CS 5330 Group Project

Spring 2025

Project Description

Application: Analyzing social media

There are a lot of research which require analysis of social media post/text. In this project, we want to build a general purpose database system that will store social media post, and the results of analysis. (Notice that you are NOT analyzing the text, just creating a database to store the text and the result of analysis)

Social Media Text

The basic unit of data for analysis is a social media text post. It can be a short text, or a long post. For this project, we only consider text posts (no image/video, or if there is, it is going to be represented by a URL – stored as text). The text can be arbitrary length.

Each post will come from a certain social media (Facebook, Instagram etc.). We assume each social media has a unique name. Each post will be posted by a user (which has a username that is a string of at most 40 characters) in that media. We assume username is unique for each media only (i.e. there may be a user name "user123" in both Facebook and Instagram, and we made no assumption that they are the same person). All this needs to be recorded.

In addition, for the post itself, we will like to store the time it was posted (year/month/day/hour/minute, second may or may not be available). Also, a post may be reposted by someone else, in such case, we want to keep track of who is reposting the post, and the time that it was reposted. We assume a user cannot post more than one post in one media at the same time, however, he/she can post multiple posts to different media at the same time.

There are other information about a post that may or may not be available, and if they are, we need to record them: location of the post (city, state, country), number of likes and dislikes (as non-negative integers) and whether the post contains multimedia component (e.g. video, audio – there is no need to distinguish among them).

For each poster, we would like to store the following information (if available): first name and last name of the poster, country of birth, country of residence; age; gender; whether the user is an "verified" user at that media.

Projects and analysis

Text are analyzed by projects. Each project has a (unique) name, a project manager (we need to store his/her last/first name), and a institute that the project is taken (we only need to store the institute's name – which is unique for each institute). We also want to record the start date and the end date (both in yy/mm/dd format) of the project (notice that the end date has to be at least the same as the start date.

Each project will analyze some of the text that is in the database. For each project, each text will be assigned a set of fields that corresponds to the result of the analysis. For example, one project may

analyze the post and return its political leaning (left/center/right). Another may return the number of objects mentioned in the text (a non-negative integer) together with the overall sentiment of the post (positive/negative).

For each project, we need to record the fields that is associated with each text. Each field have a name, which is a string. We assume field names are unique within a project.

We also will need to enable the user to record the results of the analysis. That means for each post the project analyzed, the value of each field (a string) need to be record.

Things to do

You need to design a relational database to store all the information. You need to store it in a relational database, using mySQL or MariaDB.

You will also need to develop an application that allow one to enter information to the database and retrieve information from it. Your application needs to support the following operations:

- Data Entry
    - Enter basic information about a project
    - Enter the set of posts that is associated with a project. Notice that if a post already exists, it should not be stored in the database multiple times.
    - For a project, entering the analysis result (notice that the system should allow partial results to be entered).
- Querying post. Your system should allow (at least) the following criteria (or a combination of both (by AND only))
    - Find posts of a certain social media
    - Find posts between a certain period of time
    - Find posts that is posted by a certain username of a certain media
    - Find posts that is posted by someone with a certain first/last name

  For each query, you should return the text, the poster (social media/username), the time posted, and the list of experiment that is associated with that post.

- Querying experiment: You should ask the user for the name of the experiment, and it should return the list of posts that is associated with the experiment, and for each post, any results that has been entered. Also you need to display for each field, the percentage of posts that contain a value of that field.

- (For 7330 students only). You should allow the user first query a set of posts (as above), then list all experiments that is associated with those posted (with the detail above)

CS 5/7330 Group Project

Spring 2025

Instruction

The goal of the group project is to implement a database solution to the following application, using mySQL as the backend. This handout contains information about the various requirements. The actual project will be available in a separate handout.

Project Implementation

You will be given an application that requires a database backend, and a GUI frontend. You are required to use either mySQL or MariaDB as the backend. (SqlLite is not allowed). You only need to host the database locally on your own machine. Your program should read the username/password/database name from a file that is to be included in the submission (You can name them whatever you want).

