Ekaterina Lyapina (Student ID 160984099)
Msc Big Data Science

DMSN Coursework 1: Crawling Twitter

For this coursework an application was created through Twitters app dashboard:

The crawler was written using Python. An access can be made using the following code:

```
In [1]: from __future__ import absolute_import, print_function
        import json
        import tweepy
        from tweepy.parsers import import JSONParser


        consumer_key="l6pgWANSy..."
        consumer_secret="..."
        access_token="..."
        access_token_secret="..."

        auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
        auth.set_access_token(access_token, access_token_secret)
        api = tweepy.API(auth, wait_on_rate_limit=True, parser=JSONParser())

        nickname = 'realDonaldTrump'
```

Twitter API profiles endpoints were requested:

# parsing profile of the user
user = api.get_user(nickname)
user = json.dumps(user)
print(user)

{"follow_request_sent": false, "has_extended_profile": false, "profile_use_background_image": true, "profile_text_col
or": "333333", "default_profile_image": false, "id": 25073877, "profile_background_image_url_https": "https://pbs.twi
mg.com/profile_background_images/530021613/trump_scotland__43_of_70_cc.jpg", "verified": true, "translator_type": "re
gular", "profile_location": null, "profile_image_url_https": "https://pbs.twimg.com/profile_images/87427619735759667
2/kUuht00m_normal.jpg", "profile_sidebar_fill_color": "C5CEC0", "entities": {"url": {"urls": [{"url": "https://t.co/O
MxB0x7xC5", "indices": [0, 23], "expanded_url": "http://www.Instagram.com/realDonaldTrump", "display_url": "Instagra
m.com/realDonaldTrump"}]}, "description": {"urls": []}}, "followers_count": 48492518, "profile_sidebar_border_color":
"BDDCAD", "id_str": "25073877", "profile_background_color": "6D5C18", "listed_count": 85146, "status": {"contributor
s": null, "truncated": true, "text": "The United States has an $800 Billion Dollar Yearly Trade Deficit because of ou
r \u201cvery stupid\u201d trade deals and poli\u2026 https://t.co/nFF3vEmRmY", "is_quote_status": false, "in_reply_to
_status_id": null, "id": 96999165393039361, "favorite_count": 519, "source": "<a href=\"http://twitter.com/download/
iphone\" rel=\"nofollow\">Twitter for iPhone", "retweeted": false, "coordinates": null, "entities": {"symbols":
[], "user_mentions": [], "hashtags": [], "urls": [{"url": "https://t.co/nFF3vEmRmY", "indices": [117, 140], "expande
d_url": "https://twitter.com/i/web/status/96999165393039361", "display_url": "twitter.com/i/web/status/9\u2026"}]},
"in_reply_to_screen_name": null, "in_reply_to_user_id": null, "retweet_count": 200, "id_str": "96999165393039361",
"favorited": false, "geo": null, "in_reply_to_user_id_str": null, "lang": "en", "created_at": "Sat Mar 03 17:43:26 +
0000 2018", "in_reply_to_status_id_str": null, "place": null}, "is_translation_enabled": true, "utc_offset": -18000,
"statuses_count": 37059, "description": "45th President of the United States of America\ud83c\uddfa\ud83c\uddf8", "f
riends_count": 45, "location": "Washington, DC", "profile_link_color": "1B95E0", "profile_image_url": "http://pbs.twi

```python
# parsing timeline of the user
timeline = api.user_timeline(nickname, count=1)
timeline = json.dumps(timeline)
print(timeline)
```

