CS 598: Theory and Practice of Data Cleaning Syllabus

Course Description

Data cleaning is an essential but often under-appreciated part of data science. Some surveys report that data scientists spend around 80% of their time cleaning, wrangling, or “munging” data before data is ready to be analyzed. This course is an introduction to the theory and practice of data cleaning. We will first provide a brief overview on different dimensions of data quality. We then will turn our attention to syntactic (i.e., pattern-based) data quality issues. In the first hands-on module, we will work with regular expressions to match and extract data. We will then use OpenRefine, a powerful open source tool for cleaning individual data columns (e.g., through common transformations and operations that are used to obtain a canonical form of data). Moving on from syntactic to schema and semantic issues, we will work with both Datalog and SQL(ite) to express and check high-level integrity constraints. Finally, we will learn about workflow, provenance modeling, and querying techniques that allow users to reveal important dataflow.

Course Goals and Objectives

Upon successful completion of this course, you will be able to:

- Recognize common data quality problems and associate them with data quality dimensions.
- Develop simple regular expressions for matching, extracting, and transforming data.
- Use OpenRefine to explore and profile data, to identify data quality issues, and to clean and normalize data using its powerful built-in functions (e.g., for clustering and fusing similar data strings).
- Use declarative rules in Datalog and queries in SQL to check the validity of datasets with respect to a given set of integrity constraints spanning one or more data tables.
- Use a simple annotation language (YesWorkflow) to create and visualize workflow models that reveal dataflow dependencies often hidden in data analysis scripts in Python, R, etc.
- Use Datalog queries on provenance graphs to explore the lineage and processing history of data from workflows and scripts.

Textbook and Readings

There is no required textbook for this course, but the recommended readings are listed below.

- Sadiq, S. (2013). Prologue: Research and practice in data quality management. In *Handbook of data quality* (pp. 1-11). Springer.
- Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull., 23*(4), 3-13.
- Ganti, V., & Sarma, A. D. (2013). Data cleaning: A practical perspective. *Synthesis Lectures on Data Management*, 5(3), 1-85
- Bertossi, L., & Bravo, L. (2013). Generic and declarative approaches to data quality management. In *Handbook of data quality* (pp. 181-211). Springer.
- Abedjan, Z., Chu, X., Deng, D., Fernandez, R. C., Ilyas, I. F., Ouzzani, M., Papotti, P., Stonebraker, M., & Tang, N. (2016). Detecting data errors: Where are we and what needs to be done? *Proc. VLDB Endow., 9*(12), 993–1004. doi: 10.14778/2994509.2994518

Course Outline and Schedule

This 4-credit hour course is 12 weeks long. See the table below for the detailed course schedule.

Week	Duration	Topics	Assignment
1	5/15 - 5/21	Introduction to Data Cleaning	
2	5/22 - 5/28	Regular Expressions	Regular Expressions Homework
3	5/29 - 6/4	OpenRefine	OpenRefine Homework
4	6/5 - 6/11	Relational Model and Queries	
5	6/12 - 6/18	Datalog & Integrity Constraints	Datalog Homework
6	6/19 - 6/25	Integrity Constraints & SQL	SQL Homework

7	6/26 - 7/2	Workflows	Data Cleaning Project (available)
8	7/3 - 7/9	Provenance	Provenance Homework
9	7/10 - 7/16	YesWorkflow	
10	7/17 - 7/23	Data Cleaning Project	
11	7/24 - 7/30	Data Cleaning Project (cont'd)	
12	7/31 - 8/5	Data Cleaning Project (cont'd)	Data Cleaning Project

Assignment Deadlines

For all assignment deadlines, please refer to the Course Deadlines, Late Policy, and Academic Calendar pages.

Elements of This Course

The course is comprised of the following elements:

- **Lecture Videos.** Each week, the concepts you need to know will be presented through a collection of short video lectures. You may stream these videos for playback within the browser by clicking on their titles or download the videos.
- **Homework Assignments.** There are five homework assignments in this course. Homework assignments are machine auto-graded programming assignments. You will have unlimited attempts on all homework. All homework should be submitted before the deadline (see the Course Deadlines page). The homework is designed to allow you to gain a deeper understanding of the course content using practical exercises.
- **Data Cleaning Project**. The data cleaning project is a conclusive project of the entire class. It requires you to integrate the knowledge and skills you gained from the entire course. The project will be announced in Week 7. The data cleaning project is hand-graded. *If the deadline has not yet*

passed, and you would like to change a previous submission of the final project, please email mcsds-support@illinois.edu and we'll manually remove your existing submission so you can replace it.

- **Online Forum and Office Hours**. This course will use a Piazza-based discussion forum. The instructor and teaching assistants will be participating in the online forum and also offer online office hours. See the Communications page for details.

Please note, in order to access course materials and assignments, you will need to pay the Coursera fee $158 ($79 per MOOC equivalent) for this course (a degree course equals to approximately two MOOCs) in addition to the University of Illinois tuition.

Grading Distribution and Scale

Grading Distribution

Your final grade will be calculated based on the activities listed in the table below. Your official final course grade will be listed in Enterprise. The course grade you see displayed in Coursera may not match your official final course grade.

Assignment	Occurrence	Percentage Weight of Final Grade
Homework	5	5 x 15% per HW = 75%
Data Cleaning Project	1	25%
Course Total		**100%**

Grading Scale

Letter Grade	Percent Needed	Letter Grade	Percent Needed	Letter Grade	Percent Needed
A+	**95%**	**B+**	**80%**	**C**	**65%**

A	**90%**	**B**	**75%**	**D**	**60%**
A-	**85%**	**B-**	**70%**	**F**	**Below 60%**

View Grades

You can view your grade on each assignment by clicking the **Assignments** tab on the left menu bar.