IR ASSIGNMENT -3 Utkarsh Pal 2020144

Q- ¾)

```
Total number of rows for the product category  Headphones  is  30471
Total number of reviews for the product category  Headphones  is  372167
Total number of unique products for the product category  Headphones  is  30471
Number of good reviews  312041
Number of bad reviews  60126
Average rating for the category  4.00616658650552
overall
1     31616
2     28510
3     41427
4     75024
5    195590
```

Q-5)

Preprocessing steps are mentioned below:-

```python
acronym_dictionary = {
    "NASA": "National Aeronautics and Space Administration",
    "FBI": "Federal Bureau of Investigation",
    "CIA": "Central Intelligence Agency",
    "UNESCO": "United Nations Educational, Scientific and Cultural Organization",
    "NATO": "North Atlantic Treaty Organization",
    "WHO": "World Health Organization",
    "IMF": "International Monetary Fund",
    "UNICEF": "United Nations International Children's Emergency Fund",}
```

✓ 0.0s

```python
import re
import nltk
from nltk.stem import WordNetLemmatizer

def pre_process(text):
    # lowercase
    text = text.lower()
    #remove tags
    text = re.sub("<!--?.*?-->","",text)
    # remove special characters and digits
    text=re.sub("(\\d|\\W)+"," ",text)
    # replace acronym with full form
    for key, value in acronym_dictionary.items():
        text = text.replace(key, value)
    #lemitze text
    text = text.split()
    lemmatizer = WordNetLemmatizer()
    text = [lemmatizer.lemmatize(word) for word in text]
    text = " ".join(text)

    return text

# Preprocess the reviewText column
df['reviewText'] = df['reviewText'].apply(pre_process)
```

Q-6)

a. Top 20 most reviewed brands in the category that you have chosen.

b. Top 20 least reviewed brands in the category you have chosen.

```
TOP 20 brands with most reviews
('Sony', 32955)
('Sennheiser', 21516)
('Bose', 9582)
('Plantronics', 8340)
('Skullcandy', 8316)
('JLAB', 7731)
('JVC', 7692)
('Audio-Technica', 6791)
('Philips', 6527)
('Panasonic', 6053)
('Koss', 5784)
('LG', 5624)
('Samsung', 5604)
('Mpow', 5480)
('Bluedio', 5132)
('MEE audio', 4644)
('Anker', 4290)
('Symphonized', 4284)
('TaoTronics', 4059)
('Klipsch', 4050)
```

```
Top 20 brands with least reviews
('i.VALUX', 0)
('MPF Products', 0)
('CAD', 0)
('SONCM', 0)
('W-Sound', 0)
('Rademax', 0)
('Raytek', 0)
('Welcome to Sophia shop,it fit f
('New Unbrand', 0)
('KEKH', 0)
('Link Depot', 0)
('iEazy', 0)
('Mobix', 0)
('Pugster', 0)
('ALSISK', 0)
('Boise', 0)
('Paris Business', 0)
('YAN HUA WU', 0)
('ThinkFreebies', 0)
('Ikey Audio', 0)
```

c) Most positively reviewed 'Headphone' is

```
Title:  Sony MDRZX100 Headphones (Black) Good Review Count:  2850
```

d)

```
Number of reviews for each 5 consecutive years
year
2000       600
2005     17132
2010    119733
2015    234702
```