[{"contributors": null, "truncated": true, "text": "The United States has an $800 Billion Dollar Yearly Trade Deficit
because of our \u201cvery stupid\u201d trade deals and poli\u2026 https://t.co/nFF3vEmRmY", "is_quote_status": false,
"in_reply_to_status_id": null, "id": 96999165393039361, "favorite_count": 578, "source": "<a href=\"http://twitter.
com/download/iphone\" rel=\"nofollow\">Twitter for iPhone", "retweeted": false, "coordinates": null, "entities":
{"symbols": [], "user_mentions": [], "hashtags": [], "urls": [{"url": "https://t.co/nFF3vEmRmY", "indices": [117, 14
0], "expanded_url": "https://twitter.com/i/web/status/96999165393039361", "display_url": "twitter.com/i/web/status/9
\u2026"}]}, "in_reply_to_screen_name": null, "in_reply_to_user_id": null, "retweet_count": 219, "id_str": "9699916533
93039361", "favorited": false, "user": {"follow_request_sent": false, "has_extended_profile": false, "profile_use_bac
kground_image": true, "default_profile_image": false, "id": 25073877, "profile_background_image_url_https": "https://
pbs.twimg.com/profile_background_images/530021613/trump_scotland__43_of_70_cc.jpg", "verified": true, "translator_typ
e": "regular", "profile_text_color": "333333", "profile_image_url_https": "https://pbs.twimg.com/profile_images/87427
6197357596672/kUuht00m_normal.jpg", "profile_sidebar_fill_color": "C5CEC0", "entities": {"url": {"urls": [{"url": "ht
tps://t.co/OMxB0x7xC5", "indices": [0, 23], "expanded_url": "http://www.Instagram.com/realDonaldTrump", "display_ur
l": "Instagram.com/realDonaldTrump"}]}, "description": {"urls": []}}, "followers_count": 48492539, "profile_sidebar_b
order_color": "BDDCAD", "id_str": "25073877", "profile_background_color": "6D5C18", "listed_count": 85146, "is_transl
ation_enabled": true, "utc_offset": -18000, "statuses_count": 37059, "description": "45th President of the United Sta
tes of America\ud83c\uddfa\ud83c\uddf8", "friends_count": 45, "location": "Washington, DC", "profile_link_color": "1B
95E0", "profile_image_url": "http://pbs.twimg.com/profile_images/874276197357596672/kUuht00m_normal.jpg", "followin
g": false, "geo_enabled": true, "profile_banner_url": "https://pbs.twimg.com/profile_banners/25073877/1520088304", "p
rofile_background_image_url": "http://pbs.twimg.com/profile_background_images/530021613/trump_scotland__43_of_70_cc.j
pg", "screen_name": "realDonaldTrump", "lang": "en", "profile_background_tile": true, "favourites_count": 24, "name":
"Donald J. Trump", "notifications": false, "url": "https://t.co/OMxB0x7xC5", "created_at": "Wed Mar 18 13:46:38 +000
0 2009", "contributors_enabled": false, "time_zone": "Eastern Time (US & Canada)", "protected": false, "default_profi
le": false, "is_translator": false}, "geo": null, "in_reply_to_user_id_str": null, "lang": "en", "created_at": "Sat M
ar 03 17:43:26 +0000 2018", "in_reply_to_status_id_str": null, "place": null}]

```python
#parsing followers of the user
followers = api.followers(nickname, count=1)
followers = json.dumps(followers)
print(followers)
```

