

# APME – Assignment 3

Your task in this assignment is to analyse some “big” data scrapped from the Internet using Pandas, Tweepy and Matplotlib. There are no specific requirements on the code that you will produce to solve this assignment, as long as you put your code in the file `assignment03/solution.py` and you can show that you have fulfilled correctly the specification detailed below.

You will analyse tweets about “programming” languages, with the ultimate objective of extracting some interesting links to know more about programming. Your task is the following:

1) Create your dataset:

You will have to “listen” for tweets with the keywords “python”, “java”, “c++”, “golang”, “php” tweeted around the world for a period of 3~4 hours, this should allow you to build a sufficiently big dataset<sup>1</sup>. Store these tweets in a json file<sup>2</sup>. Take note of when you have listened for tweets in your dataset.

How many tweets did you store in the file?

2) Using the tweets in the json file, create a dataframe retaining, for each tweet, only information about the language, the user location, and the actual tweet (i.e., the ‘text’)

3) Create and show a histogram of the top 5 languages used in the tweets that you have stored (see an example in `top5.png`). Make sure to embellish your histogram with all the required information (labels etc.)

4) Plot a histogram of the top 5 location from where the tweets that you have stored were issued. Make sure to embellish your histogram with all the required information (labels etc.)

5) Create a dataframe that retains only the tweets in the English language  
How many tweets are there in this dataframe?

6) From the tweets in English, create a dataframe to retain only the relevant ones. A tweet is relevant if it contains at least one of the following keywords: [tutorial, programming, software, linux, windows, mac os, data, dictionary]  
How many tweets in English are relevant?

7) Plot a histogram to show the relation between “number of relevant tweets in English” and the keywords that you have used at 1)  
(That is, how many relevant tweets in English have you found about python, java, c++, etc.?)

8) Some of the relevant tweets in English may contain links to Web pages<sup>3</sup>. You need to extract these links and store them in files named `<language>.txt`  
(for instance, `python.txt` will store all the links found in relevant tweets in English about “python”).  
Note that a Web link is a string of variable length that always starts with “`http://`” or “`https://`”.  
How many links did you find for each language?

---

1 An experiment made by the lecturer, listening for tweets using the provided keywords on a Friday between 11am and 2pm yielded about 80,000 tweets, of which about 18000 in English, of which about 400 relevant. The json file on which the files were dumped was about 490MB.

2 Sometimes, when storing large numbers of tweets, your json file may contain a special dictionary of the type `{limit : ...}`. These are not actual tweets and you should of course skip them while processing the json file.

3 To answer this question, you may need to process tweets in a dataframe one by one. To iterate on the content of a dataframe `df` row by row, you can use the `iterrows()` method. Example:

```
for index, row in df.iterrows():  
    # do something with the context of the current row of the dataframe
```

Instructions:

- You have to conduct this project **individually**.
- You have to **submit the code** that you develop **on blackboard** by the **deadline of December 9th at 10pm**.

Please zip the folder “assignment03” on your computer and upload the zip file.

**VERY IMPORTANT:**

**DO NOT SUBMIT THE JSON FILE WHERE TWEETS ARE STORED** (it can be a very big file!), BUT **BRING IT WITH YOU AT THE DEMO**.

- You have to **give a demo** of your code to the TA. Details about the schedule will be provided in due time. You are not allowed to change your code between the submission deadline and the demo (we will check!).

“Giving a demo” means to show the functionality of your software. So, you have to develop appropriate code in the main() of your application to showcase the implemented functionality. Failing to demonstrate the implementation of (some of the) functionality will lead to point deductions.

- Late submission policy:

*Up to 6 hours late:* the grade will decrease of 5 points for each ½ hour (e.g., your submission is 1.5 hour late, your grade will decrease of 15 points)

*More than 6 hours late:* these cases are critical and a decision will be taken on a case-by-case basis (note that all assignments are graded on a scale of 100 points)