Data Collection and Pre-processing

CMU Academy

AI, DATA SCIENCE, AND MACHINE LEARNING

Project Phase 1

The first phase of the project is dedicated to data integration and cleaning. Some steps will be performed manually (M) and others automatically (A) by developing your own Python scripts and using existing libraries.

1. Define a set of questions that you want to answer (they can be vague for now and can be inspired by the following step)
2. Select at least two distinct datasets to integrate and analyse that are related to the questions (This selection must be approved by the professor during a class)
    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
        iv. Google:https://datasetsearch.research.google.com/
        v. EU Open Data Portal: data.europa.eu/euodp
        vi. US Government's Open Data: data.gov
        vii. United Nations Data: data.un.org
        viii. OECD Data: data.oecd.org
        ix. Open Data Network: opendatanetwork.com
        x. 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!)
    c. Describe each original data set selected (M+A)
        i. Describe the dataset: size, number of attributes, type, etc
        ii. Identify the most important information in each data source

iii. Identify missing or incomplete data, and identify possible strategies to solve these issue

iv. Identify possible problems regarding data quality

3. Define an integrated data model (M)
    a. Again, ER diagram, relational model, graph, etc.
    b. Define for each attribute in original data sources, if any operation is need (e.g, concatenation, extraction of portion, etc)
4. Develop and implement a strategy that based on the data models defined in 3 and 4 is able to: (A)
    a. extract the data from the dataset according to the data models in 3
    b. use the data from a) and the model defined in 4) to produce a single integrated dataset.
5. For the selected attributes (the ones part of the integrated dataset):
    a. Classify the attribute as numeric, textual, categorical, or boolean. If you can't classify, discuss why (e.g., an attribute has values 1, 3, and medium, so it's neither numeric nor categorical). (A+M)
    b. Plot the distribution of numerical values. Find outliers (You can use histograms, but the report does not need to show all of them, just one or two of the most interesting ones) (A)
    c. Report the average, min and max length of textual values (A)
    d. For categorical values, report the number of unique values and their distribution. (A)
    e. Report the fraction of missing values for each attribute.
    f. Present other issues you found, such as: different formats (common with dates), synonyms, misplaced attribute values (e.g., last name included in first name attribute). (A+M)
6. Select one or more problems identified in 1 and implement a solution, e.g.:
    a. ill-in missing values following a particular strategy
    b. normalize date formats
    c. normalize string formats
7. Implement one or more strategies for entity similarity to detect and merge duplicates (A)

    a. Use string similarity to detect potential duplicates i. make sure to normalize textual data, dates, etc (A)
    b. Define rules to solve duplicates (solve one or two types of issues, even if you detect more) i. 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.

8. 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 (3 page limit excl. references and annexes): https://www.acm.org/publications/proceedings-template