Explore Emotion Expressions in GitHub Most Influential Open Source Communities

Meng Yang
Tandon School of Engineering, NYU
NY, U.S.
my1421@nyu.edu

Kaiwen Peng
Tandon School of Engineering, NYU
NY, U.S.
kp1804@nyu.edu

***Abstract* —** As the largest Git repository hosting service provider, GitHub collects huge dataset of millions of open source projects every day. There is many interesting information in these dataset, such as the popularity of programming language. What expected to dig out in this project is the emotions of developers and users of those most influential open source organizations, showed in the commit comment texts.

Keywords — GitHub, Ranking, Text Emotion, Classification, MapReduce

I. Introduction

GitHub is the largest web-based Git repository hosting service. There are millions of projects under progressing on GitHub every day. Authors are trying to do analytics with this information to tracing out the emotions expressions of open source community of the most influential open source organizations.

This problem could be separated into two steps: ranking the organization/projects by their influences; collecting text dataset of target organizations and applying classification algorithm to determine its emotion spectrum.

First, the most successful projects and organizations on GitHub will be ranked by their impact which measured by the stars and forks. One efficient and tested calculation method is academic impact index, H-index.

Moreover, the texts to be analyzed will come from the commit messages and pull requests written by developers. Exploring emotions in GitHub message is a classification problem. One naive approach will be like classic word count, with an emotional words dictionary showing 5 basic emotions: Anger, Sadness, Happiness, Fear and Anxiousness. Further improvement will introduce machine learning method to help this classification.

II. Motivation

GitHub renders the data collected at GitHub Archive, an available and easily accessible public dataset which is an exciting news for data analysts. There are many open projects working on this dataset including GHtorrent. GitHub is one of the best places to analyze up-to-date status of open source software development. By exploring the emotional expressions in the messages committed by developers on GitHub, an overview on the characteristic of each organization could be plotted. From the result, more information could be speculated: the current state of an organization project and the potential of further development.

III. Related Work

Find one efficient and accurate approach to determine the impact of an author or organization's work on academic community is a long-term topic. A simplest approach is counting the number of times an author or organization is cited. Various other measures are kept been developed to improve the accuracy. In 2005, Hirsch proposed the index h[1], defined as the number of papers with citation number >= h, as an useful index to characterize the scientific contribution of a researcher. The h-index is intended to measure simultaneously the quality and quantity of scientific contribution in an unbiased way. In 2006, Leo Egghe suggested an improvement based on h-index, called g-index[2]. This index is calculated based on the distribution of citations received by a given researcher's publications. For instance, given a set of articles ranking in decreasing order of the number of citations that they received, the g-index is the unique largest number such that the top g articles received together at least g^2 citations. It is found that the ranking g-index column resembles more the overall feeling of "visibility" or "life time achievement" than the ranked h-index column does.

Sentiment Analysis is a popular topic in Machine Learning or Natural Language Processing field. With the booming of machine learning methods in natural language processing and information retrieval, a wide works research on this field. An overviewed of sentiment analysis was given in Lee & Pang's

work [3]. There are also works [4] focusing on more detailed classification of sentiment expressions, presents a new approach to phrase-level sentiment analysis that first determines whether an expression is neutral or polar and then disambiguates the polarity of the polar expressions. With this approach, the system is able to automatically identify the contextual polarity for a large subset of sentiment expressions. However, there is few works focusing on web blog message like microblog. In 2010, Paroubek & Pak [5] published their work on sentiment analysis of Twitter data. Their paper focused on using Twitter for the task of sentiment analysis which showed how to automatically collect a corpus for the purposes of sentiment analysis and opinion mining. Using the corpus, they build a sentiment classifier, that is able to determine positive, negative and neutral sentiments for a document.

IV. Design

1. Overview

The whole project consists of two main steps, ranking organizations and analyzing emotion spectrum based on text data. Data flow pipeline is showed as following chart:

