

Summative Coursework Set Front Page

Module Title: **Programming in Python for Data Science**

Module Code: **CS3PP19**

Lecturer responsible: **Miguel Angel Sanchez Razo**

Type of Assignment (coursework / online test): **Coursework**

Individual / Group Assignment: **Group (max 2 members)**

Weighting of the Assignment: **70%**

Page limit/Word count: **NA**

Expected hours spent for this assignment: **14**

Items to be submitted: **Jupyter notebook (.ipynb file) and its HTML version.**

Work to be submitted on-line via Blackboard Learn by: **29th of November 2021**

Work will be marked and returned by: **20th of December 2021**

NOTES

By submitting this work, you are certifying that it is all your sentences, figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work except where explicitly the works of others have been acknowledged, quoted, and referenced. You understand that failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalised accordingly. The University's Statement of Academic Misconduct is available on the University web pages.

If your work is submitted after the deadline, *10%* of the maximum possible mark will be deducted for *each* working day (or part of) it is late. A mark of zero will be awarded if your work is submitted more than 5 working days late. You are strongly recommended to hand work in by the deadline as a late submission on one piece of work can impact on other work.

If you believe that you have a valid reason for failing to meet a deadline then you should complete an Extenuating Circumstances form and submit it to the Student Support Centre *before* the deadline, or as soon as is practicable afterwards, explaining why.

Forming your groups:

- You are responsible to form your own project groups.
- A group submission can have max two members each group.
- Each group is responsible for appointing a spokesperson, the spokesperson is responsible to send a list of the Student-Ids, full names of the group members, their email addresses and the list of accounts/influencers to be analysed (at least three).
- The spokesperson must submit the group member and the list of accounts/influencers in the following link by the 8th of October 2021:

https://forms.office.com/Pages/ResponsePage.aspx?id=xDv6T_zswEiQgPXkP_kOX5BNiMh6GDpJpqbEili8MBJUMehUT1hYU1BMRkFHNIFWRTNTNjAyUkpGWC4u

- **Any student who is not member of a group by the 8th of October will be assigned to groups randomly by the lecturer.**

Additional Covid-19 considerations.

Even though the social distancing rules are no longer in place, if you do not feel comfortable meeting with the team members, I suggest you using the following collaboration tools provided by the University:

- Microsoft OneDrive (file sharing and collaborative work).
- Microsoft Teams (chat and videocalls).
- Microsoft Outlook (Email).
- Blackboard (Group Forum).
- GitLab

Other Tools:

- Google Colab
- GitHub

1. Assessment classifications

First Class ($\geq 70\%$)	<p>The coursework demonstrates:</p> <p>Excellent knowledge and understanding of the concepts, evidence of independent research into methods used, and a thorough justification of methods</p>
Upper Second (60-69%)	<p>The coursework demonstrates:</p> <p>Good knowledge of the core concepts, showing understanding, with few mistakes. Good explanations and justification of the methods used</p>
Lower Second (50-59%)	<p>The coursework demonstrates:</p> <p>Demonstrates knowledge of core concepts but with some mistakes. Explanations and justifications of methods used are logical, but limited in depth</p>
Third (40-49%)	<p>The coursework demonstrates:</p> <p>Mistakes in application of knowledge, and shows some misunderstandings, explanation and justification of methods used is not clear or logical.</p>
Pass (35-39%)	<p>The coursework demonstrates:</p> <p>Gaps in knowledge and many mistakes, little evidence of understanding. Methods used are not explained or justified.</p>
Fail (0-34%)	<p>The coursework demonstrates:</p> <p>Large gaps in knowledge and significant mistakes, also showing limited understanding. Lack of logical explanations behind the methods used.</p>

2. Assignment description

Scenario: Twitter network map data extraction, pre-processing, and analysis

You have been asked to analyse information of the social media Twitter, such as the network of certain accounts, hashtags and some other data that can be extracted from it. You are required to implement a full Data Science Workflow going from the data gathering, cleaning, pre-processing, implementation of a model (network), and analysis of different statistics; you are also required to provide justification of the process, analysis of the findings, reasoning behind the design and implementation, decisions, and assumptions.

