Intro to Data Science, Week1 Assignments

Week 1 assignment: Twitter Sentiment Analysis in Python

Grade:100%

Part	Name	Last Submission	Score	Feedback	
1 / 6	output.txt	Thu 16 May 2013 3:11 PM PDT (UTC -0700)	5.00 / 5	View	Submit
2 / 6	tweet_sentiment.py	Thu 16 May 2013 4:21 PM PDT (UTC -0700)	10.00 / 10	View	Submit
3 / 6	term_sentiment.py	Sat 18 May 2013 9:19 AM PDT (UTC -0700)	10.00 / 10	View	Submit
4 / 6	frequency.py	Sat 18 May 2013 10:20 AM PDT (UTC -0700)	10.00 / 10	View	Submit
5 / 6	happiest_state.py	Sat 18 May 2013 1:56 PM PDT (UTC -0700)	10.00 / 10	View	Submit
6 / 6	top_ten.py	Sat 18 May 2013 2:19 PM PDT (UTC -0700)	10.00 / 10	View	Submit
	Total Score		**55 / 55**		

Part 1:

```
ubuntu@vbox ~/dropbox/Projects/Courses/Data Science/datasci_course_materials/assignment1 (master)
782 $ head -n 2 output.txt
{"created_at":"Thu May 16 21:59:55 +0000 2013","id":335152660099502080,"id_str":"335152660099502080","text":"@LPGGusttavo @Gi
l s\u00f3 tt mesmo","source":"web","truncated":false,"in_reply_to_status_id":335152370302451713,"in_reply_to_status_id_str":"
ply_to_user_id":955990278,"in_reply_to_user_id_str":"955990278","in_reply_to_screen_name":"LPGGusttavo","user":{"id":60686802
me":"Morena do Ningo s2","screen_name":"morenadoGL_","location":"Francisco Beltr\u00e3o\/PR","url":null,"description":"Nem to
00e9 capaz de chegar ao tamanho do nosso amor@IgorGusttavoL.Seguida por @FrontJR e @Mois\u00e9s","protected":false,"followers
t":1129,"listed_count":0,"created_at":"Wed Jun 13 01:18:30 +0000 2012","favourites_count":138,"utc_offset":-10800,"time_zone"
true,"verified":false,"statuses_count":60536,"lang":"pt","contributors_enabled":false,"is_translator":false,"profile_backgrou
e_background_image_url":"http:\/\/a0.twimg.com\/profile_background_images\/851596826\/e387a600e341b5aeb2649ae5ec507ab1.gif","
rl_https":"https:\/\/si0.twimg.com\/profile_background_images\/851596826\/e387a600e341b5aeb2649ae5ec507ab1.gif","profile_back
e_image_url":"http:\/\/a0.twimg.com\/profile_images\/3663366317\/8bcbdfad678366600c185a476d497b22_normal.png","profile_image_
twimg.com\/profile_images\/3663366317\/8bcbdfad678366600c185a476d497b22_normal.png","profile_banner_url":"https:\/\/pbs.twimg
868029\/1367361235","profile_link_color":"EB17E4","profile_sidebar_border_color":"000000","profile_sidebar_fill_color":"E5507
62720","profile_use_background_image":true,"default_profile":false,"default_profile_image":false,"following":null,"follow_req
ions":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"retweet_count":0,"favorite_count":0,"entities":{"
"urls":[],"user_mentions":[{"screen_name":"LPGGusttavo","name":"L P G  @Gusttavo ","id":955990278,"id_str":"955990278","indic
:"GirlGusttavete","name":"- Thay Te Amo -","id":457649658,"id_str":"457649658","indices":[13,28]},{"screen_name":"essesorriso
11,"id_str":"426200011","indices":[29,43]}]},"favorited":false,"retweeted":false,"filter_level":"medium","lang":"pt"}
{"created_at":"Thu May 16 21:59:55 +0000 2013","id":335152660103692288,"id_str":"335152660103692288","text":"@yungboiDC I alw
003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e","truncated":fa
":335151832231989249,"in_reply_to_status_id_str":"335151832231989249","in_reply_to_user_id":379897845,"in_reply_to_user_id_st
o_screen_name":"yungboiDC","user":{"id":58954526,"id_str":"58954526","name":"Laur Monteith ","screen_name":"LaurenMonteith_",
rio","url":null,"description":"IG: laurmonteith #DTA","protected":false,"followers_count":984,"friends_count":174,"listed_cou
l 21 23:07:46 +0000 2009","favourites_count":1769,"utc_offset":-25200,"time_zone":"Mountain Time (US & Canada)","geo_enabled"
atuses_count":8913,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_color":"352726","profil
ttp:\/\/a0.twimg.com\/profile_background_images\/630924520\/9jw53odphwy9fz1rui1i.jpeg","profile_background_image_url_https":"
rofile_background_images\/630924520\/9jw53odphwy9fz1rui1i.jpeg","profile_background_tile":true,"profile_image_url":"http:\/\/
es\/3668793189\/5fbadb6175034979f9675c43f0e4925f_normal.jpeg","profile_image_url_https":"https:\/\/si0.twimg.com\/profile_ima
5034979f9675c43f0e4925f_normal.jpeg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/58954526\/1360708552","p
","profile_sidebar_border_color":"EEEEEE","profile_sidebar_fill_color":"EFEFEF","profile_text_color":"333333","profile_use_ba
ult_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null
":null,"contributors":null,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentio
oiDC","name":"Darnell Curtin","id":379897845,"id_str":"379897845","indices":[0,10]}]},"favorited":false,"retweeted":false,"fi
":"en"}

ubuntu@vbox ~/dropbox/Projects/Courses/Data Science/datasci_course_materials/assignment1 (master)
783 $ ls -lah output.txt
-rwxrwx--- 1 root vboxsf 199M 2013-05-19 01:43 output.txt
```

