2/10/2023

Data Integration and Large Scale Analysis

Exercise (100 Points)

Description: The goal is to apply the concepts of data cleaning and entity matching learned in this course.

This exercise could be completed in a group of maximum 03 students. Use the benchmarked dataset of Restaurants1 (Yelp-Zomato) for entity resolution [1]. These datasets come with “Labeled Data” 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 (50 Points)

1. Prepare data (apply necessary cleaning, feature encoding, transformations and features)
2. Implement a blocking scheme (do not use python libraries, create your own scheme)
3. Identify and delete the duplicates with in each block
4. Find the perfect matches across blocks of Yelp and Zomato datasets and compare them against the ground truths (Labeled Data) and report accuracy of your pipeline
5. Create a readme to reproduce the results

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

1. Train a machine learning classifier on labeled data.
2. Try different hyper-parameters to improve validation accuracy and report cross validation accuracy for k=3.
3. Predict the instances in PredictX.csv and report accuracy using goldY.csv
4. Create a readme to reproduce the results

[Task 03]: Apply data cleaning (10 Points)

1. Download the yelp_err.csv file (where errors are introduced randomly) apply data cleaning primitives to fix the quality of data and report the accuracy using the original Yelp dataset. Report the number of corrupt instances, type of errors in each tuple, and number of fixed instances and error detection and correction techniques applied.

Submission Deadline: January 12, 2024

[1]: https://sites.google.com/site/anhaidgroup/useful-stuff/the-magellan-data-repository