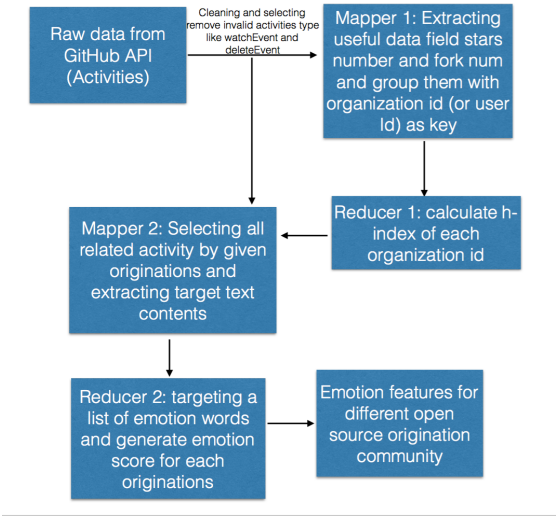

Figure 1. Overview pipeline of analytics

2.Organization Ranking

Ranking of organization is based on their influences in the open source community. In order to quantitatively measure the influence of each organization, two important indices are taken into count: Stars number and the fork number of a repository. Stars is a number of developers who watched and followed the progress of a repository, it in certain terms reflects the popularity of a project. Fork shows the number of developers who have forked the repository to their own GitHub, this number usually indicates the number of developers who are working on this project or are interested in this project.

An approach of ranking is counting the sum of these two indices. One problem of doing ranking on Hadoop cluster is that the ranking process is undertaken separately on several working nodes and each worker node is handling only a part of whole data set. This model will cause a mistake when a very active organization is evenly allocated to all worker nodes and therefore it is supposed to lose the ranking information when combing results from worker nodes.

3. Emotion Quantification Model

The essential problem in this project is how to quantify the emotion in a way that could be analyzed in a numerical approach. There are a massive of studies working on this issue. Human emotion is complicated and subtle, which is naturally hard to quantify.

One simple approach is given every text a label of emotion of positive or negative. By this binary label, the information of basic sentiment of a text file could be obtained. This is a widely-used approach. However, the binary separation of emotion would not give enough details for the analysis of the characteristic of certain community. This approach is more suitable for evaluation of products and feedback messages. In these two situations, simple classification is already good enough to get the conclusion.

A more natural and precise way is using a more specific emotional category. The basic emotions are Anger, Sadness, Happiness, Fear and Anxiousness. This category can be further specified into 87 emotional concepts. But in this project, only five basic emotions are designed here. In textual context, emotion could be categorized by denoting certain emotional keywords. A dictionary corresponding to the emotional concepts was built. It consists of five categories of emotion-denoting words.

Emotion	Key Words Dictionary
Anger	angry, annoyed, appalled, bitter, boiling, etc.
Sadness	agonized, heartbroken, sad, shamed, shocked, etc.
Happiness	great, enthusiastic, excited, pleased, proud, etc.
Fear	afraid, aghast, alarmed, fearful, fidgety, frightened, etc.
Anxiousness	concerned, distrustful, doubtful, dubious, etc.

Table 1. Emotion Dictionary with five field

Considering the characteristic of GitHub, some special words have been added to our emotional dictionary. In a typical GitHub comment, the emotion of happiness is usually expressed through the praise of a brilliant idea or celebrating the successful debugging of a tricky bug, like "Amazing awesome super bug fix hurray!" or "Yay, comments in code!". Therefore, keywords like "fixed" and "awesome" will take huge counts in defining happy emotion. Similar works have been done in other emotion fields like sadness and anger.

4.Text Sentiment Analysis

1.) Text Pre-Processing

Pre-processing of text sample is an important work to get a good performance in following analysis. The large amount of available text data collected are unstructured, massive and various. Thus, text is needed to be adjusted to be more analyzable.

The typical procedure has following steps:

a. Convert all character to lower case.
b. Remove all special characters and numbers.
c. Stop words removal.
d. Tokenization.

2.) Term frequency

Term frequency is a naive method in text analysis which is based on the assumption that, if a term appears more frequently in a document, this document has a higher correlation with this term. Thus, it deserves to receive a higher score of this term.

In Term frequency, the original frequency of a term in a document is used: the number of times that term *t* occurs in document *d*. If denote the original frequency of *t* by $f_{t,d}$, then the simple tf scheme is $\mathrm{tf}(t,d) = f_{t,d}$.

In this analysis, pre-processed comment message is the document d, while terms t is chosen from the emotional keywords of emotion dictionary.

3.) TF-IDF

TF-IDF is short for *term frequency - inverse document frequency*, which is a numerical statistic intended to reflect how important a word is to a document in a collection or corpus. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus. This method helps to adjust for the fact that some words appear more frequently in general.

The inverse document frequency(IDF) is a measure of how much information the word provides, that is, whether the term is common or rare across all documents.

