Group Report
Cat People VS Dog People

Contents

1 Problem Identification

'Cat people' and 'dog people' are respectively used to describe people love cats or dogs in short, as putting specific tags to other people to make people into groups can lead things easier to understand. We get the idea from the Facebook's research paper[3] which talks deeply about the difference between those two kinds of people.

Based on their research material, we make an assumption that cat people and dog people have some differences in the respect of movie type preference, book type preference as well as the degree of activity. The objective of our project is to find out what differences are exactly. Besides, we also want to identify whether other aspects have some diversities, such as people's location and their sentiment to express on social media which beyond their research.

2 Data Collection

To find out what are the differences between these two groups in these aspects, we choose the largest cat people's group and dog people's group respectively. The number of members in cat group is 337,238 and the number in dog group is 156,360, which is enough for our analysis.

After that, we randomly pick 80,000 people from each group with location information to analyze. We also randomly pick 2,000 people from each group and extract the movie and book preference as well as the number of a follower, how many people they follow and how many groups did they visit recently.

To avoid the effect of culture environment, we get 2,000 users from Twitter to compare the distribution of the data to see whether culture will affect the result or not. The users we pick has been cleaned already and there is no overlapping between these two tables.

We also crawl tweets data From Twitter API. We search only cat and dog hashtags as keywords instead of using cat and dog directly to avoid some meaningless tweets which although have cat or dog words but not related to cat or dog at all.

In conclusion, for Douban users we use the group information to identify the cat people and the dog people, whereas for Twitter users, as there is no such information, we use hashtag information to identify the cat people and the dog people.

3 Movie and Book Analysis

3.1 Data Preprocessing

In this section, we define the cat people and dog people as people who join the corresponding group. So, we use SQL to remove the users who join both of the groups. The number of this kind of users is 12,663.

3.2 Feature Engineering

Firstly, we build the feature vector for each movie and book. The attributes for movie are 37 types of movie and the attributes for book are 3,840 types of tags. The value in the vector are the binary values. Then, we sum up each users' the feature vectors and do the normalization. Because users are chosen randomly, some of them put no movie or book preference information on Douban. For this kind of users, We set their scores of each attribute with evenly number ($\frac{1}{37}$ for movie and $\frac{1}{3840}$) to indicate that they have not particular preference. The feature vectors are put into the data frame of pandas. $M(u)$ is the movie vector for user u and $B(u)$ is the book vector for user u.

$$M_u = \frac{\sum movie\_vectors}{number\_of\_movie\_vectors}$$

$$B_u = \frac{\sum book\_vectors}{number\_of\_book\_vectors}$$

3.3 Data Mining

3.3.1 For Movie Data

Because the dimension is 37 and the table is sparse, to avoid the curse of dimensionality, we need to use SVD to do the dimension reduction. We choose the top 23 dimensions to keep 90% information of the matrix (the sum of these 23 eigenvalues is the 90% of the total sum of eigenvalues). The problem in this project is classification, so we try to use six widely used data mining algorithms to build a model, they are: Logistic Regression, Decision Tree, Naïve Bayes, SVM, K nearest neighbor and Boosting method. Because there is no other testing data, we need to split the data so that 20% data to be the testing data and 80% data to be the training data.

3.3.2 For Book Data

To find out what kind of books do they prefer, we use the user-books' tag matrix to do the analysis. To avoid the curse of dimensionality, we do SVD to the matrix to reduce the dimension from 3840 to 573 and kepp 90% of the information from the original matrix.

3.4 Movie Result

To have an intuitive idea about the data, we reduce the dimension to 2 to draw the scatter plot.

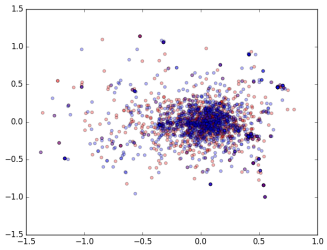


Figure 1: Movie picture

We can see from figure 1 that there is a high overlapping between red points and blue points which means they are not very different in 2 dimensions with 82.55% information lost from the original data.
After running the algorithms 10 times, the average accuracy of the results is listed below:
Logistic Regression: 0.51
Decision Tree: 0.54
SVM: 0.53
Naïve Bayes: 0.53
KNN: 0.52
Boosting: 0.51

