25/10/2024

Data Integration and Large-scale Analysis

Exercise (100 Points)

Description: The task is to create a pipeline for entity matching by following all the necessary steps of cleaning, blocking and similarity matching. Once such pipeline is ready it can be used to train a ML model for predicting new records. This exercise could be completed in a group of maximum 03 students. Use the benchmarked datasets of "Amazon-GoogleProducts" for entity resolution. The datasets can be downloaded from the link (**https://dbs.uni-leipzig.de/research/projects/benchmark-datasets-for-entity-resolution**).

[Note] The submission should be made via TeachCenter. The submission should contain all the source code files (no binaries) and a readme file (pdf/text/Word) to describe the procedure you have implemented the accuracies you achieved and a guide to reproduce the results.

[Task 01]: Entity matching (40 Points)

1. Prepare data (apply necessary cleaning/transformations or normalization)
2. Extract key features and implement a blocking scheme
3. Identify the similar records from both datasets and calculate their similarity scores. Count the number of pairs whose similarity is greater than 0.95
4. Use the PerfectMapping file to report the accuracy of your pipeline.

[Task 02]: Feature vector and ML model (50 Points)

1. Create a training dataset with binary label using the output pairs of tasks 1 (If the similarity of pairs is equal to 1.0 label them as "1 or matched" otherwise "0 or unmatched").
2. Compute atleast six new features from the given features (i.e., Levenshtein, Jaccard or cosine similarity between descriptions of Google and Amazon features)
3. Preprocess your training dataset (impute missing values, handle class imbalance)
4. Train a machine learning model using k=3 cross validations and report precision, recall and F1-score
5. If needed perform hyper-parameter optimization and feature engineering to achieve results equal or better than the below baselines
   SVM (P: 0.79, R: 0.73, F1: 0.76)
   Random Forest (P: 0.82, R: 0.76, F1: 0.79)

[Task 03]: Reporting and Reproducibility (10 Points)

1. Report your results and explain your choices for data preparation, blocking scheme and optimizations and make your scripts reproducible. Submission of Jupiter or Colab notebooks are encouraged. If you are submitting a Python project please add a setup script to setup a virtual environment and install all necessary packages.
   Submission Deadline: January 13, 2025