Using R and Text Mining to Identify Twitter Bots With Limited Data

Danny Cohen

Webster University

Problem Description

Often times, while browsing social media websites such as Twitter and Tumblr, the userbase at large will receive unwarranted messages from spam bots through direct messages or mentions in their automatically generated posts. These bots can be fast and far reaching bots designed to post one message over and over again, or they can be specific and guided in their approach to contacting users. For instance, certain bots may be created with the idea of promoting gaming equipment or youtube channels and will target users that an algorithm identifies based on attributes such as: the accounts they follow, the tweets they like, or the shopping pages they've viewed. This can either be for the purpose of selling items, or promoting other social media pages for the purpose of attaining disingenuous views.

One question often asked by disgruntled users is: "how difficult it is to set up automatic filtering for spam bots based off commonly used words, phrases, or posting patterns?" A study performed by researchers from the University of California and Indiana University estimated that "up to 15% of accounts being bots" (Varol). Based on the total userbase present on Twitter of "currently […] 319 million monthly active users, that translates to nearly 48 million bot accounts" (Newberg). Twitter noticeably handles this sort of issue much better than most social media websites, usually being able to stop spam bots after a certain number of tweets or being able to instantly delete spam bots as they pop up. Others, such as Tumblr, often times let many obvious bots stay active for months on end while they contact users with ads and social engineering scams. Tumblr appears to handle the process of removing bots based on reports generated by their users, while Twitter almost certainly makes use of an automated algorithm designed to single out and remove bots as quickly and efficiently as possible. With this in mind, the goal of this experiment is to ascertain how difficult it is to identify bots based on their possible common typing patterns in a short period of time, as well using this information to differentiate a real person from a bot.

Background

Text mining is a concept based upon the idea of deriving some sort of information from a text source, which could be any kind of information, ranging from hidden text files within digital applications, long numeric information gathered from web pages, or just pulling out important information from structured surveys. For this particular experiment, the process of text mining will be applied to twitter in an attempt to derive a possible uniform pattern between multitudes of spam bots. In common practice, the "twitteR" package in R is used to text mine twitter, but traditional methods of crawling through webpages or obtaining a personal copy of one's own tweet archive, the entirety of all the tweets that have been made on that same individual's account. The "twitteR" package is a bit more flexible, though, as it comes standard with a variety of commands used to gather and obtain information from: user accounts, search pages, trending tags, and specific tweets, and more. In general, most information obtained through user accounts is very generic, detailing numbers of followers or tweets, while the info obtained from more specific commands can retrieve swaths of tweets that a user or users have interacted with. These commands can be particularly useful when working with specific twitter users.

Methodology

The process for this research begins with collecting data from twitter archives, specifically those of six easily identifiable spam bots, five of which will be from different areas of potential interest. These areas of interest will range from topics such as: gaming, financial investments, travel, monetary scams, and automatic shopping notifications. The sixth archive will be from one of the areas above so it can be compared individually to the other, matching spam bot's twitter archive. This will be done in R using the package “twitteR,” which will then be used to gather a variety of tweets and analytics regarding each account. These will then be formatted and analyzed using a variety of R packages. These packages include: the “lubridate” package, the “tidytext” package, the “tidyr” package, and other packages that handle a number of smaller purposes.

These twitter archives will be analyzed for statistics such as: how many followers they have, how many friends, or the concept of following a user that also follows you, they have, and other information obtained from their public profile. Their individual tweets will be loaded into containers and analyzed by listing their most commonly used words or phrases. These tweets will not be scrubbed clean of common stop words, such as “the,” “a,” and “or.” This is due to the limited pool of data being pulled at just 20 tweets. Removing these common stop words would almost eliminate most words being posted. Once these lists are built, they will be loaded into multiple graphs to search out common patterns. Additionally, these accounts can all be loaded into multiple charts tracking word frequency to compare how often popular words are being repeated across each account. Ideally, a pattern would begin to show up with key, identifying phrases that could be used to more easily identify spam bots in an automated manner. More likely than this possibility will be the idea that most individuals creating spam bots will already be avoiding the use of these phrases. As a result, common typing patterns will

have to be compared, such as the overuse of mentioning other users, commonly grouped hashtags, or other possibilities of longer phrases with swapped words.

After identifying common words, phrases, or typing patterns, the next section of the methodology involves analyzing the twitter archives of five individuals who are easily identified as real. Their commons words, phrases, and typing patterns will be compared against that of the bots in order to see the potential overlap where an automatic algorithm might let spam bots slip through the cracks. The process stated above for comparing the tweets of each bot will be used here. Each real individual will be compared against each bot in order to ascertain how often their most popular words or phrases may overlap. Additionally, their profile information such as number of followed accounts and followers will be compared as well to see if there's any sort of correlation. The selection of these five users will be based on their different areas of interest, as well as their locations in the country. By attempting to spread out the location of the real individuals while still sticking to American users, there's a potential chance for mixing dialects and increasing the strength of the collected data. All users and spam bot accounts will be taken from English-speaking accounts that appear to be based out of America in order to keep the collected data as clean as possible.

Assumptions

The most major assumption being made is that a spam bot account can be easily identified simply from looking at it. This would require that spam bots meet the criteria of having repeated or incredibly similar posts, or there is clear evidence of a record of automated harassment of users. This is crucial to the experiment, as the concept of using verifiably real accounts for the spam bot comparison would completely ruin the data. In line with this thought, making sure that all individuals running the non-spam accounts are those that live in America is important, as obtaining the accounts of people living abroad in other countries could feasibly offset data comparison due to the differences in dialects or languages. Finally, one assumption that is fairly safe to make is that Twitter will be functional and available during the period where data is being collected. Without access to Twitter, this experiment falls apart as there would be no way to gather information in the first place.

Experimental Design

The first step to this experiment, before collecting any data, was to locate and select six spam bots to be studied. For this experiment, I selected the following users based upon the appearance of seemingly automated posts on their timelines: “HotelsScanner” for a spam bot relating to traveling, “Make_money88” for a spam bot relating to monetary scams, “LootTootGames” for a spam bot related to video games, “MerryKicks” for a bot related to fashion, and “BestReviewsFree” and “strader55” for two bots both related to brand boosting. In order to see useful data in the comparison between the fifth and sixth bot, ensuring that they were similar in the nature of their account’s content was vital.

After locating spam bots, the next step was to collect their information using the “twitteR” package in R. Each user’s info was loaded into separate containers through the use of the “getUser” function, which detailed basic info, such as their follower and following counts. Each spam bots follower and following counts were recorded, making special note of the results of the two bots based around brand boosting. From there, the same process of recording was repeated for each bot’s following count. Following that, the most recent 20 posts that each spam bot had favorited was collected and analyzed for frequent words or phrases and the account to which each post belonged. Finally, the most recent 20 tweets from each of their own timelines was collected and analyzed for frequent words or phrases. The reason for selecting only 20 posts stems from the desire to identify bots based off limited amounts of information. The ability to determine whether or not an account exhibits behavior of a spam bot based off of hundreds or thousands of tweets seem ineffective in the long run, especially considering how numerous spam bot accounts are.

The next step was to locate five real individuals and compare them to the data retrieved from the first five spam bots. For this section of the experiment, I selected the following users based upon the appearance of seemingly real individuals living in America: “TheRock” for his status as a celebrity,

“xD1x” for his job and hobby involving video games, “Marc_Perrone” for his use of twitter as a professional platform to inform individuals on financial matters, “MalMaiMar” for his status as a Webster student, and “gloomyhome” for his interest and focus on digital art. The same processes used for the spam bots in order to record account information, follower counts, and following counts were also used for each of the aforementioned individuals. The information collected in this way was compared to that of the spam bots to ascertain if any similar patterns in account information or typing patterns arose.

Results and Discussion

The initial results pulled back from the basic details of the spam bot accounts are very inconclusive. While certain spam bot accounts, such as the “BestReviewsFree” and “MerryKicks” account have only a middling number of followers (13 and 691, respectively), accounts such as the “LootTootGames” account has 241,094 followers, as seen in Figure 1.

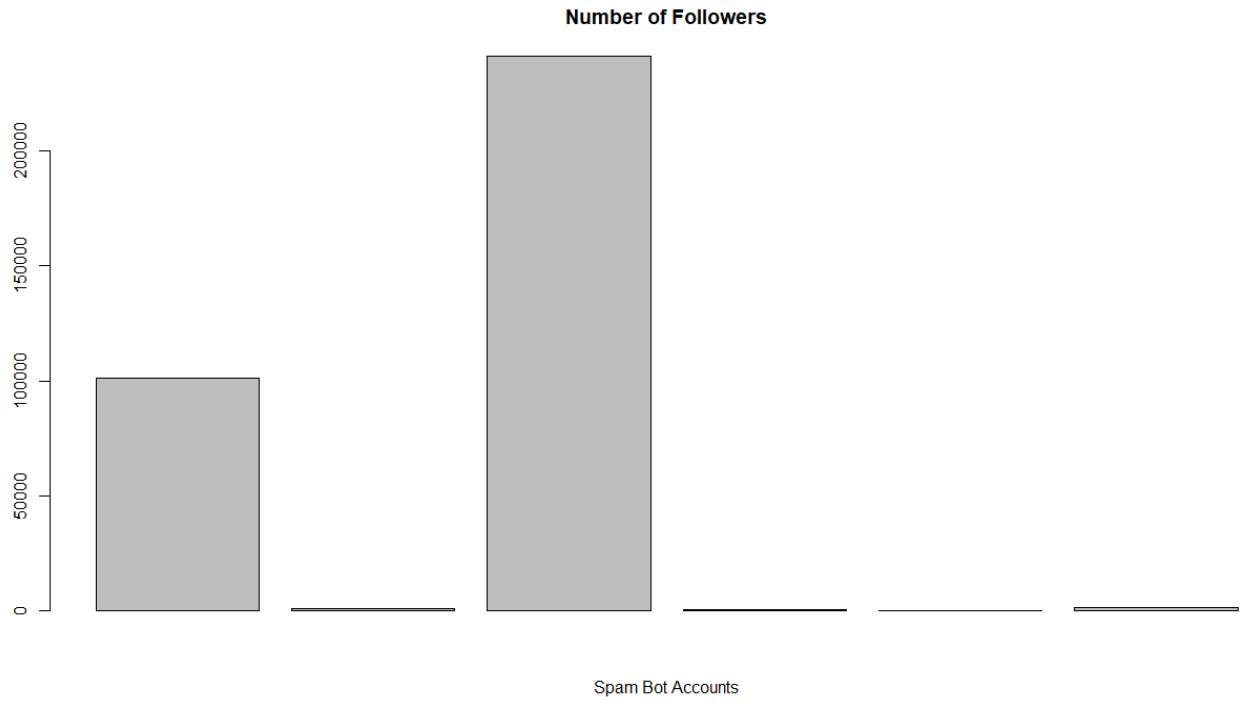

Figure 1. Number of Followers for Each Spam Bot Account

There doesn’t appear to be any consistent pattern to be identified based on their follower numbers, and this can also be seen in their following numbers. These numbers overall seem to be lower on average, usually topping out around a few thousand, but even then there are still wildly offset following numbers, such as that of the “HotelsScanner” account which is following 59,931 different accounts, seen in Figure 2.

