Data Wrangling

Sine Gov

December 30, 2017

## Data Sources

The data for this project was gathered from three sources: Reddit, the Official Overwatch Forums, and Twitter. Since we are performing sentiment analysis For this project, we are only interested in collecting the text comments from each of these sites. Each section below outlines how data was gathered and cleaned for each site.

## Twitter

Using the twitteR and ROAuth libraries, the hashtag #Overwatch was used to collect tweets. After authenticating the session with Twitter, unique tweets in English were targeted for collection. Since the twitteR package outputs the data in a clean form, no additional data wrangling was needed.

### Code to scrape tweets from twitter

...  
  
# Grab latest tweets  
print("Grabbing tweets!!!!! Will take a second.")  
tweets\_OW <- searchTwitter("#overwatch, -filter:retweets", n=14586, lang = "en")  
  
df\_tweets\_OW <- twListToDF(tweets\_OW)  
write.csv(df\_tweets\_OW, file = "OWT\_12-22.csv")

### Tweets scraped

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ï.. | text | favorited | favoriteCount | replyToSN | created | truncated | replyToSID | id | replyToUID | statusSource | screenName | retweetCount | isRetweet | retweeted | longitude | latitude |
| 1 | Merry (early) Christmas to me. #overwatch @PlayOverwatch <https://t.co/LZRLBOEylQ> | FALSE | 0 | NA | 12/22/2017 19:36 | FALSE | NA | 9.44290e+17 | NA | Twitter for iPhone | realBOBtheBOB | 0 | FALSE | FALSE | NA | NA |
| 2 | Going live in 10minutes!! Starting off the holidays with some video games! Come join me? What will we play tonight?â€¦ <https://t.co/BtjNJtzrEL> | FALSE | 0 | NA | 12/22/2017 19:35 | TRUE | NA | 9.44290e+17 | NA | TweetDeck | twistylc | 0 | FALSE | FALSE | NA | NA |
| 3 | On the set w/ @mrdubbsie @bhanford #gwne #gwnenow #esports #overwatch #esportsfranceâ€¦ <https://t.co/6Vh6cC2utT> | FALSE | 0 | NA | 12/22/2017 19:35 | FALSE | NA | 9.44290e+17 | NA | Instagram | heygirlheyLA | 0 | FALSE | FALSE | -118.243 | 34.0522 |
| 4 | Slid back into gold…..SR quicksand?? #Overwatch #OverwatchWinterWonderland <https://t.co/ToVg2WNYOv> <https://t.co/JPGLSQXQSJ> | FALSE | 0 | NA | 12/22/2017 19:34 | FALSE | NA | 9.44290e+17 | NA | Twitter Web Client | aimbothero1 | 0 | FALSE | FALSE | NA | NA |
| 5 | <https://t.co/A0OFHcuUqJ> come watch me play #overwatch #OverwatchWinterWonderland #stream #twitch #smallstreamersâ€¦ <https://t.co/CXGuUGF5pW> | FALSE | 0 | NA | 12/22/2017 19:30 | TRUE | NA | 9.44289e+17 | NA | Twitter for iPhone | Forsythiagrace | 1 | FALSE | FALSE | NA | NA |

## Reddit

A custom script was written to gather data from reddit. The script first gathers all of the hyperlinks of all the topics on the first page of the subreddit /r/Overwatch. Once it has all the necessary links, it visits each link and copies each comment, the date of the comment, and the user name. The libraries listed in the code below were used to scrape this website.