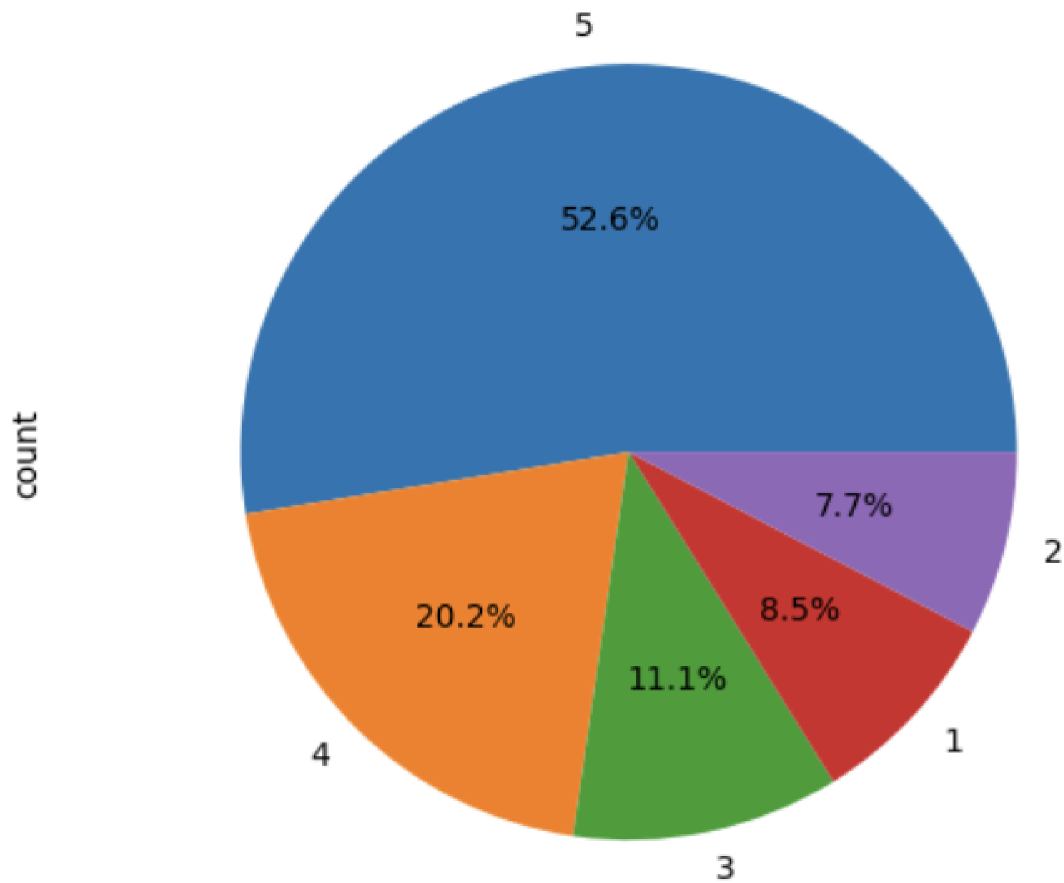
e) Word cloud for good reviews

Most common words include:- headphone, quality,love,great

The bigger words have a larger frequency

World cloud for bad reviews
Most common words include:- headphone, broke,sound,broke
The bigger words have a larger frequency

Q-6) f) Plot a pie chart for Distribution of Ratings vs. the No. of Reviews.



g)

year	
2015	88899
2016	85594
2014	54989
2017	43527
2013	31878
2018	16682
2012	15092
2011	10447
2010	7327
2009	6808
2008	4837
2007	3049
2006	1545
2005	893
2004	333
2003	123
2002	75
2001	45
2000	24

2015 had the highest reviews

h)

```
Year:  2000 Number of customers:  17
Year:  2001 Number of customers:  42
Year:  2002 Number of customers:  67
Year:  2003 Number of customers:  113
Year:  2004 Number of customers:  300
Year:  2005 Number of customers:  778
Year:  2006 Number of customers:  1337
Year:  2007 Number of customers:  2646
Year:  2008 Number of customers:  3611
Year:  2009 Number of customers:  4893
Year:  2010 Number of customers:  5682
Year:  2011 Number of customers:  8297
Year:  2012 Number of customers:  11991
Year:  2013 Number of customers:  24992
Year:  2014 Number of customers:  42386
Year:  2015 Number of customers:  66205
Year:  2016 Number of customers:  66197
Year:  2017 Number of customers:  36391
Year:  2018 Number of customers:  14770
```

2015 also had the highest customer

Q-7

Text Vectorization: It uses **TF-IDF** (Term Frequency-Inverse Document Frequency) vectorization to convert the text data into numerical features suitable for machine learning algorithms. The TfidfVectorizer from sklearn.feature_extraction.text is used for this purpose. TF-IDF represents the importance of a word in a document relative to a collection of documents. It assigns higher weights to words that are more unique to a particular document and less frequent across all documents.
We set max_features to 20,000 to handle a large vocabulary size.

Q-8
The overall score is then convert into 3 classes and here are the count of each class

```
overall_class
good       270614
bad         60126
average     41427
Name: count, dtype: int64
```

Q-9
Data Splitting: It divides the data into training and testing sets in a 75:25 ratio using the train_test_split function from sklearn.model_selection. This allows for training the model on a portion of the data and evaluating its performance on unseen data.

