Midterm I Material

CMSC 320

Preliminaries

- Data Analysis Cycle: acquisition -> preparation -> modeling -> communication

References

- Lecture Notes Ch. 2-4

Measurement types

- categorical
- ordered categorical (ordinal)
- discrete numerical
- continuous numerical
- text, datetime
- the importance of units

References

- Lecture Notes Ch. 5
- HW 1

Data Manipulation Operations

- single table operations (subsetting attributes, subsetting entitites)
- more single table operations (sorting, creating new attributes, summarization, grouping entities *group by*)
- operation pipelines
- the multiple types of joins

References

- Lecture Notes, Ch. 6,7,13
- HW 1, 2

Basic plotting

Level 1

- The data/mapping/geometry definition of data visualizations
- Frequently used plots: scatterplot, bar graph, histogram, boxplot

References

- Lecture Notes, Ch. 8

Best practices

- the importance of reproducibility
- tools to improve reproducibility (debugging data science)
- data science ethics and responsible conduct of research (informed consent, privacy and anonimity)

References

- Guest lecture by John Dickerson, posted on calendar

Tidy Data and Data Models

- Components of a Data Model
- Basics of the Entity-Relationship and Relational Data Models
- The components of an ER diagram
- The relationship between tidy data, the ER and the Relational models
- Keys/Foreign Keys in the Entity-Relationship data model
- How an ER diagram is converted into a set of Relations (data tables)
- Integrity and consistency: uniqueness constraints, relationship multiplicity constraints, referential constraints

References

- Lecture Notes, Ch. 11, Lecture slides

SQL and Database Systems

- the difference between declarative and procedural representation of data operations
- the Select-From-Where SQL query
- Joins in SQL
- Database query optimization principles
- JSON

References

- Lecture Notes, Ch. 12, 15
- HW 2

Data scraping

- The hierarchical structure of HTML documents
- Basic CSS selector syntax: type, class, id, attribute

References

- Lecture Notes 16.2, Lecture slides

Data cleaning

- Common problems in data tidying
- The gather and spread data tidying operations (data values as headers)
- Normalizing data tables (More than one entity in a table)
- Regular expression basics

References

- Lecture Notes Ch. 17, 18

Entity Resolution

- The Entity Resolution problem
- Calculating similarity between categorical attribute values
- Calculating similarity between numeric attribute values
- Calculating similarity between entities
- Solving the one-to-many resolution problem

References

- Lecture Notes Ch. 19

Network Data

- Using graphs (nodes, edges) to represent data (entities, relationships)
- Derived attributes from graphs (degree, betweeness)

References

- Lecture Slides

Midterm Structure

The midterm will consist of three sections: ~8-10 multiple choice questions, ~5-7 short questions, and 1 or 2 longer questions. Multiple choice will test concepts and definitions along with problems similar to written exercises in class. Short questions will be similar to written problems done in homework, along with concept questions where longer written answers are required. Longer questions are for problem solving (e.g., design a data pipeline or SQL queries to carry out a specific task).

You can bring 1 double sided 8.5x11in sheet of notes to the exam.