### Code to scrape hyperlinks of topics on the front page

library(rvest)  
library(tidyr)  
library(RCurl)  
library(dplyr)  
library(curl)  
  
#Read URL  
ow.forums <- read\_html("https://www.reddit.com/r/Overwatch/")  
  
#scrapes links from read URL  
links <- ow.forums %>% html\_nodes("a") %>% html\_attr("href")  
  
#Makes scrapped links a data frame  
links <- as.data.frame(links, row.names = NULL)  
  
#Find the links for the forums (since links are /forums) and keep these.   
keep<- grepl("^https://www.reddit.com/r/Overwatch/comments/", links$links)  
#print(keep)  
keep <- as.data.frame(keep, row.names = NULL)  
  
#bind the answers to the new dataframe  
links <- cbind(links, keep)  
  
#remove the rows of unneeded links  
#links now contains all of the links for the forum post to be scrapped  
#paste to strip out unnecessary rows  
links <- links[links$keep == "TRUE",]  
links <- paste(links$links)  
  
#new df ow.total  
owreddit.total <- data.frame("link", "title", "time", "text", "user", stringsAsFactors = FALSE)  
print(owreddit.total)

### Code to scrape data from each topic

#While loop iterates through scraped links for comments.  
n <- 1  
while(n <= length(links)) {  
 read\_single\_links <- read\_html(links[n])  
   
 #scrapes the link's title  
 link\_title <- gsub("https://www.reddit.com/r/Overwatch/comments/", "", links)  
 link\_title <- substr(link\_title,8, nchar(link\_title)-1)  
 link\_title <- gsub("\_", " ", link\_title)  
   
 #scrapes comments  
 comments <- read\_single\_links %>% html\_nodes(".md")  
   
 #convert from xml to usable format  
 comments <- trimws(comments)  
 comments <- gsub("<.\*?>", "", comments)  
   
 #Cleans comments by replacing \n html code   
 comments <- gsub("\n", "", comments)  
   
 #scrapes the .tagline node which has username data  
 reddit\_user <- read\_single\_links %>% html\_nodes(".tagline")  
   
 #Scrapes to the user name  
 reddit\_user <- gsub('.\*user/', "", reddit\_user)  
   
 #Scrapes the stuff after class out  
 reddit\_user <- gsub('class.\*', "", reddit\_user)  
   
 #clean up of user name  
 reddit\_user <- gsub(" ","",reddit\_user, fixed=TRUE)  
 reddit\_user <- gsub("\"","",reddit\_user, fixed=TRUE)  
 reddit\_user <- gsub("<p","",reddit\_user, fixed=TRUE)  
   
 #scrapes comment timestamp by date and cleans it  
 comment\_time <- read\_single\_links %>% html\_nodes(".tagline, time")  
 comment\_time <- gsub('.\*datetime=', "", comment\_time)  
 comment\_time <- gsub('T.\*', "", comment\_time)  
 comment\_time <- gsub('\"', "", comment\_time)  
 comment\_time <- gsub('<p.\*', "", comment\_time)

### Code to write all the scraped data into a csv file

#adds topic to a vector with link, title using a for loop  
 for(i in 1:length(comments)) {  
 owreddit.total <- rbind(owreddit.total, c(links[n], link\_title[n], comment\_time[i], comments[i+2], reddit\_user[i+1]))  
 }  
   
 print(n)  
 n<-n+1  
 #Sys.sleep(2)  
}  
  
write.csv(owreddit.total, file = "reddit\_comments.csv")

### Overwatch subreddit scraped

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | X.link. | X.title. | X.time. | X.text. | X.user. |
| 2 | <https://www.reddit.com/r/Overwatch/comments/7jd5bw/its_the_most_wonderful_time_of_the_year_and_we/> | its the most wonderful time of the year and we | 12/12/2017 | I want my bohrd | zumoro |
| 3 | <https://www.reddit.com/r/Overwatch/comments/7jd5bw/its_the_most_wonderful_time_of_the_year_and_we/> | its the most wonderful time of the year and we | 5/24/2012 | What are you talking about? This is your bird. | WildSeaturtle |
| 4 | <https://www.reddit.com/r/Overwatch/comments/7jd5bw/its_the_most_wonderful_time_of_the_year_and_we/> | its the most wonderful time of the year and we | 12/12/2017 | I WANT MY BOHRD | BigShield |
| 5 | <https://www.reddit.com/r/Overwatch/comments/7jd5bw/its_the_most_wonderful_time_of_the_year_and_we/> | its the most wonderful time of the year and we | 12/12/2017 | Okay, nothing’s impossible. Are we talking about- uh, is this a bird back in Russia? | SassyAssAhsoka |
| 6 | <https://www.reddit.com/r/Overwatch/comments/7jd5bw/its_the_most_wonderful_time_of_the_year_and_we/> | its the most wonderful time of the year and we | 12/12/2017 | Omg. Ivan Vanko? | BigShield |

