Interesting Data Analysis based on IAmA

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

IAmA is a relatively “formatted” subreddit comparing with the other counterparts. in this project, we try to do some simple analysis on the data crawled from IAmA by presenting several questions in advance. generally speaking, we don’t focus on a series of logically continuous problems. instead, these questions are comparatively “isolated” with each other but give us a better understanding of what’s happening and give shed light on the features behind the data of this community.

1 INTRODUCTION

r/IAmA is a subreddit for question-and-answer interactive interviews termed "AMA" (short for "Ask Me Anything"). AMA interviewees have ranged from various celebrities to everyday people in several lines of work. Also, IAmA is one of oldest subreddits since the foundation of reddit. one of our group members is a patron of IAmA. In fact, he is very interested in these occupations he has no knowledge of. that’s why he fell in with IAmA once he found this place. therefore, when we try to decide the datasets of our study, he suggests us we can analyze the data from IAmA, studying something interesting and learn some unknown features of dynamics in IAmA behind the data.

2 OBJECTIVES

In this project, our objectives includes trying to do research on answering the following questions based on the data we are going to collect.

1.What are the categories or backgrounds of IAmA hosts? The hosts of IAmA that would like to answer questions are across all kinds of industries and come from different backgrounds. Even quite a few celebrities are among the participants. Thus we would like to analyze the discrepancy, the diversity and common features of the backgrounds and categories of those host.

2. What are the most popular topics and words of highest frequency on IAmA? Qustions of various topics ranging from politics sports entertainments to science, you name it, are raised by the listeners and audiences. Therefore, we would like to find out which topics are most popular and what categories of questions are mostly discussed.

3. During the time of data collection, in which time slot there are most users involved as the question is raised to the host? As our data collection will last in a continuous time span, we would like to dig into the time slot of users’ involvement and find out the busiest range of time of IAmA during our data collection time.

4. Whether there is a connection between IAmA and twitter? Since we collected 1% stream data from twitter as the coursework requires, we would like to see if there is any connection between the popular key words in twitter and those in twitter. Do the two data sources share any common trends? We would like to find out whether or not there exists an answer.

Table 1: Twitter Stream Database twitte_collection Table

Field Name	Description
text	tweet content
like count	like number of this tweet
retweet count	retweet number of this tweet
time stamp	UNIX time stamp of tweet create time
hashtags	any hashtags included in this tweet
preview_image_url	any media included in this tweet

Table 2: Reddit r/IAmA Database IAmA Table

Field Name	Description
subreddit	subreddit name prefix
id	post id in subreddit
title	title of this post
flag	r/IAmA categories
author	post author reddit name
text	additional description
score	post score
comments	total number of comments in this post
time_stamp	UNIX time stamp of post create time
url	URL of this post

3 METHODOLOGY

Table 1 and table 2 show what we have inside the database. We will use two online services and some statistical methods to finish the analysis.

To answer question 1, we will need the title of each posts in r/IAmA and IBM[1] Watson Natural Language Understanding(NLU) service. r/IAmA [2] require each host using "I am a ..." as the format to introduce themselves. We will feed this sentence to NLU service and get categories of the host. For each host we will get at least one category.

In answer to question 2, we will need the additional description of each post in r/IAmA and Microsoft Azure[3] Cognitive Services. r/IAmA [2] require each host to provide some additional information of themselves and the discussion topics under this post. We can use this part of text to generate key words by using Azure[3] cognitive services. Since the number of key words depends on the length of the text, normally we will get at least two key words.

For question 3, we will need the created time of each comment. We will count all posted comments over time, with one hour as

Table 3: Twitter stream Database

Table Name	Projection
hashtag_collection	10M
twitter_collection	100M

Table 4: Reddit r/IAmA Database

Table Name	Projection
comments	75K
CrossPosts	40
IAmA	240

the smallest unit and then scrape time period of the week with the highest user engagement during the period of data collection. Also we can calculate the ratio of total number of comments to the number of question we collected in the same post. This ratio can generally reflect the level of engagement of the host and the questioner.

Last but not least question 4, we will need the key words of each post and Twitter stream database. Normally, we will get more then 3 key words from Azure[3] cognitive services. Searching the database using these keywords, it is supposed that a tweet is relevant to this reddit post if it contains more than half of the keywords and is not too far apart in terms of posting time.

4 VALIDATION

As of the end of November, we will collect enough data for our analysis. We will not need any addition data. In r/IAmA, the discussing topics are wide enough. As we have been estimated in the report of our data collection system(Project1), the amount/count of each item is in table3 and table4.

REFERENCES

[1] https://cloud.ibm.com/apidocs/natural-language-understanding?code=python
[2] https://www.reddit.com/r/IAmA/wiki/index
[3] https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/