Programming Assignment 1: Sentiment Analysis of Twitter Data

Twitter has emerged as a fundamentally new instrument to obtain social measurements. For example, researchers have shown that the "mood" of communication on twitter can be used to predict the stock market.

In this programming assignment you will:

- Load and prepare a collected set of twitter data for analysis
- You will estimate the sentiment associated with individual tweets
- You will estimate the sentiment of a particular term

Please keep in mind the following points:

- This assignment is open-ended in several ways. You will need to make some decisions about how to best solve each of the problems mentioned above.
- **It is absolutely fine to discuss your solutions with your classmates (either in person or in Piazza) but you are not allowed to share code.**
- **Each student must submit their own solution via CANVAS.**
- You will have unlimited number of uploads to CANVAS.

Formatting of Twitter Data

Strings in the twitter data prefixed with the letter "u" are unicode strings. For example: `u"This is a string"`.

Unicode is a standard for representing a mach larger variety of characters beyond the roma alphabet (greek, russian, mathematical symbols, logograms from non-phonetic writing systems, etc.).

In most circumstances, you will be able to use a unicode object just like a string.

If you encounter an error involving printing unicode, you can use the encode (https://docs.python.org/3/library/stdtypes.html#str.encode) method to properly print the international characters. You can find more information about UNICODE and Python 3 here (https://docs.python.org/3/howto/unicode.html).

Question 1: Loading and Cleaning Twitter Data [20 points]

In this first part, you will need to load a sample of tweets in memory and prepare them for analysis. The tweets are stored in the file `tweets.json`. This file follows the *JSON* format. JSON stands for JavaScript Object Notation. It is a simple format for representing nested structres of data --- lists of lists of dictionaries of lists of ... you get the idea.

Each line in of `tweets.json` represents a message. It is straightforward to convert a JSON string into a Python data structure; there is a library to do so called `json`. Below we will show you how to load the data and how to parse the first line in the `tweets.json` file.

```
In [18]:   import json

           # Open the input file
           input_file = open('tweets.json','r')

           # Load the first few lines
           line_count = 6
           for line in input_file:
               tweet = json.loads(line)
               print(tweet)
               line_count -= 1
               if line_count < 0:
                   break


           # Close the input file
           input_file.close()
```

{'delete': {'status': {'id': 616962824619687936, 'id_str': '61696282
4619687936', 'user_id': 3061937538, 'user_id_str': '3061937538'}, 't
imestamp_ms': '1439761273410'}}
{'created_at': 'Sun Aug 16 21:41:13 +0000 2015', 'id': 6330307796190
12608, 'id_str': '633030779619012608', 'text': 'ايه الأكل 😜', 'sourc
e': 'Twi
tter for iPhone', 'truncated': False, 'in_reply_to_status_id': N
one, 'in_reply_to_status_id_str': None, 'in_reply_to_user_id': None,
'in_reply_to_user_id_str': None, 'in_reply_to_screen_name': None, 'u
ser': {'id': 1305908600, 'id_str': '1305908600', 'name': 'شخصيه دبلوِ
ماسيَة.!�’َ', 'screen_name': 'snoow3333', 'location': '', 'url': None,
'description': 'شخص# ب إسمي ربط أبْنثّم لاأنـني فقط أجِد أشبم مدئِلُةَ عشت
طبعآ بَأخَلّآمَكْ أنـا بكون ووينِ...........دبلومـاسية يـه_ّلا', 'protected': False
, 'verified': False, 'followers_count': 18816, 'friends_count': 7559
, 'listed_count': 25, 'favourites_count': 14468, 'statuses_count': 1
33539, 'created_at': 'Tue Mar 26 23:22:14 +0000 2013', 'utc_offset':
10800, 'time_zone': 'Nairobi', 'geo_enabled': False, 'lang': 'ar', '
contributors_enabled': False, 'is_translator': False, 'profile_backg
round_color': 'C0DEED', 'profile_background_image_url': 'http://abs.

