Wrangle Report

December 30, 2018

Introduction

In this document is presented the implemented procedure to gather, assess and clean data accordingly with the Udacity Wrangle and Analyse Data project for the Data Analysis Nanodegree program. The steps follow for the implementation of the project can be summarize as Gathering of the Data, Assessing of the Data and Cleaning of the Data, each of these related with specific methods described in the Udacity program.

Gathering Data

The assessing of the data was straight forward, following the same procedure for al obtained datasets. The procedure was to visually inspect the datasets for any obvious problem, then to use pandas isna function to look for undefined values and then to look for inconsistencies about the columns data types using the info pandas function. Following these procedures allowed the finding of different quality and tidiness issues for all gathered datasets.

Nevertheless, as found it later when analysing the data, an important method was missed in order to implement a better assessing of the data, the use of the pandas describe function. It was not possible to identify a problem related with ratings numerator and denumerator because of that, a problem that can be appreciated when using the mentioned function.

Assessing Data

The assessing of the data was straight forward, following the same procedure for al obtained datasets. The procedure was to visually inspect the datasets for any obvious problem, then to use pandas isna function to look for undefined values and then to look for inconsistencies about the columns data types using the info pandas function. Following these procedures allowed the finding of different quality and tidiness issues for all gathered datasets.

Nevertheless, as found it later when analysing the data, an important method was missed in order to implement a better assessing of the data, the use of the pandas describe function. It was not possible to identify a problem related with ratings numerator and denumerator because of that, a problem that can be appreciated when using the mentioned function.

Cleaning Data

The cleaning of the data involved the use of pandas, numpy and bs4. Pandas was the main library used, but in specific circumstances was less complex to rely on other libraries instead of using

pandas. Moreover, the presence of html code in the gathered data caused the use bs4 library.

Most of the quality issues related with the gathered datasets were about the finding of undefined values or to irrelevant information on different columns, for example repetitive information in the same rows. Furthermore, there were cases when the row values were not bad but if changed it could increase their understanding when analysing the data. For example, changing the use of True and False for 0's and 1's.

The procedure mentioned in this section was implemented in order to solve the found problems described in the previous section.

Mergin Data

The merging of the data was implemented to combine all acquired information into a single dataframe. However, some additional quality issues appeared after that. This was because not all processed dataframes possessed the same information, even when all the merged data was related with the same Twitter IDs. Then, undefined values appeared and consequently the type of the column's values changed because of that, especially columns composed by integers values. As consequence, it was required to remove the undefined values and to change the types of the column values as before. Moreover, some columns were dropped, and the rest rearranged for the dataframe to be easily to read.

```
In [ ]:
```