Client: Dr. Daniel Golberg

Project Members: Sapir Dooley, Chibuike Imo

Project Title: Instagram Web Scraper

1. Introduction

The purpose of this project is to create a Python script that will read an Instagram feed and collect information based on user-inputted criteria. This script is intended to solve the challenge of collecting customer-related information through Instagram at a large scale. With this script, the user should be able to simply run the program; the program will log in to a specified account and collect information, in this case: location data of users that fit a specific set of criteria. In this proposal we shall lay out the specific functions of the script as well as the criteria that will drive data collection.

1. Motivation

For this specific instance of this code, we are interested in assisting the marketing team for the company TK Paddles, LLC. This company produces custom wooden wall decorations, or “paddles”, for college students in the Fraternity and Sorority community. There are thousands of Greek organizations in the United States and each organization may have tens to hundreds of chapters. In many of these organizations, it is common to give a paddle to either big/little brother/sister or the organization's new member educator each semester.

1. Functions

The first function of this script will be to open a browser and log in to an instagram account. This will be achieved using a simple web driver that will allow us to simulate a web browser. Next the code will load the feed by scrolling down the page a set number of times. After this the code will read the caption of each individual post, looking for specific keywords. If a post is found, the script will like that post and search it for a location tag. If no location tag is found the script will open the users profile and search for terms that the user can infer location from. After this is completed, the script will save the URL for the post.

1. Criteria

While each organization is different, generally the terms used are the same. When a class of new members is presented to the public, the terms “crossing”, “presentation”, “probate”, “rollout”, and “Neophyte” are used just to name a few. When most of the followers of the account are Greek, it is unlikely that these terms would be used elsewhere. This criteria will help the script determine what posts are most important as they will likely relate to a class that has just joined and therefore have not purchased a paddle yet. The terms that will help us determine location will be university names or acronyms as it is highly likely that is where a paddle would be shipped were it purchased.

1. Conclusion

The purpose of this project is to scrape an instagram feed for specific keywords and save those posts so that someone can go back and easily identify potential customers and their general locations. This information is valuable to the company so that the marketing team can know who needs to be reached out to, what regions are most likely to benefit from advertising, and what regions are not being touched.