Mobile Application Development

Data Science Capstone Final Project

Prof. Wright

Phoebe Wu

May 1, 2019

George Washington University

I. Introduction

According to the research, the smartphone usage has almost reached 3 billion in 2018. Moreover, in late 2015, mobile web traffic surpassed desktop for the very first time. Due to the rapidly increasing market share of smartphone usage nowadays, mobile application development plays a significant role for smartphone users. Therefore, without a doubt, the smartphone application market is now highly competitive. So how to make an app downloaded by more users becomes an important issue for the smartphone app companies. Thus, I would like to do the analysis to see what makes an app downloaded by more people, which leads the top trending smartphone app, and what influence the app's rating the most. After getting the result, hopefully, I can give some advice to the app companies.

Since in the smartphone market, Android holds about 26.8% of the smartphone market, while iOS is 24%. Furthermore, according to the research, 75.33% of the mobile operating system market share is Android system and 22.4% for the iOS system. As a result, I decided to separate the project into two parts which include Google play app data for the Android system and Apple store app data for the iOS system. In the project, I used several kinds of Machine Learning Algorithms and Natural Language Processing (NLP) to find out the key points of developing the top trending and high rating apps in the Android and iOS mobile application markets.

II. Description of the Datasets

- Google play app data
  - App information
  - App reviews

The data is scraped from the Google Play Store in 2018. The app information dataset contains more than 10 thousand app information and different features, such as app category, rating review count, app size, installation count, price, version, and content rating (age group the app is targeted at). The app review dataset contains 37 thousand app reviews which are the first 'most relevant' 100 reviews for each app.

The link of the dataset: https://www.kaggle.com/lava18/google-play-store-apps

- Apple store app data
  - App information
  - App descriptions

The data was extracted from the iTunes Search API at the Apple Inc website in July 2017. The App information dataset contains about 7200 Apple IOS mobile application details with 16 variables including app size, price, rating count, rating version, content rating, prime genre, supporting device number, screenshot number, and language number. The App description dataset contains the 7200 Apple App descriptions.

The link of the dataset: https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps

III. Description of the Project

There are two main parts of my project. The first part is using different Machine Learning Methods such as K Nearest Neighbors, Random Forest, Logistic Regression and Decision Tree to modeling the App Ratings and the App Installation number to analyze the features that affect both the rating and the number of installations the most.

The second part of the project is to use the Natural Language Processing (NLP) based on the Apple App descriptions and the Google Play App user reviews. On the one hand, for the user reviews, I found the most frequently used words in all reviews, the positive, neutral and negative reviews, and also the reviews for the top 500 installation count. In addition, I also did the classification of positive, neutral and negative reviews. On the other hand, for the app description, I selected the top 500 rating count app and the top rating apps (rating higher than 4.0) to find the most common words in the app descriptions. From these frequent words, I want to find out what are the users' focus while using the mobile apps and what are the app developers want to emphasize most of the app. So that I can find out whether the developers' emphasis match users' need.

IV. Experimental Setup

- Classification

The Model for classifying and predicting the number of the trending app (installation/ download number) and the high app rating.

➢ Google Play App

After pre-processing the dataset, I used correlation heatmap to see the relationship between each feature. Moreover, I also did the feature importance to select the best features for the classification models. First, in order to do the classification, I converted the original rating which are the continuous numbers from 1.5 to 5.0 (*Figure 1*) into the categorical feature (every 0.5 as a category) (*Figure 2*). I tried Random Forest and KNN. However, the accuracies were both around 0.60 which I think were not quite good enough. Therefore, I assumed that rating that higher than 4.0 are high rating app and changed the rating target into 0 (the rating lower than 4.0) and 1(the rating higher than 4.0). For the machine learning methods, I used the Random Forest Classifier, k nearest neighbors, Logistic Regression and Decision Tree.

On the other hand, for classifying the installation count (*Figure 3*), I tried Random Forest Classifier, k nearest neighbors and Decision Tree. Nevertheless, the F1 score was around 0.15 so I arranged the installation count into new categories (*figure 4*) and add a new column to do the analysis.