twimg.com/images/themes/theme1/bg.png', 'profile_background_image_ur
l_https': 'https://abs.twimg.com/images/themes/theme1/bg.png', 'prof
ile_background_tile': True, 'profile_link_color': '0084B4', 'profile
_sidebar_border_color': 'FFFFFF', 'profile_sidebar_fill_color': 'DDE
EF6', 'profile_text_color': '333333', 'profile_use_background_image'
: True, 'profile_image_url': 'http://pbs.twimg.com/profile_images/61
5338552402948097/k0eOXDro_normal.jpg', 'profile_image_url_https': 'h
ttps://pbs.twimg.com/profile_images/615338552402948097/k0eOXDro_norm
al.jpg', 'profile_banner_url': 'https://pbs.twimg.com/profile_banner
s/1305908600/1433459409', 'default_profile': False, 'default_profile
_image': False, 'following': None, 'follow_request_sent': None, 'not
ifications': None}, 'geo': None, 'coordinates': None, 'place': None,
'contributors': None, 'retweet_count': 0, 'favorite_count': 0, 'enti
ties': {'hashtags': [], 'trends': [], 'urls': [], 'user_mentions': [
], 'symbols': []}, 'favorited': False, 'retweeted': False, 'possibly
_sensitive': False, 'filter_level': 'low', 'lang': 'ar', 'timestamp_
ms': '1439761273661'}
{'created_at': 'Sun Aug 16 21:41:13 +0000 2015', 'id': 6330307796106
64960, 'id_str': '633030779610664960', 'text': 'RT @nitishuna: @LasV
egasChicas @I_luv_reds @DimondLadies @SiliconEros @herby2121 @hughco
x69 @VxExx @Swt_Red @tthumb9090 @raggdoll_98 http:/…', 'source': '<a
href="http://twitter.com/download/iphone" rel="nofollow">Twitter for
iPhone', 'truncated': False, 'in_reply_to_status_id': None, 'in_
reply_to_status_id_str': None, 'in_reply_to_user_id': None, 'in_repl
y_to_user_id_str': None, 'in_reply_to_screen_name': None, 'user': {'
id': 3030384930, 'id_str': '3030384930', 'name': 'styles', 'screen_n
ame': 'balwinderstyles', 'location': '', 'url': None, 'description':
None, 'protected': False, 'verified': False, 'followers_count': 2463
, 'friends_count': 566, 'listed_count': 19, 'favourites_count': 7829
, 'statuses_count': 9917, 'created_at': 'Thu Feb 19 15:22:15 +0000 2
015', 'utc_offset': None, 'time_zone': None, 'geo_enabled': True, 'l
ang': 'en', 'contributors_enabled': False, 'is_translator': False, '
profile_background_color': 'C0DEED', 'profile_background_image_url':
'http://abs.twimg.com/images/themes/theme1/bg.png', 'profile_backgro
und_image_url_https': 'https://abs.twimg.com/images/themes/theme1/bg
.png', 'profile_background_tile': False, 'profile_link_color': '0084
B4', 'profile_sidebar_border_color': 'C0DEED', 'profile_sidebar_fill
_color': 'DDEEF6', 'profile_text_color': '333333', 'profile_use_back
ground_image': True, 'profile_image_url': 'http://pbs.twimg.com/prof
ile_images/599589162744840192/LxOH63HU_normal.jpg', 'profile_image_u
rl_https': 'https://pbs.twimg.com/profile_images/599589162744840192/
LxOH63HU_normal.jpg', 'profile_banner_url': 'https://pbs.twimg.com/p
rofile_banners/3030384930/1439232099', 'default_profile': True, 'def
ault_profile_image': False, 'following': None, 'follow_request_sent'
: None, 'notifications': None}, 'geo': None, 'coordinates': None, 'p
lace': None, 'contributors': None, 'retweeted_status': {'created_at'
: 'Sun Aug 16 18:38:02 +0000 2015', 'id': 632984680887812096, 'id_st
r': '632984680887812096', 'text': '@LasVegasChicas @I_luv_reds @Dimo
ndLadies @SiliconEros @herby2121 @hughcox69 @VxExx @Swt_Red @tthumb9
090 @raggdoll_98 http://t.co/IdjGgfgncz', 'source': '<a href="http:/
/twitter.com/download/iphone" rel="nofollow">Twitter for iPhone'
, 'truncated': False, 'in_reply_to_status_id': 632982908068864000, '
in_reply_to_status_id_str': '632982908068864000', 'in_reply_to_user_

id': 737103644, 'in_reply_to_user_id_str': '737103644', 'in_reply_to
_screen_name': 'LasVegasChicas', 'user': {'id': 445865965, 'id_str':
'445865965', 'name': 'Just4fun', 'screen_name': 'nitishuna', 'locati
on': '', 'url': None, 'description': '#RT | I Share What I Find #Bea
utiful | Nothing better than a touch of fun | No expectations means
no disappointments | #Follow @micelle79 @balwinderstyles', 'protecte
d': False, 'verified': False, 'followers_count': 15926, 'friends_cou
nt': 771, 'listed_count': 92, 'favourites_count': 17683, 'statuses_c
ount': 27017, 'created_at': 'Sun Dec 25 00:35:36 +0000 2011', 'utc_o
ffset': 19800, 'time_zone': 'New Delhi', 'geo_enabled': False, 'lang
': 'en', 'contributors_enabled': False, 'is_translator': False, 'pro
file_background_color': 'C0DEED', 'profile_background_image_url': 'h
ttp://abs.twimg.com/images/themes/theme1/bg.png', 'profile_backgroun
d_image_url_https': 'https://abs.twimg.com/images/themes/theme1/bg.p
ng', 'profile_background_tile': False, 'profile_link_color': '0084B4
', 'profile_sidebar_border_color': 'C0DEED', 'profile_sidebar_fill_c
olor': 'DDEEF6', 'profile_text_color': '333333', 'profile_use_backgr
ound_image': True, 'profile_image_url': 'http://pbs.twimg.com/profil
e_images/574313941012115458/vSAmzsWR_normal.jpeg', 'profile_image_ur
l_https': 'https://pbs.twimg.com/profile_images/574313941012115458/v
SAmzsWR_normal.jpeg', 'profile_banner_url': 'https://pbs.twimg.com/p
rofile_banners/445865965/1425762664', 'default_profile': True, 'defa
ult_profile_image': False, 'following': None, 'follow_request_sent':
None, 'notifications': None}, 'geo': None, 'coordinates': None, 'pla
ce': None, 'contributors': None, 'retweet_count': 12, 'favorite_coun
t': 25, 'entities': {'hashtags': [], 'trends': [], 'urls': [], 'user
_mentions': [{'screen_name': 'LasVegasChicas', 'name': 'What Happens
n Vegas', 'id': 737103644, 'id_str': '737103644', 'indices': [0, 15]
}, {'screen_name': 'I_luv_reds', 'name': 'I_luv_reds', 'id': 7189272
86, 'id_str': '718927286', 'indices': [16, 27]}, {'screen_name': 'Di
mondLadies', 'name': '⭐Dimond Ladies⭐', 'id': 2714436740, 'id_str'
: '2714436740', 'indices': [28, 41]}, {'screen_name': 'SiliconEros',
'name': 'Silicon Eros', 'id': 2848949950, 'id_str': '2848949950', 'i
ndices': [42, 54]}, {'screen_name': 'herby2121', 'name': '☆ herby212
1 ☆86K ', 'id': 2274691870, 'id_str': '2274691870', 'indices': [55,
65]}, {'screen_name': 'hughcox69', 'name': '∩UDE PHOTOGRAPHY 82k',
'id': 2767126967, 'id_str': '2767126967', 'indices': [66, 76]}, {'sc
reen_name': 'VxExx', 'name': '#VenusExchange☀43K☀', 'id': 322943281
1, 'id_str': '3229432811', 'indices': [77, 83]}, {'screen_name': 'Sw
t_Red', 'name': '✨ Sweet Red ✨', 'id': 3087022267, 'id_str': '3087
022267', 'indices': [84, 92]}, {'screen_name': 'tthumb9090', 'name':
'Mr Tasty ♻69K♻', 'id': 2665853560, 'id_str': '2665853560', 'indice
s': [93, 104]}, {'screen_name': 'raggdoll_98', 'name': 'His Girl ',
'id': 352691921, 'id_str': '352691921', 'indices': [105, 117]}], 'sy
mbols': [], 'media': [{'id': 632984655252160512, 'id_str': '63298465
5252160512', 'indices': [118, 140], 'media_url': 'http://pbs.twimg.c
om/media/CMjQMLSUAAAfMUW.jpg', 'media_url_https': 'https://pbs.twimg
.com/media/CMjQMLSUAAAfMUW.jpg', 'url': 'http://t.co/IdjGgfgncz', 'd
isplay_url': 'pic.twitter.com/IdjGgfgncz', 'expanded_url': 'http://t
witter.com/nitishuna/status/632984680887812096/photo/1', 'type': 'ph
oto', 'sizes': {'medium': {'w': 500, 'h': 750, 'resize': 'fit'}, 'th
umb': {'w': 150, 'h': 150, 'resize': 'crop'}, 'large': {'w': 500, 'h
': 750, 'resize': 'fit'}, 'small': {'w': 340, 'h': 510, 'resize': 'f

