# 1. Big Data Analytics – Lab Project

For the BDA Lab Project, we would like you to choose a data analysis problem and solve it. Any problem. For example, you can download data from Twitter containing a given hashtag, *covid* for example, retrieve the most common hashtags every day and see if there is any correlation with covid cases in a region (you will also need to download covid data). Maybe some hashtags can act as predictors.

You can use any data source: Twitter, Instagram, Spotify, etc.

Regardless of your choice, the analysis should go through the following steps:
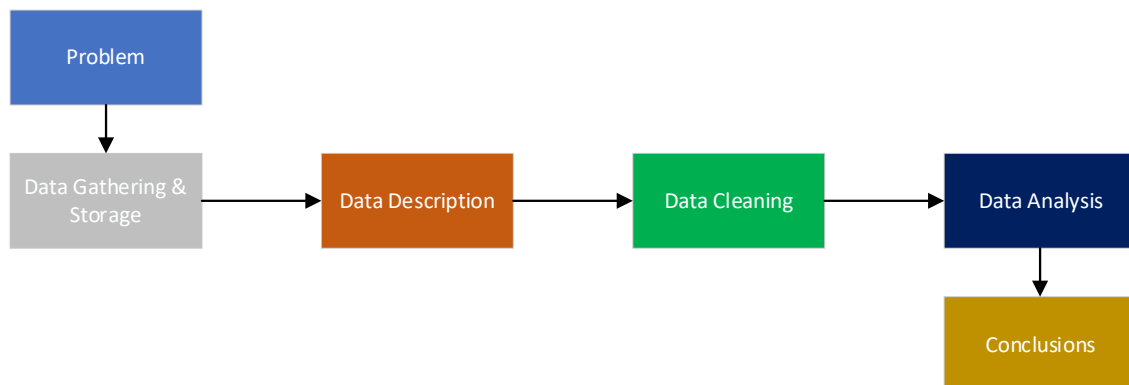
1. Describe the problem, including any required background, and explain why you believe it is important / interesting.
2. Collect the data and store it. Apply the techniques we have seen in the course. Relational vs. No-SQL, which one makes more sense?
3. Describe the data for the reader. Which fields are included by default? Which are you going to use for your analysis? For example, Twitter includes geolocation information, but just on a very small number of tweets, and the detected language may not always be correct.
4. Does the data publisher have any convention to indicate missing values? How are you going to treat them? Do you see any strange values that do not adhere to that convention (outliers)?
5. Analyse your clean data set to answer the proposed problem.
6. Conclusions.

Remember, a large dataset gives you a greater chance of extracting meaningful conclusions.

1.1 Twitter analysis example

In this example we will capture data from Twitter and store it in a local DB for analysis.
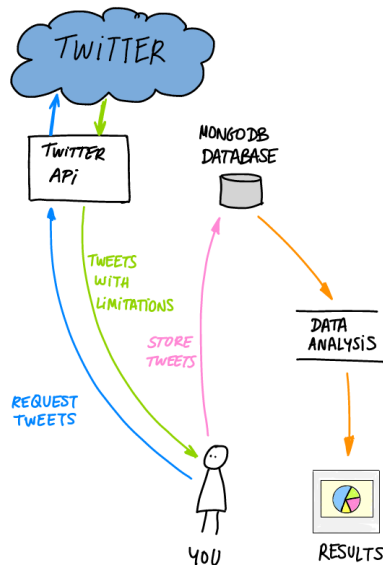
1.1.1 Prerequisites

Python packages (pip install …):

- pymongo
- tweetpy
- matplotlib
- numpy

Twitter:

- Must have a Twitter account
- Application registered (you have the keys)