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

V. RESULTS

1.Orgnizations Ranking

All public GitHub events collected from November 1, 2016 to November 15, 2016 are used as the data set, for the Organization Ranking analysis. The data is accessed by GitHub API. The whole data set is of a size over 4TB, which is stored on Dumbo HDFS. This data is in JSON format. In order to extract information out of JSON file, *org.json JAVA package* is utilized. This causes some problem when trying to run the code on Hadoop cluster since the worker nodes are not able to find this third-party JAVA package. In solving this problem, the approach of using libjars option with Hadoop is chosen.

```
$ export LIBJARS=/path/jar1,/path/jar2

$export  HADOOP_CLASSPATH=/path/jar1:/path/jar2
$ hadoop jar my-example.jar com.example.MyTool
-libjars ${LIBJARS} -mytoolopt value
```

The above operations make third-party JARs available to remote map-reduce tasks running on JVMs of worker nodes.

The ranking result of top 15 influential organizations on GitHub in the first half of November 2016 is shown as below:

6154722:Google 9001
6154722:Microsoft 5622
6128107:vuejs 4374
1961952:alibaba 2507
82592:square 2389
12002442:Bilibili 2221
1916665:udacity 2101
7378196:googlesamples 1912
6412038:reactjs 1911
4680972:LeadDyno 1885
3006190:shadowsocks 1817
16437392:WhitestormJS 1767
1778935:GoogleChrome 1738
14985020:zeit 1725
17266927:lesspass 1702

This result basically meets authors' general expectations: Google and Microsoft lead at first and second place. And It is worth mentioned that Bilibili, a Chinese video website like

Vimeo, sits at a rather high position, surprisingly. This is probably due to it has some amazing projects undergoing recently.

2.Emotion Analysis

Following five charts are depicting top10 organizations with highest mood (anger, sadness, happiness, fear, anxiousness) expression in commit messages. They are displayed by Tableau software with HIVE database executing the selecting the top 10 SQL functions.

For each emotion word, a synonymous word set with a size between 50 - 100 is applied, to catch its emotions. And each emotion word carries an equal weight of 1.

Let us start with "angry", a very frequently negative emotion when programmers meet obstinate bugs and approaching dues. The results showed organization Blue-yonder leads with a percentage of 0.14% of angry words appearing among all commit message words. And organization *hyperoslo* and *lesspass* tied at second position with 0.10%. The chart is relatively flat. And the rather low absolute values are surprising but reasonable after careful consideration.

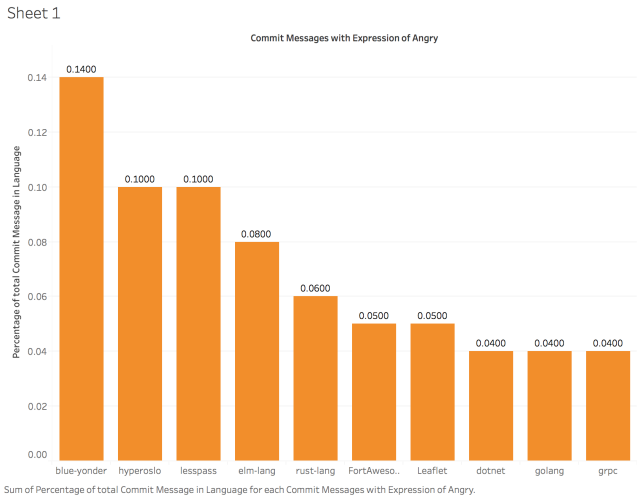

Figure 2. Top 10 organizations with high expression of anger in their comment messages

Like angry, another negative mood sad has a higher percentage. Besides blue-yongder holds the first place again, the *npm* organization surprisingly stands at second position with a 0.25% of sadness word appearance. This probably is due to differences between *Node.js* and *JavaScript* of functional programming languages and traditional programming languages: *C++, JAVA*.

