Mining Social Media Data
Fall 2012
IST 400/600

Nancy McCracken
Research Associate Professor
Contact Information: Office: 209 Hinds Hall
njmccrac@syr.edu
443-3955 (office)
Class Sessions: Mon/Wed   12:45 – 2:05pm   027 Hinds Hall

Course Description:

In this course, the Web APIs of several online Social Media platforms will be used in order to collect social media data. Scripts to use these APIs will be in the programming language Python. A large portion of the course will focus on using the Twitter API while exploring the techniques, and then will apply those techniques to other sites, which may include Facebook, YouTube comments and FourSquare. In the beginning part of the class, Python programming will be taught in the context of using the Twitter API.

The data obtained from the Social Media platforms will include people, text, and connections between them, e.g from Twitter we can collect recent tweets, get profile information of the user who sent it and see what followers there are. The goal of this class is to use the data collected from the Web APIs to answer questions such as:

Who knows whom, and what friends do they have in common?
What are they chatting about? And is it interesting?
Where are the people located?
Who are the most influential/popular people in the network?

In the class, we will connect the results of our data collection with other software packages to manipulate and view the data. These will most likely include:
• NetworkX to build networks of followers, friends, and the like.
• GraphViz to view networks and other data
• Regular Expressions in Python to find people and topics in text
• Microformats such as JSon and geo data
• OAuthorization for some types of authorized API access

The course will use the following text for Python scripts that access Social Media Web APIs.
Mining the Social Web, Matthew A. Russell, O'Reilly 2010.

We expect to also identify either an online web site or an introductory book for learning Python.
Pre-requisite: Students will be expected to know some programming before taking the course; IST 256 is an appropriate pre-requisite for undergrads or equivalent

programming knowledge gained through another avenue such as self-study. Please send email to the instructor if you have any questions about meeting the pre-requisite.

Course Organization:

The format of the course will be largely informal, viewed as an exploration of the tools in order to achieve the goals. Overall, the class format will be in the form of the instructor and students working through programming script examples together. In the early parts of the course, the instructor will have prepared a set of examples to explain and work through those parts of Python that result in using Python scripts to access the Web APIs and to manipulate the collected data.

As we progress, we will explore more of the scripts and connect to other software packages that will enable us to process and view the data. As the Social Media platforms change fairly rapidly, we will adapt our use of the software as we go through the course. Students will be encouraged to share their knowledge and discoveries for all to use.

Coursework:

Most of the work in this class will involve a mix of running existing Python scripts for the Web APIs and doing original programming to connect the data with other software. Since most of the coursework involves writing and running programs, students should be prepared to learn Python and to work in a complex software environment.

In addition to working through examples in class, homework assignments will be given in the first half of the class for doing additional examples or adapting the Python scripts to other uses. The second half will be spent in working on a project. Students may work together in small groups or may work independently.

As this is a combined undergrad/grad class, graduate students will be required to do an additional report based on the project.