Q-10) We run the following ML models and here are their metrics:-

1. RandomForrestClassifier

	precision	recall	f1-score	support
average	0.82	0.10	0.18	10256
bad	0.84	0.43	0.57	14985
good	0.80	0.99	0.89	67801
accuracy			0.80	93042
macro avg	0.82	0.51	0.55	93042
weighted avg	0.81	0.80	0.76	93042

2. DecisionTreeClassifier

	precision	recall	f1-score	support
average	0.30	0.28	0.29	10256
bad	0.55	0.54	0.54	14985
good	0.86	0.87	0.86	67801
accuracy			0.75	93042
macro avg	0.57	0.56	0.56	93042
weighted avg	0.74	0.75	0.75	93042

3. LogisticRegression

	precision	recall	f1-score	support
average	0.47	0.20	0.28	10256
bad	0.74	0.72	0.73	14985
good	0.88	0.96	0.92	67801
accuracy			0.84	93042
macro avg	0.69	0.63	0.64	93042
weighted avg	0.81	0.84	0.82	93042

4. MultinomialNB

	precision	recall	f1-score	support
average	0.53	0.00	0.01	10256
bad	0.84	0.31	0.46	14985
good	0.77	0.99	0.87	67801
accuracy			0.78	93042
macro avg	0.71	0.44	0.44	93042
weighted avg	0.76	0.78	0.71	93042

5. KNeighborsClassifier

	precision	recall	f1-score	support
average	0.32	0.05	0.09	10256
bad	0.62	0.13	0.22	14985
good	0.75	0.98	0.85	67801
accuracy			0.74	93042
macro avg	0.56	0.39	0.38	93042
weighted avg	0.68	0.74	0.66	93042

Here are the best models for each category

Class/Metric	F1	Precision	Recall
Good	LR	LR	RF/MNB
Average	DT	MNB	DT/LR
Bad	LR	RF/MNB	LR

Logistic Regression had the best accuracy.

Q-11) a)
We create a matrix with rows as reviewerID and columns as asin number, we fill the cell with overall rating and fill empty cell with the value 0.

Also we only consider the users who have reviewed more than 5 items and only the items which have been reviewed by at least 5 users for reducing the dataset size.

```
print("Number of unique products with reviews" ,df['asin'].nunique())
print("Number of unique users " ,df['reviewerID'].nunique())
✓ 0.0s

Number of unique products with reviews 2089
Number of unique users  5957
```

b) We then normalize the user-item matrix using sklearn's MinMax scaler.

