Twitter Analysis in R

Jessica Short

April 8, 2014

Background and Objective

Many insurance and financial products are sold through advisors. There is a biweekly twitter meetup for advisors to talk about relevant industry trends and strategies. These conversations are identified with 'AdvisorTalk'.

I propose to collect the tweets that compose these conversations, as well as the other twitter information of participants.

The primary objective of this analysis is to gain more information about the online presence of the advisors and to understand how they are using twitter to build their personal brand. The secondary objective is to gain an understanding of the data available from twitter, and how to analyze that data in R.

Systematic Data Collection

This section is run biweekly to acquire the advisor chat data as well as the users that participated and their other tweets. This is run bi-weekly after the chats occur, since the API only indexes 1.5 - 2 weeks of data. Collection is also affected by rate limiting as described on the twitter developers website.

```r
PFG <- getUser("ThePrincipal")
PFG_Tweets <- userTimeline(PFG)

rsTweets <- searchTwitter("#AdvisorTalk", n = 500)
tweets.df2 <- twListToDF(rsTweets)

users <- unique(llply(rsTweets, screenName))

# have to use a function so the system won't rate limit
at <- data.frame()
for (user in users) {
    # Download latest 100 tweets from the user's timeline
    tweets <- twListToDF(userTimeline(user, n = 100, includeRts = TRUE))
    at <- rbind(at, tweets)
    Sys.sleep(25)
}

saveRDS(tweets.df2, file = paste0("AT_", today(), ".Rdata"))
saveRDS(at, file = paste0("UserHist_", today(), ".Rdata"))
```

```r
collectDates <- c(collectDates, today())
```

Explore Data

First, I re-collect all the data that has been saved off after each week. Then, I create some summaries and do some data exploration and cleaning.

The dates where tweets were collected are

- 03-05-2014
- 03-20-2014
- 04-06-2014

```
collectDates <- c("2014-03-05", "2014-03-20", "2014-04-06")

# AT_test is the list of Advisor Talk tweets
AT_test <- ldply(collectDates, function(x) {
    res <- readRDS(paste0("AT_", x, ".Rdata"))
    res$ext_dt <- x
    res
})

# drop the retweets from the AT_test collection
AT_test_noRT <- subset(AT_test, !isRetweet)


# UH_test is all the tweets for users that participated in an Advisor Talk
UH_test <- ldply(collectDates, function(x) {
    res <- readRDS(paste0("UserHist_", x, ".Rdata"))
    res$ext_dt <- x
    res
})
# get user/account details
at_users <- twListToDF(lookupUsers(unique(AT_test_noRT$screenName)))
at_uids <- at_users$id
# de-dup tweets in UH, may have pulled duplicates since history was pulled 3
# times
UH_test2 <- UH_test[!duplicated(UH_test$id), ]

AT_summ <- ddply(AT_test, .(screenName), summarise, n_AT_tw = length(id), n_AT_ses = length(unique(ext_dt),
    n_AT_RT = sum(retweetCount)))

# general stats on n_following n_followers

userStats <- ddply(UH_test2, .(screenName), summarise, min_dt = min(created),
    max_dt = max(created), n_tw = length(screenName), avg_RT = (sum(retweetCount)/n_tw))

userStatsFinal <- merge((at_users[, c(13, 1:12, 14:16)]), merge(userStats, AT_summ,
    by = "screenName"), by = "screenName")
userStatsFinal <- userStatsFinal[, c(2, 1, 3:21)]

# create dataset by day/user to do a line plot

byDay <- ddply(UH_test2, .(screenName, round_date(UH_test2$created, unit = c("day"))),
    summarise, n = length(id))
colnames(byDay) <- c("screenName", "day", "n")
```

```
at_uids3 <- userStatsFinal$id[userStatsFinal$n_AT_ses == 3]
```

One of the interesting features in collecting the data was that retweets are returned in the searches. But by default, are not included in the user's history. This can be adjusted in the twitteR function UserTimeline by setting the includeRts option.

Another issue is rate limiting. I am exploring some ways to pull the data from twitter more effectively. For now, because this is an exploratory project, I am using a smaller subset of data to begin with.

I chose to eliminate retweets from the talk sessions, since the user who retweets is not generating their own content.

The derived variables created for each user include

- Number of Advisor Talk tweets
- Number of Advisor Talk sessions
- Number of retweets of AdvisorTalk tweets
- minimum date of all tweets collected
- maximum date of all tweets collected
- Number of tweets collected in user history
- proportion of tweets that are retweeted

There are also a number of variables available directly, that may be valuable.