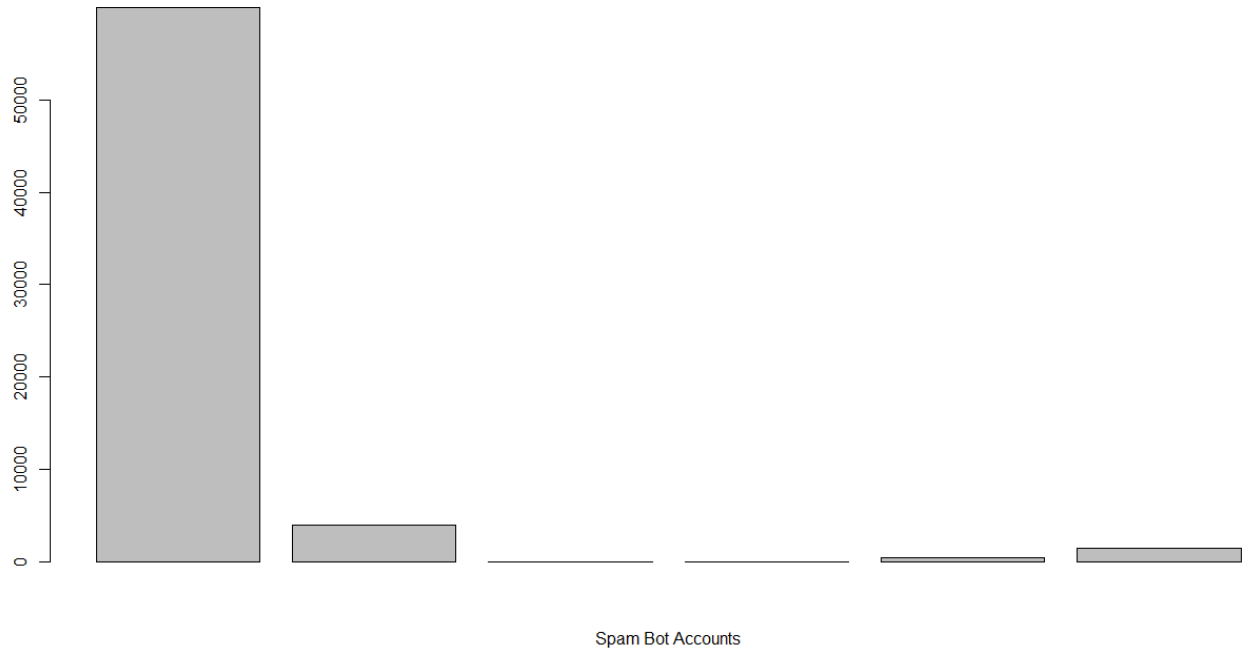

Figure 2. Number of Accounts Followed by Each Spam Bot Account

Each account's favorites proved more fruitful in terms of potential pattern recognition. For instance, many of the bots favorite the posts of other spam bot accounts, presumably run by the owner or programmer of the spam bot accounts in question. Two of the accounts, "HotelsScanner" and "BestReviewsFree," continuously favorited many of their own tweets, among other tweets that were similar to their own. Other spam bots such as the "Make_money88" and "LootTootGames" had favorited many tweets within their areas of interest, such as tweets advertising "free follows," the concept of following an account for them to follow you back, and tweets from gaming-centric companies like IGN. However, the "strader55" account had no favorites whatsoever, meaning that functionality for automatically favoriting posts was not coded into the twitter bot. These results are seen in Figure 3 through Figure 7.