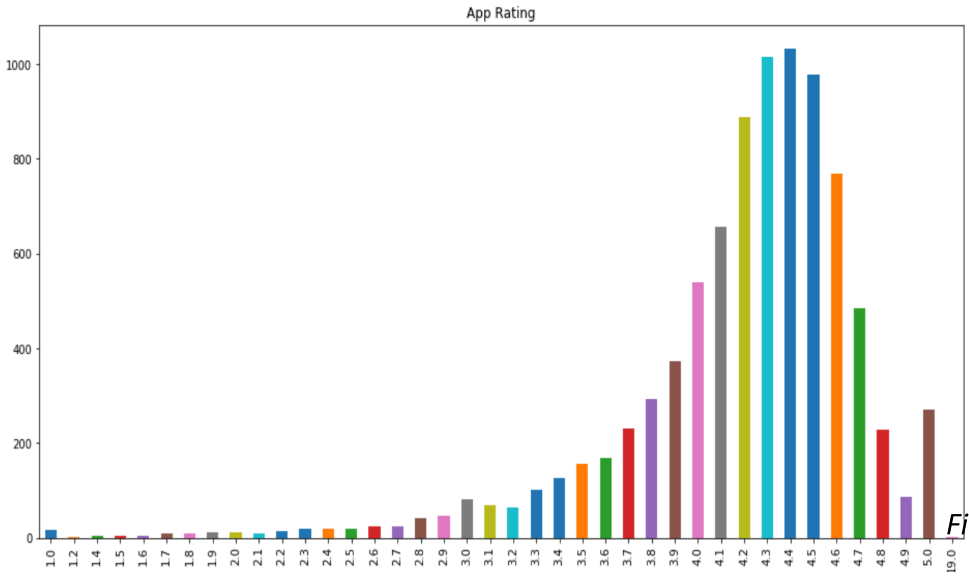


Figure 1: Google App Rating

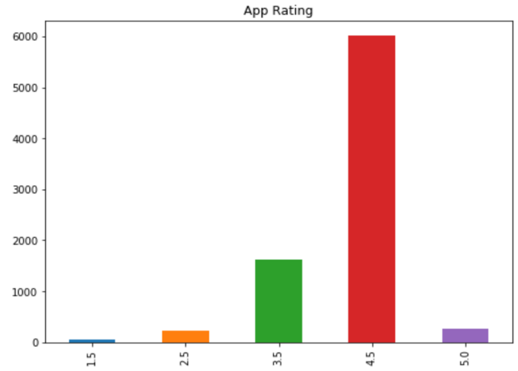


Figure 2: Google App Rating (Categorical)

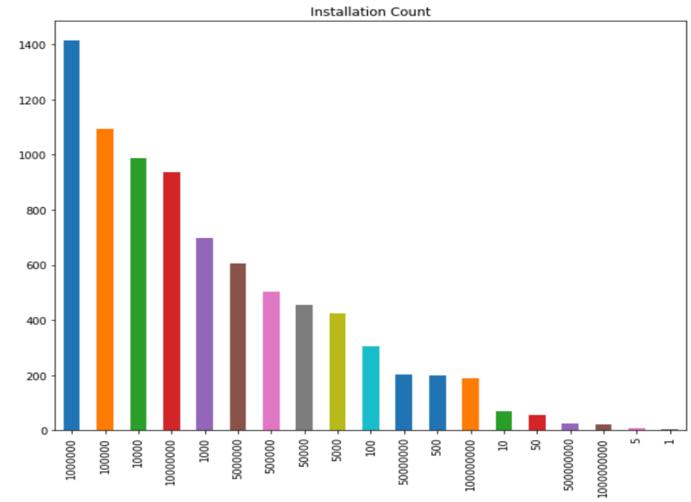


Figure 3: Installation Count

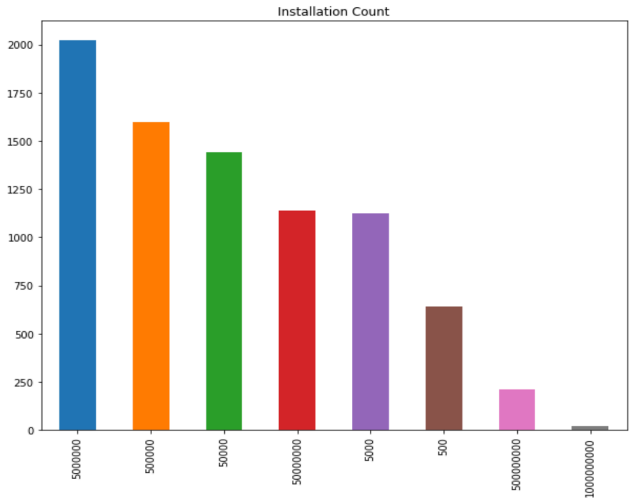


Figure 4: New Installation Count

- Apple Store App

For the IOS App, except the correlation heatmap and feature importance, I also tried the PCA Feature Selection and Cross Validation in order to get better accuracy. First of all, I tried the original ratings which were the numbers from 1.0, 1.5, 2.0 until 5.0 (*Figure 5*). but the accuracy was only 0.35. As a result, I also converted the rating target into 0 (the rating lower than 4.0) and 1(the rating higher than 4.0). For the machine learning algorithms, I used the Random Forest Classifier, k nearest neighbors, Logistic Regression and Decision Tree.