Database:
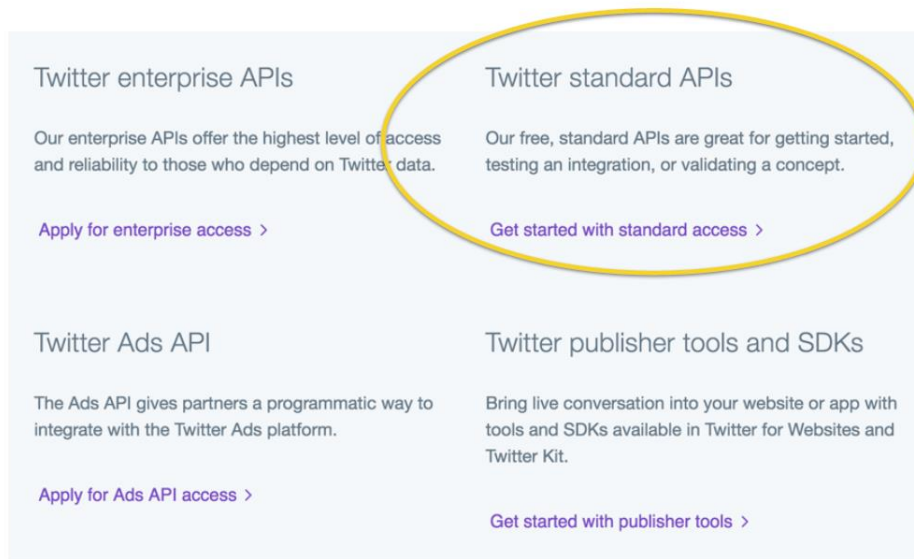
- MongoDB installed and working

1.1.2 Twitter API

Go to the developer area https://developer.twitter.com/en and apply for developer access https://developer.twitter.com/en/apply-for-access . Twitter is rolling out a v2 API, we do not recommend using it yet.

It will ask you to describe why you want access to the API. You need to make a good case, explain the assignment, this is for a masters course, send them the link to the course https://www.fib.upc.edu/en/studies/masters/master-artificial-intelligence/curriculum/syllabus/BDA-MAI , explain you do not intend to publish the results or make them public in any way, that you will make your analysis and create a report.

Once you are approved for access:

Twitter enterprise APIs

Our enterprise APIs offer the highest level of access and reliability to those who depend on Twitter data.

Apply for enterprise access >

Twitter standard APIs

Our free, standard APIs are great for getting started, testing an integration, or validating a concept.

Get started with standard access >

Twitter Ads API

The Ads API gives partners a programmatic way to integrate with the Twitter Ads platform.

Apply for Ads API access >

Twitter publisher tools and SDKs

Bring live conversation into your website or app with tools and SDKs available in Twitter for Websites and Twitter Kit.

Get started with publisher tools >

Get started: Build an app on Twitter

Twitter's API platform includes numerous endpoints to help you build an app and solution on Twitter. Our basic endpoints are available for free. As your app or solution needs grow, you'll also find enterprise APIs that include increased levels of access.

Get started with the basic REST and Streaming APIs

Twitter's basic REST and Streaming APIs enable free access to numerous endpoints. To get started, you must first create an app.

1. Create an app

To use an endpoint, you must create an app and use our OAuth-based authorization system. Visit apps.twitter.com to create one.

Application Details

Name *

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your
source attribution for tweets created by your application and will be shown in user-facing authorization screens.
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

Callback URL

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on t
given here. To restrict your application from using callbacks, leave this field blank.

Inside your newly created app you can access your keys:
- CONSUMER_KEY
- CONSUMER_SECRET
- ACCESS_TOKEN_KEY
- ACCESS_TOKEN_SECRET

Notes:
- Keys might not be immediately active
- You will need to associate your phone number to your Twitter account.

1.1.3   Collect data

Now, let's begin capturing tweets with stream.py.
1. Create a new file with your keys in the same folder (keys.txt).
2. Change the hashtags to whatever you want to capture, e.g., #ICTPSAIFRBIGDATA
3. Make sure the Mongo container is running.
4. If you run **python stream.py** you should be able to see tweets being downloaded and stored in Mongo.

You can start sending tweets with that hashtag, retweets and reply. Only tweets with the hashtag will be captured.

Notice that we set-up *stream.py* to abide by Twitter rates policy, which controls how many requests per minute we can send. Similar controls could be in place in other data sources. Beware of those and the penalties when you do not follow them.

Once you have collected enough tweets, you can stop the process.