it'}}}]}, 'extended_entities': {'media': [{'id': 632984655252160512,
'id_str': '632984655252160512', 'indices': [118, 140], 'media_url':
'http://pbs.twimg.com/media/CMjQMLSUAAAfMUW.jpg', 'media_url_https':
'https://pbs.twimg.com/media/CMjQMLSUAAAfMUW.jpg', 'url': 'http://t.
co/IdjGgfgncz', 'display_url': 'pic.twitter.com/IdjGgfgncz', 'expand
ed_url': 'http://twitter.com/nitishuna/status/632984680887812096/pho
to/1', 'type': 'photo', 'sizes': {'medium': {'w': 500, 'h': 750, 're
size': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize': 'crop'}, 'lar
ge': {'w': 500, 'h': 750, 'resize': 'fit'}, 'small': {'w': 340, 'h':
510, 'resize': 'fit'}}}]}, 'favorited': False, 'retweeted': False, '
possibly_sensitive': False, 'filter_level': 'low', 'lang': 'und'}, '
retweet_count': 0, 'favorite_count': 0, 'entities': {'hashtags': [],
'trends': [], 'urls': [], 'user_mentions': [{'screen_name': 'nitishu
na', 'name': 'Just4fun', 'id': 445865965, 'id_str': '445865965', 'in
dices': [3, 13]}, {'screen_name': 'LasVegasChicas', 'name': 'What Ha
ppens n Vegas', 'id': 737103644, 'id_str': '737103644', 'indices': [
15, 30]}, {'screen_name': 'I_luv_reds', 'name': 'I_luv_reds', 'id':
718927286, 'id_str': '718927286', 'indices': [31, 42]}, {'screen_nam
e': 'DimondLadies', 'name': '⭐Dimond Ladies⭐', 'id': 2714436740, '
id_str': '2714436740', 'indices': [43, 56]}, {'screen_name': 'Silico
nEros', 'name': 'Silicon Eros', 'id': 2848949950, 'id_str': '2848949
950', 'indices': [57, 69]}, {'screen_name': 'herby2121', 'name': '☆
herby2121 ☆86K ', 'id': 2274691870, 'id_str': '2274691870', 'indices
': [70, 80]}, {'screen_name': 'hughcox69', 'name': '∩UDE PHoToGRAPH
Y 82k', 'id': 2767126967, 'id_str': '2767126967', 'indices': [81, 91
]}, {'screen_name': 'VxExx', 'name': '#VenusExchange☀43K☀', 'id': 3
229432811, 'id_str': '3229432811', 'indices': [92, 98]}, {'screen_na
me': 'Swt_Red', 'name': '✨ Sweet Red ✨', 'id': 3087022267, 'id_str
': '3087022267', 'indices': [99, 107]}, {'screen_name': 'tthumb9090'
, 'name': 'Mr Tasty ♻69K♻', 'id': 2665853560, 'id_str': '2665853560
', 'indices': [108, 119]}, {'screen_name': 'raggdoll_98', 'name': 'H
is Girl ', 'id': 352691921, 'id_str': '352691921', 'indices': [120,
132]}], 'symbols': [], 'media': [{'id': 632984655252160512, 'id_str'
: '632984655252160512', 'indices': [139, 140], 'media_url': 'http://
pbs.twimg.com/media/CMjQMLSUAAAfMUW.jpg', 'media_url_https': 'https:
//pbs.twimg.com/media/CMjQMLSUAAAfMUW.jpg', 'url': 'http://t.co/IdjG
gfgncz', 'display_url': 'pic.twitter.com/IdjGgfgncz', 'expanded_url'
: 'http://twitter.com/nitishuna/status/632984680887812096/photo/1',
'type': 'photo', 'sizes': {'medium': {'w': 500, 'h': 750, 'resize':
'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize': 'crop'}, 'large': {'
w': 500, 'h': 750, 'resize': 'fit'}, 'small': {'w': 340, 'h': 510, '
resize': 'fit'}}, 'source_status_id': 632984680887812096, 'source_st
atus_id_str': '632984680887812096'}]}, 'extended_entities': {'media'
: [{'id': 632984655252160512, 'id_str': '632984655252160512', 'indic
es': [139, 140], 'media_url': 'http://pbs.twimg.com/media/CMjQMLSUAA
AfMUW.jpg', 'media_url_https': 'https://pbs.twimg.com/media/CMjQMLSU
AAAfMUW.jpg', 'url': 'http://t.co/IdjGgfgncz', 'display_url': 'pic.t
witter.com/IdjGgfgncz', 'expanded_url': 'http://twitter.com/nitishun
a/status/632984680887812096/photo/1', 'type': 'photo', 'sizes': {'me
dium': {'w': 500, 'h': 750, 'resize': 'fit'}, 'thumb': {'w': 150, 'h
': 150, 'resize': 'crop'}, 'large': {'w': 500, 'h': 750, 'resize': '
fit'}, 'small': {'w': 340, 'h': 510, 'resize': 'fit'}}, 'source_stat
us_id': 632984680887812096, 'source_status_id_str': '632984680887812

