Interesting Data Analysis based on IAmA Report

Cheng Cui
ccui7@binghamton.edu
Binghamton University
Binghamton, NY

Tianze Liu
tliu76@binghamton.edu
Binghamton University
Binghamton, NY

Mingjie Yan
myan28@binghamton.edu
Binghamton University
Binghamton, NY

ABSTRACT

In this project, we analyzed subreddit r/IAmA and twitter streaming data that collected by our previous project. We use language analysis APIs provided by IBM and Microsoft as well as statistical methods to analyze the data we collected. From subreddit r/IAmA itself to the link between r/IAmA and Twitter, we do find some interesting result.

1 INTRODUCTION

The objectives we intended to achieved can be categorized in the following questions as we proposed before, which are : 1. What are the categories or backgrounds of IAmA hosts? 2. What are the most popular topics and words of highest frequency on IAmA? 3. During the time of data collection, in which time slot there are most users involved as the question is raised to the host? 4. Whether there is a connection between IAmA and twitter?. To answer these questions, We implemented our solution using natural language understanding tools and plotting libraries, filtering and converting the data we needed and visualizing the results in different types of figures which will be illustrated in detail in the following sections. Finally we do get some interesting facts and conclusions.

2 BACKGROUND AND RELATED WORK

r/IAmA[1] is a subbreddit created at May 27, 2009 with 20.6M total members ranged from various celebrities to everyday people in several lines of work.

Some websites analyze the upvotes traffic or IAmA such as "Delay for Reddit"[2] to get the best time to post. Inspired by this idea, we decided to analyze the comments time traffic (question 3) and to know when is the busiest that most users are involved in discussion which is maybe also the best time for the potential host to start an Q&A on IAmA.

3 DATASETS

3.1 Reddit

From 10/21/2020 till 11/27/2020, we get 162 IAmA posts and 30 IAmA crossposts with 17,989 comments in total. Also shows in Table 1. Some crossposts are earlier than October 21, that is to make the final number of crossposts relatively high, if removed that part of the data there will be only 14 records. For each IAmA post, we collect this post's title, main body text, create time, score, total comments, IAmA category, post id, subreddit prefix and author's username. IAmA allow crosspost, which means that the real AMA discussion takes place in another reddit board. For each crosspost, we use same data structure as IAmA post to store the information. For comments, we did not collect all the replies in comments, which means we only interest in the participants' original question to the host. According to subreddit r/IAmA rule, "Top-level comments must ask a question"[3], since we only interest in the original question to host, we just ignore the replies in comments.

Table 1: Reddit r/IAmA Database

Table Name	Collected Data
comments	17,989
CrossPosts	30
IAmA	162

3.2 Twitter

The collection of twitter data from 1%sample stream started from 11/10/2020 and ended at 11/25/2020 10:25 pm. Due to MongoDB run out of our VM's memory, we do not have the last one and half hour's data. We saved the data fields of each tweet including the tweet ID, hashtags, timestamp, author ID, tweet text, like count, retweet count, media keys and image urls. Table 2 shows the detail number.

Table 2: Twitter Database

Table Name	Collected Data
twitter_collection	18.8M
hashtags_collection	64.7M
twitter_timestamp	24.1M

4 EXPERIMENTAL SETUP

4.1 Categories of IAmA hosts

To answer this question, we use the title of each posts in r/IAmA and IBM Watson Natural Language Understanding(NLU) service[4]. r/IAmA require each host using "Iama..." as the format to introduce themselves[3]. We will feed this sentence to NLU service and get categories of the host. For each host we will get at least one category.

If the reliability of the returned category is less then 0.5, we will consider to use IAmA category which is set by IAmA moderators. During our experience, this situation have not happened yet. In another word, all our result in category is based on IBM Watson NUL service.

Another limit is that IBM Watson category is based on a pre-trained 5-level taxonomy hierarchy[5], sometimes a very accurate category can be returned, but in most cases only a three level category is returned, and in some extreme cases only a one level category. For the top four categories we will statistics second level categories, others will only focus on first level.