Part 2:

```
import sys
import json
import string

def lines(fp):
    print str(len(fp.readlines()))

_
def main():
    sent_file = open(sys.argv[1])
    scores = {}

    for line in  sent_file:
        term, score  = line.split("\t")
        scores[term] = int(score)

    tweet_file = open(sys.argv[2])


    for tweet in tweet_file:
        tweet_text = json.loads(tweet)
        tweet_score = 0
        try:
            for  word in tweet_text['text'].split(" "):
                word_score = 0
                try:
                    word_score = scores[word]
                    tweet_score += word_score
                except KeyError:
                    pass #no score for this word
            print tweet_score
        except KeyError:
            print tweet_score
            pass #badly formed tweet


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

Part 3

```
import sys
import json
import string

def main():
    sent_file = open(sys.argv[1])
    scores = {}

    for line in  sent_file:
        term, score  = line.split("\t")
        scores[term] = int(score)

    tweet_file = open(sys.argv[2])

    new_sents = {}
    #new_word_scores[word]=[count,score_sum]

    # Given a list of json formatted tweets, take the valid ones
    for line in tweet_file:
        # Don't consider deleted tweets
        if line[2:8] == "delete":
            continue

        # Convert to acual JSON format
        tweet = json.loads(line)

        # Ignore badly formatted tweets
        if not "text" in tweet.keys():
            continue

        # Ensure they are unicode encoded
        tweet_text = tweet['text'].encode('utf-8')

        # First, calculate the sentiment of the tweet based on they
        # terms we have sentiment scores for already
        tweet_score = 0
        for word in tweet_text.split(' '):
                    try:
                        tweet_score += scores[word]
                    except KeyError:
                        pass #no score for this word

        # Second, use the tweet sentiment score to assign a sentiment
        # for the terms we don't have sentiment scores for
        for word in tweet_text.split(' '):
            data = new_sents.get(word)
            if data == None:
                # New term found - assign its sentiment from the tweet sentiment
                new_sents[word] = [1, tweet_score]
            else:
                # Update existing term sentiment
                new_sents[word] = [data[0]+1, data[1] + tweet_score]

    # end for

    for term in new_sents:
        print term + " " + str(new_sents[term][1] / new_sents[term][0])

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

Part 4:

```
import sys
import json
import string

def main():

    tweet_file = open(sys.argv[1])

    terms = {}
    #terms[term]=count

    # Given a list of json formatted tweets, take the valid ones
    for line in tweet_file:
        # Don't consider deleted tweets
        if line[2:8] == "delete":
            continue

        # Convert to acual JSON format
        tweet = json.loads(line)

        # Ignore badly formatted tweets
        if not "text" in tweet.keys():
            continue

        # Ensure they are unicode encoded
        tweet_text = tweet['text'].encode('utf-8')

        for word in tweet_text.split(' '):
            # check if we know this term already, if so increment, otherwise record its
            data = terms.get(word)
            if data == None:
                # New term found - add it to the dict
                terms[word] = 1
            else:
                # Update existing term score
               terms[word] = terms[word] + 1

    # end for

    # Calculate term frequencies
    for term in terms:
        if term.strip() == "":
            pass
        else:

            print term.strip() + "  %f" % (float(terms[term]) / float(len(terms)))

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

Part 5