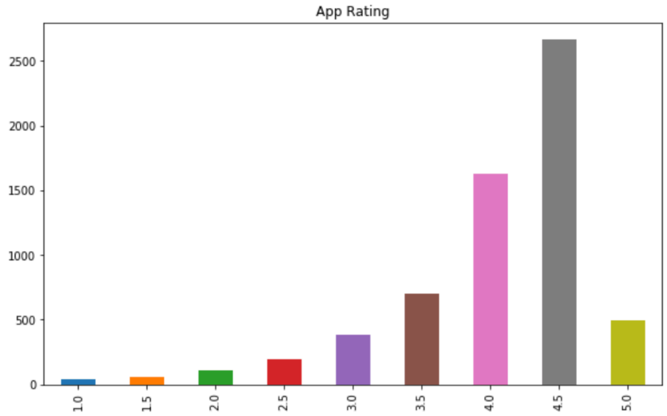


Figure 5: Apple App Rating

Moreover, because there is no installation count but only the reviews count in the Apple App dataset, I used the review count (*Figure 6*) to represent the installation count and add a new feature, rating count before (total rating count – current version rating count), to classify the trending apps.

- Natural Language Processing

Basically, I used the NLTK package in this project. I preprocessed the reviews, including removed the punctuations, got the bag of words, lowered the case and removed the stop words, and finally, found out the most frequent words to do the analysis.

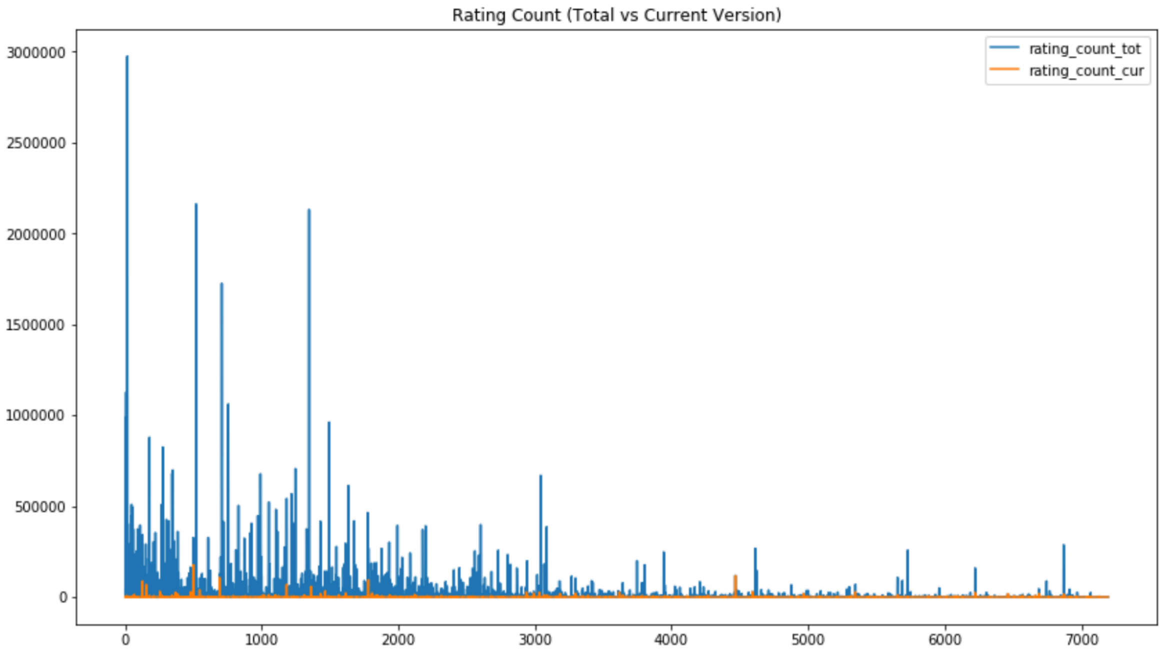


Figure 6: Rating Count

➢ Google Play App – App Reviews

For the analysis of the app reviews, I focused on finding the most frequent words that appear in the reviews to realize that what users may care the most while they are using the apps. I separated the dataset into all app reviews, the positive, neutral and negative, and the top 500 installation apps' reviews to do the analysis.

