Data Science Portfolio

Wenhang Bao
204 W 108th ST | New York, NY 10025 | (917)714-6682 | wb2304@columbia.edu
Github: https://github.com/HolyMonarch Linkedin: https://www.linkedin.com/in/wenhangbao

This is my data science portfolio where I present some results from my previous projects. I reproduced these project independently in a different way. If I previously used R, I would use Python in this portfolio, except for those functions that could not be substituted like CNN in Python.
These projects represent my skills, but not all of them. My skill set is well listed in my resume.

1. Image Classifier

1.1.Abstract

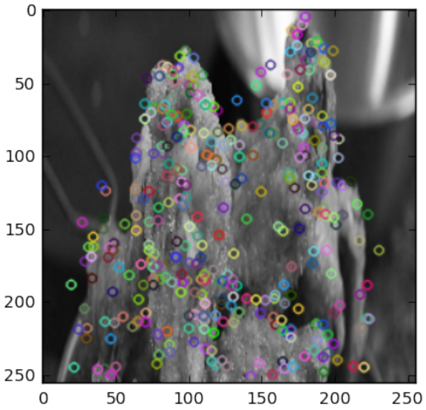
The dataset is a set of 2000 images of poodle and fried chicken. The objective of this project is to create an AI program that accurately distinguish pictures of poodle dogs from pictures of fried chicken. The portability of this AI program (holding storage and running memory cost) and the computational efficiency (test running time cost) are of great concern. So the trade-off between training time and accuracy is of great significance.

Github: https://github.com/TZstatsADS/Fall2016-proj3-grp3
Portfolio:https://github.com/HolyMonarch/Data-Science-Portfolio/blob/master/image%20clasifier.ipynb

1.2.Methods

1.2.1. feature engineering

In the previous project, the instructor provided the SIFT features. But as feature engineering is usually the first step of a data science project, I reproduced it by using SIFT algorithm to extract features from these images. After I got over 700,000 descriptors, I selected k-means clustering as the dimensionality reduction strategy. I clustered these features into 5000 groups and use bag-of-words methods to obtain the final SIFT features.



1.2.2. Model

I used SIFT features and different Machine Learning models to build the AI program. In the course, we built SVM, random forest, GBM, etc. with R. We used GBM with SIFT features as our baseline model.

In this portfolio, I constructed only Random Forest model with Python as an illustration, but I used the result we obtained in the course.

After tuning model parameters, we compared the performance of ML models (SVM, Random forest, GBM, CNN) and select CNN as our final model.

Classifier	GBM	SVM	RanForest	CNN
Training Time(s)	7000	1100	1400	240
Accuracy(%)	67	64	64	94

1.3.Results

We successfully built model which could improve the accuracy to 94%, which also could be better with more sophisticated deep learning structure. The time limitation is a problem. Feature engineering and parameters tuning take 70% of the total time. The trade-off between time and accuracy is of great significance not only because the requirement of this assignment which asked us to shorten the test time, but also because that even though we have enough time for this project, we will not have so much time allows us to tune parameters repeatedly when we are facing a real industry task.

2. Trees in New York

2.1.Abstract

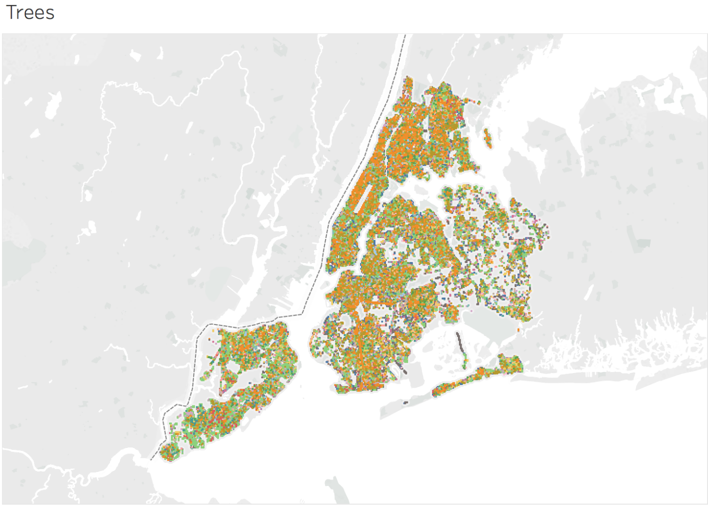
The objective of this project is to create an application which could help tree lovers to understand and locate all kinds of trees planted besides the street all over New York City. The data were downloaded from NYC open data with over 460,000 records. This application also provides statistical analysis of problems these trees are facing.