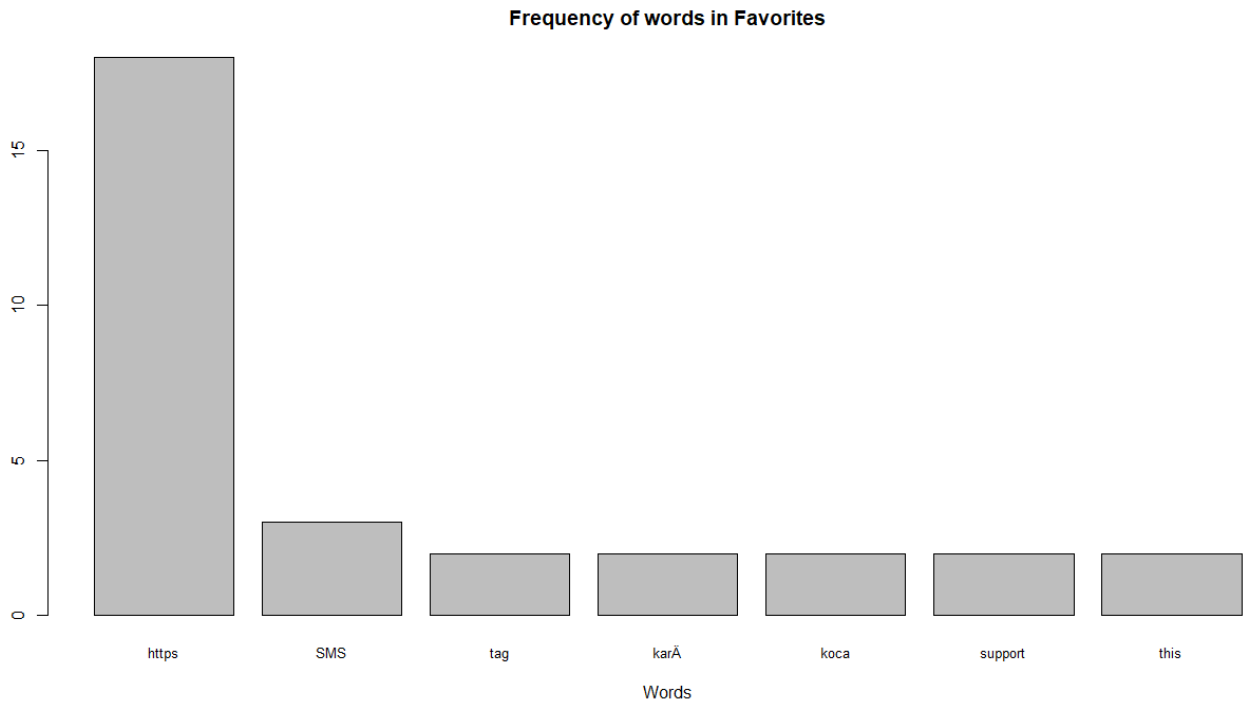

Figure 3. Word Frequency in Favorited Tweets of the "BestReviewFree" Account

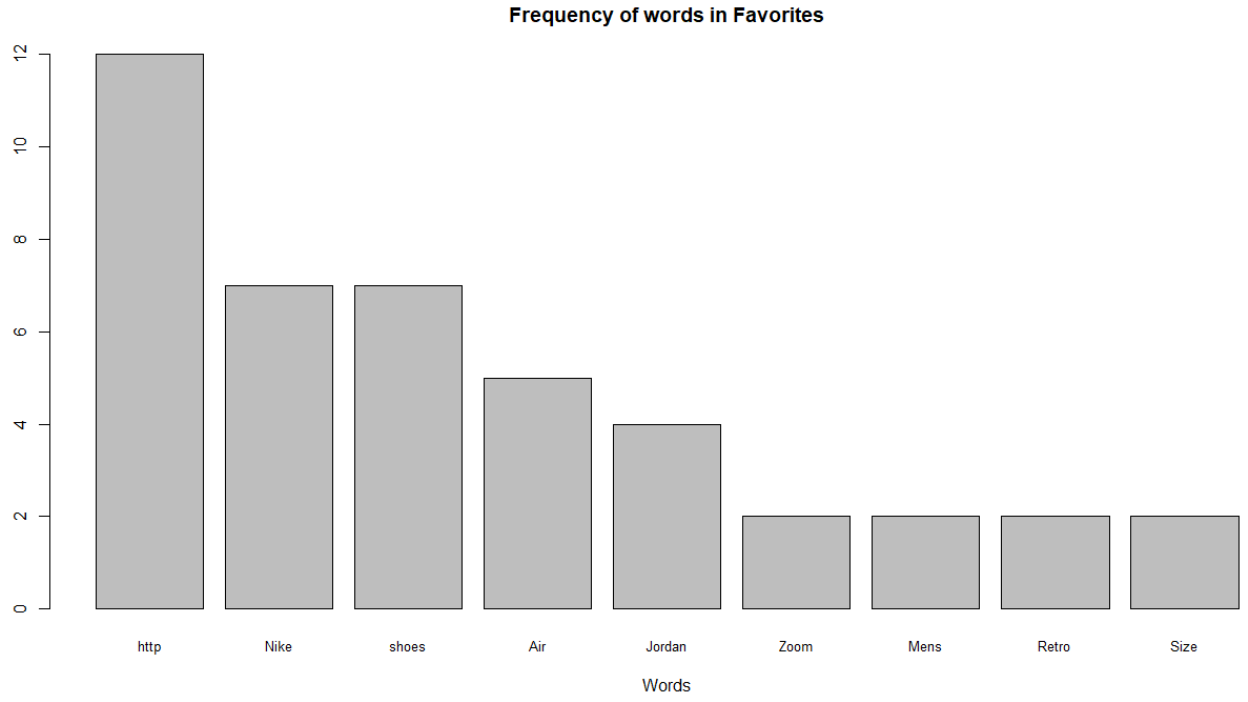

Figure 4. Word Frequency in Favorited Tweets of the "MerryKicks" Account

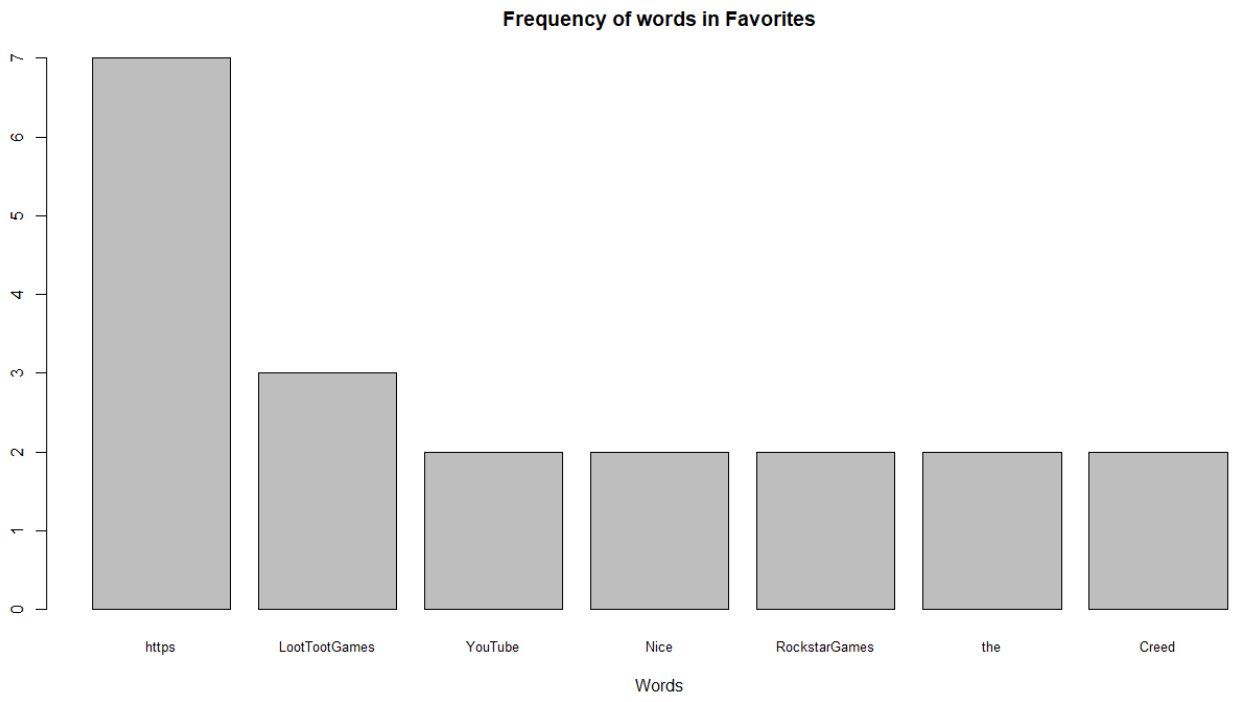

Figure 5. Word Frequency in Favorited Tweets of the "LootTootGames" Account

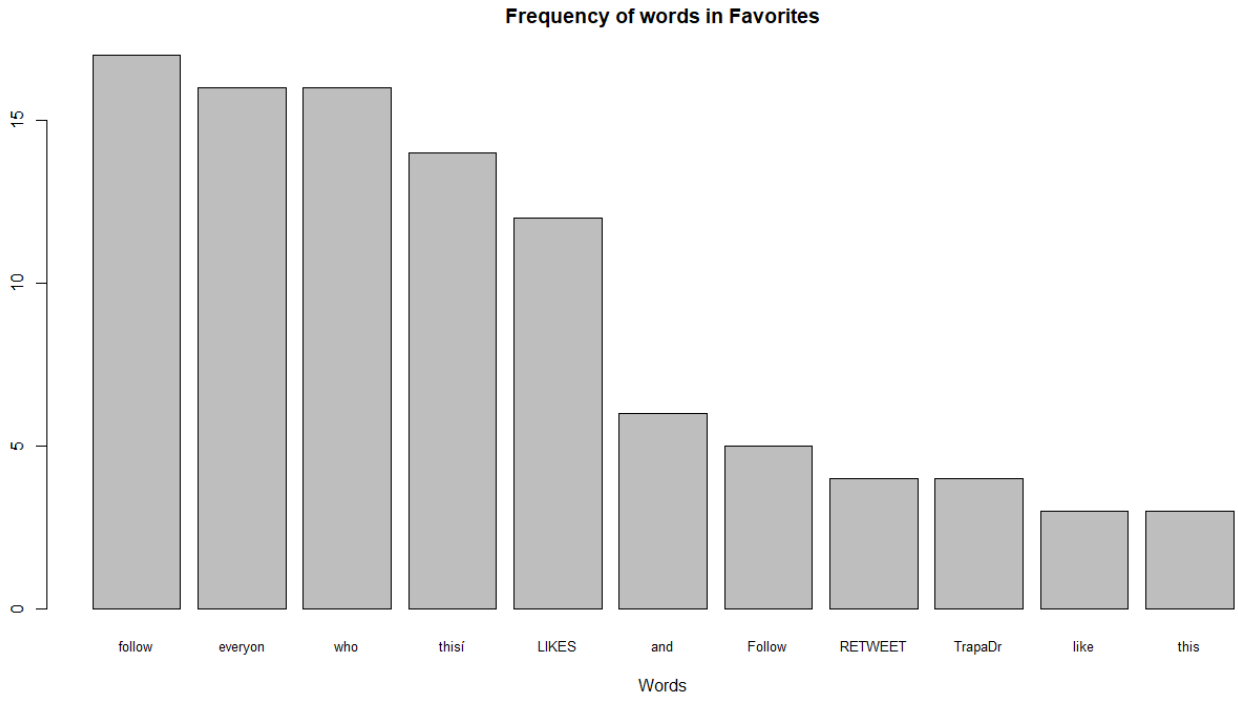

Figure 6. Word Frequency in Favorited Tweets of the "Make_money88" Account

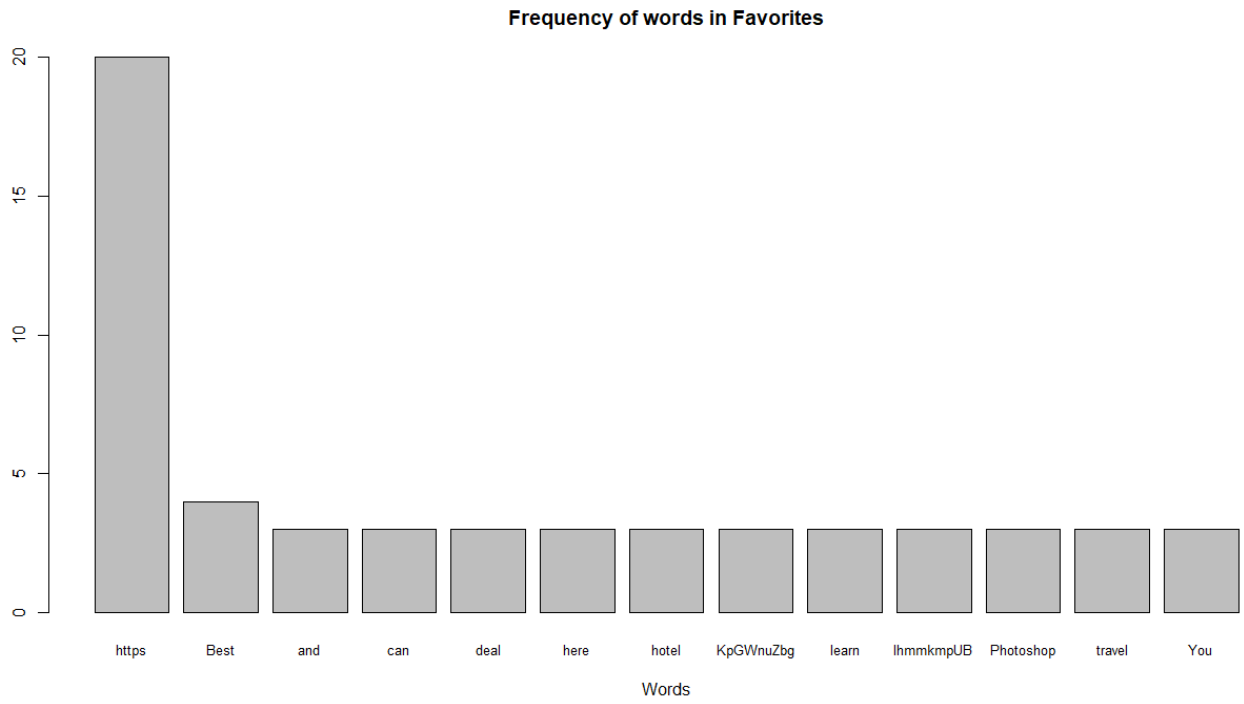

Figure 7. Word Frequency in Favorited Tweets of the "HotelsScanner" Account

The timelines from each spam bot account have even more in common, as most of them incorporate posting links out to external websites and making use of words such as “free,” “easy,” or “cheap.” Most tweets are also relatively short, often coming in well underneath 140 characters. These common words are graphed in Figure 8 through Figure 13.