Additionally, I also did the positive, neutral and negative classification for the reviews. After preprocessing the reviews, I used the CountVectorizer in Scikit-learn package to convert texts into encoded vectors and transform the vectors back to arrays by using toarray() for the classification methods, Random Forest Classifier, KNN and Decision Tree.

- Apple Store App – App Descriptions

For the analysis of the app descriptions, I focused on finding the most commonly used words in the descriptions to see what are the most app developers' emphasis. Furthermore, I wanted to know if the developers' emphasis and the users' concerns matched.

I separated the dataset into all app descriptions, the top 500 installations of apps' descriptions, and the top rating apps' (higher than 4.0) descriptions to do the analysis.

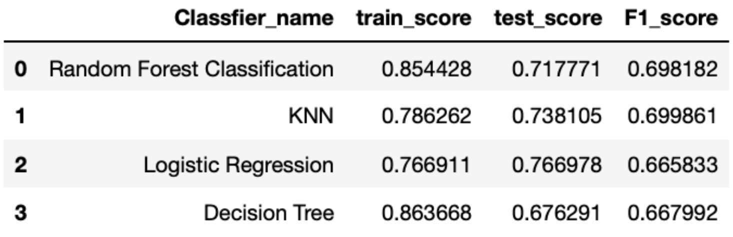
V. Result

- Google Play App

  - App information

- App Rating Classification (rating >= 4.0 & rating < 4.0)

For the Google Android App, I used Random Forest Classifier's feature importance and found out the three features, 'Size', 'Category' and 'Price', are the most important features for the Google app rating. I tried Random Forest Classifier, KNN, Logistic Regression, and Decision Tree. According to the F1 score, Random Forest Classifier and K Nearest Neighbors have the best result (*Figure 7*).

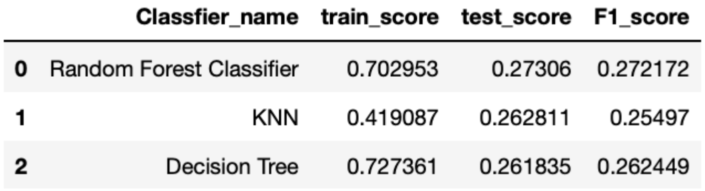


	Classfier_name	train_score	test_score	F1_score
0	Random Forest Classification	0.854428	0.717771	0.698182
1	KNN	0.786262	0.738105	0.699861
2	Logistic Regression	0.766911	0.766978	0.665833
3	Decision Tree	0.863668	0.676291	0.667992

Figure 7

- App Installation Number (*Figure 4*) Classification

After the feature importance, the result showed that 'Size', 'Category', 'Price' and 'Content Rating' are the best combination for the installation classification. I tried Random Forest Classifier, KNN, and Decision Tree. The results are about 0.26 (*Figure 8*) and the accuracy did not improve even after using the cross validation and PCA.



	Classfier_name	train_score	test_score	F1_score
0	Random Forest Classifier	0.702953	0.27306	0.272172
1	KNN	0.419087	0.262811	0.25497
2	Decision Tree	0.727361	0.261835	0.262449

Figure 8

- App reviews

  - Most commonly used words

In all reviews, the most frequent words are ‘game’, ‘news’, ‘free’, ‘ads’, ‘easy’, ‘light’, ‘weather’, ‘English’ and ‘baby’ (*Figure 9*). From these words, I assumed that game, news, weather are the three popular categories in Android application. Moreover, free or not, ads disruption, easy to use and the English version are also important for users.

Furthermore, I separated the Positive, Neutral and Negative reviews in order to obtain more details about the concerns from the users, especially the negative reviews. In the plot (*Figure 10*), except the emotional words, ‘time’, ‘ads’, ‘useless’, ‘slow’, ‘money’, ‘pay’ and ‘waste’ are the most commonly appeared words in the negative reviews. From these words, I infer that time consuming, app functions, app speed and money costing are the most concerns while users are applying the apps.

