bmABA89RQ", "created_at": "Tue Jun 24
  16:02:10 +0000 2008", "time_zone": "Central Time (US & Canada)",
  "protected": false}, "retweet_count": 66, "id": 429129916446031872,
```