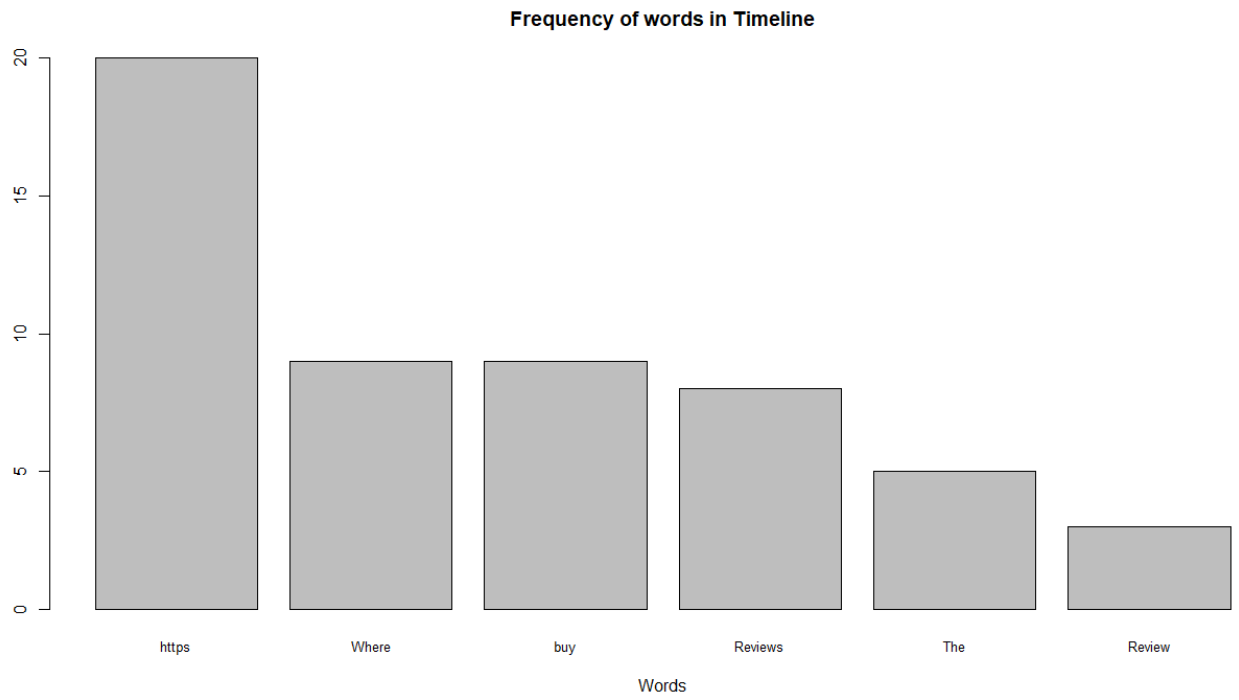

Figure 8. Word Frequency in Posted Tweets of the "BestReviewsFree" Account

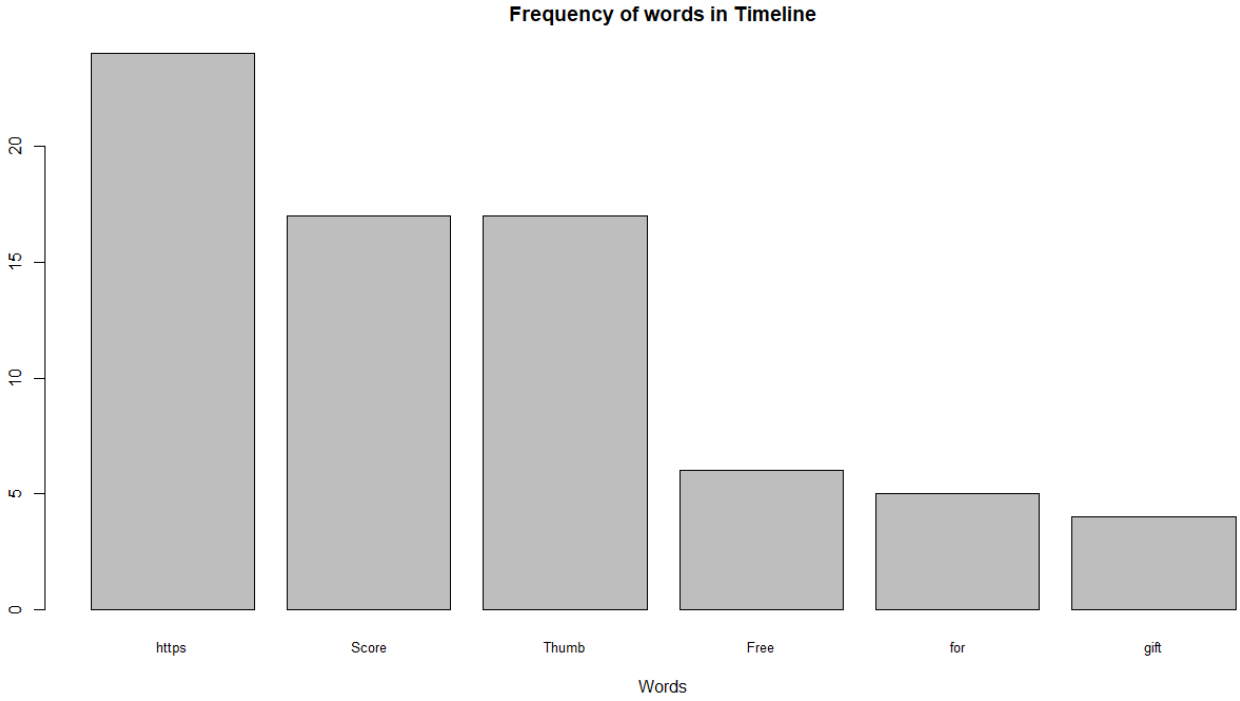

Figure 9. Word Frequency in Posted Tweets of the "strader55" Account

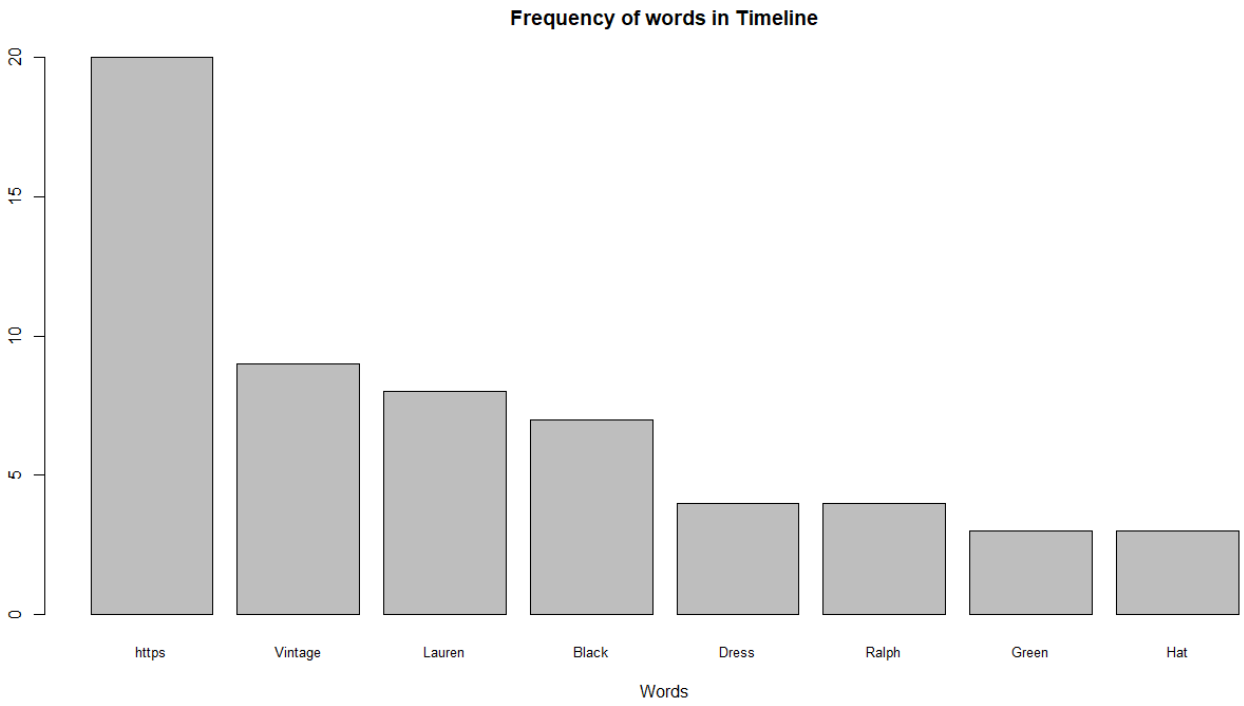

Figure 10. Word Frequency in Posted Tweets of the "MerryKicks" Account

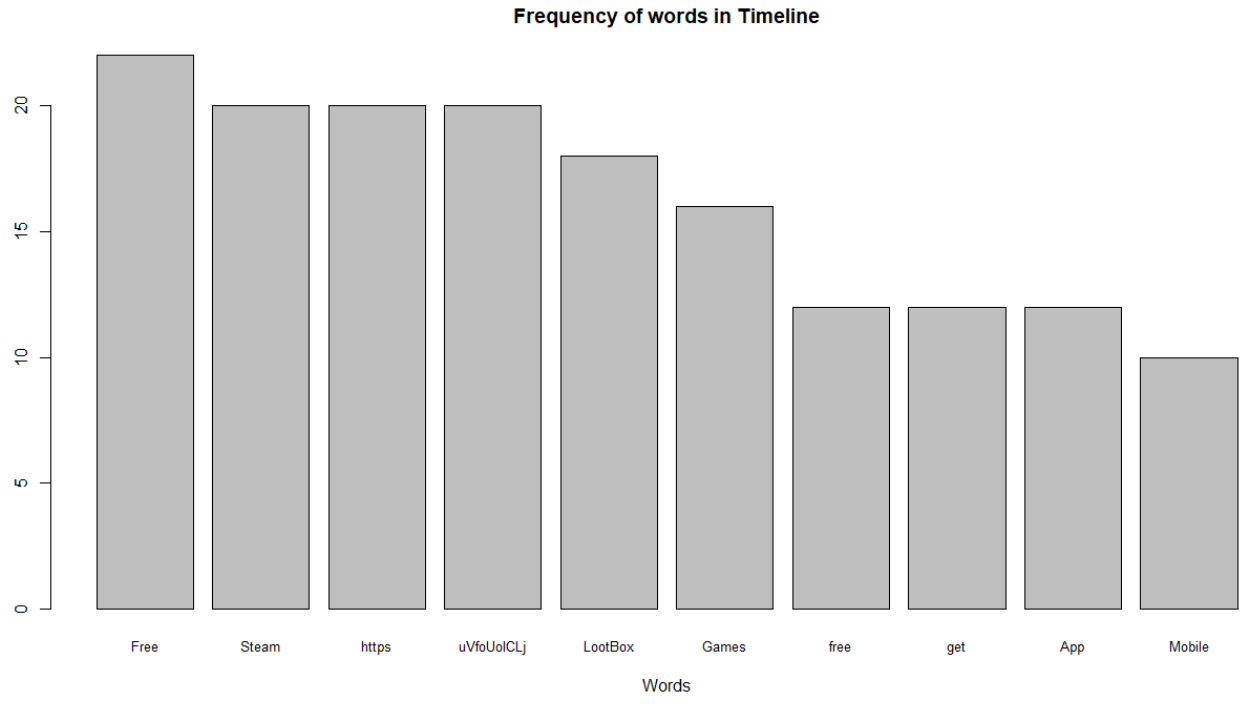

Figure 11. Word Frequency in Posted Tweets of the "LootTootGames" Account

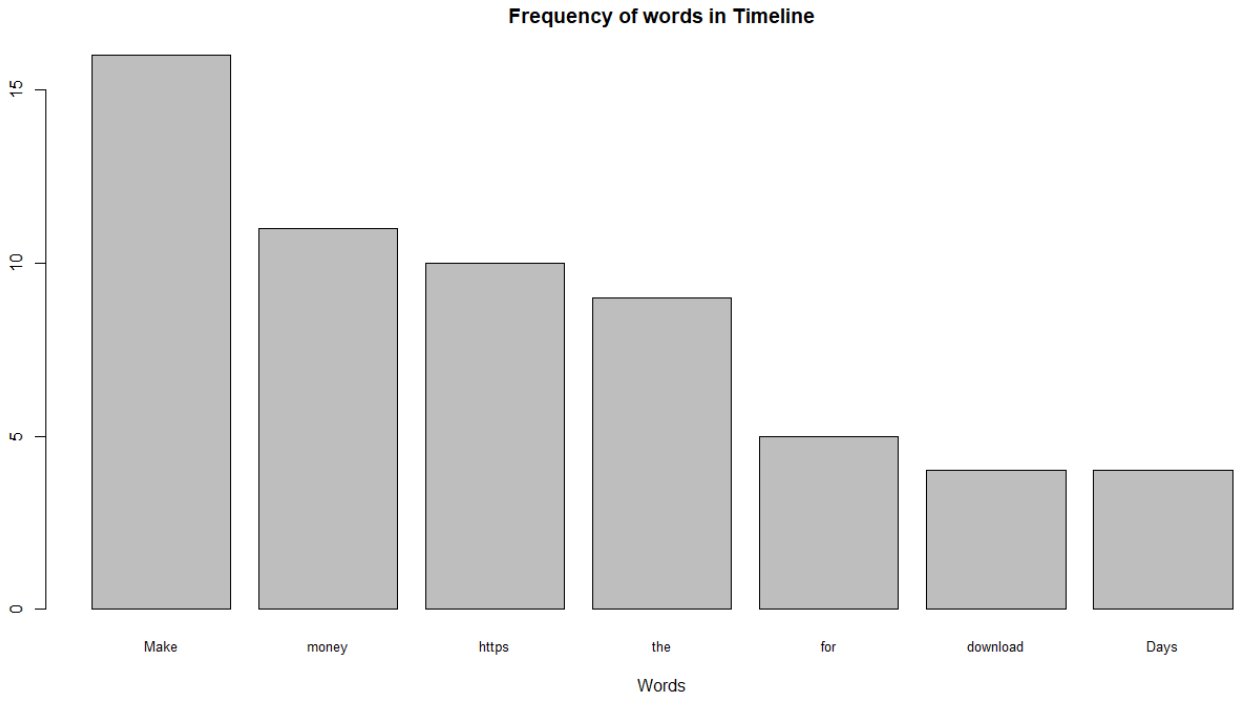

Figure 12. Word Frequency in Posted Tweets of the "Make_money88" Account

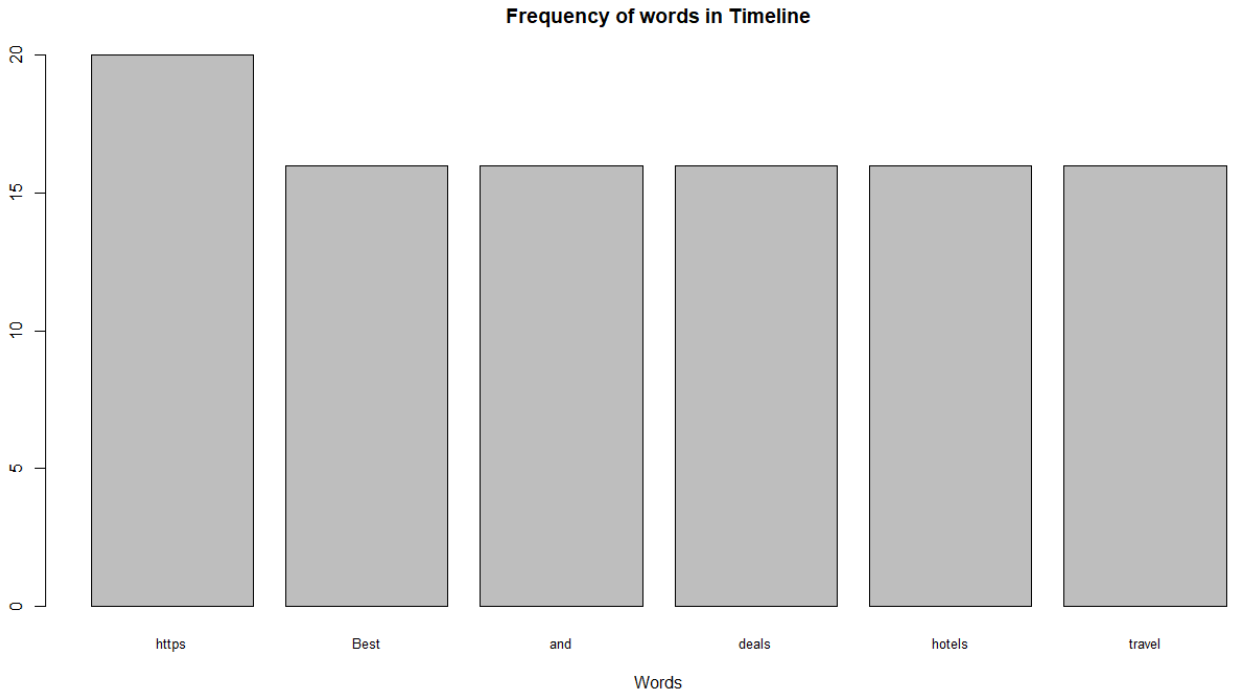

Figure 13. Word Frequency in Posted Tweets of the "HotelsScanner" Account

These comparisons become more obvious when viewed between the two spam bot accounts from similar areas of interest, “strader55” and “BestReviewsFree.”