Your Tasks

Your overall task as a group is to implement the data science process on data collected from Twitter of at least three accounts and three hundred tweets (last tweets) of each account.

Task 1 – Data Gathering and Pre-processing

Implement a process/workflow to extract information from Twitter. Your solution must consider:

- API connection and data extraction from the data source
- Data Pre-processing from the data source to transform the original data into a Pandas dataframe.
- Perform a data cleansing activity considered relevant for the process (if any).
- Provide the explanation of the process, the justification behind it, lessons learned and findings.

For more details of the data extraction from Twitter please review below in this document section [5. Additional Considerations.](#)

Task 2– Exploratory Data Analysis (EDA)

Merge the information of the accounts and perform an EDA on the conjoint data to explain the nature of it. Some of the information you can analyse (not limited) is:

- Comparison between accounts, e.g. number of followers, are the accounts producing original tweets or mostly retweeting, etc.
- What users/accounts are more mentioned? Are there any favourite accounts among the influencers/accounts selected? What are the tweets that have the most likes? What tweets have been retweeted the most? Which one is the most successful account considering these two variables? At least three points should be taken for the analysis.
- Any other analysis that you consider relevant.
- Provide the explanation of the process, the justification behind it, lessons learned and findings.

Task 3 – Network analysis

The goal of this task is to create a network that represents the area of influence of the accounts/influencers selected. For this you need to consider the network as bidirectional, there are two ways to do it: you can extract the accounts that the influencer is following and/or create the links from the accounts that were retweeted. You need to provide the following:

- Explanation and justification of the data pre-processing implemented (if any).
- Provide a sample (max 10 records) of the edge list and the neighbour list of the network.
- Produce a visualisation of the network topology and discuss the output.
- Calculate statistics of the network, plot them where relevant, and discuss the results, explaining the meaning of any statistics you have calculated.
 - Statistics of the network such as
 - Degree Distribution
 - Cluster coefficient
 - Betweenness Centrality
 - Assortativity
- Conclusions and lessons learned.

Use Networkx (Python library) to calculate statistics of the network, rather than implementing your own Python code to do so. The visualisation may be hard to interpret at first, experimenting with different settings for the layout may help.

3. Assignment submission requirements

- You must create a Python 3.6 or above Jupyter notebook; when possible, use the packages included in Anaconda, Python 3.6 or above versions in your notebook. If you have a good reason to use a Python package not included in Anaconda, please contact the lecturer (m.sanchezrazo@reading.ac.uk) first to check before using it (excepting the libraries mentioned in [5. Additional considerations](#)).
- Before submitting, please remove the Twitter API connection credentials that you used to extract the data as it is confidential data.
- Your notebook should be submitted on Blackboard Learn, under the Assignments section, as one archive containing:
 - A .ipynb notebook file
 - A .html HTML version of the notebook. (These can both be saved from the Jupyter interface under File -> Download as).
- At the beginning of the submission, please add the following (in a markdown cell in the notebook):
 - Module Code:
 - Assignment report Title:
 - Student Number (e.g. 25098635):
 - Date (when the work completed):
 - Actual hrs spent for the assignment:
 - Assignment evaluation (3 key points):
- Include your student ID number in the name of the file containing your work.

4. Marking scheme

Task	Marks Available
Task 1 - Data Gathering and Pre-processing Demonstrates understanding of data extraction and data pre-processing technics.	20
Task 2 - EDA Demonstrates understanding of exploratory data analysis.	20
Task 3 - Network Analysis Demonstrates understanding of network data analysis.	20
Application of best practices such as implementation of functions, clear data manipulation code with Pandas/Lists/Dictionarys or any other data structure defined during the analysis.	15
Appropriate visualisation methods chosen to answer questions about the data.	15
Report structure (Format, completeness, readable, coherent).	10
• Total	100

5. Additional considerations

To extract the tweets from the accounts/influencers you have selected, one of the team members would have to do the following:

