Analyzing Big Data with Python PANDAS

This is a series of iPython notebooks for analyzing Big Data – specifically Twitter data – using Python's powerful PANDAS (Python Data Analysis) library. Through these tutorials I'll walk you through how to analyze your raw social media data using a typical social science approach.

The target audience is those who are interested in covering key steps involved in taking a social media dataset and moving it through the stages needed to deliver a valuable research product. I'll show you how to import your data, aggregate tweets by organization and by time, how to analyze hashtags, how to create new variables, how to produce a summary statistics table for publication, how to analyze audience reaction (e.g., # of retweets) and, finally, how to run a logistic regression to test your hypotheses. Collectively, these tutorials cover essential steps needed to move from the data collection to the research product stage.

Prerequisites

I've put these tutorials in a GitHub repository called PANDAS. For these tutorials I am assuming you have already downloaded some data and are now ready to begin examining it. In the first notebook I will show you how to set up your ipython working environment and import the Twitter data we have downloaded. If you are new to Python, you may wish to go through a series of tutorials I have created in order.

If you want to skip the data download and just use the sample data, but don't yet have Python set up on your computer, you may wish to go through the tutorial "Setting up Your Computer to Use My Python Code".

Also note that we are using the iPython notebook interactive

computing framework for running the code in this tutorial. If you're unfamiliar with this see this tutorial "Four Ways to Run your Code".

For a more general set of PANDAS notebook tutorials, I'd recommend this cookbook by Julia Evans. I also have a growing list of "recipes" that contains frequently used PANDAS commands.

As you may know from my other tutorials, I am a big fan of the free Anaconda version of Python 2.7. It contains all of the prerequisites you need and will save you a lot of headaches getting your system set up.

Chapters:

At the GitHub site you'll find the following chapters in the tutorial set:

Chapter 1 – Import Data, Select Cases and Variables, Save DataFrame.ipynb
Chapter 2 – Aggregating and Analyzing Data by Twitter Account.ipynb
Chapter 3 – Analyzing Twitter Data by Time Period.ipynb
Chapter 4 – Analyzing Hashtags.ipynb
Chapter 5 – Generating New Variables.ipynb
Chapter 6 – Producing a Summary Statistics Table for Publication.ipynb
Chapter 7 – Analyzing Audience Reaction on Twitter.ipynb
Chapter 8 – Running, Interpreting, and Outputting Logistic Regression.ipynb

I hope you find these tutorials helpful; please acknowledge the source in your own research papers if you've found them useful:

Saxton, Gregory D. (2015). *Analyzing Big Data with Python.* Buffalo, NY: http://social-metrics.org

Also, please share and spread the word to help build a vibrant community of PANDAS users.

Happy coding!

How Many Tags is Too Much?

Including a hashtag in a social media message can increase its reach. The question is, what is the ideal number of tags to include?

To answer this question, I examine 60,919 original tweets sent in 2014 by 99 for-profit and nonprofit member organizations of a large US health advocacy coalition.

