# 1st Lab Project - Neo4J

This is a hands-on lab project on Neo4j. You will practice how to create, manage and query/process graphs. Each group will consist of 3 students (one with 4) which will experiment on a task that contains Twitter data, Python and Neo4j. **The group will have to prepare a consistent, detailed hands-on tutorial as a Medium[1] article**, focused on the general topic of "Analysis of Twitter Data with the help of Neo4j Graph Database and Python"; no further report will be required. An indicative example of a Medium tutorial article can be found here or here. The dataset that you will use is composed of tweets on a specific topic and will be available until March 21st. Download Link: https://we.tl/t-sx9BFlcXQk . We will also provide a read-access user to a MongoDB hosted by us, that includes this data as well (you will receive an email).

The tutorial will have the following sections:

A. Installing and setting up Neo4j
B. Population of the Neo4j graph model with the tweets that have been collected in Python
C. Perform the queries that will produce answers to the questions of Table 1 in Python. The groups with the **odd** serial number (WDM1 data) will handle queries for the queries with **odd** index (1,3,5,..), while the groups with **even** serial numbers (WDM2 data) will handle queries with all the even numbers (2,4,6,...). All the groups will need to provide answers to **2 (two)** more queries of their choice (not in the list).
D. Document the process and **your conclusions** in your Medium-like article
E. Share your code with Jupyter Notebook
F. You will have to upload/send your solutions until **05/04/2023**

The Data Model:
An example of how Twitter data can be represented in Neo4j:

**Nodes:**

- User nodes – represents a Twitter user (handle and number of followers)
- Tweet nodes – represent a tweet (text, number of likes)
- Hashtag nodes – represent a hashtag
- URL nodes - represent a URL

**Relationships:**

- TWEETED relationship – in between a User and a Tweet; indicates that this user is the author of the tweet; also indicates the date at which it was tweeted
- RETWEETED relationship – in between a User and a Tweet; indicates this user retweeted this tweet; also indicates the date at which it was retweeted
- HAS_HASHTAG relationship – in between a Tweet and a Hashtag
- HAS_URL relationship – in between a Tweet and a URL
- USED_HASHTAG relationship – in between a User and a Hashtag
- USED_URL relationship – in between a User and a URL
- MENTIONED relationship – in between two Users

[1] https://medium.com/

Table 1

| # | Question |
|---|---|
| 1 | Get the total number of tweets |
| 2 | Get the total number of retweets |
| 3 | Get the total number of hashtags (case insensitive) |
| 4 | Get the 20 most popular hashtags (case insensitive) in descending order |
| 5 | Get the 20 most popular URLs in descending order |
| 6 | Get the total number of URLs (unique) |
| 7 | Get the followers count of each user |
| 8 | Get the 20 users with most followers in descending order |
| 9 | Get the number of tweets & retweets per hour |
| 10 | Get the hour with the most tweets and retweets |
| 11 | Get the user with the most replies |
| 12 | Get the users, in descending order, that have been mentioned the most |
| 13 | Get the top-20 hashtags that co-occur with the hashtag that has been used the most |
| 14 | Get the top 20 tweets that has been retweeted the most and the persons that posted them |
| 15 | Get the most “important” user in the dataset (use Graph algorithms: Pagerank, Betweenness centrality, etc. ). You will apply these algorithms in the mention network (which includes retweets) |
| 16 | Same as 15 |
| 17 | For the 5th most important user, get the list of hashtags and URLs that have been posted (if no hashtags or URLs - check another user e.g. 6th, 7th , etc..) |
| 18 | Get the users that post tweets with hashtags most similar to those used by the most important user |
| 19 | Get the user communities that have been created based on the users’ interactions and visualise them (Louvain algorithm) |
| 20 | Same as 19 |

**Neo4J Tutorials:**

- https://neo4j.com/developer/get-started/
- https://www.tutorialspoint.com/neo4j/index.htm
- https://www.youtube.com/watch?v=ou2st6FYxR8&ab_channel=Neo4j
- https://www.youtube.com/watch?v=IShRYPsmiR8&ab_channel=ChrisHay
- https://vladbatushkov.medium.com/learn-neo4j-cypher-basics-in-30-minutes-94d68a52544
- https://medium.com/neo4j/hands-on-with-the-neo4j-graph-data-science-sandbox-7b780be5a44f
- https://towardsdatascience.com/how-to-get-started-with-the-new-graph-data-science-library-of-neo4j-3c8fff6107b
- https://py2neo.org/2021.1/index.html
- https://medium.com/smith-hcv/graph-databases-neo4j-and-py2neo-for-the-absolute-beginner-8989498ebe43