1.1.4 Analysis

```
##################################################
# Plot of Languages (autodetected by Twitter)
##################################################

langsList = []
for t in my_tweets:
    langsList.append(t['lang'])
D = Counter(langsList)

# ----------- Bar Plot ------------------------
plt.bar(range(len(D)), D.values(), align='center')
plt.xticks(range(len(D)), D.keys())
plt.title('Languages spoken in the tweets captured')
plt.show()
```
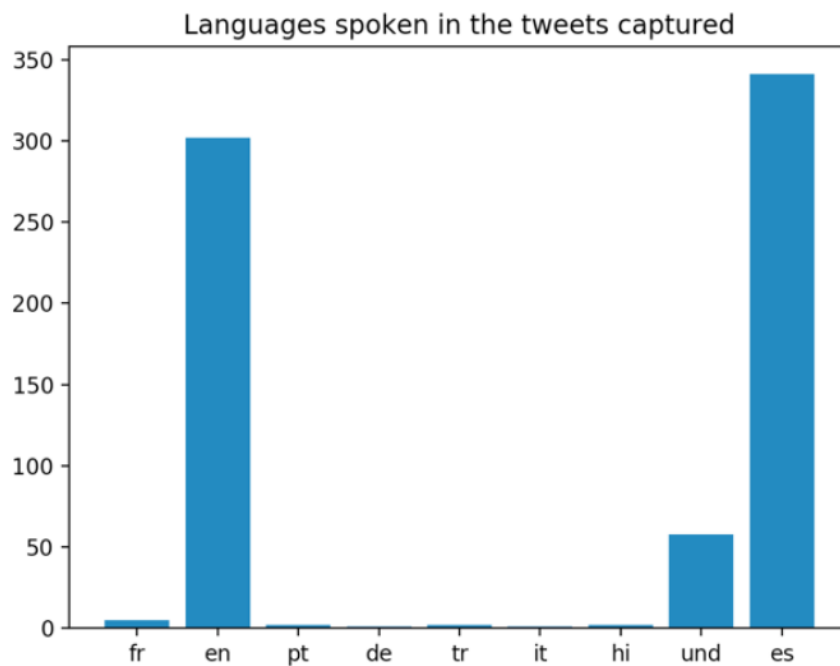

Tweets captured with hashtag #MondayMotivation and #FelizLunes
on Monday, Feb 26 2018

```
##############################################################################
# Plot how many of them are retweets, replies, quotations or original tweets
##############################################################################
my_tweets.rewind() #Reset cursor
retweets = replies = quotations = originals = 0
for t in my_tweets:
    if t.get('retweeted_status') is not None:
        retweets=retweets+1
    elif t['is_quote_status'] is not False:
        quotations = quotations+1
    elif t.get('in_reply_to_status_id') is not None:
        replies = replies+1
    else:
        originals = originals+1
# ----------- Pie Chart -----------------------
labels = 'Original Content', 'Retweets', 'Quotations', 'Replies'
sizes = [originals, retweets, quotations, replies]
frequencies = [x/numTweets for x in sizes]
colors = ['gold', 'yellowgreen', 'lightcoral', 'lightskyblue']
explode = (0.1, 0, 0, 0)  # explode 1st slice
# Plot
plt.pie(sizes, explode=explode, labels=labels, colors=colors,
        autopct='%1.1f%%', shadow=True, startangle=140)
plt.axis('equal')
plt.title('Percentage of Tweets depending on how the content is generated')
plt.show()
```
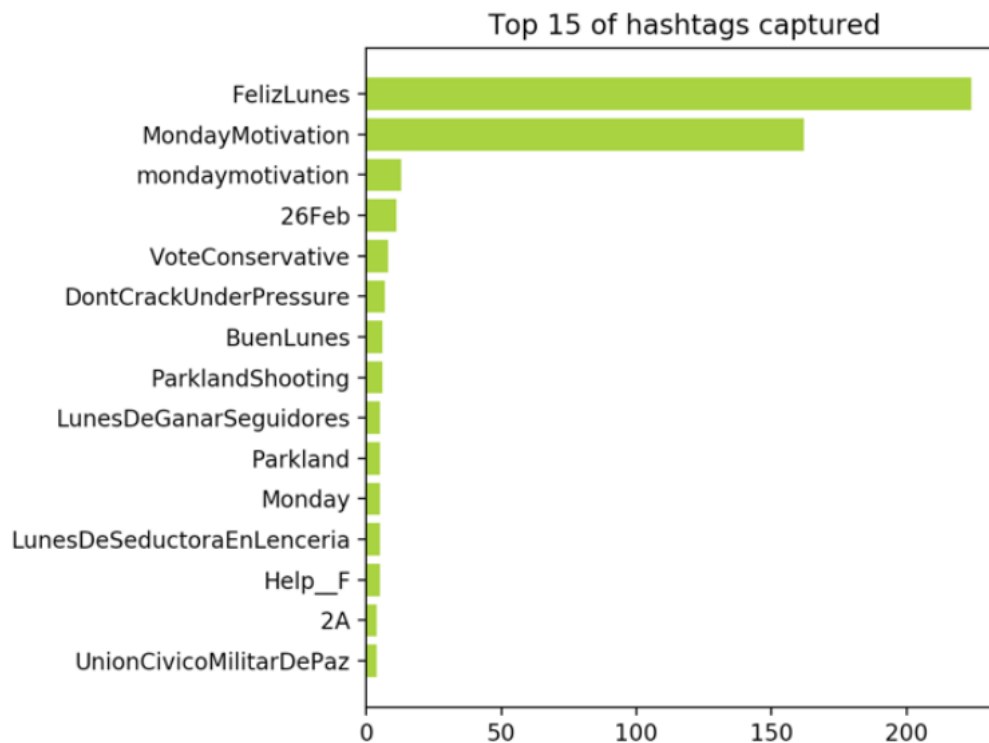

