Request for Proposals
Computer Science Capstone Projects – COS 397, Fall 2020 through Spring 2021
University of Maine, School of Computing and Information Science

Proposals are due to Terry Yoo, Associate Professor of Computer Science, by Noon, Tuesday, September 15, 2020 at terry.yoo@maine.edu.

FALL 2020 – COS 397 COMPUTER SCIENCE CAPSTONE PROJECT PROPOSALS

Project Title: Designing a user-friendly scraper tool for social media sites

By: Judith E. Rosenbaum, Associate Professor, Communication & Journalism

Brief Description (approximately 500 words)

One of the biggest issues faced by research into social media today is how to gain access to the data. In order to make any kind of statements about how people use social media, the impact social media have on specific social issues, the outcome of marketing campaigns, the ways in which discussions on social media evolve and change, one needs to be able to analyze large volumes of social media posts. While qualitative research, where researchers dive into a few hundred posts to investigate processes of meaning making and collaboration, can usually be carried out with a simple search in the social media platform of choice, any kind of large-scale quantitative research has to rely on a scraping tool. While scraping tools are available, these require a moderate-to-high level of understanding of computer codes and systems, something most researchers in the social sciences and business do not possess. This means that research into behaviors on social media is limited to those researchers with a background in math and/or computer science, leaving behind those whose focus might be equally relevant (e.g., psychology, sociology, marketing, risk and health communication) but not as computer-oriented. What is needed is a scraping tool that can be used across social media platforms, that allows for the collection of large data sets (and not just links to posts, as they can suffer from link rot), and that is, above all, user-friendly.

Therefore, this proposal calls for the design of a user-friendly, robust scraping tool that can be used across social media platforms and that collects the actual (publicly available) social media posts into a spreadsheet that is easily accessible for people with little to no knowledge of computer science.

Goals for the project (approximately 50 words)

To design a user-friendly, robust scraping tool that can be used across social media platforms and that collects the actual (publicly available) social media posts into a spreadsheet that is easily accessible for people with little to no knowledge of computer science.

Total Duration / Elapsed Time [in weeks]:

External

24 weeks

- 12 weeks of planning, requirements definition, design, and prototyping – to be completed by December 15

- 4 weeks of initial implementation – to be completed early March
- 8 weeks of testing and deployment – to be completed early April

Schedules / Deadlines [if any]:

The second testing stage (to be carried out by the lab of graduate students at CMJ – see below – should be completed by mid-March. The final testing stage should be completed by mid-April.

Learning Objectives for student teams:

Student teams who select this particular project will learn how to:
- Communicate with clients outside the field of Computer Science
- Determine what "user friendly" means depending on skill level of clients
- Determine needs of clients and translate these into actionable processes
- Design functional tool to collect social media posts
- Create instruction manual for scraping tool
- Educate users on tool
- Create write up and presentation outlining the tool and its functions

Expected Project Experiences (select from the list):

This project will involve the following experiences:
- Problem definition
- Project scope definition
- Workflow analysis
- Development of functional specifications
- Identification of and negotiation for needed project resources
- Examination of an unfamiliar technical area
- Identification of others' technical expertise
- Identification and evaluation of alternatives
- Development and presentation of recommendations
- Responsibility and accountability for a discrete product
- Role definition in a task group and participation in group dynamics

Recommended experience (What operating system is required? What programming language? Other skills?):

The team must have some experience using various social media platforms. They must have some basic understanding of how the platforms' API operates, and possibly be familiar with python (this is what most currently available scraping tools rely on, however, it is possible this team can come up with alternatives). In addition, it is also important that the team members know how to talk to people outside their field and have a feel for the capabilities (and limits thereof) of non-computer-science experts.

Expected Outputs/Products and likely requirements (specific programming language, operating system, integration with existing software, web-based requirements, *etc.*):

A tool, preferably web-based so easily accessible, that allows researchers to collect data from various social media platforms while using search terms related to: hashtags, people, key terms, location, time, date, and other relevant

factors.
This tool would ideally work regardless of changes to the platform's API or include clear instructions on how to adjust the tool to work after such changes have been implemented, if at all possible.
A file with clear instructors on how to use the tool and how to troubleshoot any issues / bugs that may arise.
The tool would ideally produce a CSV file (or something similar) with the posts (text and pictures), time and location (if available), poster name, and any replies / comments / likes.

Past experiences by the client (If software already exists, what is wrong? What has worked in previous versions, and what has not?):

There are currently several scraping tools out there. These are mostly python-based tools that require some moderate to advanced understanding of coding on the part of the researchers (which most don't have). The tools are often supported by communities on forums that will help with resolving bugs and other issues. While this is great, this still requires a high level of understanding of programming and computer systems that most researchers don't have. If they do, they often don't have the time to spend on resolving such issues. Most of these tools only allow a researcher to collect 50 posts at a time, which means the researcher needs to dedicate a lot of time to collecting a large corpus of posts and they also have no idea when they have collected all the posts with a certain tag, hashtag, or keyword. In addition, many tools geared toward Instagram and other image-based platforms only collect a link to the post, which means the researcher still needs to click on 10k links to get their data.

Proposed Testing Plan (How will the team test their product? Do you have recommended/required testing strategies? What resources are available (test platform, stand-alone network, *etc.*)? Is test data available?):

The product should move through several testing stages:

1. The team itself will test their tool using the most popular social media platforms (Facebook, Twitter, Instagram, Snapchat – the exact scope of which platforms can realistically be included is negotiable)
2. The team will then present their tool to a group of myself and my lab of graduate students in CMJ who work with social media data and who will test the tool for user-friendliness, bugs, etc. This stage should produce several recommendations for improvement which should be implemented in the final tool
3. Final testing stage to be carried out by both the team, myself, and my lab.

Benefits to U Maine:

Social media research is a "hot topic" in a wide variety of fields; communication, media studies, psychology, sociology, homeland security, marketing, and business, just to name a few. The ability to analyze large datasets of social media posts to understand who does what on social media with what intention and what outcome is incredibly important in understanding and influencing today's society. Designing a tool that allows researchers with skill and expertise in any of these areas but not computer science would open up this particular area of research to far more researchers from a much wider variety of backgrounds than is possible today. If such a tool were to be developed at UMaine, it would put the university at the forefront of groundbreaking social media research.

Project Sponsor(s):

Judith E. Rosenbaum

Other Resource People:

N/A

Software/server access required:

None that I am aware of.