

CONCEPT

Create a program that allows users to input a city and find the most "instagrammable" spots in the area. The program will scrape Instagram for the most liked posts in a city and filter out external factors, in order to return the best specific locations to take an Instagram photo and get the most likes per post.

DATA COLLECTION

<https://github.com/rarcega/instagram-scraper>

We plan to use the above Instagram scraper and its --location flag to find the top posts in a city, as well as its --latest tag to find the most recent posts in a city. We'll then use the --user tag on the posts we've found to seed more post locations in the city to analyze.

DATA CLEANING

We plan to store the post data we collect in a SQL database. We want two tables - users and posts. The user table will store user data (handle, number of followers, average number of likes per post, post ids, if they're an influencer). The post table will store post data (link to post, number of likes, location, user id, time).

Posts will only include posts that have been created 2 or more days ago – we only want posts that have had enough time to accumulate a representative number of likes.

DATA ANALYSIS

At a minimum, we'll use the database described above to get an ordered list of where to go to get the most likes. We'll look at all the stored posts for a location and find its mean number of likes, also considering median and variance and filtering out outliers. The intent is to identify locations that draw attention for the location itself and not for external factors.

We'll then sort the results based on what locations have the highest average number of likes.

RECENT DEVELOPMENTS

Since our last check-in, we've set up our Google Cloud platform and started collecting sample data.

DATA COLLECTION PIPELINE

