Stat 222: Twitter Project

1 Data Description

For this project, your primary data source will be Twitter. And you will be expected to work with the Twitter API using Python to access this data. On Wednesday (1/20), I will briefly review Python, introduce JSON (Javascript Object Notation), and demonstrate how to interact with the Twitter API using the Python Twitter Tools. However, you will need to do outside reading to get up to speed with these tools.[1]

An important goal for this project is to provide an opportunity for you to get more practice using Python. In particular, for this project I expect you to gain more experience working with basic Python structures (lists, dictionaries, tuples, and strings) and work with JSON and CSV using Python. You will also be using Python's string processing and text mining capabilities to process the data. While you may wish to use some of the packages in the scientific Python stack, the focus will be on core Python language features and a few specialized libraries.

Python Twitter Tools is one of several Python packages[2] for interacting with the Twitter API. It is fairly minimal and is the package used in chapter 1 and 9 of *Mining the Social Web*.[3]

- Chapter 1: Mining Twitter: Exploring Trending Topics, Discovering What People Are Talking About, and More
- Chapter 9: Twitter Cookbook
- Mining the Social Web notebooks

2 Your Assignment

Each group is responsible for creating a slide and poster presentation that visually answers a set of questions, which you will determine for yourselves. You will first need to decide on a

[1] I am holding a 1-day Python bootcamp on Saturday, January 23rd. If you don't feel comfortable with Python, you should attend the bootcamp.

[2] http://www.danielforsyth.me/analyzing-a-nhl-playoff-game-with-twitter

[3] https://github.com/ptwobrussell/Mining-the-Social-Web-2nd-Edition

set of question that you can use Twitter data to answer. Once you determine the questions you wish to address, you will need to use the Python Twitter package to download the data from Twitter that you will use to answer the question. You are free to use Python to analyze the data and create plots. Since you've had limited practice with Python, you are welcome to use R for some of the analysis and for creating your figures. However, even if you decide to use R for some of the analysis and plotting, you must use Python to retrieve the data as well as most of the preprocessing. For example, you may decide to use Python to create a term-document matrix (or a term frequency by inverse document frequency matrix) and save the data as a CSV file, which you can then read with R.

In terms of number of plots in your poster, a *minimum* number to aim for is 10, although it might be natural for some plots to be grouped together into a single figure. You will probably want to create more plots for your slide presentations. Last year, Cari Kauffman and I created a slide presentation as an example of the kind of questions and visual answers you might use. You should look over our example, although you should not simply replicate our plots. I've posted our example slides (http://jarrodmillman.com/stat222-spring2016/projects/twitter/senators/senators.pdf) along with the transcript of our talk (http://jarrodmillman.com/stat222-spring2016/projects/twitter/senators/senators-talkscript.txt) on the course website.

Which questions you answer is up to you, but think about telling a story. The story will be more interesting if the questions you address are related to each other in some way. Here are a few example topics:

- investigate the relation of breaking news on Twitter versus traditional news sources
- compare stop words usage on Twitter versus NY Times
- chart how the ratio of positive versus negative words used in tweets involving some event (or issue) change over time
- relate tweets about a TV show/movie/book to their viewers/ticket sales/sales over time
- examine tweets containing a specific hashtag
- compare tweets relating to competing atheletic teams
- look at tweets from academics perhaps relating to a conference or journal article

Here are a few things to keep in mind as you think about what questions you want to ask and how you might address them.

- You will be limited in the amount of historical data you can retrieve, so questions about how things evolve over time may difficult to investigate. I will show you how

to use cron jobs running on a server to run your queries automatically on a regular schedule. This will allow you to somewhat overcome the data limitations imposed by Twitter policy. For example, if Twitter only provides responses for the last two weeks, then using a cron job you can automatically retrieve two weeks of data every week (or two).

- Since tweets are short texts, text mining is a natural approach to take. However, it isn't the only one. For instance, you might create a graph where the nodes are twitter accounts and the edges represent which accounts follow which accounts. Using this graph you could try to identify hubs, authorities, and communities.
- While the focus of your project will involve Twitter data, you should consider incorporating external data sources. For example, you may want to relate your findings to current events (e.g., a spike in activity may be associated with major news events occuring just before the spike). Or, perhaps, you will want to get additional data about certain Twitter users (e.g., biographical details about a Twitter user from Wikipedia or their homepage) or organization (e.g., recent press releases or news articles).

Timeline and logistics

Here is the tentative schedule:

Monday	Wednesday
	(1/20) Start Twitter project
(1/25) Text mining I	(1/27) Text mining II
(2/1) Graphics I	(2/3) Graphics II
(2/8) Slide presentations	(2/10) Group work
No class	(2/17) Group work
(2/22) Poster presentations	

You should spend some time exploring Twitter (`http://www.twitter.com`) as soon as you can to familiarize yourself with how it works. Explore the Twitter interface to discover who uses it and how. You may wish to begin following a few accounts. Pay attention to how people use `#hashtags` and `@mentions`. In addition to individuals, look at what organizations (e.g., companies, non-profits, political campaigns) use it and how. While doing this, you should also begin thinking about what kind of questions you may want to address.

