Can We Predict Whether a User Will Retweet the Personalvictories Hashtag?

Alexander Egorenkov

Monday, July 27, 2015

(Sorry that this is very rough. The most useful repository files are scraper.py and fit_text_model.py . The text model files should run with minor modification to the directory variable)

Summary

We use the Twitter API to collect information about users who use the #personalvictories hashtag, including posters and retweeters. From this information we expect to be able to predict whether a given user will retweet the hashtag. Furthermore, we can find characteristic among users that make them more likely to retweet the hashtag.

Description of your data set and how it was obtained

Retrieval Process

- Use manual twitter search to find tweets, also refered to as statuses that contains the #personalvictories hashtag in varying letter cases.
- From the manual search we can extract user IDs, status IDs, and the text of the status. As well as how many times the status has been favorited or retweeted.
- We can extract the user IDs of retweeters by using our collection of status IDs and searching for related retweeters through the Twitter search API
- Once we have all the IDs of random users, poster, and retweeters, we can collect the most recent statuses posted by each user.
- From these statuses we can extract keywords and other useful features for classifying users.

Description of any pre-processing steps you took

- Made sure that none of the randomly selected users coincided with retweeters.
- Made sure that the #personalvictory hashtag was not left in the feature matrix.
- Created a document term matrix both that keeps track of both word count and presence.

What you learned from exploring the data, including visualizations

- Generally, the keywords associated with retweeters are intuitive, such as diet, fitness, and smoke.
- There are a handful of associated keywords that I didn't expect such as extra and girlposts.
  - I suspect extra comes up due to people celebrating about getting extra items for free, which also suggests that I am accidently including retweets in my sample of statuses.
  - girlposts is a twitter handle, could the hashtag be associated with this account in some way?

How you chose which features to use in your analysis

- Much of the choice is driven by wha tis available in a reasonable amount of time through the Twitter API
- Since my core interest is the text data, much of the choice also comes down to simple what seems to work.

Details of your modeling process, including how you selected your models and validated them

- The only model I am currently testing is Naive Bayes on tweet data predicting a binary class
  - I valided with test/train split using accuracy and AUC as metrics
  - So far I've found that n-grams are helpful
  - It only takes a handful (50) of keywords to make good predictions
  - Too many keywords worsen the model
  - Testing for keyword presence is more effective than testing for keyword count

Your challenges and successes

- Taking full advantage of the Twitter API is a concurrent/multi-threaded process, there's lot going on at once.
  - My work is a bit scattered, but I approach the problem by building several individual web scrapers and use a single script to coordinate them with an event based message passing method rather than running simultaneous threads.
- Text data in csv form can be unpleasant due to punctuation
- I'm working on this, but all my json files parse the data well, maybe I should just work off the json rather than take an intermediate step in csv
- Since the response variable is based on status information and the features are based on status information, it's possible to include too much of the same information. +Not a big problem overall, but I may have a few issues to resolve.

Possible extensions or business applications of your project

- Identify properties of likely retweeters to infer demographic or psychographic information for readers of personalvictory.com
- Predict whether a given twitter handle may find a personalvictory message interesting and target gentle advertising

Conclusions and key learnings