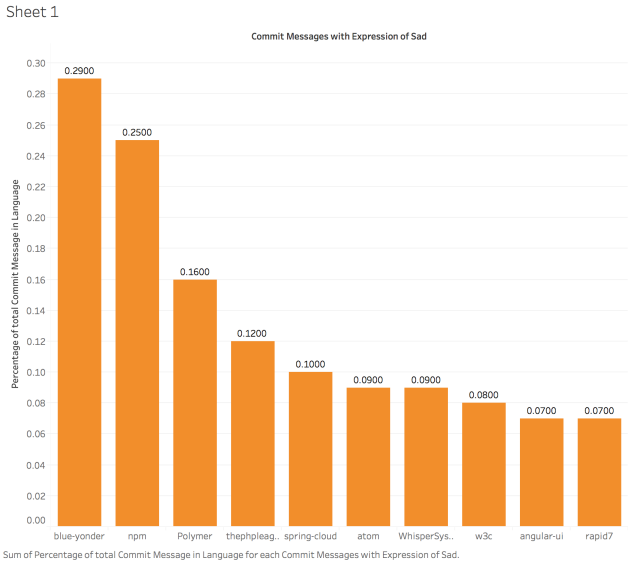

Figure 3. Top 10 organizations with high expression of sadness in their comment messages

For similar reasons, one can experience angry one can experience joyful. The emotion of happiness like spending a long time featuring out a bug often lasts for a while when developers commit on GitHub. It is obvious that the percentage of happiness is much higher than the other mood, which is a pleasant thing. The statistic shows that the leading one, *Udacity*, has a 1.31% value. It is more than five times compared to previous two negative moods and following two: fear and anxiousness. Moreover, values of other companies here also follow *Udacity*, with a rather flat bar chart presented here.

This is mainly caused by the fact that most developers are rational as adults. They are used to take the thing into consideration that: this is a job and a team. Being professional. Write commit comments rationally, focusing on the job. And just reveal good mood if need.

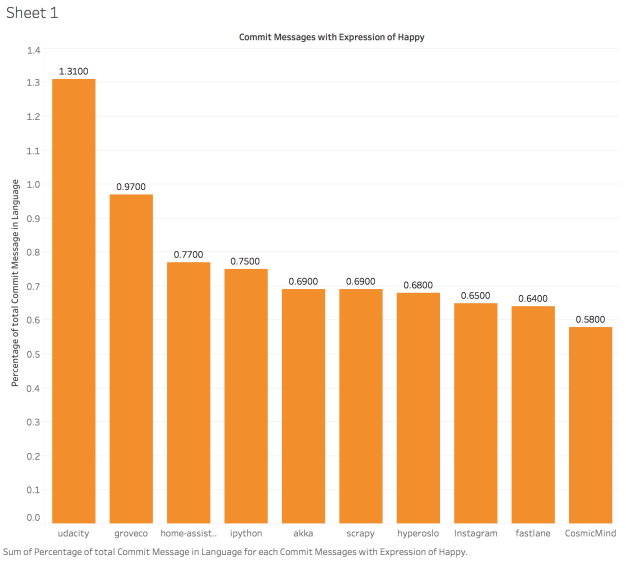

Figure 4. Top 10 organizations with high expression of happiness in their comment messages

In the end, the last two bar charts of Fear and Anxiousness is similar to previous two negative moods, just these two are rather cliffy, beyond *yeoman* and *alexa* with *ipython* leads with 0.13%, 0.26% and 0.24% respectively in Fear and Anxiousness statistics, other companies all demonstrate very low values below 0.05%.

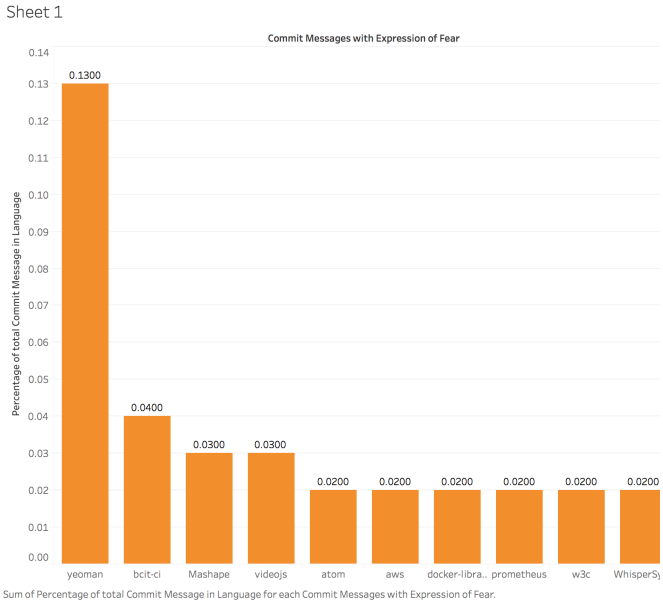

Figure 5. Top 10 organizations with high expression of fear in their comment messages