{"previous_cursor": 0, "previous_cursor_str": "0", "next_cursor": 1593939388844547156, "users": [{"follow_request_sen
t": false, "has_extended_profile": false, "profile_use_background_image": true, "live_following": false, "default_pro
file_image": true, "id": 1383501900, "profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme
1/bg.png", "translator_type": "none", "verified": false, "blocked_by": false, "profile_text_color": "333333", "mutin
g": false, "profile_image_url_https": "https://abs.twimg.com/sticky/default_profile_images/default_profile_normal.pn
g", "profile_sidebar_fill_color": "DDEEF6", "entities": {"description": {"urls": []}}, "followers_count": 0, "profile
_sidebar_border_color": "C0DEED", "id_str": "1383501900", "profile_background_color": "C0DEED", "listed_count": 0, "i
s_translation_enabled": false, "utc_offset": null, "statuses_count": 0, "description": "", "friends_count": 46, "loca
tion": "", "profile_link_color": "1DA1F2", "profile_image_url": "http://abs.twimg.com/sticky/default_profile_images/d
efault_profile_normal.png", "following": false, "geo_enabled": false, "blocking": false, "profile_background_image_ur
l": "http://abs.twimg.com/images/themes/theme1/bg.png", "screen_name": "cotachell", "lang": "en", "profile_background
_tile": false, "favourites_count": 1, "name": "Richelle Cota", "notifications": false, "url": null, "created_at": "Sa
t Apr 27 04:15:03 +0000 2013", "contributors_enabled": false, "time_zone": null, "protected": false, "default_profil
e": true, "is_translator": false}], "next_cursor_str": "1593939388844547156"}

```python
friends = api.friends(nickname, count=1)
friends = json.dumps(friends)
print(friends)
```

{"previous_cursor": 0, "previous_cursor_str": "0", "next_cursor": 1579256327759506458, "users": [{"follow_request_sen
t": false, "has_extended_profile": false, "profile_use_background_image": true, "live_following": false, "default_pro
file_image": false, "id": 818927131883356161, "profile_background_image_url_https": null, "translator_type": "none",
"verified": true, "blocked_by": false, "profile_text_color": "333333", "muting": false, "profile_image_url_https": "h
ttps://pbs.twimg.com/profile_images/893600318025474048/CbCbznYp_normal.jpg", "profile_sidebar_fill_color": "DDEEF6",
"entities": {"url": {"urls": [{"url": "https://t.co/wyOVgSLgBV", "indices": [0, 23], "expanded_url": "http://www.Whit
eHouse.gov", "display_url": "WhiteHouse.gov"}]}, "description": {"urls": [{"url": "https://t.co/eVVzoBb3Zr", "indice
s": [105, 128], "expanded_url": "http://wh.gov/privacy", "display_url": "wh.gov/privacy"}]}}, "followers_count": 2723
088, "profile_sidebar_border_color": "C0DEED", "id_str": "818927131883356161", "profile_background_color": "F5F8FA",
"listed_count": 6494, "status": {"contributors": null, "truncated": false, "text": "RT @FoxNews: .@PressSec: \"We jus
t passed the biggest tax cuts in history. ISIS is on the run. We're rebuilding our military. We're rebuildi\u2026",