## Overwatch Forums

Simliar to the reddit, a custom script was written to gather data from the official Overwatch forums. The script first gathers all of the hyperlinks of all the topics on the first page of the subreddit /r/Overwatch. Once it has all the necessary links, it visits each link and copies each comment and and the user name. Note that there is code to scrape date data, however it works in rare instances. Since date is not necessary for the analysis later, no further effort was made to clean this up and will be removed in the final version of this script. The libraries listed in the code below were used to scrape this website.

### Code for gathering links

library(rvest)  
library(stringr)  
library(tidyr)  
library(RCurl)  
library(dplyr)  
library(curl)  
  
  
#Read URL  
ow.forums <- read\_html("https://us.battle.net/forums/en/overwatch/22813879/")  
  
#scrapes links from read URL  
links <- ow.forums %>% html\_nodes("a") %>% html\_attr("href")  
  
#Makes links scrape a data frame  
links <- as.data.frame(links, row.names = NULL)  
  
#Find the links for the forums (since links are /forums) and keep these.   
keep<- grepl("^/forums", links$links)  
keep <- as.data.frame(keep, row.names = NULL)  
  
#bind the answers to the new dataframe  
links <- cbind(links, keep)  
  
#remove the rows of unneeded links  
#links now contains all of the links for the forum post to be scrapped  
links <- links[links$keep == "TRUE",]  
  
#links only contains the tail end of the links. creating front\_url to be concatenated later  
front\_url <- "https://us.battle.net"  
  
  
#concatenate to create a full link and remove duplicates  
full\_links <- paste(front\_url,links$links, sep = "")  
#remove duplicate full\_links  
clean\_links <- unique(full\_links)  
  
  
#new df ow.total  
ow.total <- data.frame("link", "title", "time", "text", stringsAsFactors = FALSE)

### Code for scraping data from each link and writing it to file

#For loop to access URL from full\_links. Scrapes titles and topic  
  
n <- 1  
while(n <= length(clean\_links)) {  
 read\_single\_links <- read\_html(clean\_links[n])  
   
 #scrapes the title of the thread  
 title\_results <- read\_single\_links %>% html\_nodes(".Topic-title")  
 title <- xml\_contents(title\_results) %>% html\_text(trim = TRUE)  
 #print link and title below. line for to see new threads easier  
 print("------------------------------------------------------------------------------")  
 print(clean\_links[n])  
 print(title)  
   
 #pulls content of thread  
 topic\_results <- read\_single\_links %>% html\_nodes(".TopicPost-bodyContent")  
 topic <- trimws(topic\_results)  
   
 #cleanup of bodycontent text.   
 clean\_topic <- gsub("<.\*?>", "", topic)  
 clean\_topic <- gsub("\n", "", clean\_topic)  
   
  
 #pulls date  
 topic\_time <- read\_single\_links %>% html\_node(".TopicPost-timestamp")  
 clean\_time <- gsub('<.\*content=\\"', "", topic\_time)  
 clean\_time <- gsub('\">.\*', "", clean\_time)  
 clean\_time <- as.POSIXct(clean\_time, format = "%m/%d/%Y %I:%M %p")  
 clean\_time <- paste(clean\_time)  
 print(clean\_time)  
   
 #adds topic to a vector with link, title using a for loop  
 for(i in 1:length(clean\_topic)) {  
 ow.total <- rbind(ow.total, c(clean\_links[n], title, clean\_time[i], clean\_topic[i]))  
 }  
   
 print(n)  
 n<-n+1  
   
}

### Additional Cleanup of Overwatch Forum Data

#A problem was found in the text in that some of it contained dates in the comment. Additional code was needed to remove the dates from the text.  
  