Implementation is in `Q1.py`, results are in `Q1result.txt`.

4.2 Most Frequent Words

In this part, we use the main body text of each post in r/IAmA and Microsoft Azure Cognitive Services[6]. r/IAmA require each host to provide some additional information of themselves and the discussion topics under this post[3]. We use this part of text to generate key words by using Azure cognitive services. After get all the key words, we statistic word frequency and then use wordcloud library[7] to generate the wordcloud plot.

The number of key words depends on the length of the text. During our experimental, all posts have over five key words, except one crosspost r/LockdownSkepticism/jvtpwz[8] has one single key word "question". Because the host only says: "Here to answer your questions!"

We also found that some of these posts are poisoning our keyword frequency statistics by repeat some words again and again. For example in r/IAmA/jmafy5[9], the host repeat F word up to 15 times, this post is the only source of F word in our database. Therefore, we added an additional indexing feature so that the same keyword in a post counts at most twice.

Implementation is in `Q2.py`, results are in `Q2result1.txt`. Here we have also generated the data needed in the question 4, stored in folder named "`keyword1`".

4.3 IAmA Commnents Traffic Analysis

In this section we collected the time stamps of all IAmA comments through 10/21/2020 to 11/27/2020 and do the analysis on which time range is the busiest time of the IAmA subreddit usually. We count all the sum of number of comments that is posed on each hour in weekdays and normalize the count by dividing the number of each weekday during this period.

For example, during our data collection, the most popular time slot of IAmA is during 18:00 to 19:00 on Thursday as shown in Figure 3, in which more than 200 comments on average are made. The brighter the color, the more average comments are made during that hourly time slot. Basically, during our collection time, IAmaA has most users involved in late afternoon and evening. It would be the best time to participate in related discussion. Implementation is in `q3_time_analysis.py` and the results of the matrix of count numbers is in `time_count_data.txt`

4.4 Relevance between r/IAmA and tweet

In this section, we take the outcome of qestion 2 as input, which is a list of text files including corresponding subreddits IDs and key words (under the folder named "keyword1"), then, extract all the tweets collected during 11/20/2020 to 11/25/2020 from MongoDB. Here, we only care the id and text fields of each tweet and leave out the other for unnecessary overheads. We compare each tweet with key words list of each IAmA posts, try to examine whether or not there exist some keywords in this tweet. If the number of key words occurring in one tweet is greater than half of the total number of one post's, we think the that tweet is connected with this IAmA post. For convenience (see a list of text files in the directory named 'result_Q4'), we offer another text file to show which IAmA post get matched by certain tweet. (see text file 'result_index.txt').The detailed implementation of this answer is the Python file,'Q4.py'.

4.5 Twitter Sample Stream

We continuously collected data of 1% sample stream from twitter through 11/20/2020 to 11/25/20 using the twitter official API and calculated the number of tweets received in each hour during that time. The result is shown in Figure 4. The implementation is in `tw_time_analysis.txt`

5 RESULT

5.1 Categories of IAmA hosts

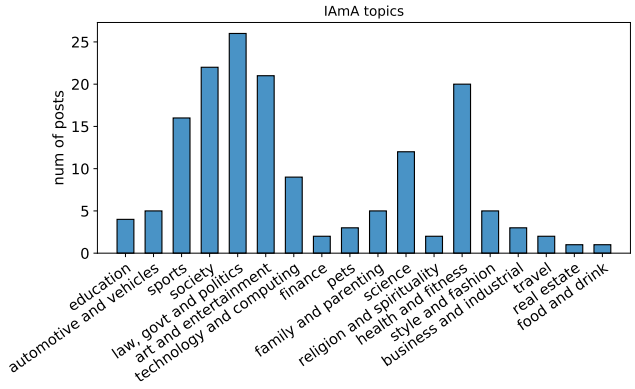

Figure 1: categories of IAmA hosts

Based on our experiment, the top four categories of IAmA topics are: "law, govt and politics"(26), "society"(22), "art and entertainment"(21) and "health and fitness"(20). We make a bar chart for this result (Figure 1). For these four categories, we also have a second level categories bar charts (seen in attachment Figure). And for the rest result, can seen in file "`Q1L1.csv`".

