Stat 222 Project 3: Twitter

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 Monday (2/23), we 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.

Python Twitter Tools is one of several Python packages[1] 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.*[2]

- 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 presentation that visually answers a set of questions, which you will determine for yourselves. You will first need to decide on a set of question that you can use Twitter data to answer. Once you determine the questiona 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 the some of the analysis and plotting, you must use Python to retrieve the data as well as most of the preprocessing. Using Python save the data as a CSV file, which you can then read with R.

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

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

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

Timeline

Your final presentations are due in three weeks. Here is the tentative schedule of the next several classes:

Monday	Wednesday
(2/23) Start Twitter project	(2/25) Poster presentations for airline data
(3/2) Text mining	(3/4) Pecha Kucha
(3/9) Group work	(3/11) Practice presentations
(3/16) Final presentations	

Note that you will have a practice presentation on the Wednesday, March 11th. Your group will need to have already prepared and practiced your presentations within your group prior to the 11th. We expect a fairly polished talk for your practice presentation. After your practice presentation you will receive feedback on how to improve your presentation. Part of your final grade for this project will involve how respond to feedback from your practice presentation in your final presentation.

3 Initial Guidance (to do for first data debrief)

Since this is the first 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.*

An important goal for this project is to provide an opportunity for you to get more practice using Python. In particular, for this project we 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.

We will reserve 20 minutes on Wednesday (2/25) for you to meet with your new group and discuss possible topics/datasets. In preparation for that, you should each come up with three ideas between Monday and Wednesday. In addition to the Twitter data, your group should also discuss what other data sources you may need to use.

4 Next Steps

Next Monday (3/2), we will discuss text mining in Python using the Natural Language Toolkit (nltk).[3] By this point, your group should have downloaded the Twitter data and whatever other data you think you will need.

5 Presentation Details

All presentations will be given in the **Pecha Kucha**[4] style. Pecha Kucha presentations have a very strict format. A Pecha Kucha presentation consists of 20 slides that are automatically advanced 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 will need to practice your talks before you present in class.

Note that your group will need to have an official in-class practice presentation on Wednesday (3/11). After your practice presentation, you will receive feedback, which you should incorporate in your final presentation on Monday (3/16).

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

[3] http://www.nltk.org
[4] `http://en.wikipedia.org/wiki/PechaKucha`

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