The two brand boosting spam bot accounts pointed out unique contrasts and comparisons that get somewhat muddled when looking at the sheers numbers of the other aforementioned spam bot accounts. For instance, their follower and following counts are not too similar or proportional (seen in Figure 14 and Figure 15), and their favorited tweets also share nothing in common since “strader55” had no favorited tweets.

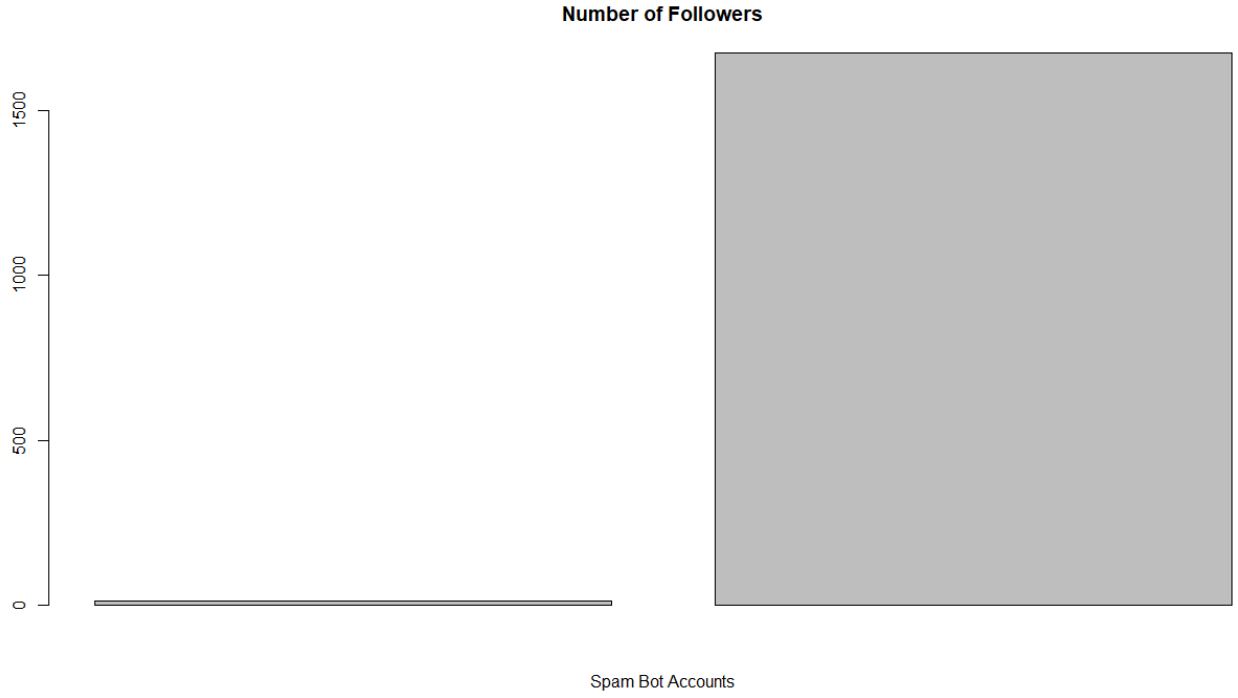

Figure 14. Number of Followers for the "BestReviewsFree" and "strader55" Spam Bot Accounts

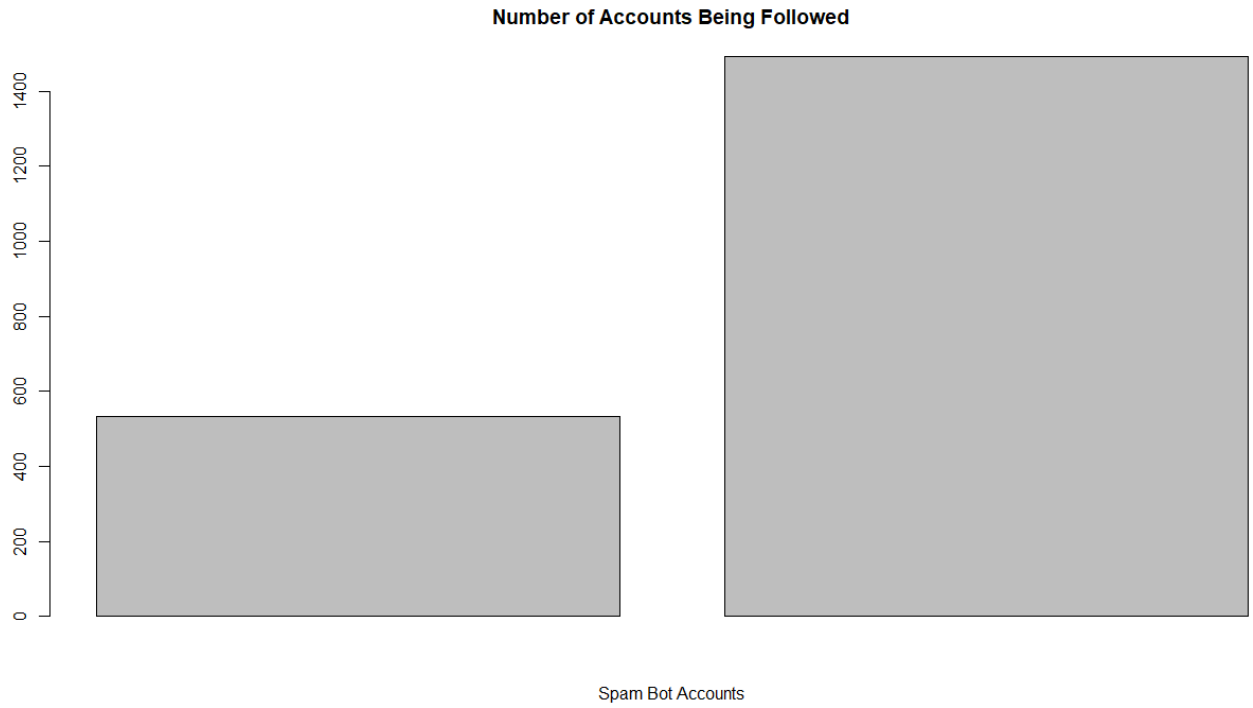

Figure 15. Number of Followed Accounts for the "BestReviewsFree" and "strader55" Spam Bot Accounts

However, the tweets they posted to their user timelines had two major concepts in common, posting links to external websites and having many frequently repeated words like “free” or “best.” While their frequent words had no overlap, the typing pattern of each one appears similar in nature, specifically the pattern of discussing a certain object or product and then attaching a link at the end of each tweet. However, when this behavior is compared to that of real individuals, there are some striking similarities to be seen.

The initial results for each real individual’s account is fairly varied, similar to that of the bots. For instance, “The_Rock” had nearly 13 million followers due to his status as a celebrity, where every other individual did not even have 200,000 followers, seen below in Figure 16.

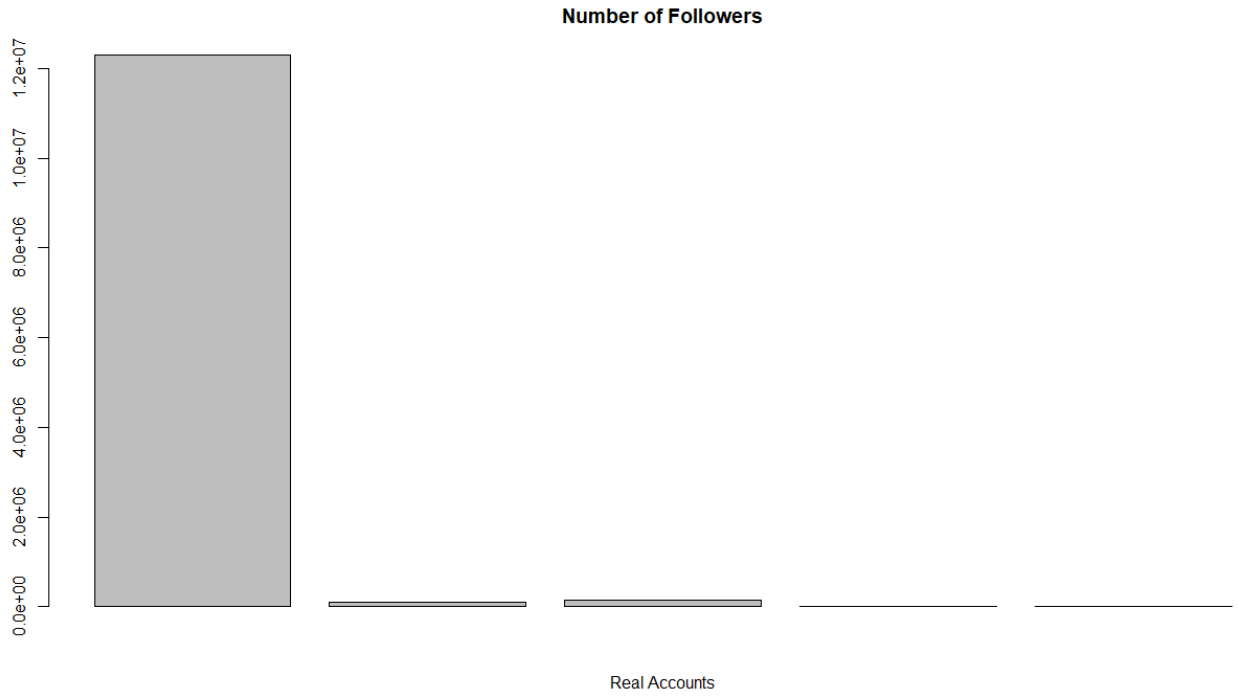

Figure 16. Number of Followers for Each Real Individual

On the flip side of that, “The_Rock” had the smallest following count of all the real individuals at just 178 accounts followed. The highest following count amongst the real individuals was the “Marc_Perrone” account, seemingly due to a desire to spread his twitter brand by following large masses of accounts, seen in Figure 17.

