Exercise 1

Intelligent data analysis
DV1597

March 23, 2024

This first exercise consists of five different tasks. The tasks are extracted from Steven Skiena's material [1, 2].

The exercise should be performed **individually**, i.e., no group cooperation. Please hand in your written as a **PDF** file via Canvas. The report should be in **English** and include your name and the answers to the five tasks below. The exercise is graded with G/Ux/U.

1. Identify **three** different datasets that are available online, e.g., using, for instance, the following sources (or any other public sources of your choice):

   (a) Google Dataset Search

   (b) DATA.GOV

   For each of the three datasets, write a brief description of its content, e.g., number of columns, type of data in columns, etc.

2. For each dataset, write down the **two** most interesting questions (according to you) that it is possible to answer using the data. **Avoid Yes/No questions.**

3. Why does data cleaning play a vital role in data analysis? Motivate your answer by providing an example.

4. During data analysis, how do you treat missing values in data? Motivate your answer by providing an example.

5. Implement a function that extracts the set of hashtags from a data frame of tweets. Hashtags begin with the "#" character and contain any combination of upper and lowercase characters and digits. Assume the hashtag ends where there is a space or a punctuation mark, like a comma, semicolon, or period. For instance, consider the following example data:

```
This is an  #example  Tweet for the interesting #DataAnalysis
course  at #BTH in  #2023. The #AIStudents are taking the course
for the second year at the campus in #Karlskrona.
```

Reference:

1. Skiena, Steven S. "What is Data Science?" in The data science design manual. Springer, 2017.
2. Skiena, Steven S. "Data Munging" in The data science design manual. Springer, 2017.