Group 9 Final Project Report

1. Motivation and objectives:

   There is a website named ZhiHu(知乎) in China. At the beginning, the users there are mostly social elites, and the topics discussed in ZhiHu are mostly about serious stuffs. When Zhihu becomes larger, more and more people joined in then there is a saying that ZhiHu has become more and more boring, people there talk much more boring stuffs than before. As data analysts, we'd like to know if the same stuffs happened in American online community, so we decided to analyze Reddit, which is a very famous website in the world. We want to find out the change of the topics of Reddit, and the language pattern, such as mimics or slangs used by Reddit users. Also, we want to know if Reddit users' behavior has become ruder or more civilized as the number of users getting bigger. Finally, with the result we've got, we want to set up a model which can predict if a post will be hot in Reddit, the precise prediction can have much commercial use (for example, company can invest the hot post to propagate their products, that is to say, advertising)

2. Project introduction:

   Our project can be separated into 3 parts:
   - Scrape 10,000 posts from Reddit world news community from 1/10/2018 to 1/11/2018 and process the posts to find out the top ten topics people discussed recently. With which we can understand what do the reddit users like to read and comment now days.
   - Scrape another 10,000 posts from Reddit world news community from 1/10/2013 to 1/11/2013 to find out the reddit users' top ten favorite topics 5 years ago. And also, by processing the posts, we can find out the language pattern (what mimics do Reddit users like to use) and the civilized level of Reddit users (in this project, dirty word rate of a single comment can represent it). Then we can analyze Reddit's change by make a comparison between these two results.
   - With the data and results we found, we'll build a predicting model to predict if a post will be hot, then we'll submit 20 posts to Reddit to see if the result match our prediction.

   By analyzing the results, we may figure out how the Reddit community changes and try to predict what will be popular in the next 5 years.

3. Methodologies:

   Our project has used these methodologies:
   - To find out the topics, we used K-means and hierarchical clustering, however, because K-means is randomly choosing a center, the result will change every iteration, so we just gave up it, and for the same reason, we didn't try LDA (latent dirichlet allocation).
     To find the best cosine distance, we did the clustering by hand for the first 100 posts of both 2013 posts and 2018 posts, then just set different distance values to see if the

hierarchical clusters match the clusters we did by hand. Then we find when the cosine distance is 0.985(for the posts in 2018) and 0.955 (for the posts in 2013) performs best, then we chose them to do the clustering.

- To find out the users' attitudes towards different posts, we did sentiment analysis. 2 methods are used to find out people's attitudes, the first method is to count the active and negative words in each comment, these active/negative word dictionaries are given by professor during the class. Another method is to use NLTK package to calculate the negative/positive score of each comment. By comparing the results of the two methods, we can understand sarcastic meaning of the comments. For example, a comment may have a negative score but many active words, then it must mean irony. And we also used the dirty word list from MIT's dirty word dictionary.
- To find out the people's language pattern, we used sklearn package's vector counter, we count the single word without stop words and phrase with 2/3/4 words in each comment, the top 20 tokens are kept, to represent the popular language pattern.
- To predict whether a post will be hot or not, we do classification. We use several methods of classification like Naïve Bayes, Linear Discriminant Analysis, Logistic regression and K nearest neighbors. For the label of hot, we have two kinds of definition. First, we define the post with comment number more than 10 as the hot one. Then we improve our way to set the label. We sort the post by comment number, and we choose top 25% as the hot news. Besides, while training the model, our input variables used to be the tfidf matrix that generated from the news title, the create time, last time and the sentiment score calculated from the title. We optimize our input variables, and now we can use only tfidf matrix to predict whether a post will become hot or not. So, we totally use two sets of input and two sets of labels to train the model. And we can get four kinds of model four each classification algorithm.

4. Analysis of Experiment results
    - About clustering:

    We used different methods to process the data as clustering candidates, for the posts from 2018, we scraped the popular topics in October. But for the data from 5 years ago, we picked the data whose comments is more than 10 as the candidates. Then we did the clustering as mentioned.

    By counting the number of the clusters, we can find the top ten topics in five years ago which are: president (Obama Spying), environment, NSA international spying, Mideast teenager female rights, Palestine and Israel conflict, human conflicts, US drone kill, nuclear between US and Iran, American with Snowden, and google with NSA.