c) User -User recommender system

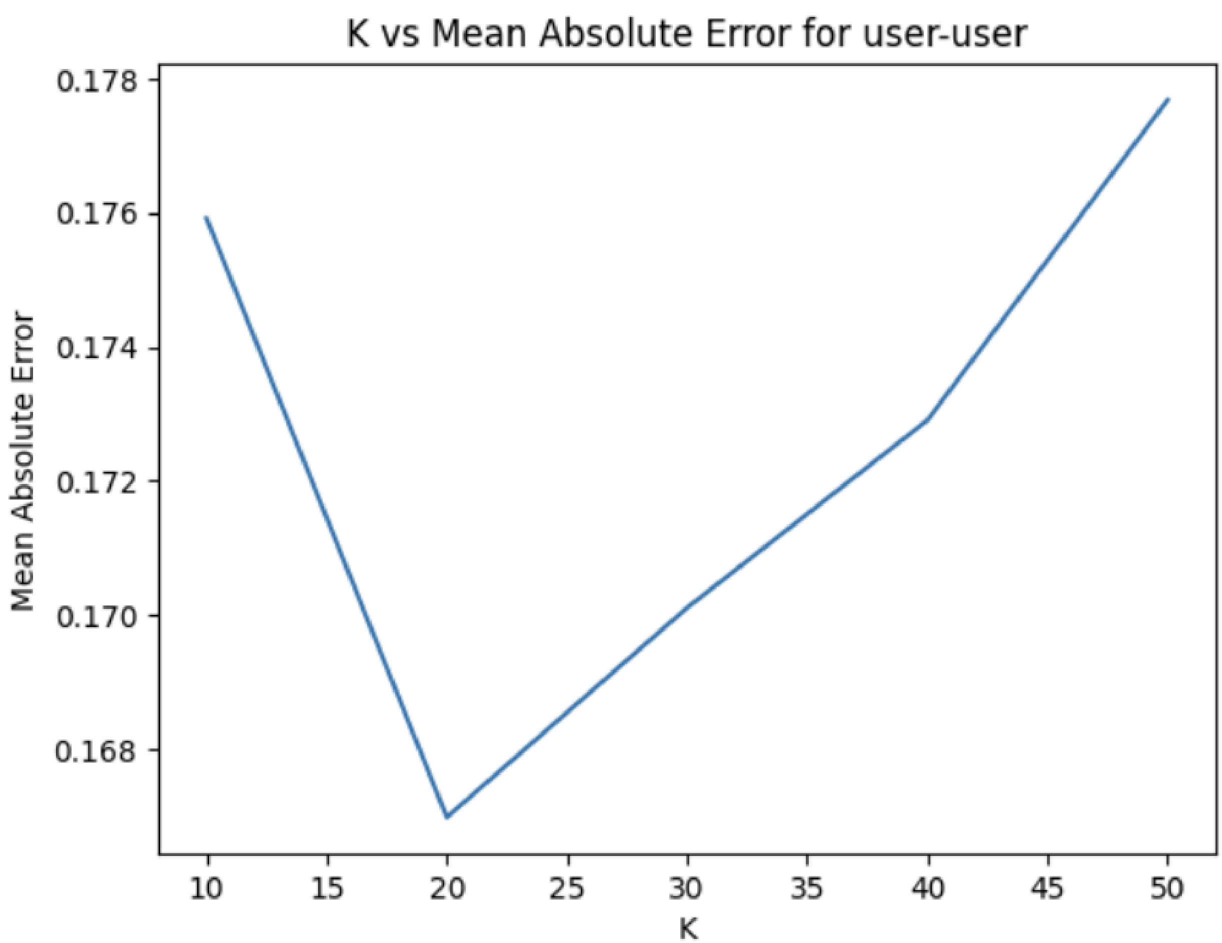
1. Cosine similarity is computed between users based on their rating vectors. This similarity metric measures the cosine of the angle between two vectors and ranges from -1 to 1, with higher values indicating greater similarity.
2. The cosine similarity matrix of users is divided into K folds, and the recommender system is trained and evaluated K times.
3. For N = [10,20,30,40,50] we do the following :-
4. In every iteration in K fold, for every user in the test set .
5. We predict its closest N - Neighbours in the train set.
6. Then for every item we calculate
   MAE = |actual rating - average of rating given by neighbors| and add the error.
7. Final return the error