096'}]}, 'favorited': False, 'retweeted': False, 'possibly_sensitive
': False, 'filter_level': 'low', 'lang': 'und', 'timestamp_ms': '143
9761273659'}
{'created_at': 'Sun Aug 16 21:41:13 +0000 2015', 'id': 6330307796315
66848, 'id_str': '633030779631566848', 'text': 'إنّ العرب إذا تغلبوا
على أوطان أسرع إليها الخراب والسبب في ذلك أنها أمة وحشية بإستحكام عو
ائد التوحش وأسبابه فصار لهم خلقة وجبلة n\n\- ابن خلدون', 'source': '<
a href="http://twitter.com/download/iphone" rel="nofollow">Twitter f
or iPhone', 'truncated': False, 'in_reply_to_status_id': None, '
in_reply_to_status_id_str': None, 'in_reply_to_user_id': None, 'in_r
eply_to_user_id_str': None, 'in_reply_to_screen_name': None, 'user':
{'id': 3110676047, 'id_str': '3110676047', 'name': 'اقتباس', 'screen
_name': 'eqtybas', 'location': '', 'url': None, 'description': 'أجمل
ماقال الحكماء على مر الأزمان', 'protected': False, 'verified': False,
'followers_count': 268, 'friends_count': 0, 'listed_count': 1, 'favo
urites_count': 2, 'statuses_count': 238, 'created_at': 'Tue Mar 24 1
5:56:17 +0000 2015', 'utc_offset': -25200, 'time_zone': 'Pacific Tim
e (US & Canada)', 'geo_enabled': False, 'lang': 'en', 'contributors_
enabled': False, 'is_translator': False, 'profile_background_color':
'000000', 'profile_background_image_url': 'http://abs.twimg.com/imag
es/themes/theme1/bg.png', 'profile_background_image_url_https': 'htt
ps://abs.twimg.com/images/themes/theme1/bg.png', 'profile_background
_tile': False, 'profile_link_color': '9266CC', 'profile_sidebar_bord
er_color': '000000', 'profile_sidebar_fill_color': '000000', 'profil
e_text_color': '000000', 'profile_use_background_image': False, 'pro
file_image_url': 'http://pbs.twimg.com/profile_images/61544279363734
7328/bRkcNeZS_normal.png', 'profile_image_url_https': 'https://pbs.t
wimg.com/profile_images/615442793637347328/bRkcNeZS_normal.png', 'de
fault_profile': False, 'default_profile_image': False, 'following':
None, 'follow_request_sent': None, 'notifications': None}, 'geo': No
ne, 'coordinates': None, 'place': None, 'contributors': None, 'retwe
et_count': 0, 'favorite_count': 0, 'entities': {'hashtags': [], 'tre
nds': [], 'urls': [], 'user_mentions': [], 'symbols': []}, 'favorite
d': False, 'retweeted': False, 'possibly_sensitive': False, 'filter_
level': 'low', 'lang': 'ar', 'timestamp_ms': '1439761273664'}
{'created_at': 'Sun Aug 16 21:41:13 +0000 2015', 'id': 6330307796064
66560, 'id_str': '633030779606466560', 'text': 'RT @moonmona7: لن ال
.فَـقُـذ رحلت لأنَـي ْ اَشُـبَـعـتك ُ حبآ حـتى أكتفيتn\... ومك على رحيلك
.. http://t.co/ceD9aTiFfI', 'source': '<a href="http://twitter.com/d
ownload/iphone" rel="nofollow">Twitter for iPhone', 'truncated':
False, 'in_reply_to_status_id': None, 'in_reply_to_status_id_str': N
one, 'in_reply_to_user_id': None, 'in_reply_to_user_id_str': None, '
in_reply_to_screen_name': None, 'user': {'id': 586818520, 'id_str':
'586818520', 'name': 'N.ALSHAHRI', 'screen_name': 'Anaalzeenn', 'loc
ation': '\u200f\u200b\u200f\u200b\u200f\u200b\u200f\u200b\u200f\u200
b\u200f\u200b\u200f\u200b', 'url': None, 'description': 'http://anaa
lzeenn.tumblr.com', 'protected': False, 'verified': False, 'follower
s_count': 332, 'friends_count': 277, 'listed_count': 0, 'favourites_
count': 153, 'statuses_count': 1467, 'created_at': 'Mon May 21 19:39
:58 +0000 2012', 'utc_offset': None, 'time_zone': None, 'geo_enabled
': True, 'lang': 'en', 'contributors_enabled': False, 'is_translator
': False, 'profile_background_color': 'C0DEED', 'profile_background_
image_url': 'http://abs.twimg.com/images/themes/theme1/bg.png', 'pro

