ISA 414 – Managing Big Data

Lecture 17 – Midterm Project

Dr. Arthur Carvalho
arthur.carvalho@miamioh.edu

MIAMI UNIVERSITY

Agenda

- Today: introduction to the midterm project problem
  - APIs, libraries, *etc*.
- Thursday, 10/21: Project consulting (optional)
  - No lecture. Why?
    - API keys will be shared among all
      - You might need extra time to complete the project because of limited API quota
      - I need extra time to monitor API usage and react to problems
    - I want to give you time to play and explore the Twitter API
      - It might (or might not) be useful in your final project
- Friday, 10/22 at 11:59 pm: Project deadline

Lecture Objectives

- Quick review of Homework 8
- Learn about social media analytics
  - Focus on Twitter API
- Introduction to the midterm project

Lecture Instructions

- Download the notebook "*Lecture 17.ipynb*" available on Canvas
  - Open the above file with VS Code

Social Media Analytics

- Recall our first class
  - Social media
    - Very rich data
      - Likes, emotions, comments, replies, photos, location, groups, …
    - Highly **unstructured** data
      - Textual data, images, geolocation, …
    - **Massive amounts** of data
      - *E.g.*, Facebook has over 2 billion users
      - Each one with her/his own photos, comments, likes, …

Social Media Analytics

- Social media analytics is the process of monitoring, collecting, and analyzing social media data
  - Usually driven by specific requirements from an application
- Goals:
  - Facilitate conversations and interactions between online communities
  - Extract useful patterns and insights to serve stakeholders
    - *E.g.*, active users, the underlying platform, external organizations, …

Social Media Analytics

- Our focus today will be primarily on monitoring social media
  - Because of social media companies have:
    - Strengthened their ability to engage in company-customer dialog
    - Created mechanisms for customer-customer dialog
    - Monitored and mediated the above dialogues
  - As a consequence, recent years have seen several new platforms focused on social media monitoring
    - E.g., Radian6 (now part of Sales Force), Pulsar, …

Social Media Analytics

- Those dialogues help companies in virtually all phases of a product or service life cycle
  1. Product design
     - Capture and understand conversations involving loyal customers, average customers, or (potential) new customers
     - "Co-creation" of new products and/or services

     - "I wish this product would do …"
     - Example: Del Monte Foods, Inc. created a new dog-food product in just six weeks by listening to dog owners
       - https://www.youtube.com/watch?v=yP_3bpCPZaQ (3 min)

Social Media Analytics

- Those dialogues help companies in virtually all phases of a product or service life cycle
  2. Product production (supply responsiveness)
     - Businesses can anticipate significant changes in demand based on changing customer tastes and behavior
       - Adjust accordingly by ramping production up or down
     - Example: Popeyes' Chicken Sandwich in 2019
       - Went viral on Twitter
       - Drove sales up 16%, rose profits 13%, and …
       - … Popeyes runs out of chicken sandwich

Social Media Analytics

- Those dialogues help companies in virtually all phases of a product or service life cycle
  3. Product utilization
     - Who are the consumers buying the product?
       - Info from their social media profile
     - What do the consumers think of the product/brand?
     - Consumer support
     - Our focus in the midterm project
       - What do “clients” think of #ISA414?

Social Media Analytics

- Those dialogues help companies in virtually all phases of a product or service life cycle
  4. Product disposal
     - Build a “green” brand
       - Companies can engage in online conversations about disposal
       - Track consumer concerns
       - Suggest disposal places…
       - … and replacement items

