Foundations of Data Science 2023/2024

Activity 2: data quality and data cleaning

This activity is dedicated to data quality. Some steps will be performed manually (M) and others automatically (A) by developing your own Python scripts and using existing libraries.

1. Define a question that you want to answer (they can be vague for now and can be inspired by the following step)
2. Select at one dataset to analyse that is related to the question
    a. You can search online or use one of the following sources:
        i. Portal Nacional de Dados Abertos: www.dados.gov.pt
        ii. Lisboa Aberta: dados.cm-lisboa.pt
        iii. Amazon AWS: http://aws.amazon.com/datasets
        iv. Kaggle: https://www.kaggle.com/datasets
        v. Google:https://datasetsearch.research.google.com/
        vi. EU Open Data Portal: data.europa.eu/euodp
        vii. US Government's Open Data: data.gov
        viii. United Nations Data: data.un.org
        ix. OECD Data: data.oecd.org
        x. Open Data Network: opendatanetwork.com
        xi. World Bank Data Catalog: datacatalog.worldbank.org
    b. Select structured data (do not select text corpora - let's focus on tabular data as much as possible!)
3. Describe the original data set selected (M+A). Use
    a. Describe the dataset: size, number of attributes, type, etc
    b. Identify the most important information in each data source
    c. Identify missing or incomplete data, and identify possible strategies to solve these issues
    d. Identify possible problems regarding data quality
    e. Define for each attribute in original data source, if any operation is need (e.g, concatenation, extraction of portion, etc)
4. Develop and implement a strategy for data cleaning that addresses missing data and entity duplicates:
    a. Apply a strategy to solve missing data
    b. Implement one or more strategies for entity similarity to detect and merge.
        i. Use string similarity to detect potential duplicates
        ii. make sure to normalize textual data, dates, etc
        iii. Define rules to solve duplicates (solve one or two types of issues, even if you detect more) e.g., if names of actors have string similarity >80% and the set of movies they act in overlaps by more than 90%, then merge the actors.
5. 6. Write a report detailing each step
    a. Use tables/charts to report any results that can fit a table/chart

b. Include excerpts of the original data, extracted data and integrated data as annexes (not counting towards page limit)
c. Discuss choices and decisions
d. Describe the open source tools and libraries employed in your project, describing briefly how they were used.
e. Include an estimation of hours each student contributed to the project as an annex.
f. Use the ACM template (2 page limit excl. references and annexes): https://www.acm.org/publications/proceedings-template