COMP8430 - Data Wrangling – 2019
Data wrangling project
u6551675

1. Data sets, problem description, and overall strategy

1.1 Data sets, problem description

Android is one of the two most popular operating systems. While iPhone users may purchase and download applications from Apple App Store, Android users are equipped with Google Play Store. Although there are many researchers doing research with Apple App Store data (since there are lots of public App Store data sets), fewer researchers do research on Google Play Store's application due to the difficulty of data scraping.

Personally, I'm interested in the correlation between user's review and application's features. In other words, the factors which may affect user's attitude towards apps - frequently updating, application's size and so on. Hence, this report will focus on the linkage between user's review towards different applications and the features which may affect user's review.

1.2 Overall strategy

Firstly, we need to explore the data. We need to know the meaning of each data set, how many records are collected, how many features are extracted, and the meaning of these features.

Secondly, we need to assess data quality and clean the data. Duplicate data need to be removed, while missing values need to be handled with respect to specific conditions. Outliers and unexpected values also need to be considered.

Thirdly, data linkage is required. We need to figure out how to merge these data sets. In this case we take application name as the key and merge data sets by intersection. In other word, record that only occurs in single data set would be dropped.

Fourthly, transformation should be applied. As it would discuss later, in this case, most features are in string format. Since we are trying to figure out the correlation between different factors and user review, data transformation may help us figure out the correlation between them.

Finally, we'll build correlation table to discover the correlation among these features.

2. Data description and data exploration

2.1 Data description

Both data sets can be accessed from https://www.kaggle.com/lava18/google-play-store-apps. The data was collected by Lavanya Gupta, an Indian software engineer. It was created on Sept 5th, 2018 and updated on Feb 4th, 2019. The purpose of scratching these data is to challenge the sophisticated modern-day techniques using JQuery, which is different from iTunes App Store and make scraping more challenging.

2.2 Data exploration

2.2.1 googleplaystore.csv

googleplaystore.csv is a csv file which describe. There are 10,841 records and 13 different features as shown below.

No	Name	Type	Description
1	App	string	application name
2	Category	string	category the app belongs to

3	Rating	float	overall user rating of the app
4	Reviews	string	number of user reviews for the app
5	Size	string	size of the app
6	Installs	string	number of user downloads/installs for the app
7	Type	string	paid or Free
8	Price	string	price of the app
9	Content Rating	string	Content Rating
10	Genres	string	an app can belong to multiple genres
11	Last Updated	string	date when the app was last updated on Play Store
12	Current Ver	string	current version of the app available on Play Store
13	Android Ver	string	min required Android version

Figure 1. googleplaystore.csv description

As we're going to explore the correlation between user review and application' feature, all features listed above are considered important. Therefore, I would keep all of them for later analysis. Due to the number of features and the limitation of the report length, I would explain some features here, together with their outliers and unexpected values – and leave the rest part to section 3, which would discuss data quality assessment issues.

1. App:

   Key attribute to merge two data sets.

   Duplicate values: There're 9660 unique values. In other word, there're duplicate records and we need to drop them later.

2. Category:

   Records are classified into 33 categories.

   Incorrect values: There is a record which is classified as '1.9'.

   Reason: Values are in the wrong attribute.

3. Rating:

   User's review might be affected by the rating given by Google Play. Rating should within 1 to 5.

   Missing values: There are 1,474 missing values

   Exception: There is a record which is classified as '1.9'.

   Reason: Values are in the wrong attribute. (Same record as we've mentioned)

4. Reviews:

   'Reviews' here refers to the total number of reviews.

5. Size

   Users might keep negative attitude towards applications which has a large size.

6. Installs

   With number of user downloads/installs for the app and the reviews, we can figure out how likely a user would write a review.

……

All 13 features are considered important and would be kept for later analysis.

2.2.2 googleplaystore_user_reviews.csv

googleplaystore.csv is a csv file which reflects user's attitude towards . It contains 64,163 records. There are 5 different features as shown below:

No	Name	Type	Description

1	App	string	application name
2	Translated_Review	string	user review
3	Sentiment	string	Positive/Negative/Neutral
4	Sentiment_Polarity	float	sentiment polarity score
5	Sentiment_Subjectivity	float	sentiment subjectivity score

Figure 2. googleplaystore_user_reviews.csv description

1. App:
   Key attribute to merge two data sets.
2. Translated_Review
   In this case, we use Translated_Review as a reference. Record would be dropped if the reviewer wrote nothing.
3. Sentiment
   Feature which directly reflects reviewer's emotion. One of our goals is to find the correlation between other features and sentiment.
4. Sentiment_Polarity
   Sentiment polarity score can be regarded as a result of data transformation.
5. Sentiment_Subjectivity
   Similar to Sentiment subjectivity score can be regarded as a result of data transformation.

All 5 features are considered important and would be kept for later analysis.