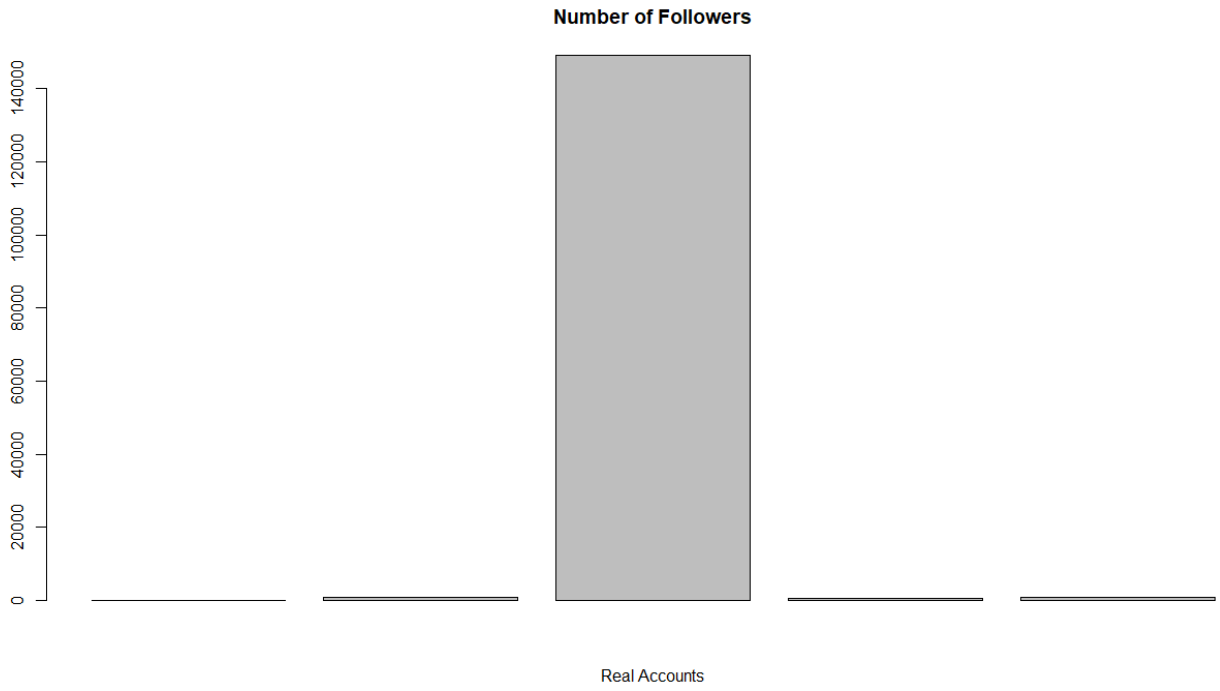

Figure 17. Number of Followed Accounts for Each Real Individual

This info appears to have no indication of any real person's status due to the wide variety of followers or followed accounts. There are no connections to be made among their favorited tweets, as every real individual had a widespread numbers of interests and followed accounts from which these tweets were pulled. The word frequency of these favorites can be seen in Figure 18 through Figure 22.