5.2 Most Frequent Words

During our experiment, the most frequent words are: "Proof"(110), "question"(103), "time"(68), "year"(66), "new"(63), "Reddit"(59), "people"(45), "election"(40). We use a wordcloud plot to shows this result (Figure 2). For the rest key words' frequency, seen in file "`Q2frequency.csv`".

5.3 IAmA Commnents Traffic Analysis

The X axis from 0 to 7 indicates the weekdays from Monday to Sunday correspondingly. The Y axis the range of 24 hours of a day from 00:00 to 23:59. Each color block indicates the average number of comments of that time slot. As the color bar shown in Figure 3, the brighter the color the larger the average number of IAmA comments are made during the hour of the weekday. The most popular time slot of IAmA is during 18:00 to 19:00 on Thursday.

5.4 Relevance between r/IAmA and tweet

As of the evening on 11.29, our program has run for nearly 10 hours. Due to the large size of tweet collections, at this moment we haven't got any matched tweet ids. Even though, as expected, we did do an experiment that we probably need another 30 hours to iterate the entire database. Therefore, we will constantly check the output file

Figure 2: most frequent words in IAmA

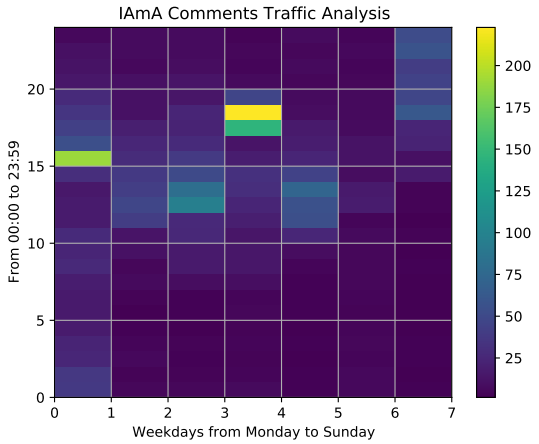

Figure 3: IAmA comments traffic distribution in hours (EDT)

and the program operation state. As mentioned, no single tweet id appears in result_index text file yet, meaning no definitive evidence to show the relevance between r/IAmA and 1% streaming tweets.

5.5 Twitter Sample Stream

Because MongoDB run out of our VM's memory at 11/25/2020 10:25pm, we did not have the required last one and half hour's data. The number of tweets received is counted hourly over the X axis through 11/20/2020 00:00 to 11/25/2020 11/25/2020 23:00. Basically, from 120,000 to 220,000 tweets come in per hour, shown in Figure 4.

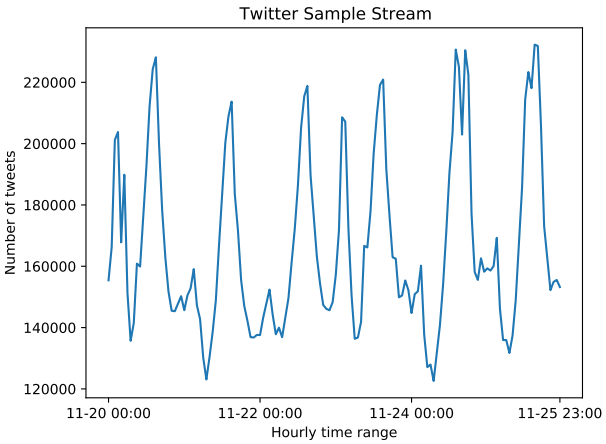

Figure 4: number of tweets through sample stream (EDT)

6 DISCUSSION AND CONCLUSION

6.1 r/IAmA is a diverse discussion community

In the month that we have been crawling the data, two important things have happened: the 2020 presidential election, and the continuing spreading of COVID-19. With this background, we still see a diversity of discussion topics in r/IAmA. Like in r/IAmA/jkub3n[10], the host is a research scientist in nuclear fusion. In r/IAmA/jj4fk2[11], the host is a journalist digging into CIA mind-control experiments.