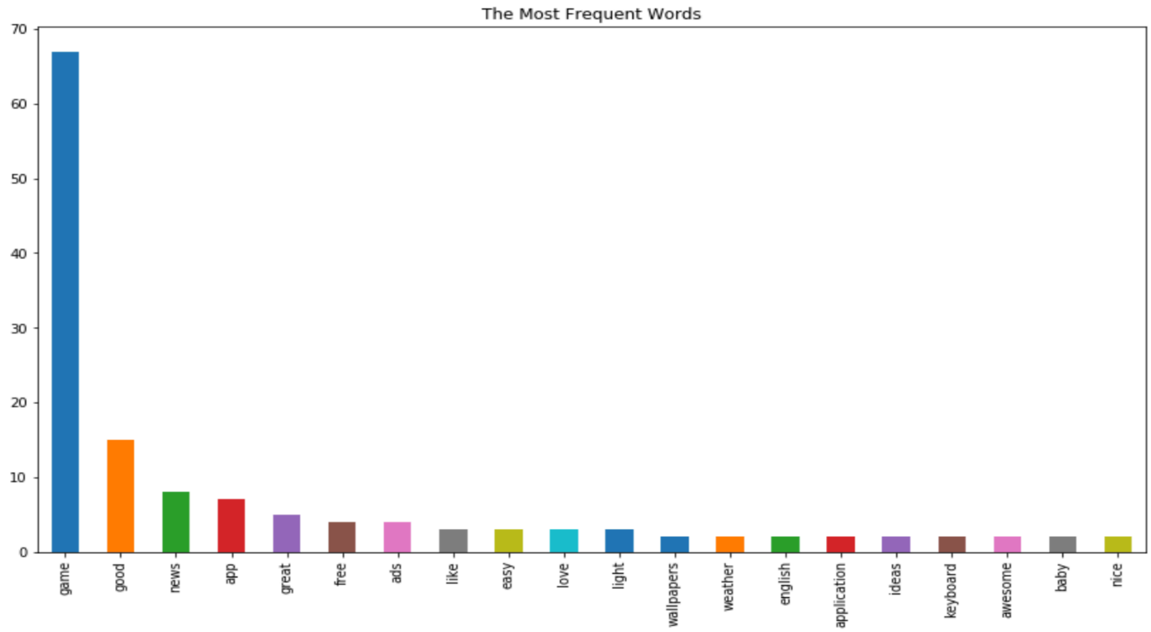


Figure 9: The most frequent words in all Google App reviews

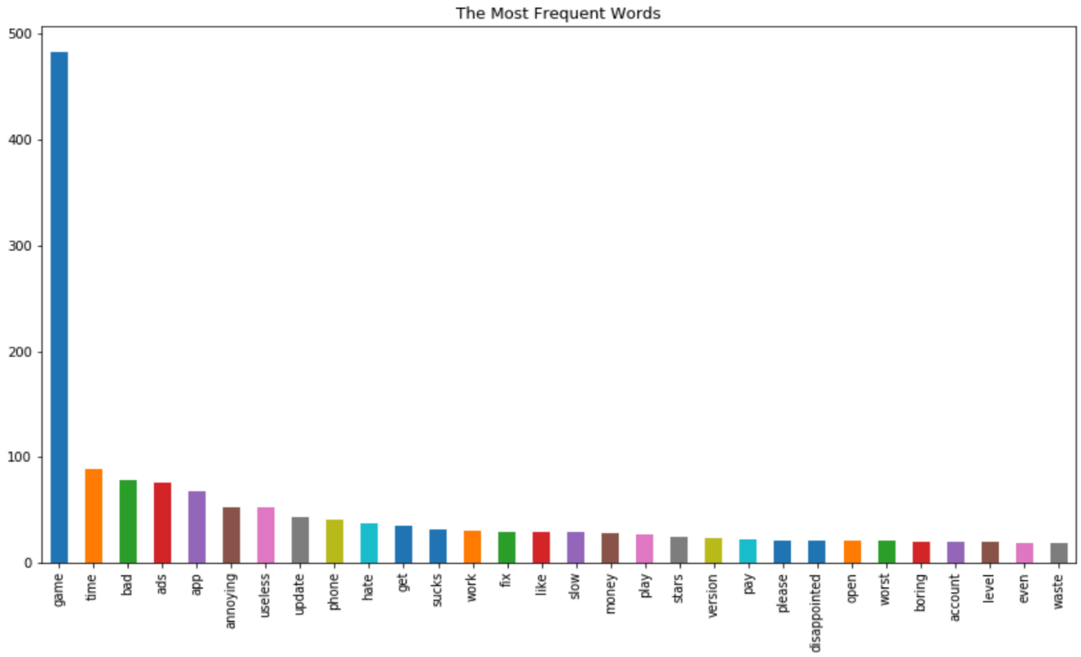
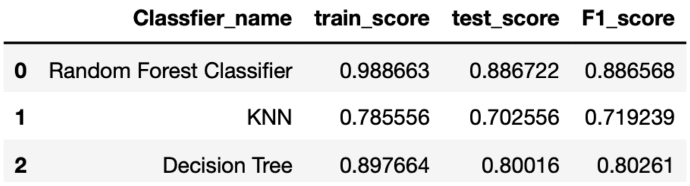


Figure 10: The most frequent words in Negative reviews

- Positive, Neutral and Negative Classification

For the review classification, after the preprocessing, I try Random Forest Classifier, KNN, and Decision Tree. In these three models, Random Forest Classifier shows the best F1 score, 0.88 (*Figure 11*).

