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

Professor: Nik Bear Brown

Due: Wednesday January 31, 2018

Gathering, Scraping, Munging and Cleaning Data

In this assignment, you will be populating your database with 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 me.

The process is as follows:

You must find sources of data. (This can be downloads, XML files, JSON, HTML pages, data repositories, etc.) (10 points)

You must download and reformat the data to fit your database schema. (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 Validity/ Accuracy (10 points)
Audit Completeness (10 points)
Audit Consistency/Uniformity (10 points)

You must clean and insert the data into your database. (10 points)

You must come up with 5 more use-cases that involve a join between at least two tables. (10 points)

You must test that you can retrieve your data by implementing all of the use-cases in that you created in all of the assignments so far. (10 points)

SQL Schema that makes sense and is in at least in first normal form. (10 points)

A brief README document explaining all of the files, the tests and their results and code. (10 points)

Design Requirements

Your submission must include:

- Sample data from every table.
- SQL for all of your inserts and queries
- Any code and scripts you used
- A brief README document explaining all of the files, the tests and their results and code.

Scoring Rubric

- (10 points) You must find sources of data.
- (10 points) You must download and reformat the data to fit your database schema
- (10 points) Audit validity/ accuracy
- (10 points) Audit completeness
- (10 points) Audit consistency/uniformity
- (10 points) SQL to insert the data into your database. Scripts to clean your data.
- (10 points) You must come up with 5 more use-cases that involve a join between at least two tables.
- (10 points) You must test that you can retrieve your data by implementing all of the use-cases in that you created in all of the assignments so far.
- (10 points) SQL Schema that makes sense and is in at least in first normal form.
- (10 points) A brief README document explaining all of the files, the tests and their results and code.

Submission of Assignments

Your submission should be a zip file with all of the required work. You will submit your assignments via BlackBoard. 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 represents only the raw scores. Not normalized or curved grades.