Percentage of Tweets depending on how the content is generated
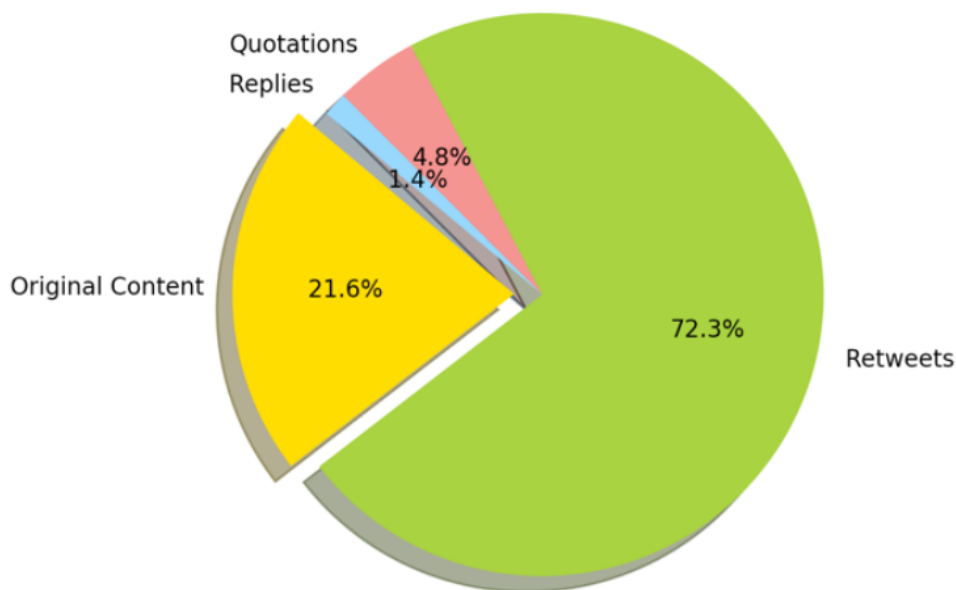
Tweets captured with hashtag #MondayMotivation and #FelizLunes
on Monday, Feb 26 2018

```
###################################################################
# Plot secondary hashtags
###################################################################
my_tweets.rewind()
hashList = []
for t in my_tweets:
    for e in t['entities']['hashtags']:
        h = e['text']
        hashList.append(h)
D = Counter(hashList)
subset = dict(D.most_common(15))
sorted_subset = sorted(subset.iteritems(), key=operator.itemgetter(1))

# ----------- Horizontal Bar Plot ------------------------
pos = range(len(sorted_subset))
plt.barh(pos, [val[1] for val in sorted_subset], align = 'center', color =
'yellowgreen')
plt.yticks(pos, [val[0] for val in sorted_subset])
plt.tight_layout()
plt.title('Top 15 of hashtags captured')
plt.show()
```


Tweets captured with hashtag #MondayMotivation and #FelizLunes
on Monday, Feb 26 2018

Projects from previous years

- Twitter sentiment analysis (pro/against vaccines)
- Spotify data: sadder songs during COVID lockdowns?
- Twitter spam identification
- Hate speech in Twitter
- Feminism analysis in Twitter
- Characterize groups of users using machine learning over Instagram metrics

Grading criteria

- Problem difficulty
- Data gathering & pre-processing
- Analysis is complete
- Problem is solved (bigger datasets are more helpful)