Linnæus University

Social Media Ecosystems, 4ME304: Fall Term 2022

Assignment #3: Analyzing and processing social media data on Twitter using Python

Deadline: 08/12/2023

Contact Persons: Ahmed Taiye Mohammed, ahmedtaiye.mohammed@lnu.se
Alisa Lincke, alisa.lincke@lnu.se

Questions about the assignment can be posted to the forum available in Moodle. This assignment can be done individually or in groups of 2 students **(last option is recommended).**

Description:

In this knowledge module of this course, we have discussed those theoretical aspects of Big DataAnalytics (BDA) in relation to social media content. In the coming asynchronous and on-line meetings, we will carry out a hands-on experiences addressing different concepts and techniques related to BigData processing and visualization with Python tools. We will use the Repl.It environment to practically illustrate these ideas. There, you will be required to conceptualize and implement the concepts and ideas presented in the videos (Theoretical concepts in BDA) as well as in the videos illustrating Hands-on Lab. A detailed list of the mentioned videos is specified later in this document under the section named “videos”.

Your task:

Your task in this assignment will be to analyze historical data from a Twitter source related to a specific event, topic, or location selected by you (The theme). The following steps include practical as well as more theoretical sections on how to collect, process, analyze and visualize that data. Please remember to save your code (practical aspects) for later documentation required as this part is central for the assignment and the content you will need to submit. You should complete the next steps:

Practical tasks (coding):

1. Select an event, topic, or location for your data retrieval
2. Create a Twitter developer account for retrieving Tweets according to the selected theme
3. Collect tweets: Create a dataset with a min of 1000 tweets using Python (tweepy library)
4. Clean the collected tweets while using the NLTK libraries with these various functions including stop word removal, word tokenization and word stemming
5. Apply Text feature extraction using the Term Frequency Inverse document frequency library (TFidfVectorizer) using the scikitlearn module
6. Perform clustering while using K-Means algorithm: please verify to gain 10 clusters documents

Report:

7. Please formulate a report that includes the documentation of your Python encompassed code with explanations, concerns, analyzes, and challenges you encountered during your coding process. Do not forget to refer to the different papers available in Moodle while providing your explanations. Additionally, include in your submission files the Python code you have generated.

8. Select one main topic mentioned in the videos addressing the theoretical aspects of BDA. Formulate a text as part of your report (max 800 words) in which you address the impact of your programmatic efforts (done in tasks 1-6 of this assignment) considering the theoretical aspects of BDA. For your analysis, you may rely on the following topics (choose at least 3 of them) presented in earlier videos: data as a valuable resource, considerations of data capturing, social data analytics, digital services ecosystem, user centric data, influence analytics and data privacy. For example, you may address on how the value of data varies and impact the results of your program and therefore it provides a better ground for data-driven decision making. Thus, the analysis of your results will have a stronger foundation leading to benefit for various purposes including productivity, businesses and even politics (depending on the topic you have selected).

Videos

Links to the five Hands-on and practical aspects of BDA videos can be found at:

1. Obtaining and Using Tokens for Using Twitter (https://vimeo.com/490227705)
2. Collection of Twitter Data Using Python (https://vimeo.com/490230307)
3. Data Clustering (https://vimeo.com/490234397)
4. Clustering Visualization(https://vimeo.com/490236608)
5. Evaluation (Elbow Method) and Clustering (K-Means)(https://vimeo.com/490239480)

Final results:

The results from your efforts (tasks 1-8) should be reported following the publication format available at http://goo.gl/OtPQ5 . Note that the results of your work on task 8 (theoretical discussion) should be presented in a document that does not exceed five-six pages including the references (you may use references from this course or additional ones).

As mentioned, this task should be conducted individually or in pairs (preferable). Please upload aZIP file named 'yourlastnames_assig3.zip' to the respective Moodle folder. The ZIP file should contain a PDF file named 'yourlastnames_assig3.pdf' and another file containing your code named as 'yourlastnames_Codeassig3.zip'.

References:

You should have a look and read all the papers posted in Moodle discussing different issues related to this topic. The papers are available at:
https://mymoodle.lnu.se/pluginfile.php/7240057/mod_assign/introattachment/0/Papers%2313_17.zip?forcedownload=1