Data Wrangling Process for Sentiment Analysis of Amazon Product Reviews Dataset

Capstone 1 Data Wrangling Report

Data Acquisition

For my capstone 1 project, I will be using Amazon product reviews of baby products, grocery food, and cellphone datasets. These review datasets are made available by Julian McAuley UCSD. The link to the datasets is http://jmcauley.ucsd.edu/data/amazon/ and the datasets are available by request. The datasets were downloaded as the JSON file and loaded into Pandas DataFrame in jupyter notebook.

Data Cleaning

After I load the datasets into Pandas DataFrame (df) I did the following steps to know and clean the dataset:

- df.shape and df.columns attribute to find out the number of rows and columns, and column names respectively
- Renamed some of the columns using df.rename() function.
- Counted the missing values using isnull() and info() methods
- Filled the missing values of the reviewer's name with 'No Name' using fillna() method
- Checked for duplicate reviews using duplicated() method and drop the duplicate values using drop_duplicates() method.
- Removed short reviews (less than 3 words)
- Removed non-English reviews using langdetect and nltk libraries
- Removed punctuations from the review text
- Removed stopwords from the review text using nltk stopwords module
- Draw random samples of 25K reviews per dataset
- Made charts for distributions of ratings and review length.