INFO 6210
Data Management and Database Design
Gathering, Scraping, Munging and Cleaning Data
Assignment 1

Professor: Nik Bear Brown

Due Thursday January 30th, 2020

Gathering, Scraping, Munging and Cleaning Data

In this assignment, you will be gathering real-world data. This process is often called data munging or data wrangling. All of your database tables must be populated with real-world data. Any substitution of simulated data for real-world data must be pre-approved by the TAs.

This assignment can be done in pairs, or individually.

The process is as follows:

You must find sources of data. (This can be downloads, XML files, JSON, HTML pages, data repositories, etc.) The data must have a thematic relation (e.g. all NBA data, all job data, etc.) (30 points)

There must be three sources:

1. A web scraper
2. A web API
3. Raw text, csv, xml, or excel data

You must create a conceptual database model (10 points)

You must download and reformat the data to fit a conceptual database model. (This involves using web scrapers, web API's, formatting scripts, parsing files, etc.) (10 points)

You must audit the quality and estimate the amount of data you'll gather. This involves auditing (i.e. testing to evaluate quality/accuracy - a systematic and independent examination of data). You will need to audit the following:

Audit (10 points)

- Audit Validity/ Accuracy
- Audit Completeness
- Audit Consistency/Uniformity

You must clean the data or show that it doesn't need cleaning (10 points)

Write a report explaining all of the files, the tests and their results and code. (30 points)

Professionalism (10 points)

- Consistent Naming Scheme
- Commenting & Documentation
- File and Folder Organization
- Separation of Code and Data
- Use Efficient Data Structures and Algorithms
- Keep Your Code Portable
- Contribution statement

Design Requirements

Your submission must include:

- Sample data from every source
- A conceptual schema explaining the data relations
- Any code and scripts you used
- A report explaining all of the files, the tests and their results and code

Scoring Rubric

- (30 points) You must find sources of data
- (10 points) Create a conceptual schema
- (10 points) You must download and reformat the data to fit your conceptual schema
- (10 points) Audit
- (30 points) Report
- (10 points) Professionalism

Submission of Assignments

You will submit your assignments via Blackboard ***and*** Github. Click the title of assignment (blackboard -> assignment -> <Title of Assignment>), to go to the submission page. You will know your score on an assignment, project or test via BlackBoard. BlackBoard only represents only the raw scores. Not normalized or curved grades. A jupyter notebook file ALONG with either a .DOC or .PDF rendering of that jupyter notebook file must be submitted with each assignment.

Multiple files must be zipped. No .RAR, .bz, .7z or other extensions.

Assignment file names MUST start with students last name then first name OR the groups name and include the class number and assignment number.

Assignment MUST estimate the percentage of code written by the student and that which came from external sources.