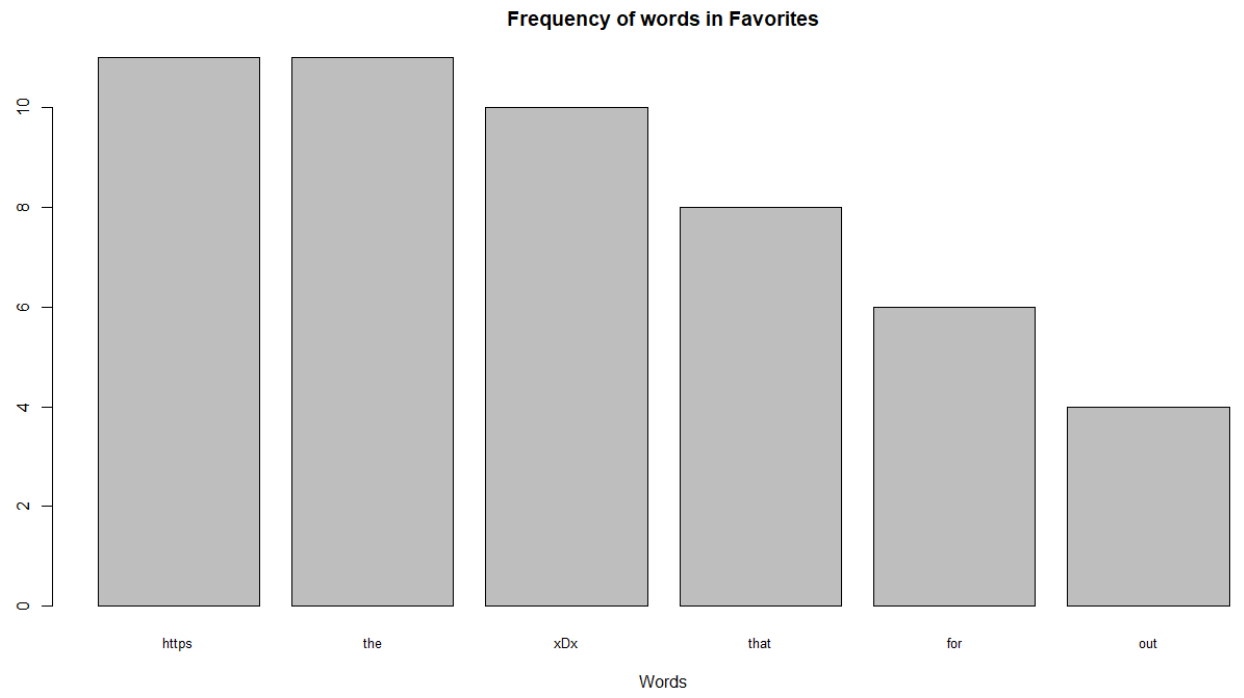

Figure 18. Word Frequency in Favorited Tweets of the "xD1x" Account

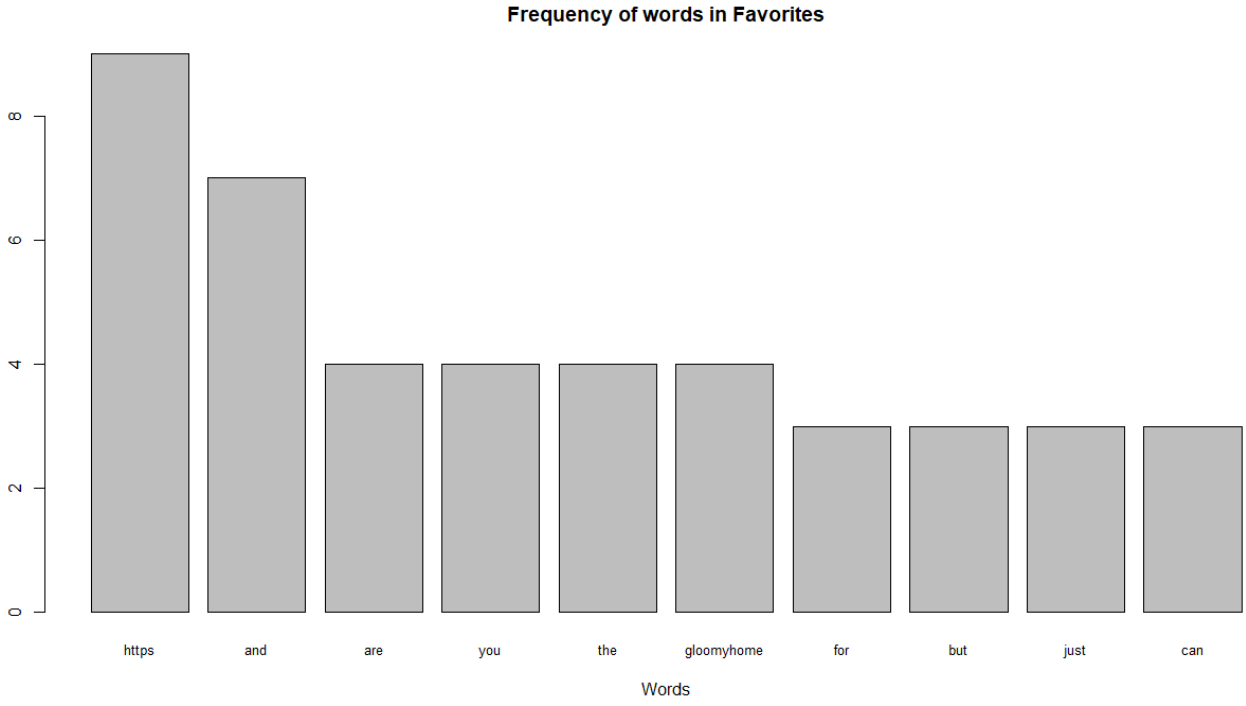

Figure 19. Word Frequency in Favorited Tweets of the "gloomyhome" Account

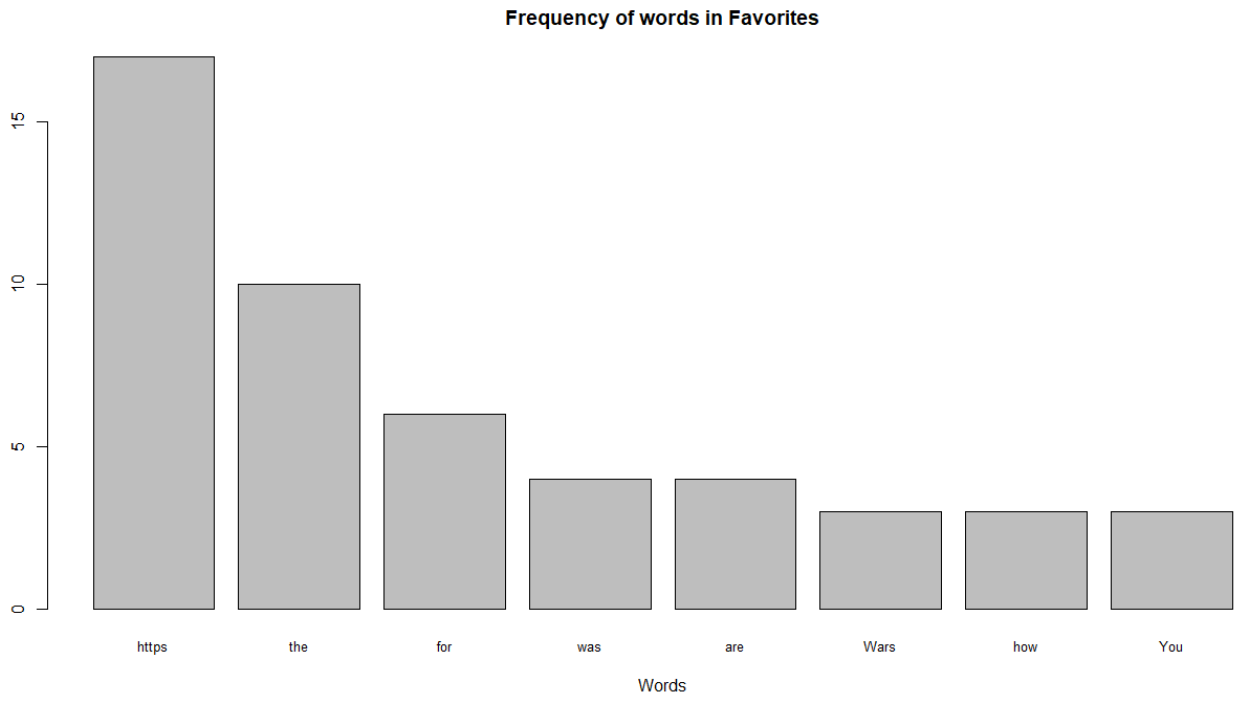

Figure 20. Word Frequency in Favorited Tweets of the "MalMaiMar" Account

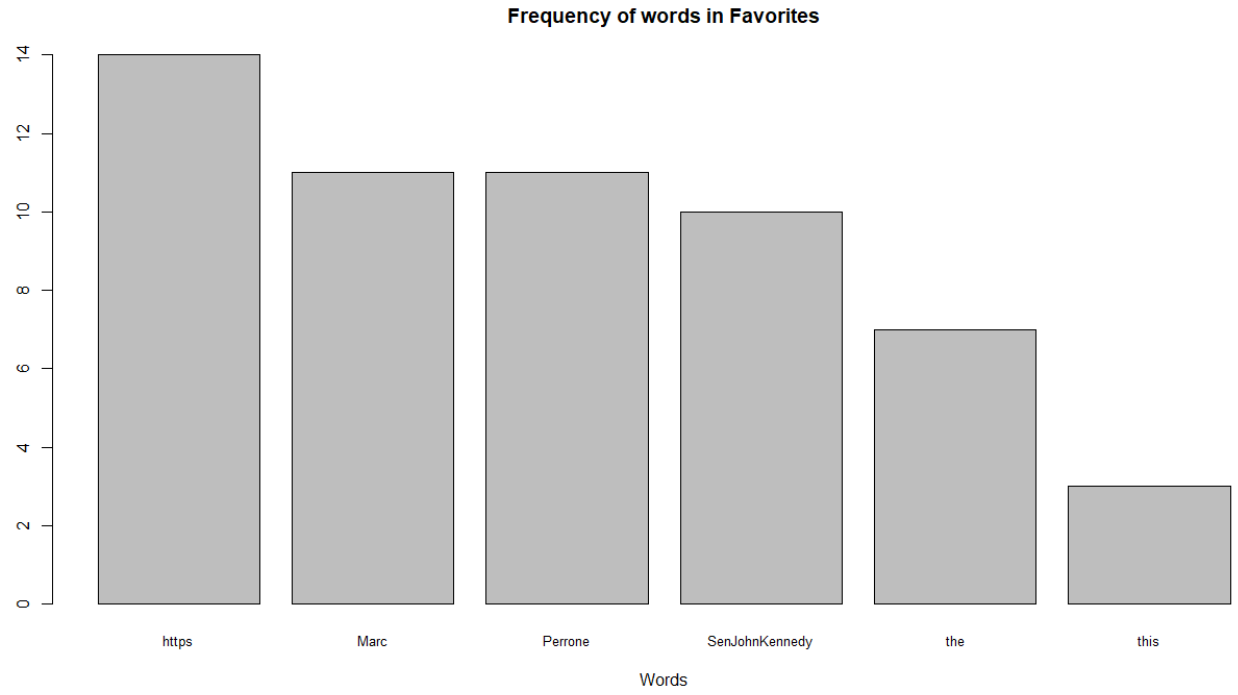

Figure 21. Word Frequency in Favorited Tweets of the "Marc_Perrone" Account

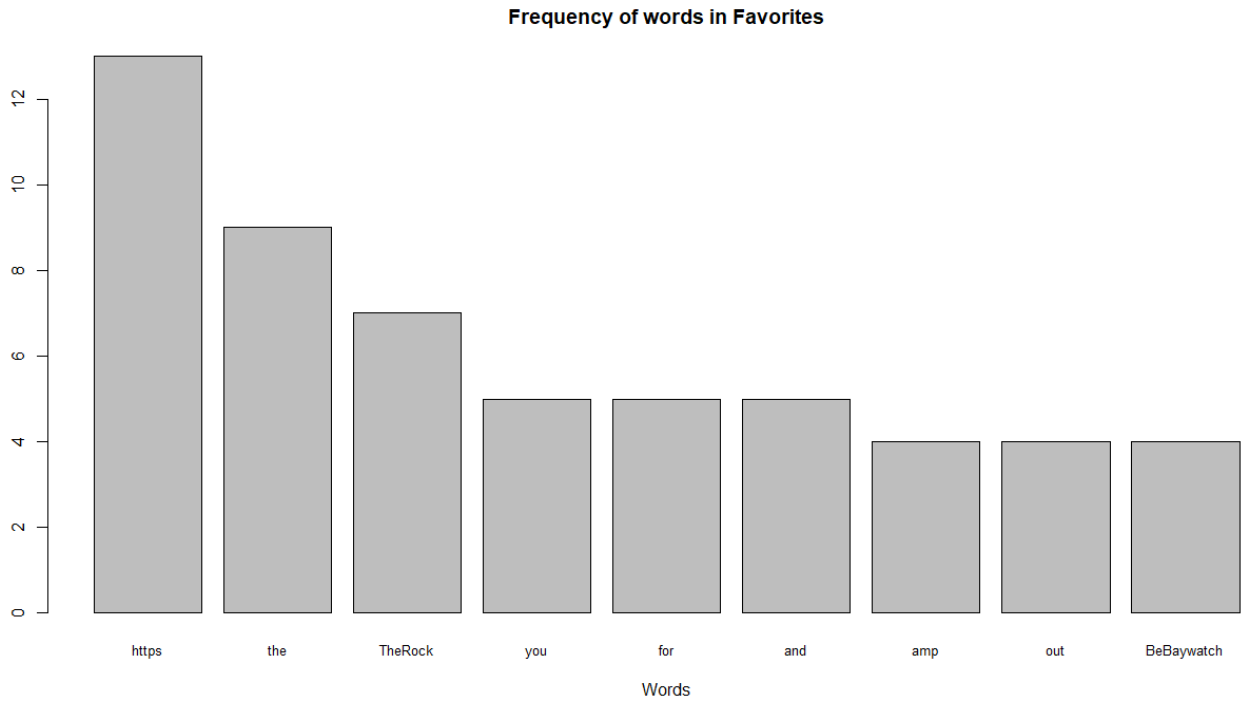

Figure 22. Word Frequency in Favorited Tweets of the "The_Rock" Account

However, their tweets reveal much more, such as a lower frequency of common words, especially when compared together, seen in Figure 23 through Figure 27.

