ITU Artificial Intelligence & Data Engineering Department
YZV 211E Intro to Data Science & Engineering, Fall 2022
Homework #1
Due October 14, 2022 11:59pm

Homework Rules

- You must submit your solution as a Jupyter (iPython) Notebook.
- Solution must include code and your analysis/comments. If the analysis is not self-explanatory and does not include adequate comments, a point penalty of up to 20 points will be applied.
- Do not share any code or text that can be submitted as a part of an assignment (discussing ideas is okay).
- Do not copy-paste code from internet sources. Do not post homework assignment problems to the internet to ask for help. If you need help, e-mail, come and talk to us.
- You may discuss the problems at an abstract level with your classmates, but you should not **share or copy code** from your classmates or the Internet. You should submit your **own and individual** homework.
- Only electronic submissions through Ninova will be accepted no later than the deadline.
- Academic dishonesty, including cheating, plagiarism, and direct copying, is unacceptable.
- If you have any questions about the homework, you can send an e-mail to Caner Özer (ozerc@itu.edu.tr).
- Note: The **submitted solutions WILL BE CHECKED WITH THE PLAGIARISM TOOLS!**

Problem Definition

In a data analysis pipeline, the first step is to collect the data from some source. This step is generally tedious and may introduce potential errors. For these reasons, it becomes necessary to have a look at the data for any duplicates, missing features, and invalid values. For instance, during a data scraping task on Twitter, there may be multiple identical tweets posted by a swarm of bots, or as we merge from at least 2 data sources, some of the samples may have missing features.

In this homework, we will be dealing with how to correctly collect data from a database, and preprocess the data in a Jupyter Notebook by removing the duplicates and missing data. In this way, we will make the data ready to be deployed for any further downstream task.

While we will provide the necessary guidance for this homework, we want you to satisfy a couple of requirements mentioned below.

Requirements

- Use Wikidata SPARQL to query from Wikidata for generating the dataset. Link: `query.wikidata.org`
- You must generate your own query for some data. You should write your own query for the items and data you will determine yourself. Although you can use the example queries on query.wikidata.org, you need to add on to something to customize it. Direct copy-paste of an example query is not a valid solution.
- See the example given sample query (query.txt) below for creating the dataset of humans born in Turkey and their Twitter info available on Wikidata.

- You must not use the exact same query items. In other words, do not include "humans born in Turkey" and "Twitter data" both at the same time. Failure to comply with this requirement would end up receiving 0 from the first part.
- Download the CSV file that is generated by your query and continue the analysis with Pandas. Include your query statement in your Jupyter Notebook.
- In your generated dataset, do a data cleaning analysis:
  - Are there any exact duplicates? Show the count and then remove them.
  - How many missing values are there in each column? Show the count and then remove them.
  - Are there any invalid values (e.g. a negative value for positive range, very long Twitter ID, 0s)? Show the count and then remove them.

Sample Query

Code Listing 1: query.txt

```
#Humans born in Turkey
#title: Humans born in Turkey

# Add  as many columns as you can
SELECT ?item ?itemLabel ?itemDescription ?twitterName ?twitterID
        ?twitterFollowers $twitterSnapshotDate ?sitelinks

WHERE {
    ?item wdt:P31 wd:Q5;             # Any instance of a human
          wdt:P19/wdt:P131* wd:Q43; # Who was born in any value
    wikibase:sitelinks ?sitelinks.

# that has the property of anywhere in Turkey itself.

    OPTIONAL{?item p:P2002 ?statement.
    ?statement ps:P2002 ?twitterName.}
    OPTIONAL{?item p:P2002 ?statement.
    ?statement pq:P6552 ?twitterID.}
    OPTIONAL{?item p:P2002 ?statement.
    ?statement pq:P3744 ?twitterFollowers.}
    OPTIONAL{?item p:P2002 ?statement.
    ?statement pq:P585 ?twitterSnapshotDate.}



    SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en,tr"
}
```

In Query 1, we use Wikidata database to retrieve the list of people born in Turkey, and then, we obtain the individual's Twitter account name, Twitter account ID, number of Twitter followers, and the date of Twitter follower snapshot. In this regard, we use the English and Turkish versions of Wikidata database.

Project Deliveries

- CSV file which was generated by the SPARQL Query [50 pts]
- Jupyter Notebook responsible for preprocessing the CSV file [50 pts]

Notebook Structure

You are expected to create a Jupyter Notebook similar to a report format which involves the necessary steps for constructing the dataset and how you perform the preprocessing procedures with some additional details as to how many of the samples involve missing features and have at least one duplicate. For the dataset construction part, please also include your SPARQL Query as well.