There does not seem to be an absolutely dominant topic of discussion in r/IAmA, but in the first level of categories we can see that the more frequent ones are mostly topics that are relevant to everyone. Besides, from time to time, there will be some hosts with unique experiences to hold AMAs. In this case, r/IAmA moderators will give this post with "Unique Experience" tag. Like r/IAmA/k1kxz8[12], the host is a blind world-renowned gamer. Normally, this kind "Unique Experience" post will have more upvote(score) and more participation.

In general, people like to discuss the topics relevant to every but also enjoy engaging in discussions about unique experiences. These are exactly two ways of understanding the world: "Why is that?" and "How it can be that way!"

6.2 r/IAmA looks good on the surface

In section 5.2 Most Frequent Words, we can see the most frequent word is "Proof" and it is way beyond other words. That is because the first rule in subreddit r/IAmA is "All posts must contain proof"[3]. However we have total 192 posts(162+30), "Proof" only shows 110 times. That means, at most 52 posts (about 27%) in our database do not follow this rule. Some of them are Bot Ads (r/IAmA/jy5yyu[13]), some of them are not relate to AMA (r/IAmA/jxt2bd[14]), some of them are deleted by host (r/IAmA/jwjrer[15]) and some of them are violation other rules deleted by moderators (r/IAmA/jvi4gw[16]).

These indicate that r/IAmA is a social media being regulated, being reviewed. If someone have chance to see the whole picture(like

us using crawler system), it is no longer as good as it shows on surface.

6.3 r/IAmA is an active community

Through the time analysis of IAmA comments, we can find that IAmA is a quite active community with lots of users involved in making comments through all time and the busiest time is late afternoon and evening.

6.4 The connection between r/IAmA and 1% samples of twitter is very weak

As we mentioned earlier, we checked about one third tweets from our tweet collection and compare with the keyword list from r/IAmA posts. No single IAmA post can be connected with many tweets(to be specific, according to our matching policy, it should be "one" not "many"). Even though this result is related with our matching policy, which involves large amount of details to design, still, it is plausible that the connection between this two is very weak intuitively because of the characteristics of r/IAmA, e.g.people are invited to give an introduction of one certain topic(focusing on his/her career usually), instead, twitter users can post whatever he wants.

Another thing is Twitter has 140 words limit, that makes our matching even harder and less accuracy, because some post's key words list has over 140 words.

REFERENCES

[1] https://www.reddit.com/r/IAmA/
[2] https://www.delayforreddit.com/analysis
[3] https://www.reddit.com/r/IAmA/wiki/index
[4] https://www.ibm.com/cloud/watson-natural-language-understanding
[5] https://cloud.ibm.com/docs/natural-language-understanding?topic=natural-language-understanding-categories-hierarchy
[6] https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/
[7] https://amueller.github.io/word_cloud/
[8] https://www.reddit.com/r/LockdownSkepticism/comments/jvtpwz/
[9] https://www.reddit.com/r/IAmA/comments/jmafy5/
[10] https://www.reddit.com/r/IAmA/comments/jkub3n/
[11] https://www.reddit.com/r/IAmA/comments/jj4fk2/
[12] https://www.reddit.com/r/IAmA/comments/k1kxz8/
[13] https://www.reddit.com/r/IAmA/comments/jy5yyu/
[14] https://www.reddit.com/r/IAmA/comments/jxt2bd/
[15] https://www.reddit.com/r/IAmA/comments/jwjrer/
[16] https://www.reddit.com/r/IAmA/comments/jvi4gw/