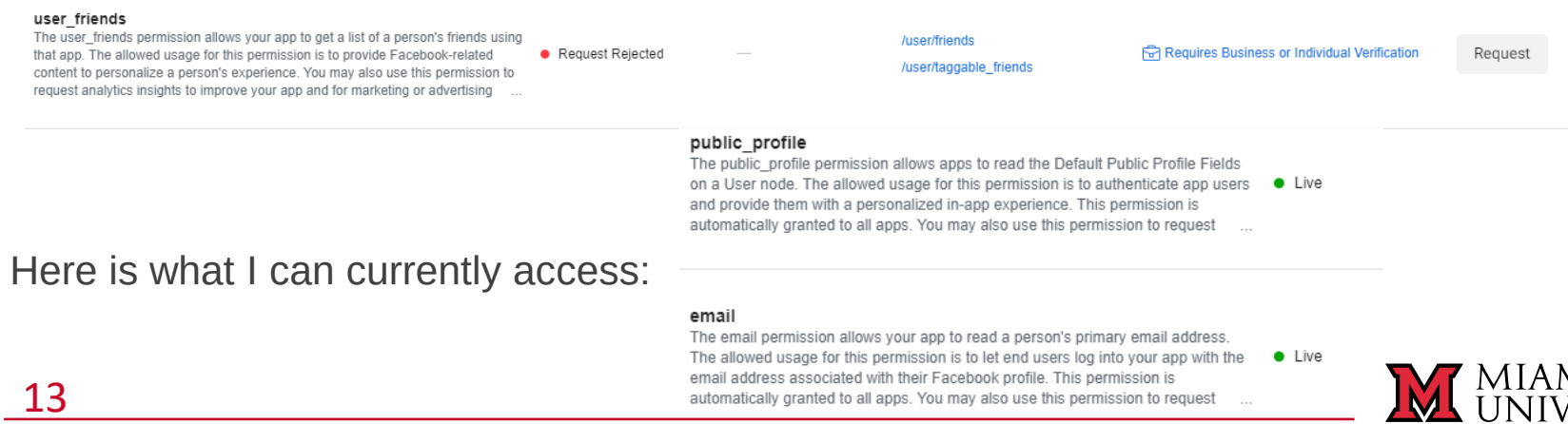
Social Media Analytics

- Traditional techniques
  - Sentiment/Emotion analysis
    - Automatically extract user sentiment or emotion
  - Topic modeling
    - Automatically detect dominant themes or topics
  - Social network analysis
    - Used on a social network graph to understand its underlying structure and properties as well as to identify the relative importance of different nodes within the network

Social Media Analytics

- Social media platforms: data collection via API
  - Facebook (Graph API): https://developers.facebook.com/
    - One must create an App and request permissions
      - Most are only granted to businesses
      - Process is incredibly cumbersome (since 2016/2017)
    - My request to access my list of friends via an API ☹

Here is what I can currently access:



MIAMI UNIVERSITY

Social Media Analytics

- Social media platforms: data collection via API
  - Reddit: https://www.reddit.com/dev/api/
    - Add .json or .xml to any Reddit page to get its content in JSON or XML

```
# posts from https://www.reddit.com/r/miamioh/
import requests
import pandas

response = requests.get("https://www.reddit.com/r/miamioh.json")
response = response.json()

df = pandas.DataFrame()

for post in response["data"]["children"]:
    json_to_row = pandas.json_normalize(post["data"])
    df = df.append(json_to_row)
```

Social Media Analytics

- Social media platforms: data collection via API
  - Reddit: https://www.reddit.com/dev/api/
    - Add .json or .xml to any Reddit page to get its content in JSON or XML

```
# posts from https://www.reddit.com/r/miamioh/comments/oygda9/email_from_president_crawford_to_all_faculty
import requests
import pandas

response = requests.get("https://www.reddit.com/r/miamioh/comments/oygda9/email_from_president_crawford_to_all_faculty.json")
response = response.json()

df_post = pandas.json_normalize(response[0]["data"]["children"][0]["data"])
df_comments = pandas.DataFrame()


for post in response[1]["data"]["children"]:
    json_to_row = pandas.json_normalize(post["data"])
    df_comments = df_comments.append(json_to_row)
```

Social Media Analytics

- Social media platforms: data collection via API
  - Twitter: https://developer.twitter.com/
    - One must create an App and answer several questions to receive an API key
      - "The core use case, intent, or business purpose for your use of the Twitter APIs."
      - "If you intend to analyze Tweets, Twitter users, or their content, share details about the analyses you plan to conduct, and the methods or techniques."
      - "If your use involves Tweeting, Retweeting, or liking content, share how you'll interact with Twitter accounts, or their content."
      - "If you'll display Twitter content off of Twitter, explain how, and where, Tweets and Twitter content will be displayed with your product or service, including whether Tweets and Twitter content will be displayed at row level, or aggregated."
      - …
    - You are highly encouraged to get your own API keys
      - But I will share mine with all you (fingers crossed)
      - https://developer.twitter.com/en/apply-for-access

Twitter API

- API limitations

GET endpoints

The standard API rate limits described in this table refer to GET (read) endpoints. Note that endpoints not listed in the chart default to 15 requests per allotted user. All request windows are 15 minutes in length. These rate limits apply to the standard API endpoints only, does not apply to premium APIs.