1. Have or create a Twitter account
2. Request a developer twitter account. You can request it in the following link:
[Apply for access – Twitter Developers | Twitter Developer Platform](#)
 - a. You might need to validate your account
 - b. You need to create a “new app” and provide information regarding the reason why you would like to have access. You can mention something related to your analysis such as sentimental, network and/or hashtag analysis, etc.
 - c. **The account can take up to a week to be validated.**
 - d. Once the account is validated, click on “Create app”
 - i. Provide the information regarding the app (name, description, etc.)
 - e. Go to “Key and tokens” and copy the credentials (API key, API key secret, Bearer token, Access token, Access token secret). **This info is shown once so please keep it somewhere safe** (you can request them later but they are going to be different).
 - f. Some reference videos:
 - i. [How To EASILY Get Twitter API KEY | Apply For Twitter Developers Account | Download REAL-TIME TWEETS - YouTube](#)

Other considerations:

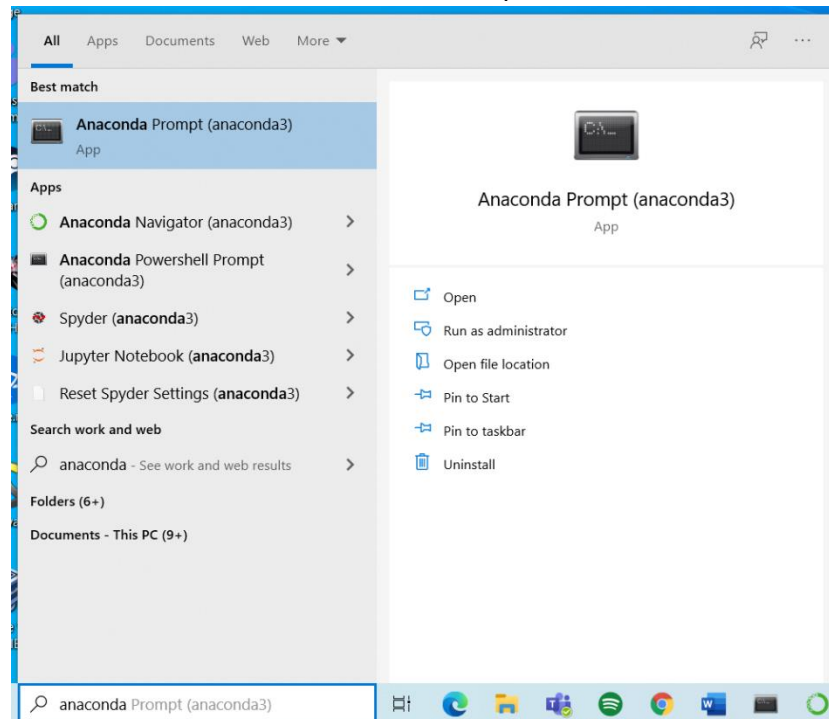
1. The Python libraries recommended (you might need to install them depending on the environment) to extract and manipulate the data are:
 - a. Tweepy (more details in [Appendix A – Tweepy installation](#))- [Tweepy Documentation — tweepy 3.10.0 documentation](#)
 - b. Twint (Good option if you struggle to get the Twitter Developer Account, more details in Appendix b – Twint installation) - [twint · PyPI](#)
 - c. Json - [json — JSON encoder and decoder — Python 3.9.6 documentation](#)
2. Please ponder the Twitter API rate restrictions, here the reference:
 - i. [Rate limits | Docs | Twitter Developer Platform](#)
3. Review the Data Dictionary of the API messages, here the reference:
 - i. <https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet>

Do not forget to contact your lecturer in case you encounter any problem or if you have any question.

Good luck!

