UTAH STATE UNIVERSITY, COMPUTER SCIENCE DEPARTMENT

CS5890 Data Science - Project Proposal
Random Acts Of Pizza

Team **Pizza Hackers**: Tam Nguyen and Hung Pham

October 27, 2015

1 Introduction

In this proposal we will discuss our plan to perform data analysis on a social interaction dataset, where requester ask for free pizza on a Reddit community "Random Acts of Pizza"[1]. We will discuss the social interaction problems, the dataset provided by Kaggle, preliminary of our analysis plan, which data science toolkits we plan to use.

2 The team: Pizza hackers

We name our team "Pizza hackers" in direct references to the task that we want to tackle, and our Computer Scientist's root. "Pizza hackers" consists of two members: Tam Nguyen and Hung Pham. Tasks such as implementations of data analysis experiments, report writing, and visualization will be divided among team members. All members will be involved in analysis of experiments results as well as any decision in term of project direction.

3 The problems

Internet is a newly discover virtual land, where humans interact with each others on multiple levels. Interaction with a new person on-line is only a few mouse clicks away. It is an active researching field to study the interaction between humans in a virtual space. Study how human response to request from a complete stranger is an interesting research question. Crown-funding has been one of the best way for startup companies and inventors to acquire necessary funding.

Althoff *et al* have investigate which factors effect the successfulness of a request [1], specifically requests for free pizzas. They used topic modeling and logistic regression to determine which narrative topic have the highest success rate and in turn determine factors that effect the outcome of a request such as: who is asking, how the request was asked. We decided to extend this research by using a different set of techniques that are more

[1] `https://www.reddit.com/r/Random_Acts_Of_Pizza/`

Table 1: Important attributes used in our analysis

Important Attributes	Explanations
Request text	
request_title	Title of the request
request_text_edit_aware	Request text after removing comments indicating the success of the request
Requester information	
requester_account_age_in_days_at_request	The age of requester (in days) at time of request
requester_number_of_comments_at_request	The number of comments on Reddit by requester at time of request
requester_number_of_posts_at_request	The number of posts on Reddit by requester at time of request.
requester_subreddits_at_request	The number of subreddits that requester had posted in at the time of request.
requester_upvotes_minus_downvotes_at_request	Difference of upvotes and downvotes of requester at time of retrieval.
requester_upvotes_plus_downvotes_at_request	Sum of upvotes and downvotes of requester at time of request.

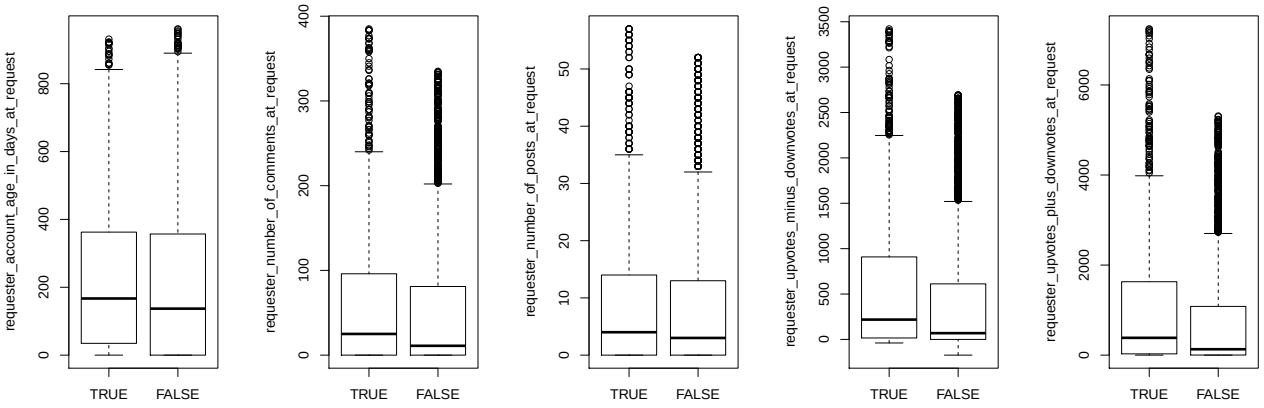

Figure 1: Boxplots of some attributes based on the success of requests

up-to-date such as distributed vector representation (Word2Vec[2]) and Support Vector Machine. We will discuss the details of our plan in The plan section.

4 The dataset

We downloaded data from Kaggle[2]. The dataset includes 4040 requests collected from the Reddit community Random Acts of Pizza between December 8, 2010 and September 29, 2013. Each data object is a request for a free pizza. There are 994 requests received a pizza while 3046 requests did not.

The dataset is in `JSON` format. Each `JSON` entry represents one pizza request. Data fields includes information about requests such as id, text, requester name, etc. and meta-data such as: time of the request, activity of the requester, community-age of the requester, etc. There are several fields are collected after the time a request was posted, the values of those fields doesn't affect whether a request receive pizza, thus, we removed those fields out of the dataset. Table 1 shows important attributes (and explanations that) we consider in our analysis.

5 The plan

To analyze our data we will perform three critical steps:

- General analysis of simple information about requesters.
- General analysis of textual attributes (request subject and body).
- Apply prediction model to measure effectiveness of each factor.

For example, we could perform comparison between successful and unsuccessful requests to see if an older account holder will have more chance at receiving a free pizza or not using a simple box plot of the attribute *requester_account_age_in_days_at_request*. Table 1 is showing such analysis for the five most interesting attributes containing requester information. We will also apply prediction model such as logistic regression using different attribute to see which one have the best predictive power. Such analysis will give us some idea of how the information available in Reddit about the requesters could effect the outcome.

The two most interesting and valuable attributes in our dataset are *request_title* and *request_text_edit_aware*. These attributes contain the title and textual body of each request. Most of decision will be based upon the contents of these two attributes. We will first performing some preprocessing including: tokenizing, stop-word removal, stemming... Then we will use a simple frequency based phrase builder to create phrases that occurred enough but not too much or too little (The details will be included in check point report). We could then investigate the relationship between terms and phrases that co-occur the most with successful and unsuccessful requests. We could also use simple logic regression to see which terms or phrases have the most predictive power.

Finally, we will apply Word2Vec on these textual attributes to extract vector representation of terms and phrases, then by using clustering we could be able to find cluster of terms and phrases that have similar syntactic and semantic meaning. We then extract the "bag-of-cluster" vector for each request and feed this to SVN so that the model can learn to predict the outcome. By investigate the weight vector of SVN model we will be able to tell which cluster (factor) have the most effect on the outcome.

6 The Toolkits

We plan to use `R` as the main programming language in our analysis. We have used `jsonlite` library for importing dataset (in `JSON` format) to dataframe in `R`. We also used `R` to produce boxplots in our preliminary analysis. All attributes about requester are quantitative, thus, using `R` as the analysis language is sufficient. For analyzing textual data, we plan to use `word2vec` library[3]. For building classifier to predict whether a request receive a pizza we plan to use `R` interface of `libsvm`[4] library.

References

[1] Tim Althoff, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky. *How to Ask for a Favor: A Case Study on the Success of Altruistic Requests*, Proceedings of ICWSM, 2014.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, Jeff Dean. *Distributed representations of words and phrases and their compositionality*, Advances in neural information processing systems, 2013.

[2] `https://www.kaggle.com/c/random-acts-of-pizza/download/train.json.zip`
[3] `https://code.google.com/p/word2vec/`
[4] `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`