DATA CLEANSING

Collecting the raw data from the website <http://jmcauley.ucsd.edu/data/amazon/> . The Reviews dataset had a total of 24 files that were ‘gzipped’ and in the JSON format. The Ratings dataset had a total of 24 files that were in CSV format. These files were segregated by ‘Product Category’ i.e. 24 and they are as follows:

|  |  |
| --- | --- |
| **Product Categories** | |
| Amazon Instant Video | Health and Personal Care |
| Apps for Android | Home and Kitchen |
| Automotive | Kindle Store |
| Baby | Movies and TV |
| Beauty | Musical Instruments |
| Books | Office Products |
| CDs and Vinyl | Patio Lawn and Garden |
| Cell Phones and Accessories | Pet Supplies |
| Clothing, Shoes and Jewelry | Sports and Outdoors |
| Digital Music | Tools and Home Improvement |
| Electronics | Toys and Games |
| Grocery and Gourmet Foods | Video Games |

DATA MASSAGING

Abstraction:

* Function to GUnzip a File
* Function to return a Dataframe from a JSON - Gzip file
* Function to Calculate Character Count in the ReviewText
* Function to Calculate Word Count in the ReviewText

Data Munging:

* Reading the Reviews into DataFrames
* Adding the Product Type Column to each DataFrame
* Adding individual DFs to AmazonReviews DataFrame List
* Labelling: Add Column Names to Ratings Dataframe since it is missing the Header
* Reading the Ratings into DataFrames
* Adding the Product Type Column to each DataFrame
* Adding individual DFs to AmazonRatingsDataFrame List
* Use Value\_counts() to understand how many columns are of type object
* Printing Reviews and Ratings Dataframes
* Dataframe df\_AzReviews[23] and df\_AzRatings[23] Before dropna
* Dataframe df\_AzReviews[23] and df\_AzRatings[23] After dropna - Shows no change for Ratings row numbers
* Dataframe df\_AzReviews - running the dropna method on the list of DFs
* Dataframe df\_AzRatings - running the dropna method on the list of DFs
* Printing Reviews & Rating Dataframe Heads to check data integrity
* Printing Reviews & Rating Dataframe Tails to check data integrity
* Indexing: Setting the index for both dataFrame Lists i.e. Reviews and Ratings as 'reviewerID'

Sampling:

* For both datasets of Reviews and Ratings we would like to now take only a sample set of the entire DataSet. I have decided on a fraction= 0.5 i.e. 5 % of each Product category and appending them into one dataframe of Sample\_AzReviews and Sample\_AzRatings

to give a total count of 9020370 and 40368768 records respectively.

Feature Engineering:

* Adding reviewertext features such as character\_count and word\_count to the Reviews dataframe list
* Adding helpful column features such as helpful\_numerator, helpful\_denominator and helpful\_percentage to the Reviews dataframe list
* Adding time series features such as reviewYear, reviewMonth, reviewDate, reviewDayofWeek and ReviewWeekofYear to the Reviews and Ratings DataFrame lists.