####-----Cleaning up the Text column since it has dates merged into some of them  
newdate <- c()  
newtext <- c()  
  
#for loop that looks for the date text andd removes it  
for(i in 1:length(ow.total$X.text.)) {  
 if(grepl('\\d\\d/\\d\\d/.\*?M', ow.total$X.text.[i]) == FALSE) {  
 print("CLEAN")  
 } else {  
 print("UNCLEAN! CLEANING")  
 m <- regexpr('\\d\\d/\\d\\d/.\*?M',ow.total$X.text.[i])  
 owfdates <- regmatches(ow.total$X.text.[i], m, invert = FALSE)  
 owcleantext <- gsub('\\d\\d/\\d\\d/.\*?M', '', ow.total$X.text.[i])  
 newdate[i] <- owfdates  
 newtext[i] <- owcleantext  
 }  
}  
  
#this loops replaces only the comments with date information with the cleaned up version  
for(i in 1:length(ow.total$X.text.)) {  
 if(is.na(newtext[i])) {  
 print("Do not write over old value")  
 } else if(newtext[i] != 'NA') {  
 #print("Copy over new data")  
 ow.total$X.text.[i] <- newtext[i]  
 ow.total$X.time.[i] <- newdate[i]  
 }  
}  
  
#remove duplicate values  
str(ow.total)  
newow.total\_clean <- ow.total[!duplicated(ow.total[4]),]  
  
#write the file  
write.csv(newow.total\_clean, 'OWFORUMS\_manualcheck\_12\_30.csv')

### Table for Overwatch Forum Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ï..link | title | time | text |
| 6 | <https://us.battle.net/forums/en/overwatch/topic/20752619458> | How do I know if someone I reported gets banned? | NA | It would be quite unethical to do so, you reported it and blizzard is the neutral party between you, and the accused. Anything that happens beyond that point is entirely between the accused and blizzard. Apart from that I believe blizzard would not ban/punish for a single game where that occured, unless the person was using a hack that was registered by ingame detections as well, and that is a good thing as anyone can have a moment where their composure slipped and by murphy’s law push to talk was on during that time. KYS is a stretch and sounds like something they would do on a regular basis if they have no qualms with saying that though. Therefore any reports from within the same game should probably count less than from 2 seperate instances. If that doesn’t satisfy your vigilante nature, think of it this way: If they are cheating, toxic, or any other blizzard “offense”, by the sheer amount of players that player encounters while playing if he is doing something wrong, he will get punished in due time. |
| 7 | <https://us.battle.net/forums/en/overwatch/topic/20752619458> | How do I know if someone I reported gets banned? | NA | Nice public shaming OP. Read the rules before posting. You’re no better than that guy you’re talking about.Edit: That said, I would really appreciate it if you got a message when logging into the main menu saying: “A player you reported for [Report Reason (aka Spamming, Griefing)] has been suspended for XX hours/days/weeks” similiar to how you get the “Leaving Several games in succession will result in a penalty” when leaving too many QP matches. |
| 8 | <https://us.battle.net/forums/en/overwatch/topic/20752619458> | How do I know if someone I reported gets banned? | NA | OP, why do you care?If you’ve reported them, then there’s nothing more for you to do…. you’re done. You get to go on playing the game and now the reported player is in Blizzard’s hands.What’s the point in continuing to focus on some random internet troll? Yes, severe transgressions should be punished… however… the sooner you learn to let go, the better your experience will become. |
| 9 | <https://us.battle.net/forums/en/overwatch/topic/20752616810> | McCree’s Race/Nationality? | NA | Canadian duh |
| 10 | <https://us.battle.net/forums/en/overwatch/topic/20752616810> | McCree’s Race/Nationality? | NA | Chinese, duh. |