6. Appendix A - Tweepy installation and Troubleshooting (under Anaconda)

1. Create a new anaconda environment:
 - a. Open a new instance of “Anaconda Prompt”



- b. Execute the following command: `conda create --name myclone --clone root`
 - i. Change “myclone” to the name of your preference

```
Anaconda Prompt (anaconda3)

(base) C:\Users\Miguel>conda create --name tweepy_sandbox --clone root
Source: C:\Users\Miguel\anaconda3
Destination: C:\Users\Miguel\anaconda3\envs\tweepy_sandbox
The following packages cannot be cloned out of the root environment:
- defaults/win-64::conda-4.10.1-py38haa95532_1
- defaults/win-64::conda-build-3.21.4-py38haa95532_0
- defaults/win-64::conda-env-2.6.0-1
- defaults/noarch::conda-token-0.3.0-pyhd3eb1b0_0
- defaults/win-64::anaconda-navigator-2.0.3-py38_0
- defaults/win-64::console_shortcut-0.1.1-4
- defaults/win-64::powershell_shortcut-0.0.1-3
- defaults/win-64::anaconda-2021.05-py38_0
```

- c. Activate the environment by: `conda activate myclone`
 - i. Change “myclone” to the name you selected before

```
(base) C:\Users\Miguel>conda activate tweepy_sandbox
(tweepy_sandbox) C:\Users\Miguel>
```

2. Install Tweepy by executing the following command In the Anaconda Prompt under the environment you defined: `pip install tweepy`

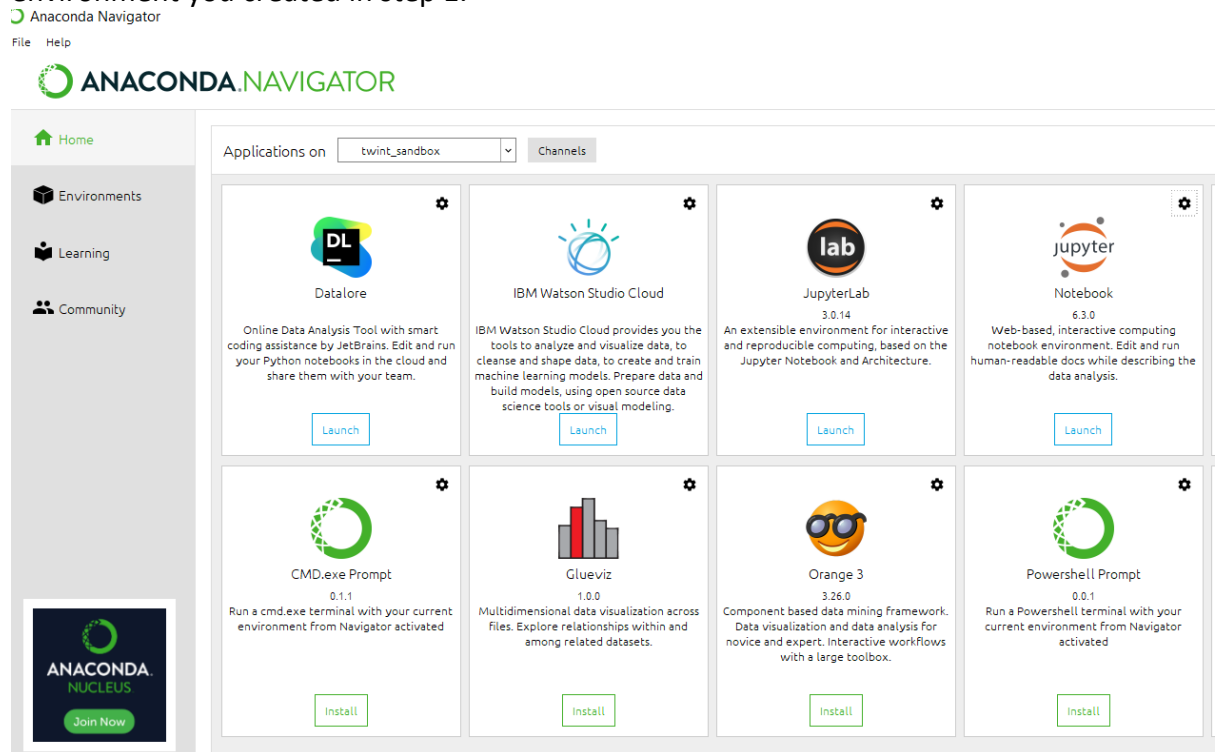
```
(tweepy_sandbox) C:\Users\Miguel>pip install tweepy
Collecting tweepy
  Downloading tweepy-3.10.0-py2.py3-none-any.whl (30 kB)
Collecting requests-oauthlib>=0.7.0
  Downloading requests_oauthlib-1.3.0-py2.py3-none-any.whl (23 kB)
Requirement already satisfied: six>=1.10.0 in c:\users\miguel\anaconda3\envs\tweepy (1.15.0)
Requirement already satisfied: requests[socks]>=2.11.1 in c:\users\miguel\anaconda3\envs\tweepy (2.27.1)
```