3.5 Book Result

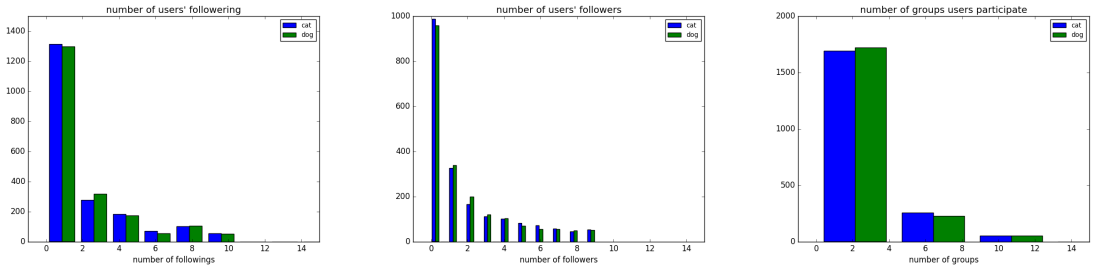
After running the same algorithms as in the Movie analysis part for 10 times, we get the average accuracy for each one to evaluate the performance. The results are listed below:
Logistic Regression: 0.52
Decision Tree: 0.52
SVM: 0.49
Naïve Bayes: 0.52

KNN: 0.51
Boosting method: 0.50

4 Activity Analysis

4.1 Douban Activity

First, we visualize the distribution of the data in the histgram format to have an intuitive idea. The distributions are shown in figure 2.



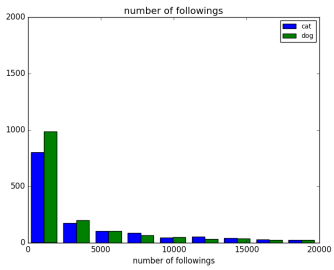
(a) Douban Following (b) Douban Followers (c) Douban Groups

Figure 2: Douban Activity

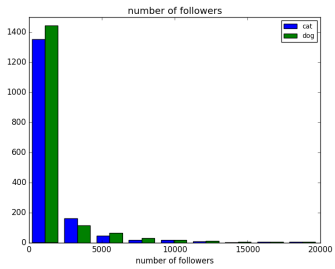
Note that the green bars are for dog and blue bars are for cat. From the graphs we can see that the two groups of people share the same distribution.
After that, we use the same data processing technighes as before, and we also use the six algorithms to test this data and the average accuracy of the results is listed below:
Logistic Regression: 0.55
Decision Tree: 0.50
SVM: 0.53
Naïve Bayes: 0.50
KNN: 0.54
Boosting score: 0.55

4.2 Twitter Activity

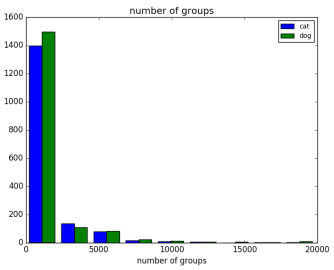
To avoid the effect of culture or communities, we also explore the data from twitter users. First, we draw histgrams once again to see the distribution of the data and compare them with the data from Douban. The graphs are shown in figure 3.



(a) Twitter Following



(b) Twitter Followers



(c) Twitter Groups

Figure 3: Twitter Activity

After that, we test the six algorithms on this dataset and the average accuracy are listed below:
Logistic Regression: 0.54
Decision Tree: 0.53
SVM: 0.50
Naïve Bayes: 0.50
KNN: 0.50
Boosting score: 0.52
With the results above, we can have a conclusion that the culture , which the community belongs to, won't make the cat people so different from the dog people, which is the result in the research from Facebook.

5 Province Analysis

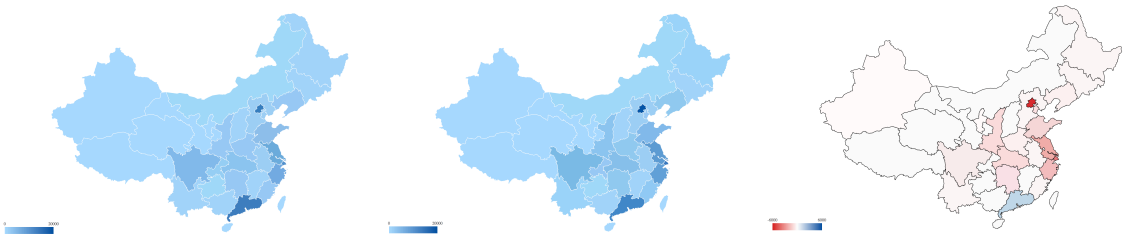
We want to identify whether people living different province have a specific interest in cat or dog, that is, whether the number of cat people and dog people is similar or not in geographically.

