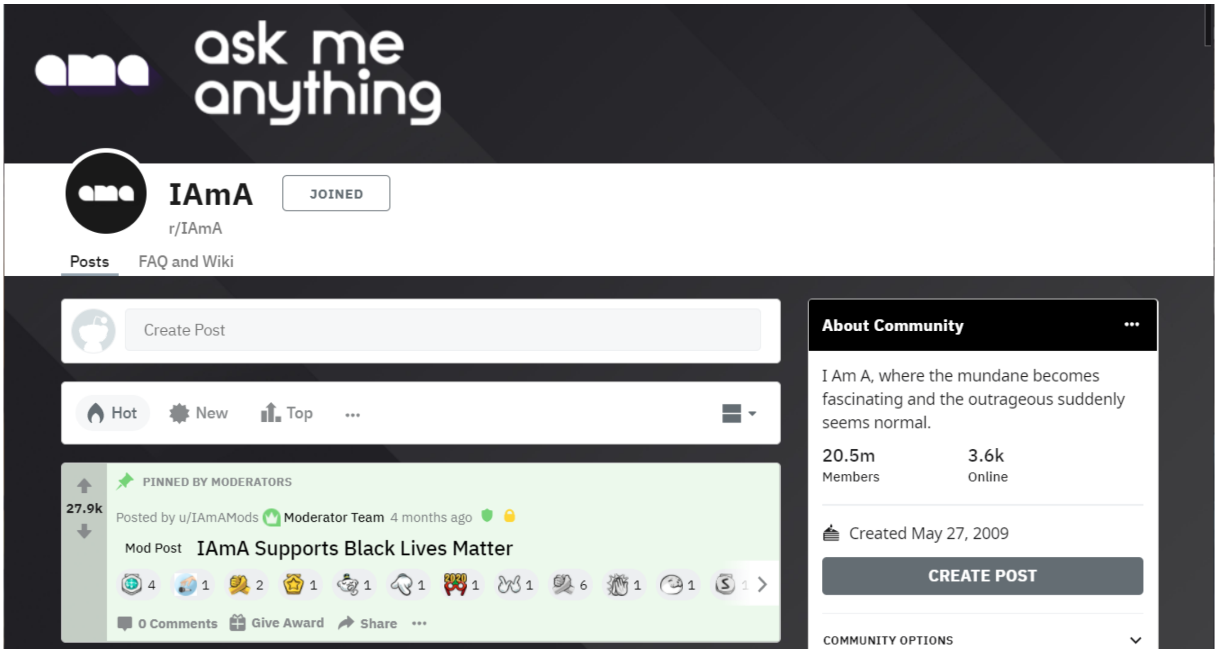
# CS 580N Project Proposal

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



Posted by u/IAmAMods Moderator Team 4 months ago

IAmA Supports Black Lives Matter

I Am A, where the mundane becomes fascinating and the outrageous suddenly seems normal.

Created May 27, 2009

**Figure 1: Screen capture of subreddit IAmA front page.**

## 1 INTRODUCTION

In these project, we are going to build a data collection system that can collect data of the subreddit IAmA continuously during the time. Using provided APIs of Reddit in our system, we are going to collect, store and analyzed the data. To align with the data we are going to handle, the database we intend to select is MongoDB which has good features of supporting JSON format data storage and query. Based on the above, a pipeline will be built that enables us to measure and mine the data further in the remaining part of the project.

## 2 DATA SOURCE

the data in our project comes from two sources. One is from approximately 1% samples of Tweets in real time as required and the other from IAmA.

IAmA is a subreddit for Q&A, with the voluntary interviewees posting threads whose topics are characterized by a unique start “I am..” to indicate one’s identity or job. It is open to any Reddit users to ask questions on any topic.

Interviewees are required to prove that they are who/what they claim to be. Comments happening in one thread are in Q&A format. Like other threads in reddit, and users can also upvote and downvote potential candidates (therefore, the more favoured a post is, the more conspicuous (upper) position a post will appear, which in turn will attract more comments).

We choose IAmA as our ideal data source because of its diversity and flexible APIs.

## 3 COLLECTING DATA

We will use the official reddit API (https://www.reddit.com/dev/api/) to collect data and at the same time the rules and terms stated by the reddit official website will be followed to avoid any ethical or legal issues. With this API, we are able to collect most data and information that we want with great convenience.

For example, if we would like to get the comments, a handy method "GET [/r/subreddit] /comments/article" offered by official reddit API documentation is available. We are going to further use these methods and make our own crawler client.

## 4 MEASUREMENTS AND ANALYSIS

The scope of our analysis and measurements will be only limited to the Ask Me Anything subreddit itself.

We plan to analyze the most popular topics and classify them into different categories. Further, since we have two types of data source, we will try to compare sentiments through the tweets and comments of the similar topics that are both present on Twitter and IAmA. Techniques of NLP(natural language processing) and statistical analysis are expected to be used.

## 5 ESTIMATES OF DATA

The average number of comments in one thread is based on rough estimation. Nearly 20,000 comments are produced weekly, but there might be some fluctuation resulting from the cross-subbreddit posts where only exist a link leading to a discussion in another subreddit. The data will be updated roughly every 5 minutes by our system.

To collect and store the data continuously, we choose MongoDB as our database system which supports JSON data storage and query.

## 6 POSSIBLE PROBLEMS

(1) the first one is cross-subreddit problem as mentioned above, so we need some extra work dealing with those crossposts.
(2) The comparison between two data sources will be a challenge. They have different types of JSON formats and user habits. Most text comments on IAmA are not subject to a word limit. Meanwhile, Twitter has a strict word limit sometimes even comes with photos.