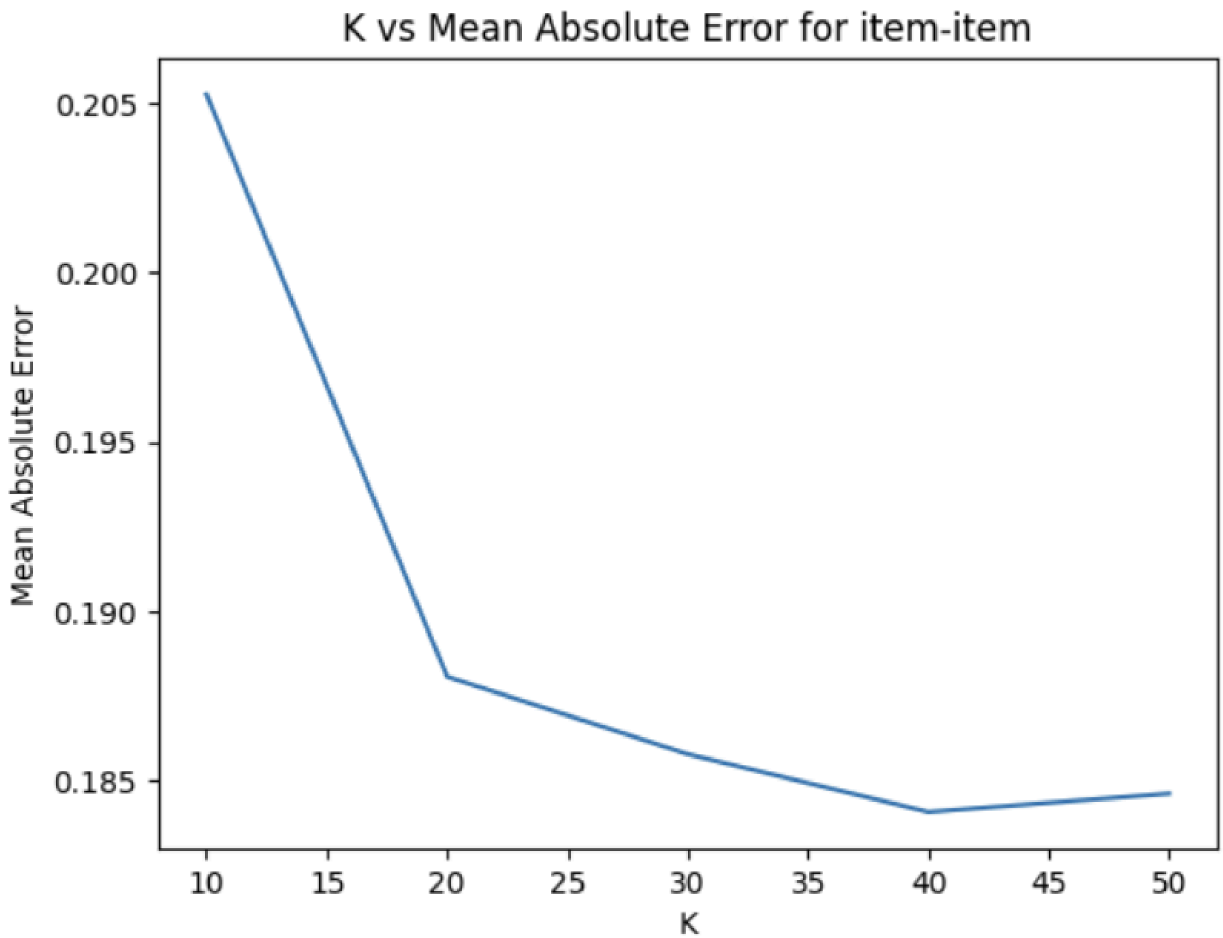
d) For item-item recommender systems we follow the same pipeline but with the inverse of the user-item matrix.

e)

```
User based collaborative filtering
Mean Absolute Error for  10  similar users is  0.1759167611095936
Mean Absolute Error for  20  similar users is  0.1669834686175381
Mean Absolute Error for  30  similar users is  0.1701075569402108
Mean Absolute Error for  40  similar users is  0.17290340538012405
Mean Absolute Error for  50  similar users is  0.17768087439108357
```

```
Item based collaborative filtering
Mean Absolute Error for  10  similar items is  0.20524742693883144
Mean Absolute Error for  20  similar items is  0.18807160287871058
Mean Absolute Error for  30  similar items is  0.1857834328440021
Mean Absolute Error for  40  similar items is  0.18407792826599004
Mean Absolute Error for  50  similar items is  0.18461724037542737
```





Q-12)
TOP 10 products by User Sum Ratings.

```
Top 10 products by User Sum Ratings
Title:  Toysdone Wireless Headphones Stereo Earbuds Wireless Sport Earphones for Running with Mic (6
Title:  Xbrn Dual Ports Adapter Splitter, 2 in 1 Headphone Jack Aux Audio &amp; Charger Adapter Cabl
Title:  Sony MDRZX100 Headphones (Black) Sum of Ratings:  1078.0
Title:  Sony MDRZX100 ZX Series Stereo Headphones (Blue) Sum of Ratings:  1078.0
Title:  Koss Porta Pro On Ear Headphones with Case, Black / Silver Sum of Ratings:  659.0
Title:  Sony MDR7506 Professional Large Diaphragm Headphone Sum of Ratings:  563.0
Title:  Sennheiser HD 202 II Professional Headphones (Black) Sum of Ratings:  478.0
Title:  Clip Style Headphone Black Lightweight and Comfortable Ear Clip. Splash Proof Water resistar
Title:  Sony MDRV6 Studio Monitor Headphones with CCAW Voice Coil Sum of Ratings:  460.0
Title:  V-MODA Crossfade LP Over-Ear Noise-Isolating Metal Headphone (Rouge) Sum of Ratings:  444.0
```