3. Data quality assessment

Data quality includes six core dimensions: completeness, consistency, uniqueness, validity, accuracy and timeliness. This section would

3.1 Completeness

a). googleplaystore.csv

No	Name	Missing values	Completeness
1	App	0	100%
2	Category	0	100%
3	Rating	1474	86.40%
4	Reviews	0	100%
5	Size	0	100%
6	Installs	0	100%
7	Type	1	99.99%
8	Price	0	100%
9	Content Rating	1	99.99%
10	Genres	0	100%
11	Last Updated	0	100%
12	Current Ver	8	99.93%
13	Android Ver	3	99.97%

Figure 3. Completeness of googleplaystore.csv

b). googleplaystore_user_reviews.csv

No	Name	Missing values	Completeness
1	App	0	100%

2	Translated_Review	26868	58.21%
3	Sentiment	26863	58.22%
4	Sentiment_Polarity	26863	58.22%
5	Sentiment_Subjectivity	26863	58.22%

Figure 4. Completeness of googleplaystore_user_reviews.csv

3.2 Consistency

The consistency of two data sets is not considered in this case. As it will be discussed later, one application may have multiple reviews from different users, and it's possible that user would leave same sentiment for different applications. Therefore we cannot calculate consistency of googleplaystore.csv and googleplaystore_user_reviews.csv.

3.3 Uniqueness

a). googleplaystore.csv

The only attribute which need to be considered is App. Since different applications may have same features, uniqueness of other features doesn't make any sense.

As it mentioned above, there are 10,841 records and 9,660 unique values. Hence, the uniqueness of application name is 89.11%

b). googleplaystore_user_reviews.csv

Uniqueness is not considered when it comes to user review table. One application may have multiple reviews from different users, and it's possible that user would leave same sentiment for different applications.

3.4 Validity

a). googleplaystore.csv

As it mentioned in 2.2.1, there is a record where values are in the wrong attribute. Hence the validity of App is 100%, while the validity of other features is 99.99%.

b). googleplaystore_user_reviews.csv

The validity of googleplaystore_user_reviews.csv's features is 100%.

3.5 Accuracy

a). googleplaystore.csv

As it mentioned in 2.2.1, there is a record where values are in the wrong attribute. Hence the accuracy of App is 100%, while the accuracy of other features is 99.99%.

b). googleplaystore_user_reviews.csv

The accuracy of googleplaystore_user_reviews.csv's features is 100%.

3.6 Timeliness

Timeliness is not considered in this case. We cannot identify whether a record is latest or not.

4. Data integration or record linkage

As it mentioned above, we're going to explore the correlation between user review and application' feature. Therefore, we need to build the linkage between user's review towards different applications and the features which may affect user's review.

There are lots of method which can be applied to merge two data sets. In this case, we use inner join to merge two data sets.

Instead of left join or outer join, we use inner join here because we need to make sure that records which only occur in single data set would be dropped. As we've mentioned above, the consistency of two data set is not 100%. Some applications only occur in

googleplaystore.csv, while some applications are only contained in googleplaystore_user_reviews.csv. If we use other methods, namely, outer join, then the new data set would look like this:

Key_A	A	…	NaN
Key_B	NaN	…	B
…	…	…	…

Suppose record begin with Key_A is an application that only contained in googleplaystore.csv, while record begin with Key_B is a review of application B that only contained in googleplaystore_user_reviews.csv. It's impossible to study the relationship between application and user, as either application's attribute or user's review is missing.

Therefore, we take the intersection of two dataset, so that record that only occurs in single data set would be dropped and we could study the correlations between two data sets.

The merged dataset contains 35,896 records, and 16 features in total.

5. Data preparation and cleaning

5.1 Data Cleaning

Data cleaning is applied before the merging of two data sets.

As we've discussed above, there are missing values and duplicate data.

Duplicate data need to be dropped at first. Then we'll handle missing values with respect to different conditions.

1. Drop data with missing values

   Data would be dropped if there's any important feature's value missing. Namely, the application name. Application name is required to merge data sets, hence we can only drop these data for later. Other attributes like Size also need to be handled in the same way.

2. Fill with 0

   Sometimes we can fill the data with a neutral value. For example, if the price is missing, we may regard it as free and replace the missing value with 0.

3. Fill with a certain value (Mean, Median.)

   Sometimes we can fill the data with a certain value. For example, if the rating is missing, we may replace it with mean, as the density of rating generally follows Gaussian distribution.

3. Leave empty

   Features like 'Current Ver', 'Last Updated' can only b. They are not that important compared with features like application name, and we cannot replace them by a certain value due to their format.

5.2 Data Reduction and Transformation

Reduction and transformation are applied after the merging of two data sets.

Firstly, we need to drop Translated_Review. Translated_Review is previously used to judge whether the user write a review or just rate but write nothing. Since we've extract user's sentiment, the original review is useless and should be dropped. Secondly, we turn strings into integers, and convert different classes into integers, so that we could build the correlation. Finally, we build the correlation table with the new data set.