1.รหัส ชื่อ และหมู่ของนิสิตในกลุ่ม

- 6610402167 นายพงษ์ศิริ กิตติยุทธนาวิน หมู่ 1
- 6610402060 นายธนกฤต ตระการจันทร์ หมู่ 1

2.วัตถุประสงค์ของระบบต้นแบบ

- ทำนายคะแนน(Rating) ที่ผู้ซื้อให้ จากรีวิวของผู้ใช้เอง(Title + Text)

3.ลิงค์ไปยังข้อมูลที่จะใช้ในระบบต้นแบบ

- Dataset: https://amazon-reviews-2023.github.io/

4.มีค่าที่หายไป (missing values) หรือไม่ อะไรบ้าง แบบใด มากน้อยแค่ไหน แก้ไขได้อย่างไร (ถ้าเป็นรูปภาพ มีรูปที่ไม่ถูกต้องหรือสมบูรณ์หรือไม่ อย่างไร แก้ไขอย่างไร)

- ในตัวข้อมูลเองนั้นมีค่าที่หายไปแต่ค่อนข้างน้อยมากเมื่อเทียบกับจำนวนข้อมูลทั้งหมด แบ่งเป็นค่าของ title 1889 ตัว และ text 1875
- แต่เมื่อมีการแปลงข้อความเป็นตัวเลขด้วยโมเดล Sentiment Analysis ตัวโมเดลประมวลผลข้อความบางตัวไม่ได้ทำให้มีข้อมูลที่หายไปเพิ่มขึ้นประมาณ 2.84%
- แก้ไขโดยการลบตัวอย่างนั้นทิ้ง (Row deletion)

5.แต่ละคุณลักษณะ/feature ต้องมีการทำกระบวนการต่างๆ (operations) หรือไม่ อะไรบ้าง อย่างไร (ถ้าเป็นรูปภาพ ต้องมีลักษณะอย่างไร เช่น ขนาด ความละเอียด รูปแบบไฟล์ ฯลฯ)

- Handling Missing Values(Row Deletion) สำหรับทุกๆ feature
- Scaling ข้อมูล helpful_vote
- Feature Crossing

6.สำหรับปัญหาและข้อมูลนี้ สามารถป้องกันการเกิด data leakage ได้อย่างไรบ้าง

- แยกข้อมูลระหว่าง test set และ train set
- ผู้เขียนโมเดลไม่ได้ไล่ดูข้อมูลทุกตัวเพราะข้อมูลมีจำนวนเยอะ ทำให้ป้องกัน overfit ของโมเดลได้

7.วิเคราะห์ความสำคัญของคุณลักษณะ ตัวไหนควรและไม่ควรใช้ (ถ้าเป็นรูปภาพ ให้วิเคราะห์ลักษณะของรูปที่ควรและไม่ควรใช้)

- helpful_vote เป็น Feature ที่นำไปใช้งานยากและไม่มีประสิทธิภาพจึงไม่ควรใช้
- text_to_polarity หรือตัวกระทู้ข้อความ นั้นทำนายผลได้ดีกว่าหัวข้อกระทู้ title_to_polarity
- การ cross feature ระหว่าง title_to_polarity กับ text_to_polarity ให้ผลลัพธ์ที่ดีที่สุด (title_text_polarity)
- แต่การรวม 2 feature นี้เข้ากับ helpful_vote นั้นให้ผลลัพธ์ตรงกันข้ามนั่นคือแย่ที่สุด
- *ทดสอบจาก Linear Regression

8.ประเด็นอื่น ๆ ที่เกี่ยวข้องกับข้อมูลของตัวเอง

- ข้อมูลนี้อาจมี Feature ที่ดีกว่านี้จากการทำ Feature Crossing ที่เหมาะสม

9.การมีส่วนร่วมของสมาชิกแต่ละคนในกลุ่ม (แต่ละคนทำอะไรบ้าง)

- พงษ์ศิริ เตรียมข้อมูลและจัดการข้อมูล(จาก assignemet1+ทำเพิ่มเติม) ตรวจสอบค่าจัดการกับ missing value ทำ Feature Crossing
- ธนกฤต scale ข้อมูล ทำ Feature Crossing วิเคราะห์ความสำคัญของคุณลักษณะ

10.การเปิดเผยการใช้เครื่องมือปัญญาประดิษฐ์ (ใช้อะไร ใช้เพื่ออะไร ใช้อย่างไร, prompt อย่างไร)

- ไม่มีการใช้

In [ ]:
```
#from numba import jit, cuda
import pandas as pd
import numpy as np
import nltk
from nltk.corpus import stopwords
from nltk.sentiment import SentimentIntensityAnalyzer
from textblob import Word, TextBlob
from wordcloud import WordCloud
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('omw-1.4')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Pongs\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\Pongs\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     C:\Users\Pongs\AppData\Roaming\nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
```