Endpoint	Requests / window per user	Requests / window per app
GET account/verify_credentials	75	0
GET application/rate_limit_status	180	180
GET favorites/list	75	75
GET followers/ids	15	15
GET followers/list	15	15
GET search/tweets	180	450
GET statuses/lookup	900	300
GET statuses/mentions_timeline	75	0
GET statuses/retweeters/ids	75	300
GET statuses/retweets_of_me	75	0
GET statuses/retweets/:id	75	300

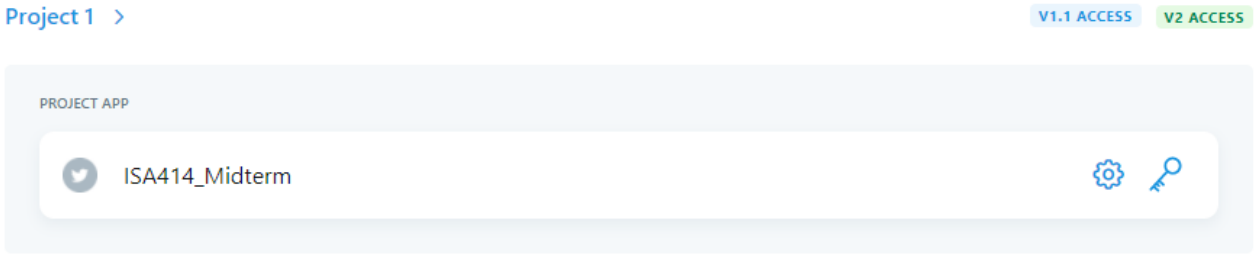
Twitter API

- Twitter authentication is slightly more complex than what we have seen so far
  - You have to sign each API request by passing several generated keys and tokens in an authorization header
  - You can generate the keys and tokens after your API access request is accepted

API key and secret:	
1. consumer_key 2. consumer_secret	Think of these as the username and password that represents the App when making API requests.
Access token and secret:	
1. access_token 2. access_token_secret	An access token and access token secret specific to the Twitter account the request is made on behalf of.

Twitter API

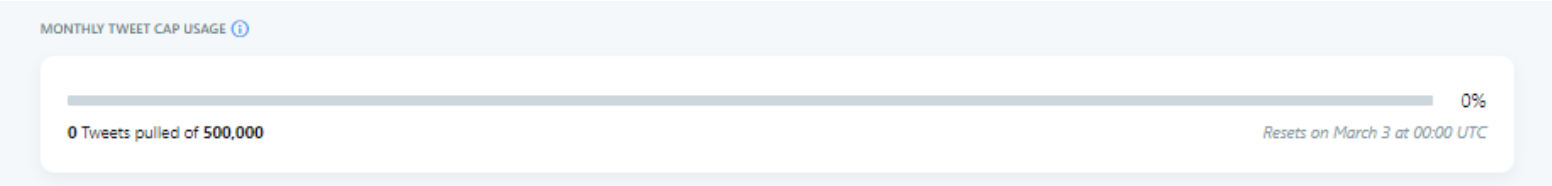
- How are we going to do in this course?
  - I have already created an app



  - We will all act as if we are a single Twitter user
    - @arthurc50170909
      - No, this is not my official account
    - This is easier than asking each of you to create a Twitter account

Twitter API

- I will be monitoring your usage
  - Individual (via IP addresses) and collective



- We will use the module tweepy to help accessing the API

  pip install tweepy

Twitter API

- Now, it is time to define our API keys
  - **YOU ARE NOT ALLOWED TO SHARE THESE KEYS WITH ANYONE OUTSIDE THE ISA 414/514 COURSE**

```
import tweepy

consumer_key = "zs0yFmnhHnfjnIS3eOXU9AtJU"
consumer_secret = "dbsGwYqw4jxgpiuY1kQH5Yf4dECdk9JCsJVAKaYH1ExUAoVz9B"
access_token = "1357013845647101955-t5k1nAOzGIKb6oohfw99l7F2aOVIJp"
access_token_secret = "DkgJun3wp8zdVSmUDr6rEYF82deUyr9EIipjcnmFPzDzl"
```