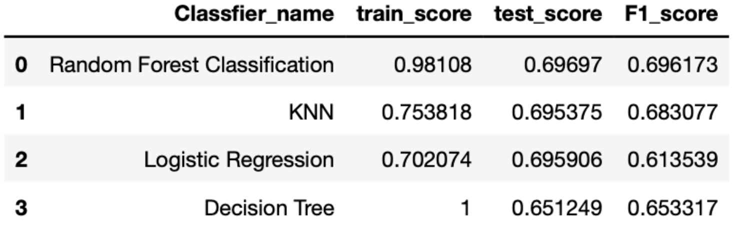


	Classfier_name	train_score	test_score	F1_score
0	Random Forest Classifier	0.988663	0.886722	0.886568
1	KNN	0.785556	0.702556	0.719239
2	Decision Tree	0.897664	0.80016	0.80261

Figure 11

- Apple Store App
  - ➢ App information
- App Rating Classification (rating >= 4.0 & rating < 4.0)

In the Apple App rating classification, after the feature importance process, I used 'size', 'prime genre', 'rating count before', 'support devices number' and 'language number' as the classification model features. I used Random Forest Classifier, KNN, Logistic Regression and Decision Tree algorithm. The results came out that Random Forest Classifier had the best F1 score, 0.70 (*Figure 12*).

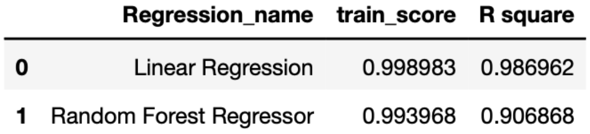


	Classfier_name	train_score	test_score	F1_score
0	Random Forest Classification	0.98108	0.69697	0.696173
1	KNN	0.753818	0.695375	0.683077
2	Logistic Regression	0.702074	0.695906	0.613539
3	Decision Tree	1	0.651249	0.653317

Figure 12

- App Rating Number (Figure 5: Rating Count) for the Top Trending App

Since the number of rating is a continuous variable, I used Linear Regression and Random Forest Regressor algorithm and ‘size’, 'rating count before', ‘content rating’, 'support devices number' and 'language number' as the model features. Linear Regression model had a better R square, 0.98, as the graph (*Figure 13*) below.



	Regression_name	train_score	R square
0	Linear Regression	0.998983	0.986962
1	Random Forest Regressor	0.993968	0.906868

Figure 13

➢ App descriptions

From the graph (*Figure 14*), we can see the words ‘game’, ‘play’, ‘Tabtale’ (a game app company), ‘music’, ‘photos’, ‘video’, ‘weather’, ‘time’, ‘free’ and ‘subscription’ are the most frequent words from all app descriptions. From these words, I think ‘game’, music’, ‘photos’, ‘video’ and ‘weather’ show the popular categories in app development and ‘time’ and ‘free’ indicate that users care about time consuming and the price.

Additionally, in the top rating (rating higher than 4.0) apps' descriptions, 'game', 'music', 'photos', 'video', 'weather', 'subscription', 'fun', 'privacy' and 'time' are the top most frequent words (*Figure 15*) which are very similar to the most frequent words from all app descriptions.