Out[ ]: True

Prepare Data (เหมือนกับ assignment1 + เพิ่มเติม)

In [ ]:
```
df = pd.read_csv("Software.csv")
```

In [ ]:
```
df.head()
```

Out[ ]:

	user_id	rating	helpful_vote	title	text	verifie
0	AGCI7FAH4GL5FI65HYLKWTMFZ2CQ	1.0	0	malware	mcaffee IS malware	
1	AHSPLDNW5OOUK2PLH7GXLACFBZNQ	5.0	0	Lots of Fun	I love playing tapped out because it is fun to...	
2	AHSPLDNW5OOUK2PLH7GXLACFBZNQ	5.0	0	Light Up The Dark	I love this flashlight app! It really illumin...	
3	AH6CATODIVPVUOJEWHRSRCSKAOHA	4.0	0	Fun game	One of my favorite games	
4	AEINY4XOINMMJCK5GZ3M6MMHBN6A	4.0	0	I am not that good at it but my kids are	Cute game. I am not that good at it but my kid...	

In [ ]:
```
df = df[df['verified_purchase']][['rating', 'helpful_vote', 'title', 'text']]
df.head()
```

Out[ ]:

	rating	helpful_vote	title	text
1	5.0	0	Lots of Fun	I love playing tapped out because it is fun to...
2	5.0	0	Light Up The Dark	I love this flashlight app! It really illumin...
3	4.0	0	Fun game	One of my favorite games
4	4.0	0	I am not that good at it but my kids are	Cute game. I am not that good at it but my kid...
5	4.0	0	good game	Made me think , variety of the puzzles kept it...

In [ ]:
```
df.isna().any()
```

Out[ ]:
```
rating          False
helpful_vote    False
title            True
text             True
dtype: bool
```

In [ ]:
```
df['title'].isna().sum(), df['text'].isna().sum()
```

Out[ ]:
```
(np.int64(1889), np.int64(1875))
```

In [ ]:
```
df.dropna(inplace=True)
df.isna().any()
```

Out[ ]:
```
rating          False
helpful_vote    False
title           False
text            False
dtype: bool
```

In [ ]:
```
df.head()
```

Out[ ]:

	rating	helpful_vote	title	text
1	5.0	0	Lots of Fun	I love playing tapped out because it is fun to...
2	5.0	0	Light Up The Dark	I love this flashlight app! It really illumin...
3	4.0	0	Fun game	One of my favorite games
4	4.0	0	I am not that good at it but my kids are	Cute game. I am not that good at it but my kid...
5	4.0	0	good game	Made me think , variety of the puzzles kept it...

In [ ]:
```
df['helpful_vote'] = df['helpful_vote'].apply(lambda x: x+1)
df.head()
```

Out[ ]:

	rating	helpful_vote	title	text
1	5.0	1	Lots of Fun	I love playing tapped out because it is fun to...
2	5.0	1	Light Up The Dark	I love this flashlight app! It really illumin...
3	4.0	1	Fun game	One of my favorite games
4	4.0	1	I am not that good at it but my kids are	Cute game. I am not that good at it but my kid...
5	4.0	1	good game	Made me think , variety of the puzzles kept it...

In [ ]:
```
# Lower Case
df['title'] = df['title'].str.lower()
```

```
df['text'] = df['text'].str.lower()
df.head()
```

Out[ ]:

	rating	helpful_vote	title	text
1	5.0	1	lots of fun	i love playing tapped out because it is fun to...
2	5.0	1	light up the dark	i love this flashlight app! it really illumin...
3	4.0	1	fun game	one of my favorite games
4	4.0	1	i am not that good at it but my kids are	cute game. i am not that good at it but my kid...
5	4.0	1	good game	made me think , variety of the puzzles kept it...

In [ ]:
```
#  Punctuations
df['title'].replace("[^a-z0-9\s]", regex=True, inplace=True)
df['text'].replace("[^a-z0-9\s]", regex=True, inplace=True)
df.isna().any()
```

```
C:\Users\pongs\AppData\Local\Temp\ipykernel_15060\919454066.py:2: FutureWarning:
A value is trying to be set on a copy of a DataFrame or Series through chained as
signment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work becau
se the intermediate object on which we are setting values always behaves as a cop
y.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.meth
od({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to pe
rform the operation inplace on the original object.


  df['title'].replace("[^a-z0-9\s]", regex=True, inplace=True)
C:\Users\pongs\AppData\Local\Temp\ipykernel_15060\919454066.py:3: FutureWarning:
A value is trying to be set on a copy of a DataFrame or Series through chained as
signment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work becau
se the intermediate object on which we are setting values always behaves as a cop
y.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.meth
od({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to pe
rform the operation inplace on the original object.


  df['text'].replace("[^a-z0-9\s]", regex=True, inplace=True)
```

