HW3

Corpus Analysis

Bing-Je Wu

IST 736 – Text Mining

Table of Contents

[Introduction 2](#_Toc37975311)

[Analysis and Models 2](#_Toc37975312)

[About the Data 2](#_Toc37975313)

[Data Processing 2](#_Toc37975314)

[Results 4](#_Toc37975315)

[Conclusion 6](#_Toc37975316)

[Reference 6](#_Toc37975317)

# Introduction

Text mining can be differed into three different perspectives, information extraction, data mining and KDD (Knowledge Discovery in Databases) process. Text mining is to discover or extract information from different written resources by computers. Resources can be various, such as websites, books, emails, reviews, articles and audios. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. To get to know the pattern through statistical approach, it usually involves the process of structuring the input text from various resources to derive patterns from the structured data, and have the evaluation and interpretation of the output. The process of structuring the input text is also called corpus construction in text mining. And this process should be taken care patiently and carefully since it is the starting point of the text mining.

Analysis and Models  
Two libraries, pandas and regular expression, were used to convert the textual data into structured format and cleaned version of dataframe.

## About the Data

Two datasets were provided for the assignment. One is the sample data from Restaurants and restaurant reviews dataset (<https://drive.google.com/file/d/11H6AbWxKsPLY3yt__OrmK0rjjYShKhig/view?usp=sharing>). The other is the sample data from Movie Review Dataset (<https://drive.google.com/file/d/17nGHPsk4RXfvRoq-ndizTzc0_0PkzAqm/view?usp=sharing>). Both datasets are labeled but with text reviews separated into several columns. The restaurants dataset contains nine reviews (documents). Five reviews are positive reviews and four reviews are negative reviews. Each text review has been divided into several columns, including the ‘sentiment’ column, also called ‘label’, at the first column. For the movie review dataset, it has five documents. Two documents are positive reviews and the other three documents are negative reviews. Each of document has been split into several columns. The ‘label’ for each of review is located at the last column of each row.

*Table 1. Initial observation of restaurant review dataset and movie review dataset*

|  |  |  |
| --- | --- | --- |
|  | Restaurants Reviews Dataset | Movie Review Dataset |
| Documents | 9 | 5 |

Data Processing  
Initially, both datasets were read into python through the ‘read\_csv()’ function from pandas library. The initial dataframes were a bit messy. The restaurant review dataframe had six columns, including the ‘label’ column at the first column. There were nineteen missing values. For the movie reviews dataset, it had fifty-seven columns, including the ‘label’ column at the end of separated review columns for each row. Eighty-one missing values existed in the dataframe.

*Table 2. Observation of the initial pandas dataframe for restaurant review dataset and movie review dataset*

|  |  |  |
| --- | --- | --- |
|  | Restaurants Reviews Dataset | Movie Review Dataset |
| Rows | 9 | 5 |
| Columns | 6 | 57 |
| Missing value | 19 | 81 |

In order to vectorize the textual data, both datasets had to be converted into the formal structured format and cleaned. There were some steps need to be done before applying the ‘CountVectorizer()’ function from sci-kit learn package. First, the missing values must be taken care of. Then, the split text for each of reviews needed to be combined as one string per document. And, last, the string for each document had to been cleaned by removing unwanted spaces, apostrophe signs, back slash …etc. After having a thorough understanding of the both datasets, a summary and strategies were planned as the Table 3 shown below.

*Table 3. Summary and strategies for restaurant review dataset and movie review dataset*

|  |  |  |
| --- | --- | --- |
|  | Restaurants Reviews Dataset | Movie Review Dataset |
| Summary | * 9 instances * labels are located at first column for each row * text data contain newline * back slash and apostrophe have to be removed | * 5 instances * labels are located at last column for each row * text data contain newline * back slash and apostrophe have to be removed |
| Strategies | * fill NAs with spaces * use nested for loops for each row and use string method to join text in to one string for each row * extract label and review * use string method to get rid of leading and trailing spaces and apostrophe * rebuild a dataframe | * split ‘label’ column from the dataframe * fill NAs with spaces * use nested for loops to iterate each row and join text into one string for each row * clean the stored string as review * rebuild a dataframe |

In order to successfully convert the dataframes into desired format for the further analysis, two user-defined functions, ‘df\_string\_to\_list()’ and ‘clean\_str()’, were built in aid of the transformation pipeline. The ‘df\_string\_to\_list()’ function takes a pandas dataframe as input. It replaces the NA values with spaces. Nested for loops were implemented within the defined function to store a list of strings for each document. And the function returns the list of strings as an output. For the ‘clean\_str()’ function, it takes a string as input. The function removes the new line character, back slash, leading and trailing apostrophe and back slash. And finally, it returns a string as an output of the function. The Table 4 shows the detail of the defined functions as below.

*Table 4. Customized functions*

|  |  |  |
| --- | --- | --- |
|  | df\_string\_to\_list | clean\_str |
| Expect | A dataframe with separated text in different columns | An unclean string  *Ex: "'Mike\'s Pizza High Point \NY \nService was very slow and the quality was low.'"* |
| Modify | Fill the NAs with spaces and combine the separated text from each column for each row as a document | Remove new line and back slash; remove leading and trailing back slash and apostrophe |
| Return | A list of documents | A string |

# Results

With transformation pipeline built, both restaurant review dataframe and movie review dataframe were able to convert into the desired format, restaurant\_df and movie\_df. The Figure 1-1, 2-1, and 3-1 show the changes in different phases for restaurant review dataframe. The Figure 1-2, 2-2, and 3-2 show the changes in different phases for movie review dataframe.

**Figure 1-1:** Sample of raw restaurant review dataframe

A screenshot of a social media post

Description automatically generated

**Figure 2-1:** Sample of structured restaurant review dataframe before ‘clean\_str()’ function applied

**A screenshot of a cell phone

Description automatically generated**

**Figure 3-1:** Sample of cleaned restaurant review dataframe, restaurant\_df, with two user-defined function applied

A screenshot of a cell phone

Description automatically generated

**Figure 1-2:** Sample of raw movie review dataframe (57 columns in total)

A screenshot of a cell phone

Description automatically generated

**Figure 2-2:** Sample of structured movie review dataframe before ‘clean\_str()’ function was applied

**A screenshot of a cell phone

Description automatically generated**

**Figure 3-2:** Sample of cleaned movie review dataframe, movie\_df, with two user-defined functions applied

A screenshot of a cell phone

Description automatically generated

# Conclusion

A structured data with desired format is required for any analysis process. Building a pipeline to get a desired output from the messy data has become the routine tasks for any data scientist. Although it seems mundane and depressing, it can effectively ease the pain and reduce the workload on the analysis phase and modeling phase.

There are many tools and open sources that provide the information about building pipeline for data cleansing and data wrangling. Developing an efficient workflow for getting the data to the desired format has become the basic and important skill. There is no absolute answer for what the best approach is. Each person should have their own thought for cleaning and wrangling the data. That is the sign of being an independent data scientist in the industry.

# Reference

1 Restaurant Review Dataset. [Web Link](https://drive.google.com/file/d/11H6AbWxKsPLY3yt__OrmK0rjjYShKhig/view?usp=sharing) ;  
2 Movie Review Dataset. [Web Link](https://drive.google.com/file/d/17nGHPsk4RXfvRoq-ndizTzc0_0PkzAqm/view?usp=sharing) ;   
3 Text mining, from Wikipedia. [Web Link](https://en.wikipedia.org/wiki/Text_mining)