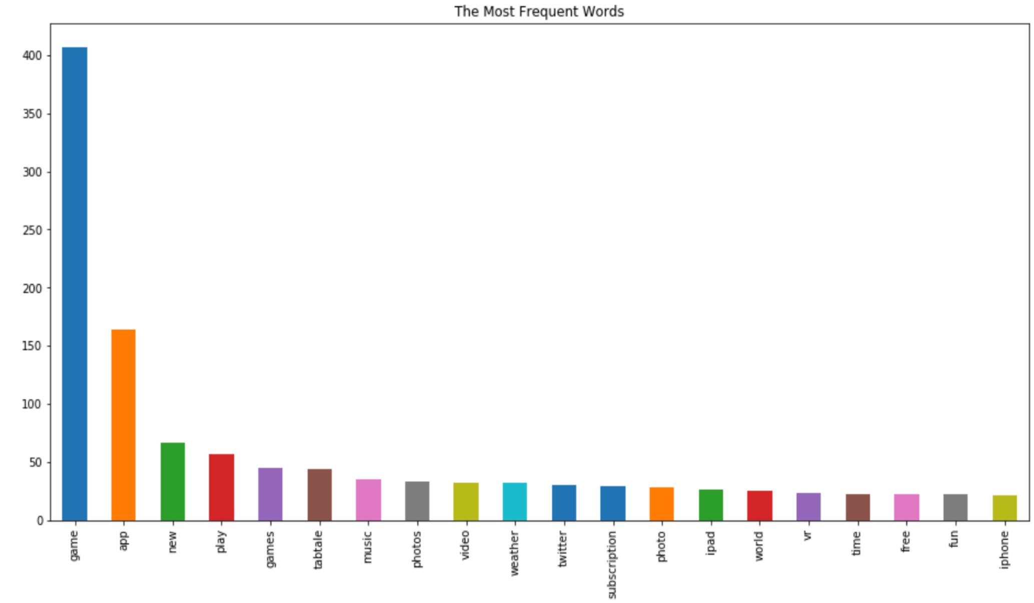


Figure 14: The most frequent words top 20 in all Apple app descriptions

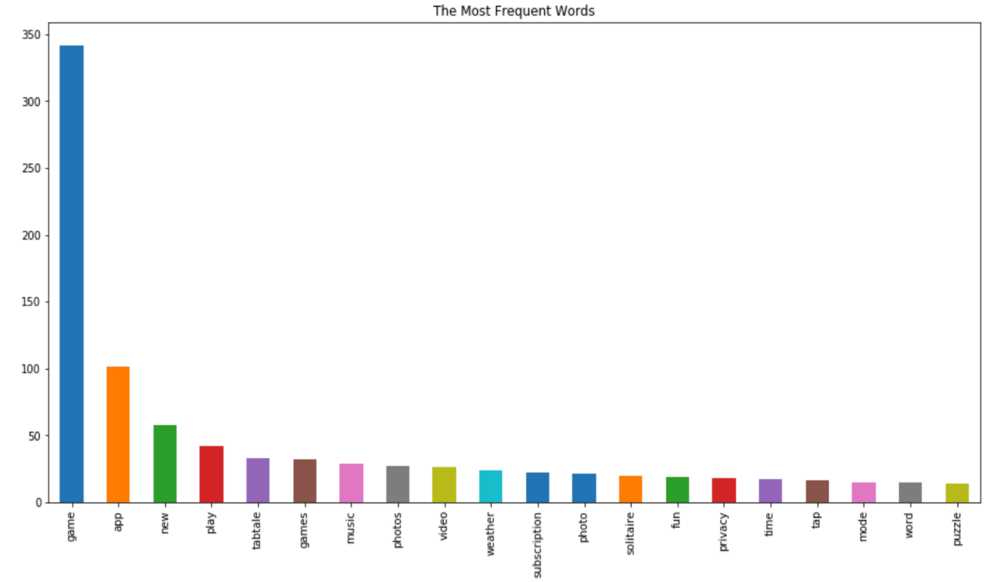


Figure 15: The most frequent words for the 'Top Rating (higher than 4.0)'

On the other hand, for the apps that are in top 500 rating count, the top commonly used words are ‘game’, ‘weather’, ‘solitaire’, ‘free’, ‘photo’, ‘subscription’, ‘music’, ‘news’, ‘friends’, ‘videos’ and ‘share’ (*Figure 16 & 17*). Most words are in the previous most frequent words except ‘friends’ and ‘share’.