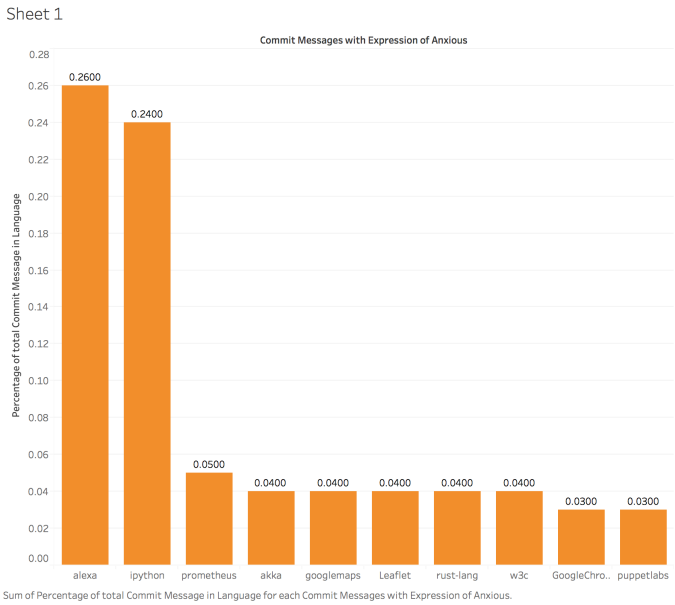

Figure 6. Top 10 organizations with high expression of anxiousness in their comment messages

3. Improvement With comprehensive Lexical Set

From above results, the appearance rate of emotional keywords in data set is very low. After analysis, it is featured out that this is due to the nature of GitHub commit messages. It always focuses on professional topics. Thus, authors did use a more comprehensive lexical set trying to get a better result.

MQPA project of Pittsburgh University provides a comprehensive dictionary - The *Subjectivity Lexicon*. This lexicon set contains over 8000 words. Each word has been given a label of polarity of positive, negative or neutral.

Part of the output of analysis is shown below:

Organization Name	Positive	Negative	Neutral
Google Chrome	3.15	1.94	1.56
Microsoft	3.72	1.81	1.76
Bilibili	0.94	0.76	1.07
Udacity	2.31	0.35	0.25
vuejs	1.51	1.01	0.86

Table 2. Selected emotion score from analysis with the Subjectivity Lexicon

VI. Future Work

In general, one drawback of this project is the commit text content of GitHub itself is weakly emotional. The performance is supposed to be much better if the analyzed content is Tweets or web blogs, as their texts are more diverse in moods.

One evaluation is building a more comprehensive dictionary of emotions with unsupervised learning, which would be hard to implement but will in certain way give a better performance. Another approach is introducing some machine learning methods into this analytic. One widely-used algorithm in text classification is *Naive Bayes*. It is easy to implement and have good performance of accuracy. To build a reliable training set, crucial for any machine learning method, campuses should be selected first, according to key emotion words, as in this project. Then, the training could be further optimized by manual sort.

VII. Conclusion

This project mainly consists two works. The open source organizations ranking result is reasonable and as expected, which indicates that using watches and stars as a measure of open source organization influence or popularity is efficient.

Emotion analysis with five dimension classifications using term frequency method does not give a perfect result. The information in this result is a little bit limited. This indicates that the naive term frequency approach with small lexical set does not work properly for weak emotional text set like GitHub comments. However, after using comprehensive lexical set, the result is significantly improved, though the dimension of emotion fields is limited to three.

Acknowledgment

We are grateful to GitHub Archive project for providing the public GitHub timeline dataset in this project, and GHTorrent, which provides easy accessible GitHub commit messages dataset.

MQPA project provides the Subjectivity Lexicon data set in the analytics part of this project.

Moreover, special thanks to Professor McIntosh for providing expertise guide in big data analytics skills and valuable suggestions.

References

1. J. E. Hirsch. Hadoop: An index to quantify an individual's scientific research output Proceedings of the National Academy of Sciences of the United States of America. Vol. 102, No. 46 (Nov. 15, 2005), pp. 16569-16572
2. Egghe, Leo (2006) Theory and practice of the g-index, Scientometrics, vol. 69, No 1, pp. 131–152. doi:10.1007/s11192-006-0144-7
3. Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Found. Trends Inf. Retr., 2(1-2):1–135
4. Baldoni, Matteo, et al. "From tags to emotions: Ontology-driven sentiment analysis in the social semantic web." Intelligenza Artificiale 6.1 (2012): 41-54.
5. Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.