Since this is a project for which you'll have to define and obtain the data yourselves, the goal for the first week is for you *to define what tweets (or other info) you want to work with, download that data, and wrangle it into a simpler format.*

Before Monday, January 25th, each student should retrieve some data from Twitter using the tools I present on the 20th. Working individually, brainstorm *three* specific

questions to use a starting point for exploring this data. The answers to these questions will almost certainly suggest other questions to you, but use these as a starting point. For each question,

- Think about what plot you could make to help answer the question.
- Describe exactly what data or data summaries you need to make the plot.
- Make an initial version of the plot. (You may want to refine it later.)
- Interpret the plot. Does it suggest any other questions/plots to consider?

These steps may sound obvious, but I know from experience that the temptation is there to jump in and start writing code before you've given much thought to what you want to learn. I think you'll find you get richer and more interesting results in an open-ended project like this if you allocate more of your time at this stage to *thinking* and less to implementation.

I will randomly assign each student to a group on Monday, January 25th. For the first half of the class, I will begin a brief introduction to text mining with Twitter data. For the second half of the class, you will meet with your group. During this initial meeting, every member of the team should present the data they retrieved over the weekend as well as the three specific questions you came up with for exploring this data.

Before class on Wednesday (1/27), your team should meet to discuss what general questions you want to ask. Once you have a general idea of what questions you want to ask, you need to decide what data you want to collect. You should immediately start collecting Twitter data with the goal of addressing your questions. In addition to the Twitter data, your group should also discuss what other data sources you may need to use and figure out how you are getting to collect it.

3 Presentation Details

For this project you will be required to create both a slide and a poster presentation. While you are expected to carefully prepare and practice your slide presentation, you should view it as a progress report.

You will have feedback from your slide presentation, which you should use while preparing your poster. You should also continue collecting data after your slide presentation.

Every group will submit their poster for potential presentation at the Berkeley Statistics Annual Research Symposium (BSTARS). All submissions will be reviewed by the Department's Industry Alliance Program (IAP; http://statistics.berkeley.edu/industry/iap) committee for possible inclusion in the program. Submissions will be considered for either a talk or poster.

The BSTARS event is scheduled on Monday, March 14.

3.1 Slide Details

The slide presentations will be given in the **Pecha Kucha**[4] style. A Pecha Kucha presentation consists of 20 slides that are automatically advanced every 20 seconds (20x20). The complete presentation lasts exactly 6 minutes and 40 seconds.

This is a very constrained format. So you will need to carefully plan and prepare your talk. Each group will have 4 members and each member will be responsible for presenting 5 slides. Since the slides will automatically advance, you **must** practice your talks before you present in class.

After your slide presentation, you will receive feedback, which you should incorporate in your poster presentation on Monday (2/22).

How to make slides

There are many ways to create slides, but make sure that you are able to save your slides as a PDF. Here are some possibilities for you to explore:

- Beamer
  http://web.mit.edu/rsi/www/pdfs/beamer-tutorial.pdf
- Pandoc
  http://johnmacfarlane.net/pandoc/demo/example9/producing-slide-shows-with-pandoc
- Powerpoint or Keynote

How NOT to make slides

- Tufte's *PowerPoint Is Evil*
  http://archive.wired.com/wired/archive/11.09/ppt2.html
- Norvig's *Gettysburg Cemetery Dedication*
  http://norvig.com/Gettysburg/sld001.htm
- Efron's *Thirteen rules*
  http://statweb.stanford.edu/~ckirby/brad/other/2013ThirteenRules.pdf

3.2 Poster Details

There are many ways to create a poster. Here are some possibilities for you to explore:

- Beamer Poster Package
  http://www-i6.informatik.rwth-aachen.de/~dreuw/latexbeamerposter.php

[4] http://en.wikipedia.org/wiki/PechaKucha

- Powerpoint (try searching for "research poster powerpoint template")
- Adobe Illustrator or InDesign (available free to students at https://software.berkeley.edu/adobe)
- Pages (recommended over Keynote for posters)

A service I can recommend is http://gif.berkeley.edu/services/printing.html. Note that they require an appointment, which I recommend you go ahead and schedule now.

If you use this printing service, your poster should be 36" or 42" along one side. (These are the paper sizes they stock.) 36" high x 48" wide is fairly common poster size.

When designing your poster, you may find it helpful to draw the layout before you try to implement it in software. Another piece of advice is to be careful with the resolution of your figures. After designing your poster, you may need to regenerate them at the actual size they will occupy. Taking a small figure and enlarging it to fit the poster will cause the image quality to be poor.

The poster should contain a title, your names, data source(s), the plots, and text to tie everything together. Someone should be able to understand the main findings simply by reading it, but be careful not to include so much text that the poster becomes difficult to skim. It's typical to "present" the poster and give more details verbally.