"is_quote_status": false, "in_reply_to_status_id": null, "id": 969735751989710848, "favorite_count": 0, "source": "<a
href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone", "retweeted": false, "coordinate
s": null, "entities": {"symbols": [], "user_mentions": [{"id": 1367531, "indices": [3, 11], "id_str": "1367531", "scr
een_name": "FoxNews", "name": "Fox News"}, {"id": 818927131883356161, "indices": [14, 23], "id_str": "818927131883356
161", "screen_name": "PressSec", "name": "Sarah Sanders"}], "hashtags": [], "urls": []}, "in_reply_to_screen_name": n
ull, "in_reply_to_user_id": null, "retweet_count": 1399, "id_str": "969735751989710848", "favorited": false, "retweet
ed_status": {"contributors": null, "truncated": true, "text": ".@PressSec: \"We just passed the biggest tax cuts in h
istory. ISIS is on the run. We're rebuilding our military. We'\u2026 https://t.co/jQIzX7zeUI", "is_quote_status": fal
se, "in_reply_to_status_id": null, "id": 969652217132277761, "favorite_count": 5885, "source": "<a href=\"https://stu
dio.twitter.com\" rel=\"nofollow\">Media Studio", "retweeted": false, "coordinates": null, "entities": {"symbol
s": [], "user_mentions": [{"id": 818927131883356161, "indices": [1, 10], "id_str": "818927131883356161", "screen_nam
e": "PressSec", "name": "Sarah Sanders"}], "hashtags": [], "urls": [{"url": "https://t.co/jQIzX7zeUI", "indices": [11
7, 140], "expanded_url": "https://twitter.com/i/web/status/969652217132277761", "display_url": "twitter.com/i/web/sta
tus/9\u2026"}]}, "in_reply_to_screen_name": null, "in_reply_to_user_id": null, "retweet_count": 1399, "id_str": "9696
52217132277761", "favorited": false, "geo": null, "in_reply_to_user_id_str": null, "possibly_sensitive": false, "lan
g": "en", "created_at": "Fri Mar 02 19:14:38 +0000 2018", "in_reply_to_status_id_str": null, "place": null}, "geo": n
ull, "in_reply_to_user_id_str": null, "lang": "en", "created_at": "Sat Mar 03 00:46:34 +0000 2018", "in_reply_to_stat
us_id_str": null, "place": null}, "is_translation_enabled": false, "utc_offset": -28800, "statuses_count": 1917, "des
cription": "@WhiteHouse Press Secretary. Proudly representing @POTUS Trump's Administration. Tweets may be archived:
https://t.co/eVVzoBb3Zr.", "friends_count": 237, "location": "Washington, D.C.", "profile_link_color": "1DA1F2", "pro
file_image_url": "http://pbs.twimg.com/profile_images/893600318025474048/CbCbznYp_normal.jpg", "following": false, "g
eo_enabled": false, "profile_banner_url": "https://pbs.twimg.com/profile_banners/818927131883356161/1501886832", "blo
cking": false, "profile_background_image_url": null, "screen_name": "PressSec", "lang": "en", "profile_background_til
e": false, "favourites_count": 320, "name": "Sarah Sanders", "notifications": false, "url": "https://t.co/wyOVgSLgB
V", "created_at": "Tue Jan 10 21:06:57 +0000 2017", "contributors_enabled": false, "time_zone": "Pacific Time (US & C
anada)", "protected": false, "default_profile": true, "is_translator": false}], "next_cursor_str": "15792563277595064
58"}