  - Next, authenticate

```
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth)
```

Twitter API

- Checking the API limitations
  - We are primarily concerned with searches

```
import time
response = api.rate_limit_status()
remaining_searches = response['resources']['search']['/search/tweets']
```

https://developer.twitter.com/en/docs/twitter-api/rate-limits

Twitter API v2 rate limits

The following table lists the rate limits for the Twitter API v2 with Standard and Academic Research product tracks at the Basic access level. These rate limits are also documented on each endpoint's API Reference page and also displayed on the developer dashboard.

Resource	Endpoint	Requests per 15-minute window unless otherwise stated	
		Per App	Per user
Tweets	Tweet lookup	300	900
	Timelines - User Tweet timeline - User mention timeline	 1500 450	 900 180
	Search Tweets - Recent search - Full-archive search	 450 300 Full-archive also has a 1 request / 1 second limit	 180

Twitter API

- Searching tweets
  - It returns a number of tweets based on a specified term
  - **It only returns data from the past 7 days**
    - One must have a premium (*i.e.*, paid) account to go beyond that
  - Example: searching for the latest 100 tweets having Microsoft

    ```
    search_results = api. search_tweets(q="Microsoft", lang="en", count=100)
    ```

  - Note that the results are in JSON
    - The lecture's notebook explains how to convert JSON to data frame

Twitter API

- What else can we do with the API?
  - Really a lot
    - See https://docs.tweepy.org/en/stable/api.html
  - Example: retrieving the IDs of ‘miamiuniversity’ followers

    search_results = api. get_followers(screen_name = "miamiuniversity", count = 100)
  - Note that the results are in JSON
    - The lecture’s notebook explains how to convert JSON to data frame

Twitter API

- What else can we do with the API?
  - Trending topics
    - One can collect the trending topics per region
      - Only certain regions are available
    - Obtaining the locations

      api. available_trends()

    - Obtaining the trend topics in Cincinnati

      api. get_place_trends(2380358)

Twitter API

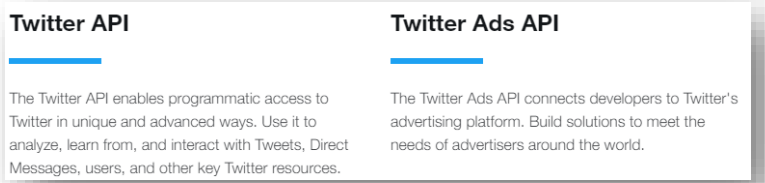
- What else can we do with the API?
  - Obtaining tweets from a user's timeline
    - *E.g.*, Microsoft

```
search_results = api.user_timeline("microsoft", count = 1000, exclude_replies = True)
```

  - Note that the results are in JSON
    - The lecture's notebook explains how to convert JSON to data frame

Twitter API

- What else can you do with Twitter API?
  - Twitter ads API
    - Not covered in this course
    - Programmatically create, schedule, and manage ad campaigns to engage people on Twitter
    - Create and manage tailored audiences using Twitter, web or mobile data
    - Granular insights of ad campaigns by a full range of metrics



Twitter API

- One last word about Twitter API
  - The tweepy package is incredibly useful
    - Manages crazy OAuth authentication
    - Handles API requests and parses API responses
  - … but can you see the information security issue?
    - Behind the scenes, what is tweepy doing with our keys and data?
    - Is it theoretically possible for the package to be collecting and sending that information to a third-party?
      - Yes, but you can always check the package's source code

Project Description

- Goal: monitor social media
  - What are the students saying about ISA 414?
    1. Continually search for tweets having term #ISA414
    2. Calculate the sentiment behind new tweets in real-time

- For more details, see the project description on Canvas

- Groups are now available on Canvas

Summary

- We learned about social media analytics
  - Focus on Twitter API

- Next lecture: in-class project work (optional)