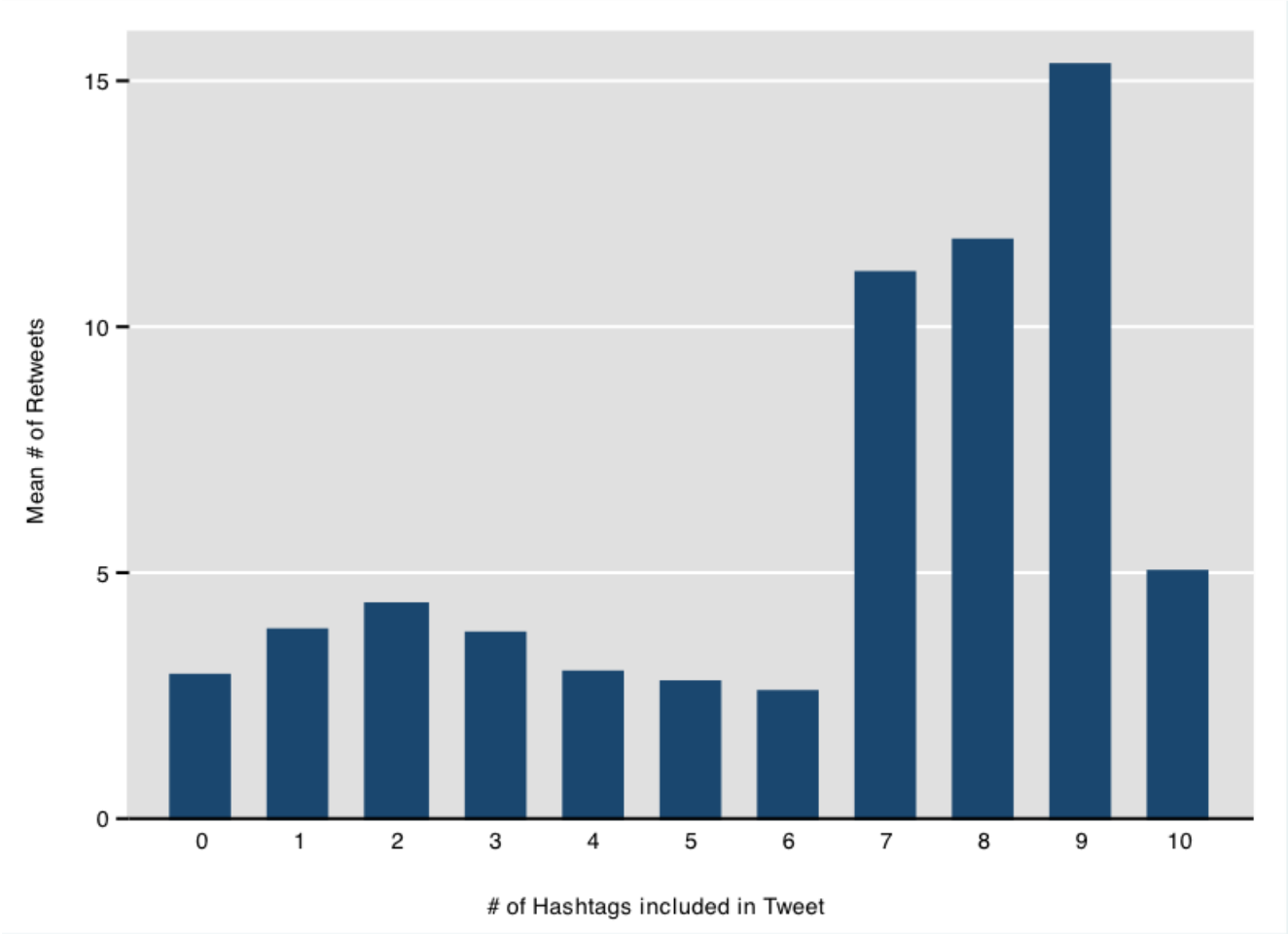
First, the following table shows the distribution of the number of hashtags included in the organizations' tweets. As shown in the table, almost a third (n = 19,747) of tweets do not have a hashtag, almost 39% (n = 23,493) have one hashtag, 19% include two hashtags (n = 11,836), 7% include three (n = 4,381), and 2% (n = 1,161) include 4. Few tweets contain more than 4 tags, though one tweet included a total of 10 different hashtags.

Frequency of Hashtags in 60,919 Original Tweets

# of Hashtags	Frequency
0	19,747
1	23,493
2	11,836

# of Hashtags	Frequency
3	4,381
4	1,161
5	227
6	49
7	13
8	4
9	7
10	1
Total	60,919

Now let’s look at the effectiveness of messages with different numbers of hashtags. A good proxy for message effectiveness is retweetability, or how frequently audience members share the message with their followers. The following graph shows the average number of retweets received by tweets with different numbers of hashtags included.



What we see is that more hashtags are generally better, but there are diminishing returns. Excluding the 25 tweets with more than 6 hashtags, the effectiveness of hashtag use peaks at 2 hashtags, with more than 3 hashtags being only as effective or less effective than no hashtags.

The evidence isn't conclusive – especially given the anomalous findings for the few tweets with 7-10 tags – but there is strong support here that, if you want your message to reach the biggest possible audience, limit your tweets to 1-2 hashtags.