1. Testing.

Before this please make sure you have a Twitter developer account and a project validated with the API connection credentials needed.

Open an “Anaconda Navigator” and change the “Application on” selecting the

environment you created in step 1.



- a. Open a new Jupyter notebook and type the following:

Defining the connection credentials:

```
import tweepy
api_key = #Your api_key
api_secret_key = #Your api_secret_key
access_token = #Your access_token
access_token_secret = #Your access_token_secret

auth = tweepy.OAuthHandler(api_key, api_secret_key)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)
user = api.get_user('twitter')

print(user.screen_name)
print(user.followers_count)
```

```
In [2]: import tweepy
```

```
In [3]: ## Defining the connection credentials:
api_key = '...'
api_secret_key = '...'
access_token = '...'
access_token_secret = '...'

auth = tweepy.OAuthHandler(api_key, api_secret_key)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)
```

```
In [4]: user = api.get_user('twitter')
```

```
In [26]: print(user.screen_name)
print(user.followers_count)
```

```
Twitter
59533582
```

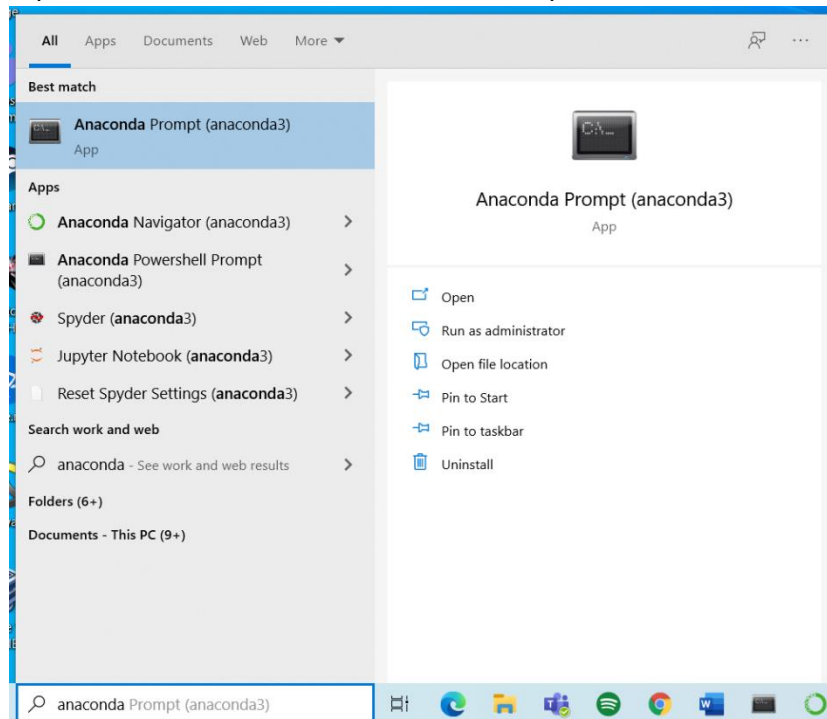
Now you can extract data from twitter!