file_background_image_url_https': 'https://abs.twimg.com/images/themes/theme1/bg.png', 'profile_background_tile': False, 'profile_link_color': '0084B4', 'profile_sidebar_border_color': 'C0DEED', 'profile_sidebar_fill_color': 'DDEEF6', 'profile_text_color': '333333', 'profile_use_background_image': True, 'profile_image_url': 'http://pbs.twimg.com/profile_images/512611052958138368/IZvuvMgu_normal.png', 'profile_image_url_https': 'https://pbs.twimg.com/profile_images/512611052958138368/IZvuvMgu_normal.png', 'profile_banner_url': 'https://pbs.twimg.com/profile_banners/586818520/1402684145', 'default_profile': True, 'default_profile_image': False, 'following': None, 'follow_request_sent': None, 'notifications': None}, 'geo': None, 'coordinates': None, 'place': None, 'contributors': None, 'retweeted_status': {'created_at': 'Sun Aug 16 16:01:06 +0000 2015', 'id': 632945185866018817, 'id_str': '632945185866018817', 'text': 'لن الـومك على رحيلك ...\ .. فَـقْـدْ رحلت لأنّـَي ْ اَشْـبَـعـتك ُ حبآ حـتى أكتفيتn http://t.co/ceD9aTiFfI', 'source': 'Twitter for Android', 'truncated': False, 'in_reply_to_status_id': None, 'in_reply_to_status_id_str': None, 'in_reply_to_user_id': None, 'in_reply_to_user_id_str': None, 'in_reply_to_screen_name': None, 'user': {'id': 196273015, 'id_str': '196273015', 'name': 'منــــــــــى الـهاشمي', 'screen_name': 'moonmona7', 'location': 'ابوظبي_❤_abudhabi #UAE', 'url': None, 'description': '\u200fأجمل حب فيf\u200fالـوجود حب (الله) ... الخاص مـمنوع -(مـا يـنبض بـه قـلبي تجسده حروفي – ومـا يـروق لـي أقـتبسه هنـا)', 'protected': False, 'verified': False, 'followers_count': 135426, 'friends_count': 113260, 'listed_count': 117, 'favourites_count': 6808, 'statuses_count': 11768, 'created_at': 'Tue Sep 28 18:52:17 +0000 2010', 'utc_offset': 14400, 'time_zone': 'Abu Dhabi', 'geo_enabled': True, 'lang': 'ar', 'contributors_enabled': False, 'is_translator': False, 'profile_background_color': 'C0DEED', 'profile_background_image_url': 'http://abs.twimg.com/images/themes/theme1/bg.png', 'profile_background_image_url_https': 'https://abs.twimg.com/images/themes/theme1/bg.png', 'profile_background_tile': False, 'profile_link_color': '0084B4', 'profile_sidebar_border_color': 'C0DEED', 'profile_sidebar_fill_color': 'DDEEF6', 'profile_text_color': '333333', 'profile_use_background_image': True, 'profile_image_url': 'http://pbs.twimg.com/profile_images/548572098344398848/7sErNY7Y_normal.jpeg', 'profile_image_url_https': 'https://pbs.twimg.com/profile_images/548572098344398848/7sErNY7Y_normal.jpeg', 'profile_banner_url': 'https://pbs.twimg.com/profile_banners/196273015/1439723717', 'default_profile': True, 'default_profile_image': False, 'following': None, 'follow_request_sent': None, 'notifications': None}, 'geo': None, 'coordinates': None, 'place': None, 'contributors': None, 'retweet_count': 15, 'favorite_count': 13, 'entities': {'hashtags': [], 'trends': [], 'urls': [], 'user_mentions': [], 'symbols': [], 'media': [{'id': 632945184238649344, 'id_str': '632945184238649344', 'indices': [86, 108], 'media_url': 'http://pbs.twimg.com/media/CMisSqTWwAARmiH.jpg', 'media_url_https': 'https://pbs.twimg.com/media/CMisSqTWwAARmiH.jpg', 'url': 'http://t.co/ceD9aTiFfI', 'display_url': 'pic.twitter.com/ceD9aTiFfI', 'expanded_url': 'http://twitter.com/moonmona7/status/632945185866018817/photo/1', 'type': 'photo', 'sizes': {'large': {'w': 442, 'h': 442, 'resize': 'fit'}, 'thumb': {'w': 150,