7 ATTACHMENT

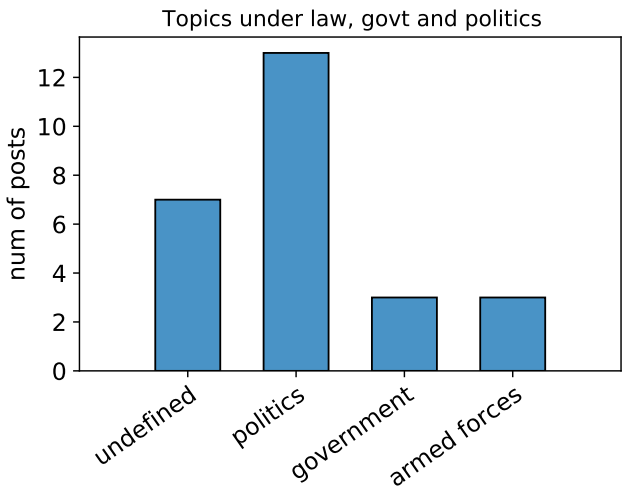

Figure 5: level 2 categories of "law, govt and politics"

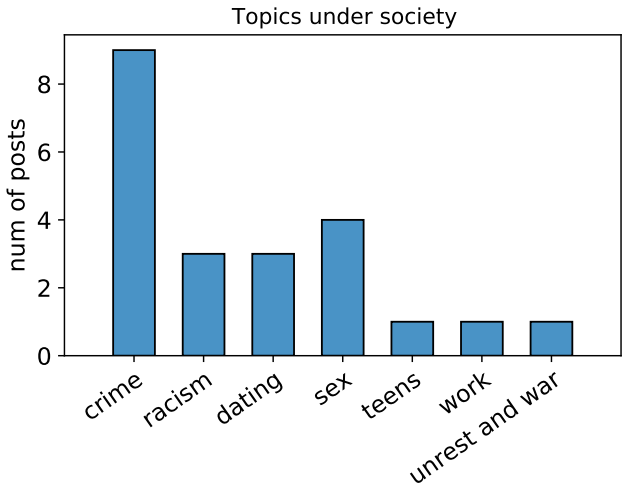

Figure 6: level 2 categories of "society"

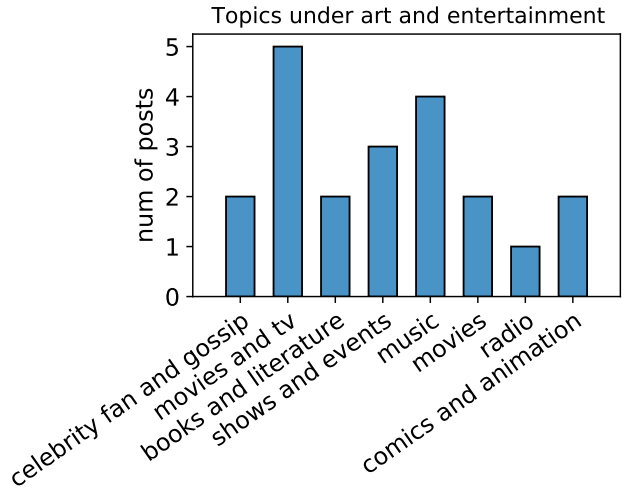

Figure 7: level 2 categories of "art and entertainment"

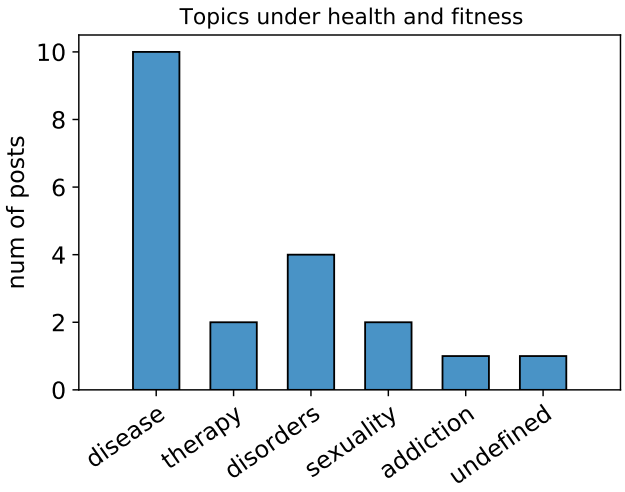

Figure 8: level 2 categories of "health and fitness"