Do not forget to review the documentation to access the data you are looking for (methods like “user_timeline”): <https://docs.tweepy.org/en/stable/index.html>

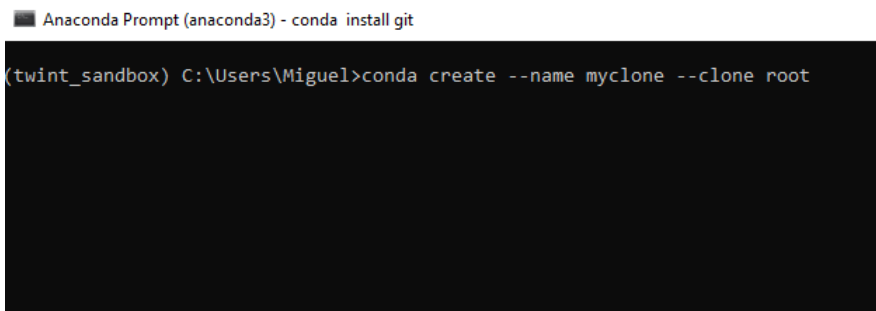
Contact your lecturer in case you have any questions.

6. Appendix B- Twint installation and Troubleshooting (under Anaconda)

1. Create a new anaconda environment:
 - a. Open a new instance of “Anaconda Prompt”



- b. Execute the following command: `conda create --name myclone --clone root`
 - i. Change “myclone” to the name of your preference



- c. Activate the environment by: `conda activate myclone`
 - i. Change “myclone” to the name you selected before
 2. Installing Twint Requirements/dependencies. According to the documentation, the following packages need to be installed beforehand.

<https://pypi.org/project/twint/>

Requirements

- Python 3.6;
- aiohttp;
- aiodns;
- beautifulsoup4;
- cchardet;
- elasticsearch;
- pysocks;
- pandas (>=0.23.0);
- aiohttp_socks;
- schedule;
- geopy;
- fake-useragent;
- py-googletransx.

- a. In the Anaconda Prompt under the environment you defined, execute the following command for **each library listed as requirement, exempt Python and Pandas:**
`pip install googletransx`

Anaconda Prompt (anaconda3) - conda install git

```
(twint_sandbox) C:\Users\Miguel>pip install aiohttp
```

Once you finish with the installation of the requirements, execute the following command: `conda install git`

Anaconda Prompt (anaconda3) - conda install git

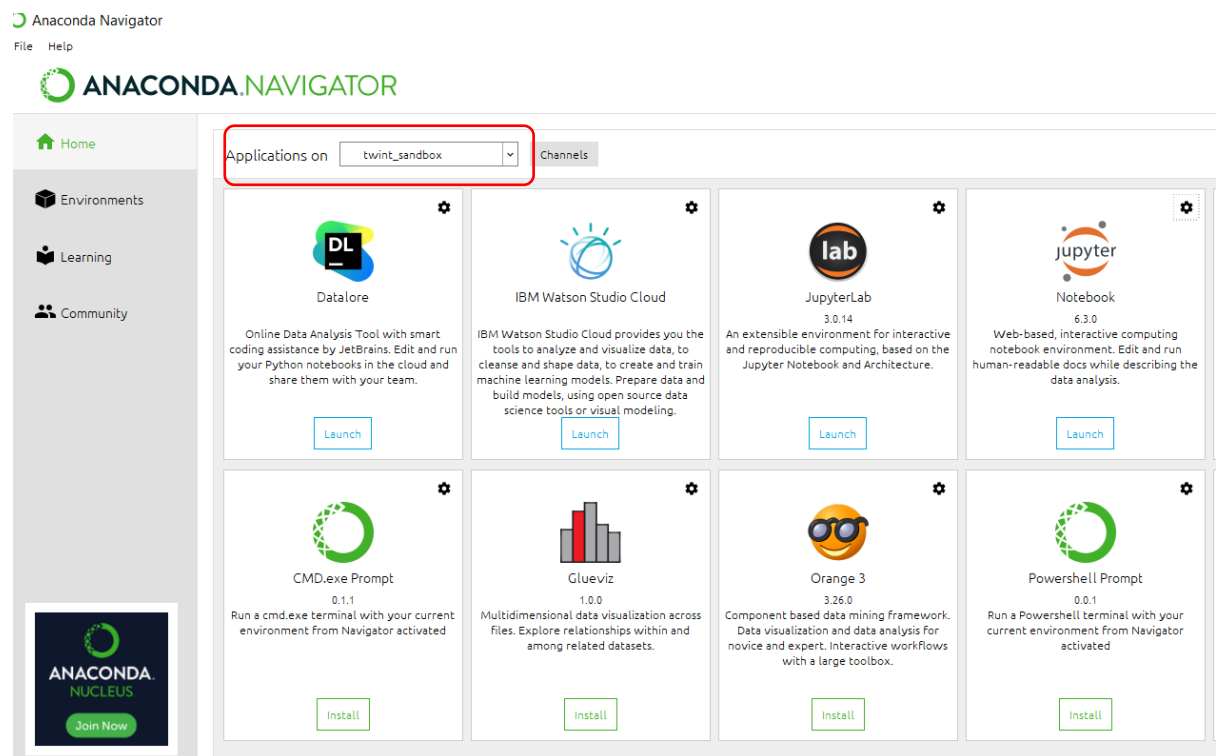
```
(twint_sandbox) C:\Users\Miguel>conda install git
```