'h': 150, 'resize': 'crop'}, 'small': {'w': 340, 'h': 340, 'resize': 'fit'}, 'medium': {'w': 442, 'h': 442, 'resize': 'fit'}}}]}, 'extended_entities': {'media': [{'id': 632945184238649344, 'id_str': '632945184238649344', 'indices': [86, 108], 'media_url': 'http://pbs.twimg.com/media/CMisSqTWwAARmiH.jpg', 'media_url_https': 'https://pbs.twimg.com/media/CMisSqTWwAARmiH.jpg', 'url': 'http://t.co/ceD9aTiFfI', 'display_url': 'pic.twitter.com/ceD9aTiFfI', 'expanded_url': 'http://twitter.com/moonmona7/status/632945185866018817/photo/1', 'type': 'photo', 'sizes': {'large': {'w': 442, 'h': 442, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize': 'crop'}, 'small': {'w': 340, 'h': 340, 'resize': 'fit'}, 'medium': {'w': 442, 'h': 442, 'resize': 'fit'}}}]}, 'favorited': False, 'retweeted': False, 'possibly_sensitive': False, 'filter_level': 'low', 'lang': 'ar'}, 'retweet_count': 0, 'favorite_count': 0, 'entities': {'hashtags': [], 'trends': [], 'urls': [], 'user_mentions': [{'screen_name': 'moonmona7', 'name': 'م نــــــى الـهاشمي', 'id': 196273015, 'id_str': '196273015', 'indices': [3, 13]}], 'symbols': [], 'media': [{'id': 632945184238649344, 'id_str': '632945184238649344', 'indices': [101, 123], 'media_url': 'http://pbs.twimg.com/media/CMisSqTWwAARmiH.jpg', 'media_url_https': 'https://pbs.twimg.com/media/CMisSqTWwAARmiH.jpg', 'url': 'http://t.co/ceD9aTiFfI', 'display_url': 'pic.twitter.com/ceD9aTiFfI', 'expanded_url': 'http://twitter.com/moonmona7/status/632945185866018817/photo/1', 'type': 'photo', 'sizes': {'large': {'w': 442, 'h': 442, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize': 'crop'}, 'small': {'w': 340, 'h': 340, 'resize': 'fit'}, 'medium': {'w': 442, 'h': 442, 'resize': 'fit'}}, 'source_status_id': 632945185866018817, 'source_status_id_str': '632945185866018817'}]}, 'extended_entities': {'media': [{'id': 632945184238649344, 'id_str': '632945184238649344', 'indices': [101, 123], 'media_url': 'http://pbs.twimg.com/media/CMisSqTWwAARmiH.jpg', 'media_url_https': 'https://pbs.twimg.com/media/CMiSSqTWwAARmiH.jpg', 'url': 'http://t.co/ceD9aTiFfI', 'display_url': 'pic.twitter.com/ceD9aTiFfI', 'expanded_url': 'http://twitter.com/moonmona7/status/632945185866018817/photo/1', 'type': 'photo', 'sizes': {'large': {'w': 442, 'h': 442, 'resize': 'fit'}, 'thumb': {'w': 150, 'h': 150, 'resize': 'crop'}, 'small': {'w': 340, 'h': 340, 'resize': 'fit'}, 'medium': {'w': 442, 'h': 442, 'resize': 'fit'}}, 'source_status_id': 632945185866018817, 'source_status_id_str': '632945185866018817'}]}, 'favorited': False, 'retweeted': False, 'possibly_sensitive': False, 'filter_level': 'low', 'lang': 'ar', 'timestamp_ms': '1439761273658'}
{'created_at': 'Sun Aug 16 21:41:13 +0000 2015', 'id': 633030779623206912, 'id_str': '633030779623206912', 'text': '@16_Pirates any shots?', 'source': 'Twitter for Android', 'truncated': False, 'in_reply_to_status_id': 633030587733798912, 'in_reply_to_status_id_str': '633030587733798912', 'in_reply_to_user_id': 542222500, 'in_reply_to_user_id_str': '542222500', 'in_reply_to_screen_name': '16_Pirates', 'user': {'id': 425394759, 'id_str': '425394759', 'name': 'Braydoe ', 'screen_name': 'BraydenKade', 'location': '', 'url': None, 'description': '801 IG & SC: Braydenkade', 'protected': False, 'verified': False, 'followers_count': 321, 'friends_count': 274, 'listed_count': 0, 'favourites_count': 3160, 'statuses_count': 7013, 'created_at': 'Wed Nov 30 23:21:33 +0000 2011', 'utc_offset': None, 'time_zone': None, 'geo

_enabled': True, 'lang': 'en', 'contributors_enabled': False, 'is_tr
anslator': False, 'profile_background_color': 'C0DEED', 'profile_bac
kground_image_url': 'http://abs.twimg.com/images/themes/theme1/bg.pn
g', 'profile_background_image_url_https': 'https://abs.twimg.com/ima
ges/themes/theme1/bg.png', 'profile_background_tile': False, 'profil
e_link_color': '0084B4', 'profile_sidebar_border_color': 'C0DEED', '
profile_sidebar_fill_color': 'DDEEF6', 'profile_text_color': '333333
', 'profile_use_background_image': True, 'profile_image_url': 'http:
//pbs.twimg.com/profile_images/622558667557048320/t-18HzWQ_normal.jp
g', 'profile_image_url_https': 'https://pbs.twimg.com/profile_images
/622558667557048320/t-18HzWQ_normal.jpg', 'profile_banner_url': 'htt
ps://pbs.twimg.com/profile_banners/425394759/1402107210', 'default_p
rofile': True, 'default_profile_image': False, 'following': None, 'f
ollow_request_sent': None, 'notifications': None}, 'geo': None, 'coo
rdinates': None, 'place': None, 'contributors': None, 'retweet_count
': 0, 'favorite_count': 0, 'entities': {'hashtags': [], 'trends': []
, 'urls': [], 'user_mentions': [{'screen_name': '16_Pirates', 'name'
: 'Josh Painton', 'id': 542222500, 'id_str': '542222500', 'indices':
[0, 11]}], 'symbols': []}, 'favorited': False, 'retweeted': False, '
possibly_sensitive': False, 'filter_level': 'low', 'lang': 'en', 'ti
mestamp_ms': '1439761273662'}
{'created_at': 'Sun Aug 16 21:41:13 +0000 2015', 'id': 6330307796022
43584, 'id_str': '633030779602243584', 'text': 'RT @Cancer_gk: A #Ca
ncer may act shy and quiet, but will adamantly defend a loved one ag
ainst outsiders.', 'source': '<a href="https://mobile.twitter.com" r
el="nofollow">Mobile Web (M2)', 'truncated': False, 'in_reply_to
_status_id': None, 'in_reply_to_status_id_str': None, 'in_reply_to_u
ser_id': None, 'in_reply_to_user_id_str': None, 'in_reply_to_screen_
name': None, 'user': {'id': 2254825403, 'id_str': '2254825403', 'nam
e': 'Halima', 'screen_name': 'Halima1_', 'location': 'Khulna,Banglad
esh', 'url': 'http://www.wattpad.com/user/IamTheRainbowGirl#', 'desc
ription': 'Life is short. So Smile while you still have Teeth :D', '
protected': False, 'verified': False, 'followers_count': 13, 'friend
s_count': 76, 'listed_count': 1, 'favourites_count': 94, 'statuses_c
ount': 202, 'created_at': 'Tue Dec 31 18:23:11 +0000 2013', 'utc_off
set': None, 'time_zone': None, 'geo_enabled': True, 'lang': 'en', 'c
ontributors_enabled': False, 'is_translator': False, 'profile_backgr
ound_color': 'C0DEED', 'profile_background_image_url': 'http://abs.t
wimg.com/images/themes/theme1/bg.png', 'profile_background_image_url
_https': 'https://abs.twimg.com/images/themes/theme1/bg.png', 'profi
le_background_tile': False, 'profile_link_color': '0084B4', 'profile
_sidebar_border_color': 'C0DEED', 'profile_sidebar_fill_color': 'DDE
EF6', 'profile_text_color': '333333', 'profile_use_background_image'
: True, 'profile_image_url': 'http://pbs.twimg.com/profile_images/62
03703546245652482/Cj1XOLNZ_normal.jpg', 'profile_image_url_https': 'h
ttps://pbs.twimg.com/profile_images/6203703546245652482/Cj1XOLNZ_norm
al.jpg', 'default_profile': True, 'default_profile_image': False, 'f
ollowing': None, 'follow_request_sent': None, 'notifications': None}
, 'geo': None, 'coordinates': None, 'place': None, 'contributors': N
one, 'retweeted_status': {'created_at': 'Sat Aug 15 02:00:10 +0000 2
015', 'id': 632371171237888000, 'id_str': '632371171237888000', 'tex
t': 'A #Cancer may act shy and quiet, but will adamantly defend a lo
ved one against outsiders.', 'source': '<a href="http://twitter.com"

