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)