3. Now you can install Twint by executing the command provided in the documentation (<https://pypi.org/project/twint/>): `pip3 install --user --upgrade -e git+https://github.com/twintproject/twint.git@origin/master#egg=twint`

```
Anaconda Prompt (anaconda3) - conda install git

(twint_sandbox) C:\Users\Miguel>pip3 install --user --upgrade -e git+https://github.com/twintproject/twint.git@origin/master#egg=twint_
```

4. Testing. Open an “Anaconda Navigator” and change the “Application on” selecting the environment you created in step 1.



- a. Open a new Jupyter notebook and type the following:

```
import twint
import nest_asyncio#
nest_asyncio.apply()
```

```
# Configure
c = twint.Config()
c.Username = "now"
c.Search = "fruit"
```

```
# Run
twint.run.Search(c)
```

Important to say that nest_asyncio is needed to execute twint in a Jupyter notebook, so don't forget to import it and apply it in your code

```
In [64]: import twint
import nest_asyncio
nest_asyncio.apply()

# Configure
c = twint.Config()
c.Username = "now"
c.Search = "fruit"

# Run
twint.run.Search(c)

1368952974979039235 2021-03-08 15:53:17 +0100 <NOW> @starwarspassion "You okay hun? You've hardly touched your floating Naboo fruit"
1365632930631065600 2021-02-27 12:00:37 +0100 <NOW> Billie Holiday was all things powerful, from her voice, her soul, to her fight. Total goosebumps watching @AndraDayMusic performing Strange Fruit in #USvsBillieHoliday https://t.co/3e1xyjrp7T
584720918468956160 2015-04-05 15:15:05 +0100 <NOW> We advise against looking at all the soft fruit in your kitchen as you get ready for the last two episodes. #GoTNOWTV http://t.co/RR0YPdMKFg
503965149376819200 2014-08-25 19:00:29 +0100 <NOW> #Friends #WhoSaidIt "Goodbye, you fruit drying psychopath." http://t.co/Aq8AlliymY
323156256846147584 2013-04-13 20:30:30 +0100 <NOW> Life's too short to eat fruit. http://t.co/jil6m7KJgM http://t.co/N2WCNuvg6A
305336396028198912 2013-02-23 15:20:44 +0100 <NOW> @TheThomasSmith Great choice, Thomas! We're all different, but in the end, we're all fruit. That's Saturday sorted, but what about Sunday?
[!] No more data! Scraping will stop now.
found 0 deleted tweets in this search.
```

Now you can extract data from twitter!

Do not forget to review the documentation to access the data you are looking for (parameters as: "Retweets", "Profile_full" and "twint.run.Profile"): <https://pypi.org/project/twint/>.

Contact your lecturer in case you have any questions.