App: https://adsproj2group5.shinyapps.io/shiny/
Github: https://github.com/TZstatsADS/Fall2016-Proj2-grp5

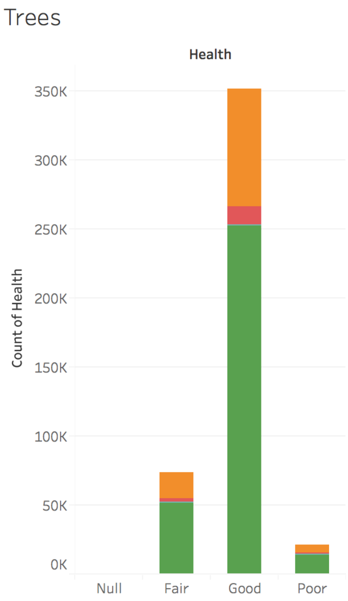
2.2.Methods

2.2.1. Tableau for EDA

I used Tableau for basic data analysis. It's of efficiency and accuracy. With the data being mapped in the graph, designer could better understand our application then design the UI site. I used tableau to understand the approximate number of each tree or analyze the proportion of different problems.



Map based on Longitude and Latitude. Color shows details about Spc Common. The view is filtered on Spc Common, which excludes Null.



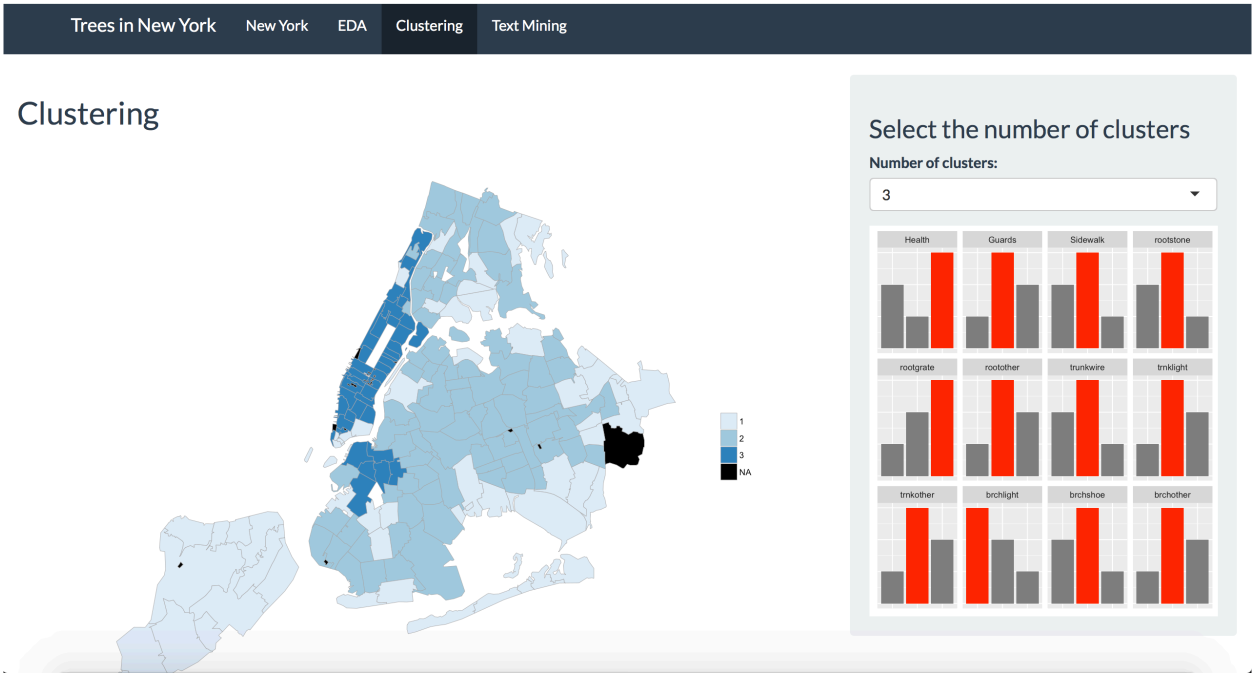
Count of Health for each Health. Color shows details about Steward. The data is filtered on Spc Common, which excludes Null.

2.2.2. RShiny

I used RShiny to develop the application. The application has 4 functions:

1. The distribution of trees in New York
2. Analysis of the negative effect
3. Clustering of geographical area
4. Text mining

So the user could use this application to know all the aspects of a specific kind of tree in New York like it’s distribution, whether if they are guarded, etc.



Clustering

Select the number of clusters

Number of clusters: 3

2.3.Results

Our application and statistical analysis can help tree lovers find one specific kind of tree. Also, it tells us the health condition of trees in different area. The clustering results told us that Manhattan and upper Brooklyn performs better than the rest area because these places provide better protection to the trees. That could help the government in city decoration and regularization.