# parsing friends ids to see the profile individually
friends_list = api.friends_ids(nickname)
friends_list = friends_list["ids"]
print(friends_list)
print(len(friends_list))

```
[818927131883356161, 22703645, 56561449, 822215673812119553, 823367015830323201, 471672239, 20733972, 322293052, 7202
93443260456960, 2325495378, 245963716, 50769180, 22203756, 729676086632656900, 14669951, 475802156, 75541946, 4163452
0, 37764422, 4121225056, 196168350, 2908170952, 34852681, 25429371, 52136185, 15513604, 108471631, 246500501, 1483914
7, 316865635, 26829777, 1222639789, 620571475, 633797941, 726414091, 82504071, 48223726, 1013672084, 29562422, 393498
94, 23970102, 16031927, 216299334, 39344374, 52544275]
45
```

# loop through each friens by id and download profile
type(friends_list)
l = []
for i in friends_list:
    # parsing profile of each user
    user = api.get_user(i)
    user = json.dumps(user)
    l.append(user)
# print the first friend's profile as an example
print(l[0])

{"follow_request_sent": false, "has_extended_profile": false, "profile_use_background_image": true, "profile_text_col
or": "333333", "default_profile_image": false, "id": 818927131883356161, "profile_background_image_url_https": null,
"verified": true, "translator_type": "none", "profile_location": null, "profile_image_url_https": "https://pbs.twimg.
com/profile_images/893600318025474048/CbCbznYp_normal.jpg", "profile_sidebar_fill_color": "DDEEF6", "entities": {"ur
l": {"urls": [{"url": "https://t.co/wyOVgSLgBV", "indices": [0, 23], "expanded_url": "http://www.WhiteHouse.gov", "di
splay_url": "WhiteHouse.gov"}]}, "description": {"urls": [{"url": "https://t.co/eVVzoBb3Zr", "indices": [105, 128],
"expanded_url": "http://wh.gov/privacy", "display_url": "wh.gov/privacy"}]}}, "followers_count": 2723051, "profile_si
debar_border_color": "C0DEED", "id_str": "818927131883356161", "profile_background_color": "F5F8FA", "listed_count":
6494, "status": {"contributors": null, "truncated": false, "text": "RT @FoxNews: .@PressSec: \"We just passed the big
gest tax cuts in history. ISIS is on the run. We're rebuilding our military. We're rebuildi\u2026", "is_quote_statu
s": false, "in_reply_to_status_id": null, "id": 969735751989710848, "favorite_count": 0, "source": "<a href=\"http://
twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone", "retweeted": false, "coordinates": null, "ent
ities": {"symbols": [], "user_mentions": [{"id": 1367531, "indices": [3, 11], "id_str": "1367531", "screen_name": "Fo
xNews", "name": "Fox News"}, {"id": 818927131883356161, "indices": [14, 23], "id_str": "818927131883356161", "screen_
name": "PressSec", "name": "Sarah Sanders"}], "hashtags": [], "urls": []}, "in_reply_to_screen_name": null, "in_reply
_to_user_id": null, "retweet_count": 1399, "id_str": "969735751989710848", "favorited": false, "retweeted_status":
{"contributors": null, "truncated": true, "text": ".@PressSec: \"We just passed the biggest tax cuts in history. ISIS
is on the run. We're rebuilding our military. We\u2026 https://t.co/jQIzX7zeUI", "is_quote_status": false, "in_reply
_to_status_id": null, "id": 969652217132277761, "favorite_count": 5885, "source": "<a href=\"https://studio.twitter.c
om\" rel=\"nofollow\">Media Studio", "retweeted": false, "coordinates": null, "entities": {"symbols": [], "user_m
entions": [{"id": 818927131883356161, "indices": [1, 10], "id_str": "818927131883356161", "screen_name": "PressSec",
"name": "Sarah Sanders"}], "hashtags": [], "urls": [{"url": "https://t.co/jQIzX7zeUI", "indices": [117, 140], "expan
ded_url": "https://twitter.com/i/web/status/96965221713277761", "display_url": "twitter.com/i/web/status/9\u202
6"}]}, "in_reply_to_screen_name": null, "in_reply_to_user_id": null, "retweet_count": 1399, "id_str": "96965221713227
7761", "favorited": false, "geo": null, "in_reply_to_user_id_str": null, "possibly_sensitive": false, "lang": "en",
"created_at": "Fri Mar 02 19:14:38 +0000 2018", "in_reply_to_status_id_str": null, "place": null}, "geo": null, "in_
reply_to_user_id_str": null, "lang": "en", "created_at": "Sat Mar 03 00:46:34 +0000 2018", "in_reply_to_status_id_st
r": null, "place": null}, "is_translation_enabled": false, "utc_offset": -28800, "statuses_count": 1917, "descriptio
n": "@WhiteHouse Press Secretary. Proudly representing @POTUS Trump's Administration. Tweets may be archived: http
s://t.co/eVVzoBb3Zr.", "friends_count": 237, "location": "Washington, D.C.", "profile_link_color": "1DA1F2", "profile
_image_url": "http://pbs.twimg.com/profile_images/893600318025474048/CbCbznYp_normal.jpg", "following": false, "geo_e
nabled": false, "profile_banner_url": "https://pbs.twimg.com/profile_banners/818927131883356161/1501886832", "profile
_background_image_url": null, "name": "Sarah Sanders", "lang": "en", "profile_background_tile": false, "favourites_co
unt": 320, "screen_name": "PressSec", "notifications": false, "url": "https://t.co/wyOVgSLgBV", "created_at": "Tue Ja
n 10 21:06:57 +0000 2017", "contributors_enabled": false, "time_zone": "Pacific Time (US & Canada)", "protected": fal
se, "default_profile": true, "is_translator": false}