- user description
- url
- name
- created
- location
- profileImageUrl
- statuses Count
- followers Count
- favorites Count
- friends Count

```
head(userStatsFinal)
```

```
##           id     screenName
## 1  556066113    _HannahMLee
## 2  531707627           _RLJ
## 3  278375993 1brentwilliams
## 4 1395576990    401k_Grande
## 5 1848775616     401Kroeger
## 6  142306693    a_t_freeman
##
## 1                                                         YP in Des Moines, currently focusing on advisor mark
## 2 Account Executive Director-CFP @ThePrincipal\nSharing my thoughts on retirement services and participant edu
## 3                 Trusted father, friend, and professional. Love life and live life to the fullest! When I giv
## 4                                           Working to help advisors help Americans prepare for a dignified re
## 5
## 6                      Account Executive - Retirement Services with Principal Financial Group. Disappointing G
##   statusesCount followersCount favoritesCount friendsCount
## 1           257             85              0           85
## 2           382            430             47          546
## 3          3222            901              4         1601
## 4           644            113            433           88
## 5           379             61              3          226
## 6            80             41              0          237
##                       url                 name             created
## 1                    <NA>           Hannah Lee 2012-04-17 13:51:03
## 2 http://t.co/IY0RydMHDF       Robert Johnson 2012-03-20 22:46:26
## 3                    <NA>       Brent Williams 2011-04-07 04:42:21
## 4                    <NA> Jefferson D Cheshier 2013-05-01 21:16:34
## 5                    <NA>     Nicholas Kroeger 2013-09-09 17:10:58
## 6 http://t.co/qUCX8DmZhg       Andrew Freeman 2010-05-10 14:38:02
##   protected verified              location listedCount followRequestSent
## 1     FALSE    FALSE      Des Moines, Iowa           9             FALSE
## 2     FALSE    FALSE      Atlanta, Georgia          12             FALSE
## 3     FALSE    FALSE West Des Moines, Iowa          21             FALSE
## 4     FALSE    FALSE            Austin, TX           6             FALSE
## 5     FALSE    FALSE                                 5             FALSE
## 6     FALSE    FALSE           Chicago, IL           0             FALSE
##                                                                                          profileImageUrl
## 1                     http://pbs.twimg.com/profile_images/2579447290/odki9v4cqvo9a8sxs5m5_normal.jpeg
## 2                         http://pbs.twimg.com/profile_images/448595991344599040/9DNkzuFV_normal.jpeg
## 3                                 http://pbs.twimg.com/profile_images/1302816045/Principal_normal.jpg
## 4         http://pbs.twimg.com/profile_images/3637348842/d7455adbf17bbbcf32c9b1de997f80b3_normal.jpeg
## 5 http://pbs.twimg.com/profile_images/378800000435024303/0b4bec9f9bd3d3851fd6956c4c6232c6_normal.jpeg
## 6                         http://pbs.twimg.com/profile_images/443061926046621696/eDwXAiPd_normal.jpeg
##                 min_dt              max_dt n_tw n_AT_tw n_AT_ses
## 1 2013-02-26 15:02:53 2014-03-07 17:05:41   94       5        1
## 2 2013-11-27 02:18:03 2014-04-05 14:37:51  132      13        3
## 3 2014-02-27 17:10:32 2014-03-20 19:10:40   90       3        1
## 4 2014-02-05 20:43:03 2014-04-04 14:55:30  100      11        2
## 5 2013-09-12 12:12:01 2014-03-28 19:04:20    3       4        2
## 6 2011-03-24 13:43:03 2014-04-04 14:01:35   65       8        1
```

Some terms that will be important throughout the analysis are defined below.

- Followers: Users that receive your updates in their feed
- Friends: Users that you follow
- Favorites: Number of Tweets favorited in account history

Social Network Graphs

In this section I explore network graphs to measure relationships among users. I am currently using friend relationships to create network edges. Replies and Retweets are also good options for describing the connection between users. I am using Sys.sleep to pause my function and avoid some rate limits.

```
# rate limited at 15 calls in 15 minutes
friends_df <- ldply(at_uids3, function(x) {
    Sys.sleep(60)
    x1 <- getUser(x)$getFriendIDs()
    x2 <- x1[x1 %in% at_uids]
    return(data.frame(cbind(node_id = x, friend_id = x2, friend = 1)))
})


library(igraph)
############## add vertex properties to graph data frame#############
c1_fullv <- graph.data.frame(d = c1, vertices = userStatsFinal)
get.edge.attribute(c1_fullv, "friend")

V(c1_fullv)
# returns a list of the IDs of each vertex in the graph.
p1 <- plot(c1_fullv)
reports_to_layout <- layout.fruchterman.reingold(c1_fullv)
```

I haven't been able to collect all the data necessary to link all the user nodes. The graph currently shows the relationships of 9 users to other participants of the talk series.

```
plot(c1_fullv, layout = reports_to_layout)

## Error:  'x' is a list, but does not have components 'x' and 'y'
```

Future Work

This is primarily an exploratory effort to understand the data that is available via twitter, and the ease of access through R and the twitteR package. There are many potential extensions of this information.

- Search Twitter for tweets regarding retirement, financial services, etc. and then scrape the profile descriptions and urls of those users to see what other web presence those users have. See if we can link twitter handles to linked in profiles, and pull any information from LinkedIn.
- Try Sentiment Analysis of the tweets collected.
- Identify what proportion of tweets among these users are 'business' related (retirement, investment, 401k, etc.) vs. personal.