5.1 Data Processing

We have the data where each category has about 100,000 items, and each item means one user and his/her specific living city.
Firstly we group those data using its city, but there are still too many cities, so we decide to group those data by provinces. We find a JSON file which lists all the province names with its secondary cities, and write a python file to automatically deal with those cities. It's not easy work because some cities have the same name and some cities use abbreviations.

5.2 Data Visualization

After data processing step, we get the data about the number of cat people and dog people for each province and use d3 to generate the data. We also count the number of cat people minus the number of dog people for each province and try to find the difference between the two categories.



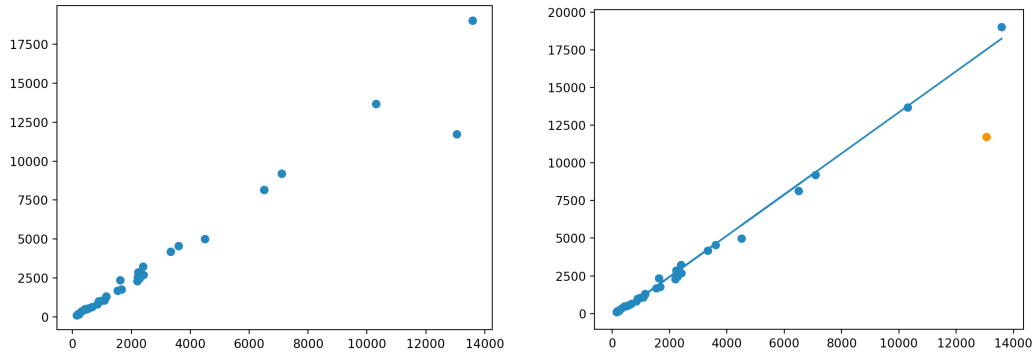
(a) The number of cat (b) The number of dog (c) The minus number

Figure 4: The province data

The result is in figure 4, the first two pictures mean the number of cat people and dog people in each province, and the last picture means the minus number (red means more dog people and blue means more cat people).

5.3 Linear Regression

It seems that those provinces have more cat people also have more dog people, which means that the two categories may have high correlation. So we use cat people's data as x axis and dog people's data as y axis and use linear regression, each node represent for a province.



(a) Dot data of each province (b) Linear regression

Figure 5: Guangdong is an outlier

Figure 5 can clearly show that Guangdong is an outlier, from the result we can see that guangdong has more cat people than predicted.

In conclusion, almost all the province has little more dog people, but Guangdong is different from other cities and has more cat people than predicted.

6 Sentiment Analysis

6.1 Data Processing

First, we crawl data From Twitter API. We search cat and dog hashtags as keywords instead of using cat and dog directly to avoid some meaningless tweets which has cat or dog word but not directly related to cat or dogs.
We then do some data cleaning works, we remove those not-words refer to CNN sentence code.[1], we also remove stopwords [2] and other non-words and just keep real words[3].

6.2 Feature Engineering

As each tweet is too short, we combine tweets with same hashtags together and regard it as one document. Then there are two documents, one is about cats and the other one is about dogs. Because the two documents have many same words, using TF-IDF is a good choice for analysis those words which are important for cats and dogs respectively.
Then we decide to group those words into some groups. Sentiment analysis dictionary is a popular way to divide words into two categories: negative and positive.
We choose the sentiment analysis dictionary[2]. The dictionary does not only have correct words, but also have many misspelled words, so it is more suitable to deal with social media content. What's more, the dictionary is highly correlated with the domain-specific orientations[4] and choosing a social media related dictionary is useful and necessary.
We count the weight of each category. After normalization, we generate the table below:

	cat	dog
positive	0.6	0.61
negative	0.4	0.39

6.3 Data Visualization

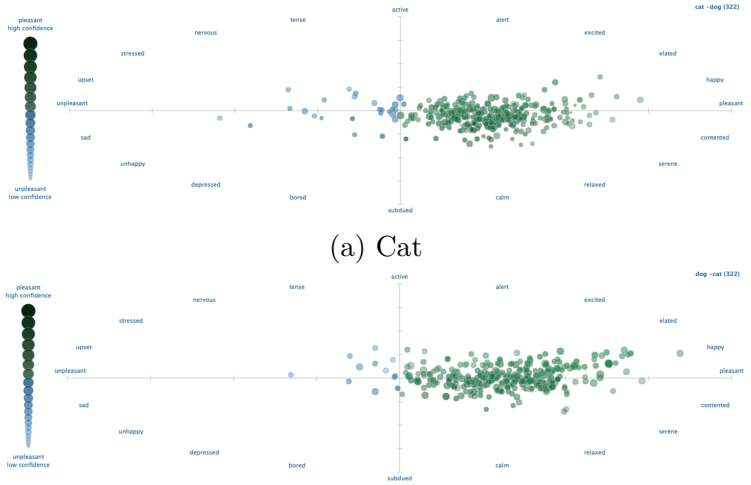
We also use some visualization tools, such as tag-cloud to visualize those words.

[1]Refer to: https://github.com/yoonkim/CNN_sentence/blob/master/process_data.py
[2]Refer to: https://github.com/amueller/word_cloud/blob/master/wordcloud/stopwords
[3]Refer to: https://www.nltk.org/book/ch02.html

(a) Cat (b) Dog

Figure 6: Tag-cloud picture show their sentiment

Figure 6 shows the sentiment of each document, green means positive color and red means negative color. The majority of words are neutral, which means they are not negative or positive, they are colored in grey.



(a) Cat

(b) Dog

Figure 7: Another sentiment analysis result

There are also some other visualization tools to analysis words, This project [1] can analysis sentiment into two dimonsions, and result can be seen at figure 7.

From the results above, we can see that although cat people seem a little more negative than dog people, they are both more positive than negative.

7 Conclusion

In this research, we want to find the difference between cat lovers and dog lovers, which are also be called as 'cat people' and 'dog people'. We crawl 'Douban' and 'Twitter' data respectively, to avoid the effect of culture or communities. Besides, crawling two social media platform gives us more types of data to analysis.

For 'Douban' data, we use their group information to define 'cat people' and 'dog people' as we believe that cat and dog lovers are willing to joining such groups which have many fellows. For 'Twitter' data, we define those 'cat people' and 'dog people' are people who are willing to send tweets with cat and dog hashtags.

From the analysis above, we can see that there is no model that can distinguish these two kinds of people in the respect of movie preference, book preference and degree of activity. Sentiment analysis also cannot distinguish this situation. From the view of the location, almost all the province has a little more dog people than cat people whereas Guangdong truly has more cat people, and we believe that Guangdong is unique and special among all provinces.

So, as far as we can know, the truth is that cat people and dog people are just so same and not as different as what has been mentioned in the research of Facebook.

The whole project can be seen at: https://github.com/hotokokoa/social-media-mining

8 Code and Data Instruction

8.1 Crawl Code

userIDCrawler.py: Crawl all the users with their user ids from each group as well as their locations and put the data into the database.

getDetails.py: Crawl the item id of his/her favorite books and movies for each individual user.

getFriends.py: Crawl the the number of fans, number of people each user follows and how many group they visit recently.

getTags.py: Crawl each movies' types and each books' tags and store them in the file system.

Tweet.py: Crawl tweets with cat and dog hashtags through twitter API.

8.2 Analysis Code

getMovieInfor.py, getBookInfor.py: Firstly, process movie and book data from douban to build feature vectors for movie and books. Secondly, build feature vectors about movie and

book preference for each user. Thirdly, run different algorithms on the matirx.
analyzeFollow.py: Build the feature vectors for activity information with 3 dimensions: following number, follower number, groups number. It also normalizes the matrix with the min-max normalization. Then, do the analysis to the normalized matrix.
tf-idf.py: Using TF-IDF to count the weight of each word in the cat and dog document and analysis the sentiment of cat and dog document.
city_data_process.py: Map each user's location to the province and count the cat and dog number of each province.
prov_analysis.py: Analysis the number of cat and dog of each province using linear regression.

8.3 Visualize Code

Tag_cloud.py: Draw tag cloud picture to visualize sentiment of cat and dog documents.
ChinaMap: Visualize cat and dog number and the difference of each province.

8.4 Data

All the data are in the data folder.
For the database file, CDPeopleDB.sqlit contains two tables, CatPeople and DogPeople, which contain all the users of the cat and dog groups and their geographic location information.