```
import sys
import json
import string
import operator

def main():

    tweet_file = open(sys.argv[2])


    sent_file = open(sys.argv[1])
    scores = {}

    for line in  sent_file:
        term, score  = line.split("\t")
        scores[term] = int(score)

    states = {
        'AK': 0,
        'AL': 0,
        'AR': 0,
        'AS': 0,
        'AZ': 0,
        'CA': 0,
        'CO': 0,
        'CT': 0,
        'DC': 0,
        'DE': 0,
        'FL': 0,
        'GA': 0,
        'GU': 0,
        'HI': 0,
        'IA': 0,
        'ID': 0,
        'IL': 0,
        'IN': 0,
        'KS': 0,
        'KY': 0,
        'LA': 0,
        'MA': 0,
        'MD': 0,
        'ME': 0,
        'MI': 0,
        'MN': 0,
        'MO': 0,
        'MP': 0,
        'MS': 0,
        'MT': 0,
        'NA': 0,
        'NC': 0,
        'ND': 0,
        'NE': 0,
        'NH': 0,
        'NJ': 0,
        'NM': 0,
        'NV': 0,
        'NY': 0,
        'OH': 0,
        'OK': 0,
        'OR': 0,
        'WI': 0,
        'WV': 0,
        'WY': '0'
    }
    #states[state]=score

    # Given a list of json formatted tweets, take the valid ones
    for line in tweet_file:
        # Don't consider deleted tweets
        if line[2:8] == "delete":
            continue

        # Convert to acual JSON format
        tweet = json.loads(line)

        # Ignore badly formatted tweets
        if not "text" in tweet.keys():
            continue

        # Ignore all non-US tweets and those without states
        if not "place" in tweet.keys() or tweet["place"] == None:
            continue
        if not tweet["place"]["country_code"] == "US":
            continue
        if not tweet["place"]["full_name"]:
            continue

        # Ensure they are unicode encoded
        tweet_text = tweet['text'].encode('utf-8')

        # Calculate the sentiment of the tweet
        tweet_score = 0
        for word in tweet_text.split(' '):
                    try:
                        tweet_score += scores[word]
                    except KeyError:
                        pass #no score for this word

        #print tweet_score

        # If we know about the state, add the score to its
        if tweet["place"]['full_name'][-2:] in states:
            states[tweet["place"]['full_name'][-2:]] += tweet_score
    # end for

    state_max_score = 0;
    happiest_state = "";
    for state in states:
        #print state + " : "  + str(states[state])
        if int(states[state]) > int(state_max_score):
            state_max_score = int(states[state])
            happiest_state = state


    print happiest_state
    #print sorted_states

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

Part 6:

```
import sys
import json
import string
import operator

def main():

    tweet_file = open(sys.argv[1])
    hashtags_popularity = {}
    # hashtags_popularity['tag'] = count

    # Given a list of json formatted tweets, take the valid ones
    for line in tweet_file:
        # Don't consider deleted tweets
        if line[2:8] == "delete":
            continue

        # Convert to acual JSON format
        tweet = json.loads(line)

        # Ignore tweets without hashtag
        if not "entities" in tweet.keys():
            continue
        if not "hashtags" in tweet['entities'].keys():
            continue
        hashtags = tweet['entities']['hashtags']
        if hashtags == None or hashtags == []:
            continue

        # Ensure they are unicode encoded
        tweet_text = tweet['text'].encode('utf-8')

        for tag in hashtags:
            if tag['text'] in hashtags_popularity:
                hashtags_popularity[tag['text']] += 1.0
            else:
                 hashtags_popularity[tag['text']] = 1.0

    # end for

    hashtags_popularity_asc = sorted(hashtags_popularity.items(), key=operator.itemgetter(1), reverse=True)

    count = 10
    for tag in  hashtags_popularity_asc:
        print tag[0] + " " + str(tag[1])
        count -= 1
        if count == 0:
            break

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

Details on each problem

Problem 0: Query Twitter with Python

If you are using the class virtual machine, run the VM, open a terminal window, and use git to make sure you have the latest class materials. To edit the file in linux, you can use vi, emacs, or gedit.

Use the urllib and json libraries in python to access the basic twitter search API and return JSON data.

To retrieve recent tweets associated with the term "microsoft," you use this url:

http://search.twitter.com/search.json?q=microsoft

To access this url in Python and parse the response, you can use the following snippet:

```
import urllib

import json



response = urllib.urlopen("http://search.twitter.com/search.json?q=microsoft")

print json.load(response)
```

The format of the result is *JSON*, which stands for JavaScript Object Notation. It is a simple format for representing nested structures of data --- lists of lists of dictionaries of lists of .... you get the idea.

As you might imagine, it is fairly straightforward to convert JSON data into a Python data structure. Indeed, there is a convenient library to do so, called json, which we will use.

Twitter provides only **partial documentation for understanding this data format**, but it's not difficult to deduce the structure.

Using this library, the json data is parsed and converted to a Python dictionary representing the entire result set. (If needed, take a moment to **read the documentation for Python dictionaries**). The "results" key of this dictionary corresponds holds the actual tweets; each tweet is itself another dictionary.

a) Write a program, print.py, to print out the text of each tweet in the result.

b) Generalize your program, print.py, to fetch and print 10 pages of results. Note that you can return a different page of results by passing an additional argument in the url:

http://search.twitter.com/search.json?q=microsoft&page=2

print.py should be executable in the following way:

$ python print.py

When executed, the script should print each tweet on an individual line to stdout.

What to turn in: Nothing. This is a warmup exercise.

Problem 1: Get Twitter Data

As always, the first step is to make sure your assignment materials up to date.

To access the live stream, you will need to install the **oauth2 library** so you can properly authenticate.

This library is already installed on the class virtual machine. Or you can install it yourself in your Python environment.

The steps below will help you set up your twitter account to be able to access the live 1% stream.

- Create a twitter account if you do not already have one.
- Go to **https://dev.twitter.com/apps** and log in with your twitter credentials.
- Click "create an application"
- Fill out the form and agree to the terms. Put in a dummy website if you don't have one you want to use.
- On the next page, scroll down and click "Create my access token"
- Copy your "Consumer key" and your "Consumer secret" into twitterstream.py
- Click "Create my access token." You can **Read more about Oauth authorization.**
- Open twitterstream.py and set the variables corresponding to the consumer key, consumer secret, access token, and access secret.

```
access_token_key = "<Enter your access token key here>"

access_token_secret = "<Enter your access token secret here>"

consumer_key = "<Enter consumer key>"

consumer_secret = "<Enter consumer secret>"
```

- Run the following and make sure you see data flowing and that no errors occur. Stop the program with Ctrl-C once you are satisfied.

```
$ python twitterstream.py
```

You can pipe the output to a file, wait a few minutes, then terminate the program to generate a sample. Use the following command:

```
$ python twitterstream.py > output.txt
```

Let this script run for a minimum of **10 minutes**.

Keep the file output.txt for the duration of the assignment, we will be reusing it in later problems.

Don't use someone else's file; we will check for uniqueness in other parts of the assignment.

What to turn in: The first 20 lines of your file. You can get the first 20 lines by using the following command: $ head -n 20 output.txt

Problem 2: Derive the sentiment of each tweet

For this part, you will compute the sentiment of each tweet based on the sentiment scores of the terms in the tweet. The sentiment of a tweet is equivalent to the sum of the sentiment scores for each term in the tweet.

You are provided with a skeleton file, tweet_sentiment.py, which can be executed using the following command:

$ python tweet_sentiment.py <sentiment_file> <tweet_file>

The file AFINN-111.txt contains a list of pre-computed sentiment scores. Each line in the file contains a word or phrase followed by a sentiment score. Each word or phrase found in a tweet, but not in AFINN-111.txt should be given a sentiment score of 0. See the file AFINN-README.txt for more information.

To use the data in the AFINN-111.txt file, you may find it useful to build a dictionary. Note that the AFINN-111.txt file format is tab-delimited, meaning that the term and the score are separated by a tab character. A tab character can be identified a "\t".The following snippet may be useful:

```
afinnfile = open("AFINN-111.txt")

scores = {} # initialize an empty dictionary

for line in afinnfile:

  term, score  = line.split("\t")  # The file is tab-delimited. "\t" means "tab character"

  scores[term] = int(score)  # Convert the score to an integer.