3. Hillary Clinton’s email analysis

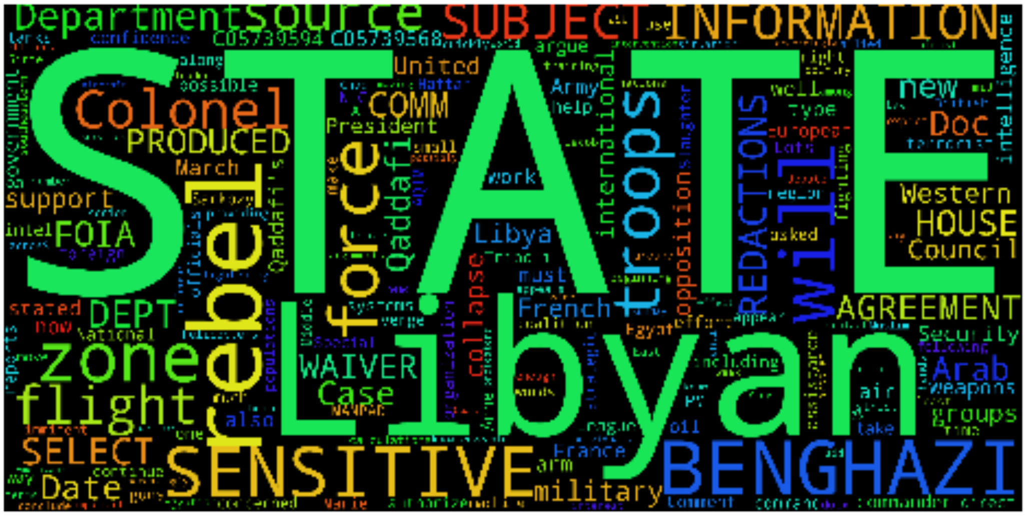
3.1.Abstract

Here I used the famous Kaggle dataset, Hillary Clinton’s email. Throughout 2015, Hillary Clinton has been embroiled in controversy over the use of personal email accounts on non-government servers during her time as the United States Secretary of State. The objective of this project is to understand the focus of her email and to categorize them.

3.2.Methods

3.2.1. Word Cloud

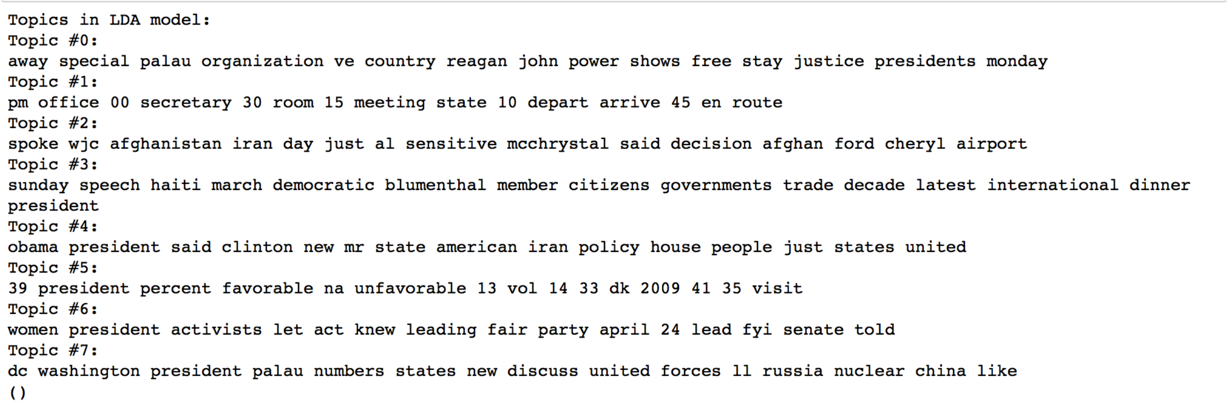
Here I extracted emails only related to president. Then I used Python package WordCloud to visualize the content of these emails, and try to depict keyword.



3.2.2. LDA model

I tried to fit LDA model to categorize these emails. We could notice that Topic 1 may be related with her meeting arrangement, Topic 2 may be related to international affairs, etc.

But also there are topics that make no sense like Topic 5.



```
Topics in LDA model:
Topic #0:
away special palau organization ve country reagan john power shows free stay justice presidents monday
Topic #1:
pm office 00 secretary 30 room 15 meeting state 10 depart arrive 45 en route
Topic #2:
spoke wjc afghanistan iran day just al sensitive mcchrystal said decision afghan ford cheryl airport
Topic #3:
sunday speech haiti march democratic blumenthal member citizens governments trade decade latest international dinner
president
Topic #4:
obama president said clinton new mr state american iran policy house people just states united
Topic #5:
39 president percent favorable na unfavorable 13 vol 14 33 dk 2009 41 35 visit
Topic #6:
women president activists let act knew leading fair party april 24 lead fyi senate told
Topic #7:
dc washington president palau numbers states new discuss united forces 11 russia nuclear china like
()
```

3.3.Results

Here I used some simple ways to analyze these text and get some insights about what the Secretary of State were doing. I could try like n-grams model in the future for further analysis.