Analysis of Reviews of Musical Instrument
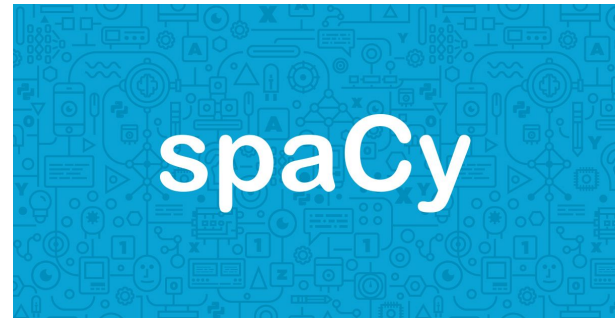
Wooyoung Jeong,
Kaggle

Design

Objective: Explore reviews of musical instruments in Amazon to figure out any unsupervised analysis that can be helpful in future.

Goal: Based on the feature, figure out the relationship and analysis between the rates and reviews of the products.

Tools Used

- Numpy & Pandas
- Scikit-learn & Statsmodels
- Spacy
- matplotlib

Data

- Reviews from Amazon between May 1996 - July 2014
- Each row represents individual review in one product
- Features
    - ID of the reviewer, ID of the product, name of reviewer, helpful, text of review, rating of the product, summary of the review, time of the review in unix and raw time

Data Cleaning

- Delete all columns that do not necessary
- There are total 10261 rows in the dataset
- Cleaned review text column to make it easily analyzed
- Separate rates to positive and negative reviews

Pos vs Neg
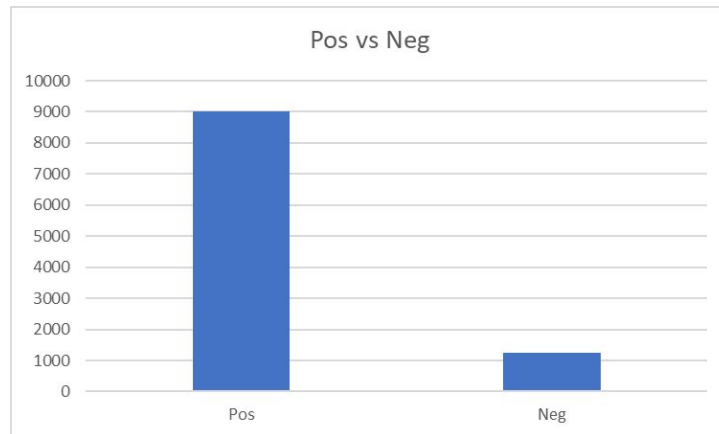
PCA

- Used Spacy to filter words
- N_components = 2
- Found the most common adjective in each side of reviews
- In top 10 words, 6 of them are overlapped.
  - To make the better analysis deleted first 5 top words
- Transformed value was
  - $-1.43e3$, $1.71e\text{-}13$ for positive
  - $1.43e3$, $9.88e\text{-}14$ for negative

Positive	Negative
great	good
good	other
other	great
little	more
more	getter
nice	little
easy	cheap
better	same
best	nice
small	few

LMF

- Set min_df for CountVectorizer to 0.001 to reduce memory usage
- Filter top 5 words in 2 groups that describes group
- Groups can be described as sound system and instrument group.

Sound	Instrument
amp	guitar
pedal	strings
sound	just
like	strings
use	like

Future Work

- Apply existing analysis to go deeper
    - Finding review types for specific instrument or instrument company
    - Finding specific adjectives for each LMF group
- Find specific pattern in the reviews to find any paid reviews or reviews that were written by one customer