Data Exploration & Preparation

Description

Dataset construction can be one of the first tasks a data scientist will have to do. In order to build that dataset, several common sources of data can be used like CSV files and JSON or XML formats (if consuming REST Services or extracting from a document based DB) and SQL Databases.

It is important for students to have knowledge in those formats. It will allow them to complete an existing dataset or building it from scratch if necessary.

Learning Objectives and Outcomes

• Understand the relational model and the RDBMS features.

• Master the SQL language in its data manipulation part.

• Extract data from a database to construct a dataset.

• Work on JSON, XML and CSV formats.

• Know how to use an ETL.

Course Schedule and Contents

• DBMS features. Relational model (tables, rows, columns, keys, foreign keys, primary key, integrity constraints). Basic requests: INSERT, UPDATE, DELETE and simple SELECT. Horizontal functions.

• Aggregations, GROUP BY, and inner joins.

• Nested queries, outer joins, HAVING and EXISTS, relational divisions.

• Introduction on other structured formats (JSON,XML, CSV), and conversion between them using python & java.

Grading

Quiz : 30%

Assignment : 70%

Policies

- I expect you to turn-in your reports on time to receive proper credit/grade.
- Any work submitted must be your own.
- I expect everyone to contribute equally to group assignments

- Attendance in every class is expected and class participation and discussion is strongly encouraged.
- Late work will not be accepted unless prior arrangements have been made directly with me.
- Cases will be decided on an individual basis.

Good Luck!