Data Science: Tools and Techniques

Assignment 1

Prepared By

Dr Muhammad Atif Tahir, Dr Abdul Aziz and Dr Nauman Durrani

Deadline: 20th Oct 2019 through Slate

Instructions:

- Only soft copy.
- You need to work as a Group of 2.
- Filename Format: studentid1_studentid2.pdf
- Only one of you need to upload on Slate.

- Download the following paper from

  https://arxiv.org/abs/1712.08971

  Human-Centric Data Cleaning

  Write a one-page summary of the above paper. Clearly mention the problems associated with Data Cleaning. [10 Points]

- According to authors from the paper "Ziawasch Abedjan et al "Detecting Data Errors: Where are we and what needs to be done?" Proceedings of the VLDB Endowment, Vol. 9, No. 12, 2016", some existing open source tools are not good enough to correct different types of data errors. You need to evaluate these different type of open source tools (2 tools) mentioned in the paper along with Python (You can also look at other sources if these tools are not free) using data sets provided data.zip plus one missing values dataset downloaded from UCI website. Is it possible to clean these data sets using open source tools? If No then why and if Yes then provide the main steps (Two page maximum). There will be demo as well of selected groups [20 Points]