rel="nofollow">iMTS', 'truncated': False, 'in_reply_to_status_id
': None, 'in_reply_to_status_id_str': None, 'in_reply_to_user_id': N
one, 'in_reply_to_user_id_str': None, 'in_reply_to_screen_name': Non
e, 'user': {'id': 1575097585, 'id_str': '1575097585', 'name': 'Cance
r Zodiac', 'screen_name': 'Cancer_gk', 'location': '', 'url': 'http:
//bit.ly/1e4gJh3', 'description': 'Not Cancer? Follow your sign – @A
ries_gk @Taurus_gk @Gemini__gk @Leo__gk @Virgo_gk @Libra_gk @Scorpio
_gk @Sagittarius__gk @Capricorn_gk @Aquarius_gk @Pisces_gk', 'protec
ted': False, 'verified': False, 'followers_count': 89117, 'friends_c
ount': 14902, 'listed_count': 138, 'favourites_count': 0, 'statuses_
count': 1718, 'created_at': 'Sun Jul 07 12:40:15 +0000 2013', 'utc_o
ffset': -18000, 'time_zone': 'Central Time (US & Canada)', 'geo_enab
led': False, 'lang': 'en', 'contributors_enabled': False, 'is_transl
ator': False, 'profile_background_color': 'DBE9ED', 'profile_backgro
und_image_url': 'http://abs.twimg.com/images/themes/theme17/bg.gif',
'profile_background_image_url_https': 'https://abs.twimg.com/images/
themes/theme17/bg.gif', 'profile_background_tile': False, 'profile_l
ink_color': 'CC3366', 'profile_sidebar_border_color': 'DBE9ED', 'pro
file_sidebar_fill_color': 'E6F6F9', 'profile_text_color': '333333',
'profile_use_background_image': True, 'profile_image_url': 'http://p
bs.twimg.com/profile_images/571741716420362240/NcZ9I7OO_normal.jpeg'
, 'profile_image_url_https': 'https://pbs.twimg.com/profile_images/5
71741716420362240/NcZ9I7OO_normal.jpeg', 'profile_banner_url': 'http
s://pbs.twimg.com/profile_banners/1575097585/1425150179', 'default_p
rofile': False, 'default_profile_image': False, 'following': None, '
follow_request_sent': None, 'notifications': None}, 'geo': None, 'co
ordinates': None, 'place': None, 'contributors': None, 'retweet_coun
t': 119, 'favorite_count': 123, 'entities': {'hashtags': [{'text': '
Cancer', 'indices': [2, 9]}], 'trends': [], 'urls': [], 'user_mentio
ns': [], 'symbols': []}, 'favorited': False, 'retweeted': False, 'po
ssibly_sensitive': False, 'filter_level': 'low', 'lang': 'en'}, 'ret
weet_count': 0, 'favorite_count': 0, 'entities': {'hashtags': [{'tex
t': 'Cancer', 'indices': [17, 24]}], 'trends': [], 'urls': [], 'user
_mentions': [{'screen_name': 'Cancer_gk', 'name': 'Cancer Zodiac', '
id': 1575097585, 'id_str': '1575097585', 'indices': [3, 13]}], 'symb
ols': []}, 'favorited': False, 'retweeted': False, 'possibly_sensiti
ve': False, 'filter_level': 'low', 'lang': 'en', 'timestamp_ms': '14
39761273657'}

Each entry in `tweets.json`, i.e., each `tweet`, corresponds to a dictionary that contains lots of information about the tweet, the user, the activity related to the tweet (i.e., if it was retweeted or not), the timestamp of the tweet, entities mentioned in the tweet, hashtags used, etc.

You can treat the `tweet` variable from above as a dicitonary and use the `.keys()` command to see the fields associated with the dictionary.

```
In [19]:  for k in tweet.keys():
              print(k)
```

```
created_at
id
id_str
text
source
truncated
in_reply_to_status_id
in_reply_to_status_id_str
in_reply_to_user_id
in_reply_to_user_id_str
in_reply_to_screen_name
user
geo
coordinates
place
contributors
retweeted_status
retweet_count
favorite_count
entities
favorited
retweeted
possibly_sensitive
filter_level
lang
timestamp_ms
```

We can select any of the aforemented values of Variable `tweet` by treating it as a dictionary. For example let's select the `text` body of the tweet, the time it was `created_at`, and the `hashtags` it contains.

```
In [20]:  body = tweet['text']
          tweet_time   = tweet['created_at']

          # tweet['entities'] is itself a dictionary with more entries. Hashtags
          is one of them.
          htags   = tweet['entities']['hashtags']

          print('Tweet body: ', body)
          print('Creation Time:', tweet_time)
          print('Hastags: ', htags)
```

```
Tweet body:  RT @Cancer_gk: A #Cancer may act shy and quiet, but wil
l adamantly defend a loved one against outsiders.
Creation Time: Sun Aug 16 21:41:13 +0000 2015
Hastags:  [{'text': 'Cancer', 'indices': [17, 24]}]
```

As you can see this tweet contains no hashtags. The body of the tweet contains several information that is not necesary for our sentiment analysis task. For example, it contains a comma, a reference to a twitter user and a link to an external website.

