T-Emoji: An automatic tweet's emoji analysis platform

Qi Yin qy652@nyu.edu
Yuan Li yl6606@nyu.edu
Victoria Yen vy406@nyu.edu
Mengxi Wu mw4355@nyu.edu

1. Introduction

Millions of tweets are generated on Twitter every day. People express their emotions and opinions on twitter's 280 characters limit. Within these characters, it is common that people put Emoji in their sentences. Emoji are ideograms used in electronic messages and web pages. However, as graphics with nuanced details, emoji may be open to interpretation. Emoji also render differently on different viewing platforms. The subtle relationship between text and emojis is intriguing. So, we would like to dig out the connections between words and emojis. Overall, we find significant potential for communication using emojis, both for word-based interpretations and for sentence-based interpretations.

2. EmojifyData-EN Database

One data source will be used in our projects: EmojifyData-EN database[2].

The EmojifyData-EN contains all the archives with tweets for January, February, Match and April of 2018 from ArchiveTeam Twitter Stream Grab. The archives are extracted from the original Tweet JSON data, while removing meaningless content like punctuation, URLs, mentions, and hashtags. Moreover, other tweets which has been written in non-English language are cleaned out. Selected tweets are insured with at least one emoji.

3. Project Architecture

The overall framework of our project can be divided into four parts: Data Acquisition, Data Storage, Data Analysis, and Data Display (front end, back end).

For Data Acquisition, we will remove unnecessary beginning and ending words, and extract emoji. We plan to use MongoDB as our database in the part of Data Storage. The Data Analysis part will be based on PySpark. The detail goals that we want to analyze in Pyspark will be discussed in the next section. We plan to deploy a website with Django for Data Display. The visualizing package we use will be D3.js.

4. Goals

Our system will analysis the hidden pattern of emoji usage. In terms of emoji, we would analysis the preference in difference kinds of sentences. And in terms of text, we would analysis the appearance of emojis, such as possibility and location. some ideas are from [1]

Here are the main functions included in our system:

Individual:

1. Find the appearance frequency of every emoji.
2. For every emoji, find the 3 other emojis that are used most frequently with it.
3. For every emoji, determine it is used more with words begin with lower case or word begin with upper case.

This part of the function allows us to understand the user's preferences for emoji and the emotions that users want to express through emoji.

Word-based:

1. For every emoji, find the top 10 words appears most when using the emoji.
2. Find which emoji used most with words begin with letter A, B, C...Z (alphabet order).
3. Find the most pair-wise frequency words and emojis.

In this way, we can define some related words for an emoji, and the implicit meaning of emoji may be discovered based on its related words.

Sentence-based:

1. Find the average of the number of emoji used in a sentence.
2. For every emoji, find the position (head, middle, end) that the emoji occurs most in a sentence.
3. Analyze the relation between the length of sentence and the number of emoji used in the sentence.
4. For every emoji, summarize the frequency that it could be used more than once in a sentence.
5. For every emoji, find the average word length in the sentences that contain it.

Provided emoji location relationship in a sentence, we can have a better understanding of emoji's part of speech. It gives us a potential to change our tweets words with some emojis, reaching a new wave of expression.

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[2] D. Larionov and Random. Emojifydata-en: English tweets, with emojis. Retrieved from =https://www.kaggle.com/rexhaif/emojifydata-en, 2019.