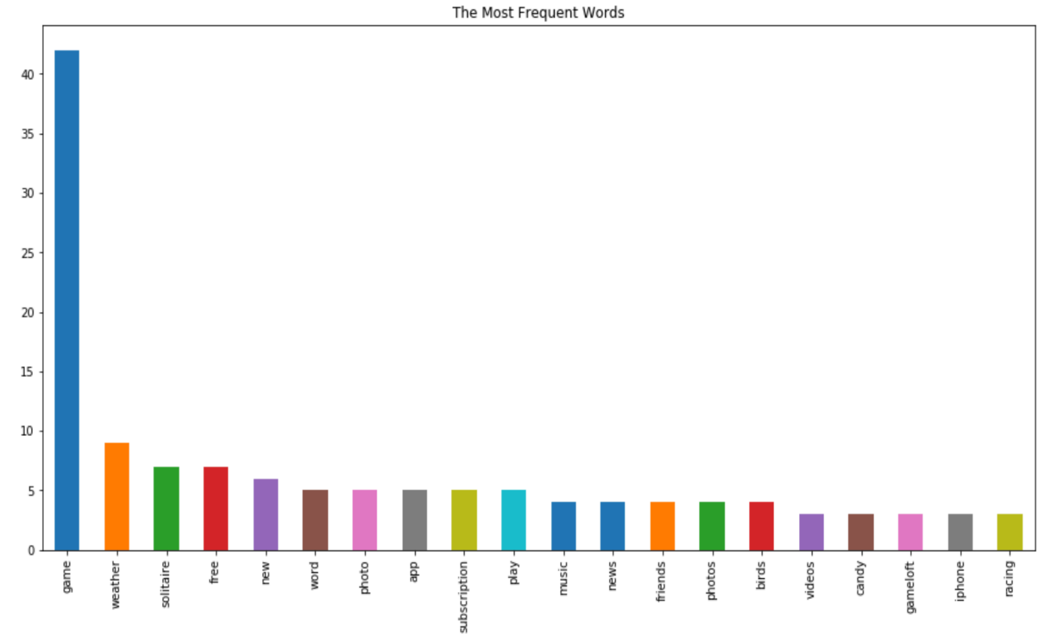


Figure 16: The most frequent words for the ‘Top 500 Rating Count’

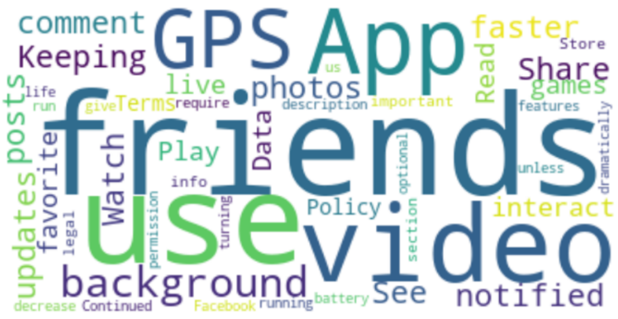


Figure 17: The word cloud for the ‘Top 500 Rating Count’

VI. Summary and Conclusion

According to my analysis, the influence of an app rating mostly depends on the app size, category, and price. For the Apple App, supporting devices and the number of supporting languages are also important features. From my point of view, this may due to the restriction of iOS Apps' applicable device and region while Google Apps are relatively global. For example, I am an iPhone user and some app that I used in my country are not able to use in the US. For the top trending app analysis, size, category and content rating are the most significate features for both Google and Apple App. Moreover, price is also crucial of leading a top trending Android app. The number of supported languages is, on the other hand a critical point for the top trending iOS app.

In the second part, app reviews and descriptions, of my project, 'game', 'news' and 'weather' are the most frequent appeared categories for both Google App reviews and Apple App descriptions, so I assumed that they are the most popular app genres for users and developers. 'Music', 'photos' and 'videos' which are also commonly mentioned in the Apple app descriptions may be the categories that app developers think are high demand in the market.

Besides, the price, ads disturbance, time consuming, easy to use, useful functions and English version are the most concern to the users. From the top rating app descriptions, except time consuming and the price, the app developers also focus on the subscription service, privacy, and fun. For the top trending apps descriptions, except free and subscription, 'friends' and 'share' come out in the top most frequent words. In my opinion, these two words may relate to the social media apps, Facebook, Twitter and Instagram which are the top 3 trending apps.

VII. Real Application & Future Improvement

According to the analysis, I suggest that Android App developers put more focus on apps' size and price, improving the operating speed and make the app easy to use. For Apple iOS developers, except the size and price, the supporting device and language of apps may be important for the higher rating and more downloads.

In the future, my analysis can be improved and more useful by keep tracking the apps and collecting more information of apps, for example, more app derails, the users' reviews and the difference between versions. Furthermore, from my point of view, the top trending apps analysis may not only require more market survey and longtime investigation for more information and the realization of social requirements.

Reference

1. Betsy McLeod (2018, Oct. 30) 75+ Mobile Marketing Statistics For 2019 And Beyond, Retrieve from https://www.bluecorona.com/blog/mobile-marketing-statistics

2. Dean Takahashi (2018, Sep. 21) Newzoo: Smartphone users will top 3 billion in 2018, hit 3.8 billion by 2021, Retrieve from https://newzoo.com/insights/articles/newzoos-2018-global-mobile-market-report-insights-into-the-worlds-3-billion-smartphone-users/

3. James Le (2018, Jan. 20) A Tour of The Top Ten Algorithms for Machine Learning Newbies, Retrieve from https://towardsdatascience.com/a-tour-of-the-top-10-algorithms-for-machine-learning-newbies-dde4edffae11

4. Mehul Rajput (2017, April 21) Top Trending Mobile App Categories in 2017, Retrieve from https://www.mindinventory.com/blog/top-trending-mobile-app-categories-in-2017/

5. Statcounter GlobalStat (2019, Mar) Mobile Operation System Market Share Worldwide – March 2019, Retrieve from http://gs.statcounter.com/os-market-share/mobile/worldwide