INF 385T - Metadata Generation and Interfaces for Massive Datasets

Instructor
Unmil Karadkar (@unmil, please do call me by first name)
Office: UTA 5.408

How to contact me:
email: via canvas or to unmil-at-ischool.utexas.edu (please include the text "INF385T" in your subject line)

with prior appointment in-person meeting or, videoconference

skype: unmil.karadkar, Google hangouts: unmilk-aT-gmail.com (please don't send email to this address)

drop by my office without an appointment—I will try to make time for you

Class Meetings
Prerequisites

- Graduate standing
- Basic knowledge of a programming language (for example: variables, constants, statements, conditionals, loops, function calling, and return values in a modern programming language such as Python, PHP, Java, or C++)
- Basic understanding of data modeling (examples: Entity-Relationship diagrams, data models, Dublin Core, METS, or data structures)
- A motivation to learn more as necessary for your project

Readings

No textbook. Assigned readings will be available online or made available via Canvas. Some online readings are only available to UT Austin students and you may access these off-campus via UT-VPN or via the libraries' web site.

Introduction

Large-scale digitization projects such as the National Digital Newspaper as well as increasing quantities of born-digital materials have put enormous collections of documents and data within our reach. This is a studio-style course designed to explore techniques that will make these massive quantities of data (although, not necessarily bigdata) useful to targeted demographics, for specific goals via the use of programmatic techniques. This course both draws upon and will enable you to contribute to the areas of digital libraries and archives (collections, digitization), computational techniques (database management, data mining), and user experience (interaction design, HCI). Participants in this course will work in small teams, crafting small projects that demonstrate the viability of the proposed solutions. Typical projects will involve the development and/or evaluation of parts of data pipelines—ingestion, transformation, storage, manipulation, and presentation.
Pedagogy and Organization

Class time will be split between short content-based lectures, reading discussions & debates, and group activities. Lectures will highlight content from assigned readings. The goal is to create a learning environment in the classroom where we raise significant questions, discuss concepts, and develop skills collaboratively. This format requires participation of all class members. Students are expected to:

- Participate actively in all activities
- Attend all class sessions; if a student misses a class, it is her/his responsibility to arrange with another student to obtain all notes, handouts and assignment sheets
- Read all material prior to class; students are expected to use the course readings to inform their classroom participation and enable them to perform the class activities and assignments
- Hand in all deliverables fully and on time. Late submissions will only be accepted as specified in the assignment. For information about addressing emergency situations and university excused absences, please contact, the student must contact the instructor as soon as possible. (see also Grading and Policies)
- Educate themselves and their peers. The successful completion of this course and their participation in the information professions depend upon the students' willingness to demonstrate initiative and creativity. Your participation in the professional and personal growth of your colleagues is essential to your success as well as theirs. Such collegiality is at the heart of professional practice. In-class group activities and discussion of the assignments is designed to encourage this kind of collaboration.
- When needed, ask for additional help from the instructor (see question-asking templates)

Learning outcomes

At the end of this course, you will be able to:

1. Describe quantitative and qualitative aspects of datasets
2. Generate or retrieve metadata from documents and Web-based collections
3. Read and interpret standards-based metadata schemas
4. Design standards-based schemas to store extracted metadata
5. Relate textual and visual user interface features for providing access to your data
6. Describe functional aspects of software for retrieving and presenting data
My Personal Goals

In addition to the content-specific objectives, I will do my best to:

- provide opportunities to think deeply and carefully about the topics being discussed in the class
- foster an environment where students all feel welcome and free to share not only their knowledge but concerns and desires about course-related topics
- help develop skills that will be useful to you in your career
- engender a deeper interest in learning that can be pursued beyond this course
- to make you proud of your achievements in this course, and,
- help you enjoy the process!

Ask good questions of your data
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).

WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).

WE FOUND NO LINK BETWEEN MAROON JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN BURGUN JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

News

GREEN JELLY BEANS LINKED TO ACNE!

95% CONFIDENCE

ONLY 5% CHANCE OF COINCIDENCE!

Scientists...

(https://www.xkcd.org/882)