    The top ten topics in 2018 are interesting news (meaningless random news and scandals), Arabic murder, president (Trump & Putin's relationship), weed legalization, bad religion, human job & life, politic news, Europe news, German/Merkel and bomb/attack/kill.

    And the hottest news is the weed legalization in 2018 and Google with NSA in 2013.

People's attitudes towards these news are mostly negative, among them, the USA drone kill in 2013 and bomb/attack/kill in 2018 triggered the most bad feelings.

- About the language pattern/sentiment analysis/civilized level:
  By comparing the average bad words rate for each single comment, we found that people tend to speak less bad words in their comment than they did in 2013 (0.47 per comment in 2018 and 0.55 per comment in 2013), also, in the extreme situation, people also speak less dirty words than before. (3 bad words per comment in 2018 and over 3.5 bad words per comment in 2013).

  About the average attitude of each post, people tend to be more active than before. Even though the average sentiment score per comment is still negative, but the values are better than 2013 not only in main comment but also sub comments. (-0.028 for main and -0.036 for sub in 2018, and -0.066 for main and -0.046 for sub) This may be because of the better economy environment, or Reddit administrators' operation (they may delete the comment with negative/dirty words.)

  But the language pattern is hard to explain, the tokens we've collected is meaningless. We guess our methods doesn't work, maybe further study needs to be done.

- About classification and prediction:

  With two sets of input variables and two ways of defining the hot label, we can get four kinds of model for each algorithm.

  When we use not only tfidf matrix, but also create time, last time and the sentiment score calculated from the title as the input to train the model, we can get a really good accuracy. We use cross-validation to test the model. If we regard the post with more than 10 comment as the hot post, the average fscore of KNN can be 0.8892. And the fscore of Linear Discriminant Analysis is about 0.5399. If we use top 25% as the hot post, the average fscore of KNN is 0.6750. And the average fscore of Linear Discriminant Analysis is 0.4481.

  The accuracy of KNN model is too high, and it is not reasonable. Because one of the input variables is last time which shows how long a post can exit on the reddit website. And this variable can directly show whether a post is hot or not. So, we shouldn't use this one as the input variable.

  When we just use tfidf matrix generated from the post title as input, we can get a more reasonable model with high precision for the hot label. If we regard the post with more than 10 comment as the hot post, the average fscore of KNN can be 76.31%. If we use top 25% as the hot post, the average fscore of KNN is 0.5573. But the precision of hot label is 58.94%.

Although, the accuracy of these models is not good enough, the precision for the hot label is better. And for all the models using Linear Discriminant Analysis and Logistic regression, they don't perform well. Maybe it is because that there is not a strong linear relationship between the news title and whether it will be hot or not.

5. Conclusion and future work

We've found the top topics users are willing to read and discuss in 2018 and 2013, people's attitude towards them, people's civilized level. The conclusions are:

- The top ten topics in 2018 are interesting news (meaningless random news and scandals), Arabic murder, president (Trump & Putin's relationship), weed legalization, bad religion, human job & life, politic news, Europe news, German/Merkel and bomb/attack/kill.
- The top ten topics in 2018 are interesting news (meaningless random news and scandals), Arabic murder, president (Trump & Putin's relationship), weed legalization, bad religion, human job & life, politic news, Europe news, German/Merkel and bomb/attack/kill.
- People's attitude towards posts are negative, but it turned more active in 2018 than 2013.
- People's civilized level is higher in 2018 than 2013.
- By using the prediction model, we can predict if a post will be hot. Then company can use them for advertising to make benefits.

There are still some shortages of our work:

- We didn't count the Reddit administrator's influence, cause some comment with negative or bad words may be removed. So, the data we scraped may not be significant. Then our result may be a 'modified' version.
- The language pattern is still demanding further study.
- As a conclusive research, our samples are not big enough, they only contain 10,000 posts in a single month of one year. We need to scrape more data from reddit, for example, 120,000 posts from a whole year. (that is, 10,000 posts for per month) Then we can find more valid results.
- The way we set the hot label can be improved. We can also the post's last time into consideration.
- Our prediction is based on the recent news. If there are some big news happen, it may influence the accuracy of out model.