Since this information is not necessary we can remove it. In other words we need to clean our input in order to prepare it for analysis. Next, we show you some basic cleaning operations using **regular expressions**. You can find more information on regular expressions here (https://medium.com/factory-mind/regex-tutorial-a-simple-cheatsheet-by-examples-649dc1c3f285).

In [21]:
```python
# Basic steps for cleaning process.
import re
# Step 1: Convert tweet to lower case
body = body.lower()
# Step 2: Find URLs and replace them with an empty string
body = re.sub(r'((www\.[\S]+)|(https?://[\S]+))', '', body)
# Step 3: Find @<user> mentions and replace them with an empty string
body = re.sub(r'@[\S]+', '', body)

print('Clean tweet body:', body)
```

Clean tweet body: rt a #cancer may act shy and quiet, but will adam antly defend a loved one against outsiders.

We are providing you with a Python script named `preprocess.py`. The script `preprocess.py` accepts one argument on the command line: a JSON file with tweets (i.e., `tweets.json`). You can run the program like this:

```
$ python3 preprocess.py tweets.json
```

There are some parts specified in this script that you need to implement. The goal of this script is to clean all the tweets in `tweets.json`. Running `preprocess.py` will generate an output file named `clean_tweets.txt` containing **one string per line** containing a clean tweet. The order of the clean tweets in your output file should follow the order of the lines in the original `tweets.json`. Basically, the first line in `clean_tweets.txt` should correspond to the first raw tweet in `tweets.json`, the second line should correspond to the second tweet, and so on. If you perform any sorting or you put the processed data in a dictionary the order will not be preserved. Once again: **The n-th line of `clean_tweets.txt` (the file you will submit) should be a string that represent the clean version of the n-the line in the `tweets.json` (the input file).**

You must provide a line for **every** tweet. If the clean tweet is the empty string then just provide a line with the empty string.

If you have implemented everything correctly, the first 8 lines of the generated output should be exactly the same as the next lines:

```
    any shots
    a cancer may act shy and quiet but will adamantly defend a loved one again
    st outsiders
```

The first 4 lines will be empty and the last 2 will have the string shown above.

Note: This is real-world data, and it can be messy! For example, not all json lines may contain valid tweets (i.e., a text field). Ask for help on Piazza if you get stuck!

What to turn in: The file `clean_tweets.txt` output by `preprocess.py` after you have implemented the missing parts in `preprocess.py`.

Question 2: Derive the sentiment of each tweet [40 points]

For this part, you will compute the sentiment of each clean tweet in `clean_tweets.txt` 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 clean tweet.

You are provided with a skeleton file `tweet_sentiment.py` which accepts two arguments on the command line: a *sentiment file* and a tweet file like the one you generated in Question 1. You can run the skeleton program like this:

```
$ python3 tweet_sentiment.py AFINN-111.txt clean_tweets.txt
```

The file `AFINN-111.txt` contains a list of pre-computed sentiment scores. Each line in the file contains a word or phrase phollowed by a sentiment score. Each word or phrase that is found in a tweet but not found 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 corresponds to the string "\t". The following snipped of code may be useful:

```python
In [ ]:  import sys
         afinnfile_name = open(sys.argv[1])
         afinnfile = open(afinnfile_name, 'r')
         scores = {} # initialize an empty dictionary
         for line in afinnfile:
             term, score = line.split("\t") # The file is tab-delimited and "\t
         " means tab character
             scores[term] = int(score) # Conver the score to an integer. It was
         parsed as a string.
         afinnfile.close()
         print(scores.items( )) # Print every (term, score) pair in the diction
         ary
```

Your script should output a file named `sentiment.txt` containing the sentiment of each tweet in the file `clean_tweets.txt`, one numeric sentiment score per line. The first score should correspond to the first tweet, the second score should correspond to the second tweet, and so on. In other words, **the n-th line of the file you submit should contain only a single number that represents teh score of the n-th tweet in the input file.**

After you have implemented everything the first 10 lines of the generated output of your script should be exactly the same as the next lines:

```
0
0
0
0
0
1
2
-4
0
0
```

What to turn in: The file `sentiment.txt` after you have verified that it returns the correct answers

Question 3: Derive the sentiment of new terms [40 points]

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

You can think about this problem as follows: We know wwe can use the sentiment-carrying words in `AFINN-111.txt` to deduce the overall sentiment of a tweet. Once you deduce the sentiment of a tweet, you can work backwards to deduce the sentiment of the non-sentiment carrying words that *do not appear* in `AFINN-111.txt`. For example, if the word *football* always appears in proximity with positive words like *great* and *fun*, then we can deduce that the term *football* itself carried a positive sentiment.

You are provided with a skeleton file `term_sentiment.py` which accepts the same two arguments as `tweet_sentiment.py` and can be executed using the following command:

`$ python3 term_sentiment.py AFINN-111.txt clean_tweets.txt`

Your script should print its output to stdout. Each line of the output should contain a term, followed by a space, followed by a sentiment. That is, each line should be in the format <u>term:string (term:string)</u> <u>sentiment:float (sentiment:float)</u>. For example if you have the pair ("foo", 54.2) in Python, it should appear in the output as: `foo 54.2`.

The order of your output does not matter.

What to turn in: The file `term_sentiment.py` after you have implemented the missing parts.

To grade your submission we will run your script on a file that contains strongly positive and strong negative tweets and verify that the terms in the strongly positive tweets are assigned a higher score than the terms in the negative tweets. Your score need not exactly match any specific solution.