print scores.items() # Print every (term, score) pair in the dictionary
```

Assume the tweet file contains data formatted the same way as the livestream data.

Your script should print to stdout the sentiment of each tweet in the file, one sentiment per line:

<sentiment:float>

NOTE: You must provide a score for **every** tweet in the sample file, even if that score is zero. However the sample file will only include English tweets

The first sentiment corresponds to the first tweet in the input file, the second sentiment corresponds to the second tweet in the input file, and so on.

Hints: The json.loads function parses a string to JSON.

Refer to the **twitter documentation** in order to determine what field to parse.

What to turn in: tweet_sentiment.py

Problem 3: Derive the sentiment of new terms

In this part you will be creating a script that computes the sentiment for the terms that **do not** appear in the file AFINN-111.txt.

Here's how you might think about the problem: We know we can use certain words to deduce the sentiment of a tweet. Once you know the sentiment of the tweets that contain some term, you can assign a sentiment to the term itself.

Don't feel obligated to use it, but the following paper may be helpful for developing a sentiment metric. Look at the Opinion Estimation subsection of the Text Analysis section in particular.

O'Connor, B., Balasubramanyan, R., Routedge, B., & Smith, N. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. (ICWSM), May 2010.

You are provided with a skeleton file, term_sentiment.py, which can be executed using the following command:

```
$ python term_sentiment.py <sentiment_file> <tweet_file>
```

Your script should print to stdout each term-sentiment pair, one pair per line, in the following format:

```
<term:string> <sentiment:float>
```

For example, if you have the pair ("foo", 103.256) it should appear in the output as:

```
foo 103.256
```

The order of your output does not matter.

What to turn in: term_sentiment.py

How we will grade Part 3: We will use a given file and make sure that your scores order the terms in roughly the same order as our solution. Your scores need not exactly match ours.

If the grader is returning "Formatting error: ", make note of the line of text returned in the message. This line corresponds to a line of your output. The grader will generate this error if line.split() does not return exactly two items. One common source of this error is to not remove the the two calls to the "lines" function in the solution template -- this function prints the number of lines in each file. Make sure to check the first two lines of your output!

Problem 4: Compute Term Frequency

Write a Python script, frequency.py, to compute the term frequency histogram of the livestream data you harvested from Problem 1.

The frequency of a term can be calculate with the following formula:
[# of occurrences of the term in all tweets]/[# of occurrences of all terms in all tweets]

frequency.py should take a file of tweets as an input and be usable in the following way:

```
$ python frequency.py <tweet_file>
```

Assume the tweet file contains data formatted the same way as the livestream data.

Your script should print to stdout each term-frequency pair, one pair per line, in the following format:

<term:string> <frequency:float>

For example, if you have the pair (bar, 0.1245) it should appear in the output as:

bar 0.1245

Frequency measurements may take phrases into account, but this is not required. We only ask that you compute frequencies for individual tokens.

Depending on your method of parsing, you may end up with frequencies for hashtags, links, stop words, phrases, etc. Some noise is acceptable for the sake of keeping parsing simple.

What to turn in: frequency.py

Problem 5: Which State is happiest?

Write a Python script, happiest_state.py, that returns the name of the happiest state as a string.

happiest_state.py should take a file of tweets as an input and be usable in the following way:

$ python happiest_state.py <sentiment_file> <tweet_file>

The file AFINN-111.txt contains a list of pre-computed sentiment score.

Assume the tweet file contains data formatted the same way as the livestream data.

We recommend that you build on your solution to Problem 2.

There are three different objects within the tweet that you can use to determine it's origin.

1 The coordinates object

2 The place object

3 The user object

You are free to develop your own strategy for determining the state that each tweet originates from.

Limit the tweets you analyze to those in the United States.

The live stream has a slightly different format from the response to the query you used in Problem 0. In this file, each line is a Tweet object, as**described in the twitter documentation**.

Note: Not every tweet dictionary will have a text key -- real data is dirty. Be prepared to debug, and feel free to throw out tweets that your code can't handle to get something working. For example, non-English tweets.

Your script should print the two letter state abbreviation to stdout. Your script will not have access to the Internet, so you cannot rely on third party services to resolve geocoded locations.

What to turn in: happiest_state.py

Problem 6: Top ten hash tags

Write a Python script, top_ten.py, that computes the ten most frequently occurring hash tags from the data you gathered in Problem 1.

top_ten.py should take a file of tweets as an input and be usable in the following way:

```
$ python top_ten.py <tweet_file>
```

Assume the tweet file contains data formatted the same way as the livestream data.

In the tweet file, each line is a Tweet object, as **described in the twitter documentation**.
You should not be parsing the "text" field.

Your script should print to stdout each hashtag-count pair, one per line, in the following format:

```
<hashtag:string> <count:float>
```

For example, if you have the pair (baz, 30) it should appear in the output as:

```
baz 30.0
```

Remember your output must contain floats, not ints.

What to turn in: top_ten.py