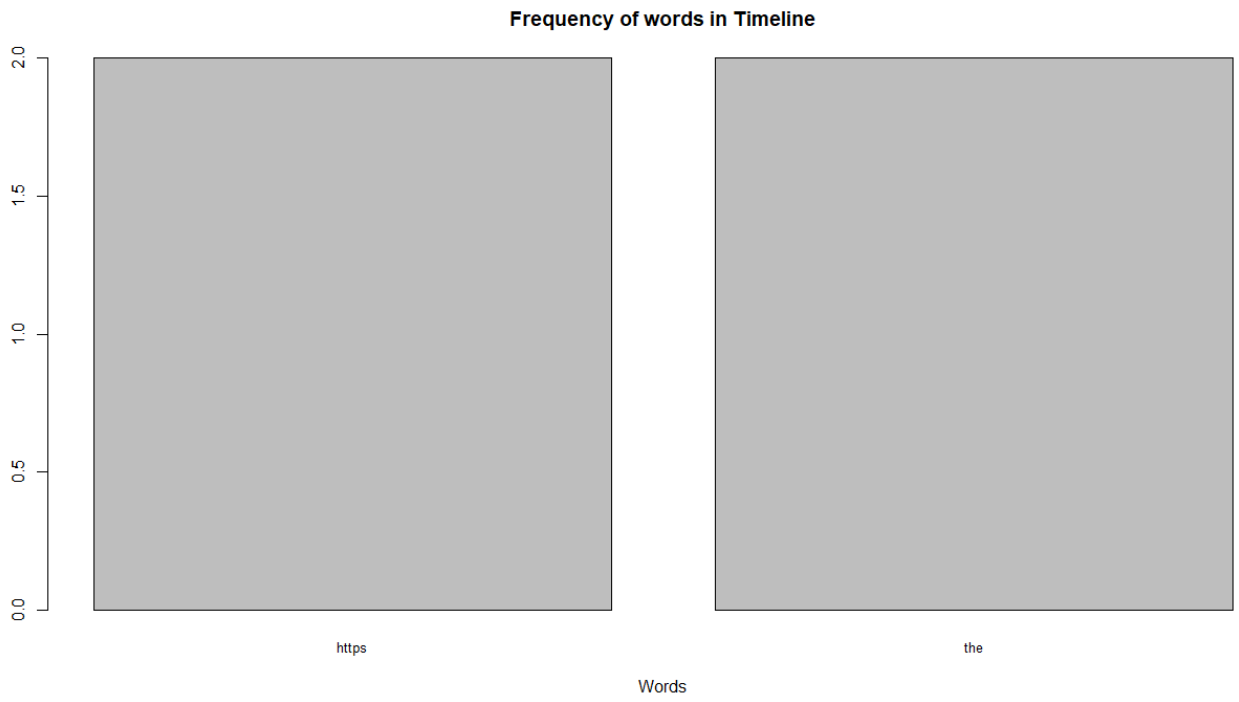

Figure 23. Word Frequency in Posted Tweets of the "xD1x" Account

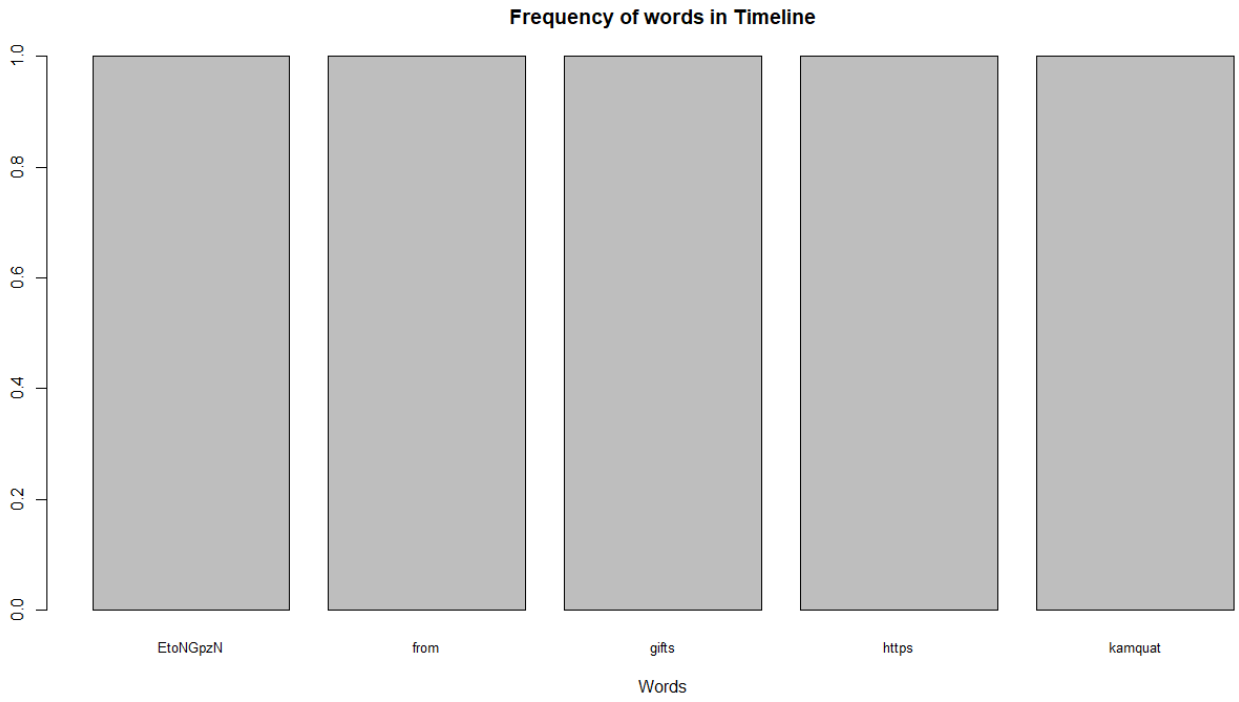

Figure 24. Word Frequency in Posted Tweets of the "gloomyhome" Account

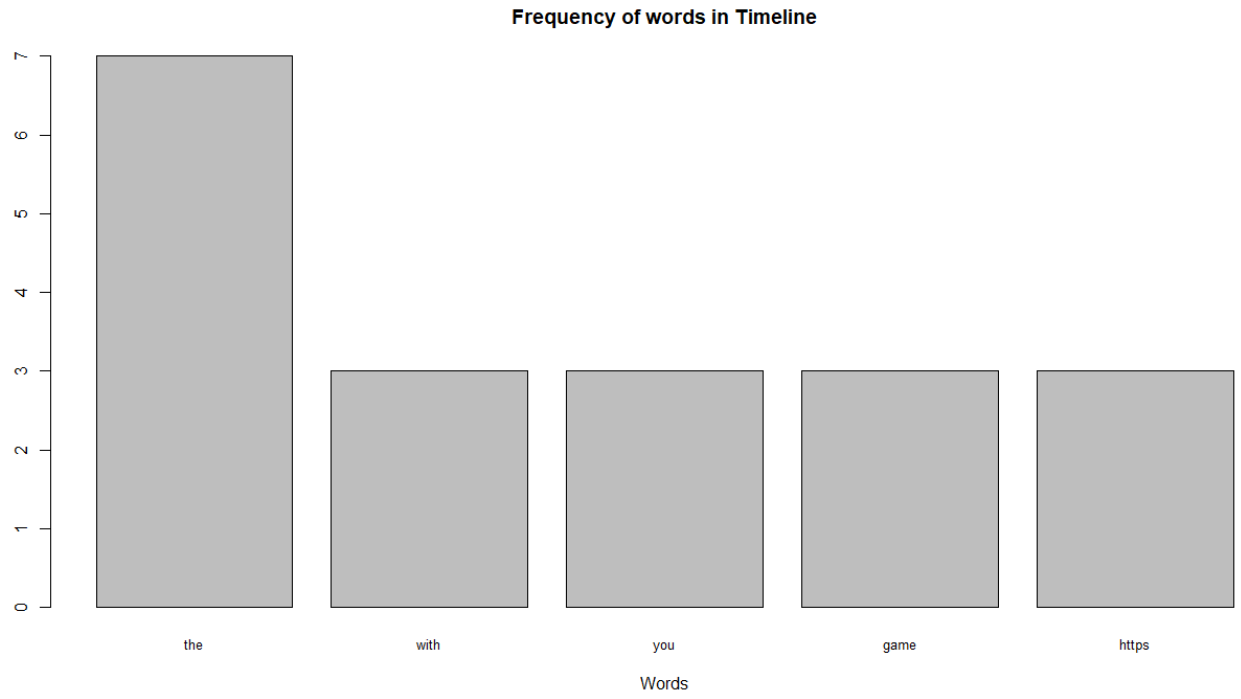

Figure 25. Word Frequency in Posted Tweets of the "MalMaiMar" Account

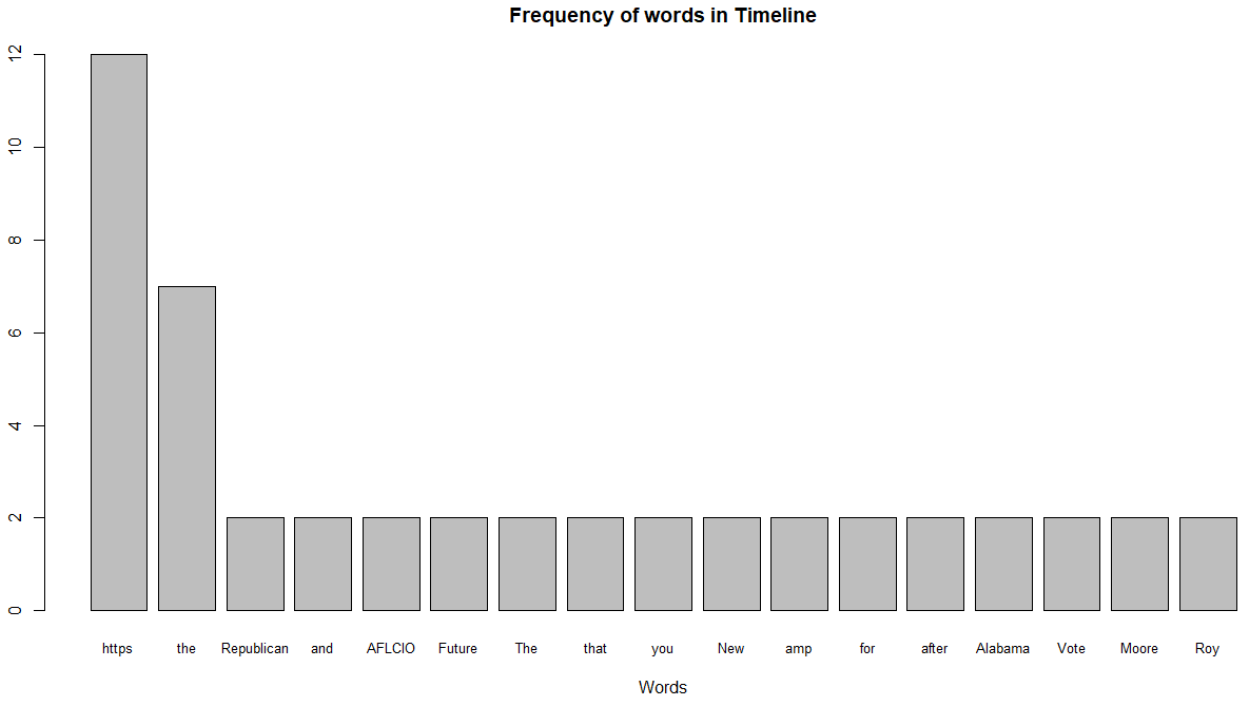

Figure 26. Word Frequency in Posted Tweets of the "Marc_Perrone" Account

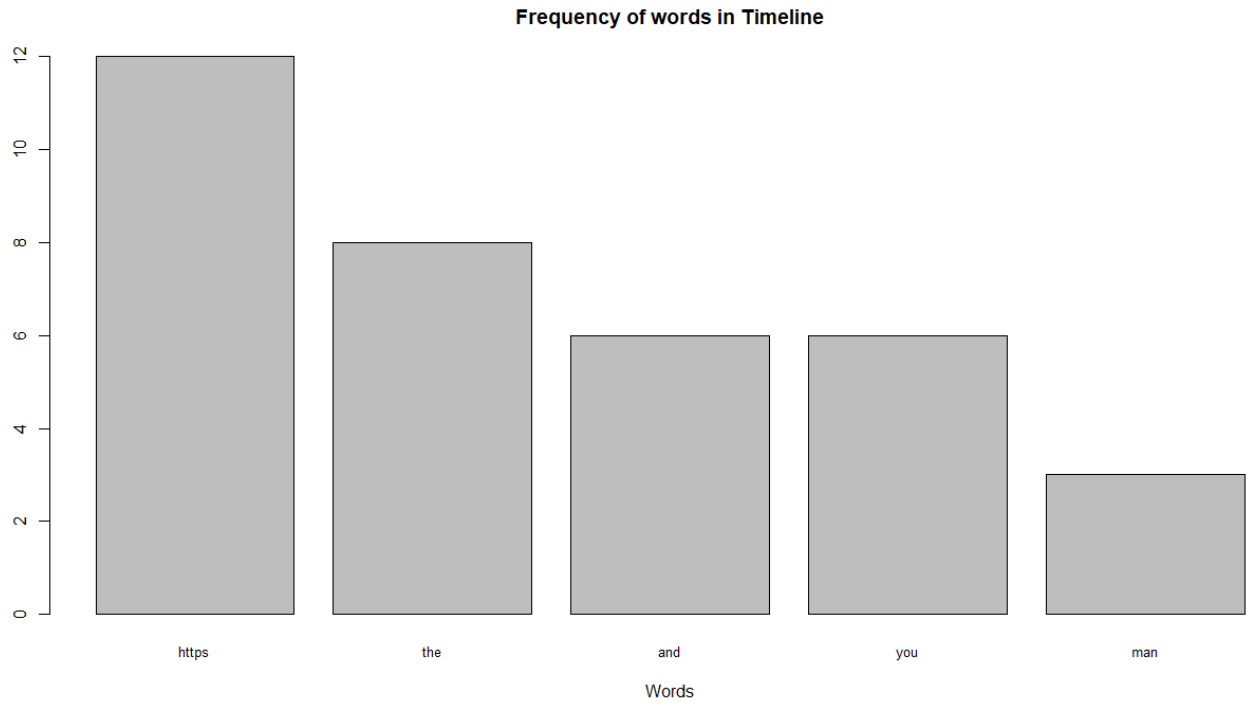

Figure 27. Word Frequency in Posted Tweets of the "The_Rock" Account

Of the five accounts studied, there were no commonly shared words besides that of common stop words. There was, however, a large amount of links out to external websites coming from each account. The frequency was lower than that of the spam bot accounts, but even then it was still present in a large majority of each user's tweets.

Issues

One major issue encountered was the lack of original tweets made by specific real users. When working with a purposefully limited data set, it also led to limited sets of results. For instance, the "gloomyhome" account had only a single tweet made by himself due to a large amount of his interactions with twitter relating to replying or retweeting others. This caused small issues when comparing the tweets of real users to individuals, but also offered a valuable point of comparison in distinguishing real users from bots by showcasing just how little content a real user could have. Additionally, there was potential for the "The_Rock" account to have shifted in numbers during the run time of the R script I created, somewhat altering the final outcome of the comparison. However, due to the nature of his account having as much attention as it does, any sort of additional followers seen since the beginning of the script running would have most likely been negligible. Finally, in the context of attempting to limit data to identify recently created spam bots, the age of these accounts differing is something that could have potentially thrown off data. However, when considering that it is difficult to find recently created accounts specifically because they are new; it becomes a challenge to find custom fit data for the problem. This can also be covered by the fact that many bot creators set out to perform email verification and other basic activities to legitimize their spam bot accounts, thereby bypassing many early determent methods that Twitter already implements.

Conclusions and Future Work

The largest difference between the spam bot accounts and the real individuals was that the spam bots consistently had links out to external sites and the potential to share frequently used words. These spam bots are built around actively advertising different websites or products and the best way to promote is to be constantly attaching links to everything they post. Real individuals generally do not do this, save for the “The_Rock” account which utilizes many links to promote his personal brand, but even he, or the individual managing his account, does not attach a link to every single tweet. Additionally, the use of words such as “free” or “cheap” spread across many different tweets appears as equally suspicious. When working with attempting to identify spam bots, especially those that may have just been created, these two facets should be utilized to create a predictive algorithm of identifying and possibly removing spam bots. However, this kind of concept comes with risks, especially if removal of these accounts is the main goal. Many individuals could potentially have their real accounts flagged or deleted by accident, potentially leading to upset among the userbase at large. Additionally, there are many useful spam bot accounts which could be removed in the process, such as accounts that post about weather warnings or shopping notifications. Automatic removal on a site as popular and large as Twitter is dangerous, but identifying and giving human review to every single flagged account is too slow for the pace that the internet has the potential to move at. This sort of delicate balancing act is hard and should be done with more data to obtain more accurate results if an automated approach to preventing spam is to be taken.

References

Newberg, M. (2017, March 10). As many as 48 million Twitter accounts aren't people, says study. Retrieved December 5, 2017, from https://www.cnbc.com/2017/03/10/nearly-48-million-twitter-accounts-could-be-bots-says-study.html

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017, March 27). Online Human-Bot Interactions: Detection, Estimation, and Characterization. Retrieved December 5, 2017, from https://arxiv.org/pdf/1703.03107.pdf