21/10/2022

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 dataset of "DBLP-ACM" for entity resolution from [1]. These datasets come with perfect-matching which you can use to compute the accuracy of your pipeline.

[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 (steps to execute your scripts so that the mentioned accuracy could be reproduced).

[Task 01]: Create an Entity Matching Pipeline with the following steps (60 Points)

1. Prepare data (apply necessary cleaning/transformations and features)
2. Implement a blocking scheme
3. Find the perfect matches and compare them against the ground truths (perfect-matchings) and report accuracy of your pipeline
4. Create a readme to reproduce the results

[Task 02]: Create an ML model for Entity Matching (40 Points)

1. Split the data into train and test sets (min 50 instances for test)
2. Create the training and validation datasets i.e., if similarity score is greater than 0.9 (or of your choice) label it a match (1) otherwise no match (0)
3. Try different similarity values/hyper-parameters to get better accuracy on validation set
4. Predict the test dataset
5. Create a readme to reproduce the results

Submission Deadline: January 13, 2023

[1]: https://dbs.uni-leipzig.de/research/projects/object_matching/benchmark_datasets_for_entity_resolution