You are free to choose how to program all the functionalities, but you should provide a standalone program to do the task. Under no circumstances should the user need/be allowed to directly enter the database by the standard database GUI.

Your program should provide a (very basic) GUI for users to enter data and queries. You can leverage a web browser to do it if you so wish (and use a web framework for it).

You are free to choose your language to implement your program, and whatever tool you need to help develop the GUI. However, all the tools you use must be freely available (free trial is NOT allowed)

Evaluation

Evaluation is done in 4 parts:

- Group check-in (4 points). There are two group check-ins (3/14, 4/14). The goal is to ensure that members of the group are communicating, and members are reasonably happy with how the group is progressing. Each check-in consists of one survey question. You will get 2 points for each check-in by answering that question -- whether you answer yes or no.

- Database design (21 points) each group should submit a ER-diagram for its database design, together with the database schema by 4/4 (Fri) 11:59pm. I will give each group feedback by 4/9 (Wed). This is worth 21 points. (BTW, you can change your design even if I give you the green light). (each group should submit a PDF file containing the information).
  Notice that if you submit the PDF by 4/1 (Tue), I will grade that with feedback by 4/4 (Fri), and you will get a 20% increase of your database design score.

- Implementation and demo (60 points). Each group will need to schedule a 30-minute time slot on between 5/5 – 5/7 for a demo. (The Monday morning/afternoon slots will be in person, others are via zoom). The time slot available are as follows:
  - 5/5 (Mon): 9:30am – 11:30 am (4 slots), 12:30pm – 3:00pm (5 slots), 8:00pm – 10:00pm (4 slots) (There will be no class on 5/5)
  - 5/6 (Tue): 9:30 am – 12:00pm (5 slots),

- 5/7 (Wed): 9:30 am – 12:00pm (5 slots), 1:30pm – 3:30pm (4 slots)

I will publish a rubric to give you a sense of how I will evaluate your work.

You will be given a rough range of your score (out of 60). If your group's expected score is 45 or less, you have the option to do a second-chance demo on 5/9 (Fri) [10:30am – 12:30pm] via zoom. Each group will have a maximum of one 10-minute slot to show that they can earn points that they have lost earlier. Notice that groups that demo early will have the first choice of time slots for second chance demos. There will not be enough slots for every group. Notice for groups that do a second chance demo, the maximum score will be 45 (out of 60)

Even though you will have to submit your project, and I may run your program to verify certain aspects of it, the implementation grade will be based on the demos.

I expect every member to be present at the first demo. I would expect at least 2 members to be present at the second-chance demo.

- Final report (15 points). You should upload your final report as a zip file by noon 5/12 (Mon), 11:59pm. Your report should contain:
  - The source code for your program.
  - A written report that contains the following:
    - The updated database schema, with comments on what each table represents (you can use the CREATE TABLE statements to describe the schema)
    - A brief installation/user manual to instruct someone (assume he/she is a junior CS student, but hasn't taken database yet) how to install and use the program

Conflict Resolution

In case of conflict within group members (for example, uneven workload), each case would be looked at individually, but here are a few guidelines.

- The second check-in is your "last" chance to express problems within the group.

- If there is at least one member that expresses problem during the second group check-in, I will set up a meeting for the whole group (most likely via zoom) on 11/19 or 11/20. Every member of the group must be present. We will try to resolve differences. However, I am open to the possibility of breaking groups (e.g. some student think others are not putting in the weight) However, since we are at the maximum capacity of groups, the groups that is broken up might have less priority in terms of scheduling for demo etc.

- If the group is happy after the response of the second check-in, I will assume the group has been working fine and all students will get very similar grades. If you feel like there is a significant discrepancy in terms of effort, you can still let me know, but I will not break the group up, even though I may test each of the group members individually.

The ability to work together is part of the assessment for the project

Propose your own project

I am open to you to propose a separate project, but any such proposal must satisfy the following conditions:

- There must be a real-life application for the project
- There must be someone else (not be another student) that can serve as the client that I can communicate with. He/She will be involved in the grading.
- The scope of the project should be like the one I propose

Contact me before 3/14 (Fri) 11:59pm if you want to do that.