9 Feedback

9.1 Meaning

Q1. Why do you want to study this topic, and what is the purpose?
A: We first read a research from Facebook[3] which talks about the differences between cat people and dog people in some aspects such as the movie and book preference, marriage status and the number of friends based on the Facebook database which is not available to the public.
So we decided to try this idea on a famous social community 'Douban', where we can get users' information without any privacy issue, to see if we can prove the ideas in the research and not and try to find the specific differences between these two groups of people.
After realizing that 'Douban' data do not have clear differences, we decided to crawl data from Twitter to identify whether data from different platforms will influence the final result.

9.2 Define

Q2. If someone like both cat and dog, ow to define him/her? If someone joins both the cat and dog group, how to group this kind of users?
A: Because we want the difference between them to make user profile, we only extract people who like cat only and dog only and remove people who like them both.
Q3. How to classify cat and dog people? Do these people join the group really raising a dog/cat? How can you confirm that people join dog group as they like the dog more or their behavior represents they like a dog?
A: We choose the largest cat and dog groups in 'Douban' and define members in these groups as cat people and dog people because we consider that people who are willing to join a community to have a discussion with other people should share hobbies. People who love to talk about cats and dogs should be cat and dog people no matter having a real cat/dog or not. For example, people who love to talk about pop music doesn't mean they have to own pop music CDs. Besides, Douban is a place for people who share common hobbies to have discussions and comments. There is no reason that people who don't like dogs or cats but joining these groups and talk to other people. But if there are some people don't like cats or dogs but still join the groups, they are the noise information and we can ignore them because of the total number of members in each group is very large, up to 150, 000. We believe that the noise information should have only a small amount.
Q4. How to select the comparison criteria? I.e. books, movie, etc.
A: We get the idea from the Facebook research which indicates that cat people and dog people have different movie and book preference and number of friends. That is why we choose to get the users' movie and book preference and friends number as the feature of each user.
Q5. Do you try if there are more than two type people (cat and dog)?
A: There are billions of people in the world and each one is a unique one, so there should be countless types of people in the world actually. But for this topic, we only focus on the cat group and dog group.

9.3 Analysis

Q6. Which attributes are used in this analysis?
A: We collect top 100 movies and books he or she like the most for each user, put them together and do normalization to make the feature vector. In detail, we use the types of movies (37 kinds of movie types in total) and tags for books (over 3000 kinds of tags) to analyze their movie and book preference.

To analyze their activity from douban and Twitter both, we use three attributes: number of followers, number of followings, and number of groups he or she visits recently.
We also collect tweets with cat and dog hashtags from tweeter, and we regarded those tweets are about users showing their interest to those pets.
Q7. Is the data depends on the culture? (Facebook / Douban)
A: Actually not. We did additional analysis on the Twitter users and find that they have the same performance as users in Douban. Because the user profile is not available on Facebook, we cannot test our ideas on it.
Q8. Studies of people's gender, leisure time, marital status, and whether sharing a residence with others, etc. can efficiently improve the accuracy of the results. But these data involve personal privacy. If only considering the three aspects of this group, the results are not very convincing.
A: First of all, too many attributes may cause the curse of dimensionality. Choosing topic related attributes only can be helpful to the study. On the one hand, some information is private that we cannot get from the Internet even though they might be very helpful. On the other hand, what we want to study here is only the three aspects of these two groups because of the Facebook research we read, and we don't know what we will find exactly in the first place.

9.4 Content

Q9. Can the same algorithm be applied for analyzing with other feature except for sentiment, book and movie features? Can we categorize people into cat and dog by using their similar behavior?
A: There are many other useful attributes in the world that might be very helpful for the analysis but not available to the public. We have gained as many attributes as we can to make the model works. To distinguish two kinds of people, we need their different behavior instead of their similar behavior. The similar behavior only makes them looks the same and unable to find the difference.

References

[1] Healey and Ramaswamy. *Visualizing Twitter Sentiment*. URL: `https://www.csc2.ncsu.edu/faculty/healey/tweet%5C_viz/`.

[2] Minqing Hu and Bing Liu. "Mining and summarizing customer reviews". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2004, pp. 168–177.

[3] Amaç Herdadelen Lada Adamic Moira Burke. *Cat People, Dog People*. Aug. 2016. URL: `https://research.fb.com/cat-people-dog-people/`.

[4] Bing Liu. "Sentiment Analysis and Subjectivity." In: *Handbook of natural language processing* 2 (2010), pp. 627–666.