Out[ ]:
```
rating          False
helpful_vote    False
title           False
text            False
dtype: bool
```

In [ ]:
```
#Delete Stop Words
sw = stopwords.words('english')
df['title'] = df['title'].apply(lambda x : " ".join(x for x in str(x).split() if
```

```
df['text'] = df['text'].apply(lambda x : " ".join(x for x in str(x).split() if x
df.head()
```

Out[ ]:

	rating	helpful_vote	title	text
1	5.0	1	lots fun	love playing tapped fun watch town grow earnin...
2	5.0	1	light dark	love flashlight app! really illuminates dark, ...
3	4.0	1	fun game	one favorite games
4	4.0	1	good kids	cute game. good kids are. love nik wallenda!
5	4.0	1	good game	made think , variety puzzles kept fun play. gl...

In [ ]:
```
df
```

Out[ ]:

	rating	helpful_vote	title	text
1	5.0	1	lots fun	love playing tapped fun watch town grow earnin...
2	5.0	1	light dark	love flashlight app! really illuminates dark, ...
3	4.0	1	fun game	one favorite games
4	4.0	1	good kids	cute game. good kids are. love nik wallenda!
5	4.0	1	good game	made think , variety puzzles kept fun play. gl...
...	...	...	...	...
4880176	5.0	1		fun addictive exciting
4880177	1.0	2	worst game ever	worst game ever toxic people bad connection wo...
4880178	5.0	3	better!!!	fabulous game 10000 times better pocket editio...
4880179	5.0	1	everything need	awesome! upgraded coreldraw 8. worried load ol...
4880180	5.0	1	huge fan	omg fun keeps playing btw let's go gaming year

4642375 rows × 4 columns

In [ ]:
```
#Lemmatization
df['title'] = df['title'].apply(lambda x: " ".join([Word(word).lemmatize() for w

df
```

Out[ ]:

	rating	helpful_vote	title	text
1	5.0	1	lot fun	love playing tapped fun watch town grow earnin...
2	5.0	1	light dark	love flashlight app! really illuminates dark, ...
3	4.0	1	fun game	one favorite game
4	4.0	1	good kid	cute game. good kid are. love nik wallenda!
5	4.0	1	good game	made think , variety puzzle kept fun play. gla...
...	...	...	...	...
4880176	5.0	1		fun addictive exciting
4880177	1.0	2	worst game ever	worst game ever toxic people bad connection wo...
4880178	5.0	3	better!!!	fabulous game 10000 time better pocket edition...
4880179	5.0	1	everything need	awesome! upgraded coreldraw 8. worried load ol...
4880180	5.0	1	huge fan	omg fun keep playing btw let's go gaming year

4642375 rows × 4 columns

In [ ]:
```
df['text'] = df['text'].astype(str).apply(lambda x: " ".join([Word(word).lemmati
df
```

Out[ ]:

	rating	helpful_vote	title	text
1	5.0	1	lot fun	love playing tapped fun watch town grow earnin...
2	5.0	1	light dark	love flashlight app! really illuminates dark, ...
3	4.0	1	fun game	one favorite game
4	4.0	1	good kid	cute game. good kid are. love nik wallenda!
5	4.0	1	good game	made think , variety puzzle kept fun play. gla...
...	...	...	...	...
4880176	5.0	1		fun addictive exciting
4880177	1.0	2	worst game ever	worst game ever toxic people bad connection wo...
4880178	5.0	3	better!!!	fabulous game 10000 time better pocket edition...
4880179	5.0	1	everything need	awesome! upgraded coreldraw 8. worried load ol...
4880180	5.0	1	huge fan	omg fun keep playing btw let's go gaming year

4642375 rows × 4 columns

In [ ]:
```
df.to_csv('software-prepare.csv')
```

In [ ]:
```
def polarity_score(text):
    result = SentimentIntensityAnalyzer().polarity_scores(text)['compound']
    return result
df['title_to_polarity'] = df['title'].apply(polarity_score)
df['text_to_polarity'] = df['text'].apply(polarity_score)
df
```

Out[ ]:

	rating	helpful_vote	title	text	title_to_polarity	text_to_polarity
1	5.0	1	lot fun	love playing tapped fun watch town grow earnin...	0.5106	0.9403
2	5.0	1	light dark	love flashlight app! really illuminates dark, ...	0.0000	0.9159
3	4.0	1	fun game	one favorite game	0.5106	0.4588
4	4.0	1	good kid	cute game. good kid are. love nik wallenda!	0.4404	0.8858
5	4.0	1	good game	made think , variety puzzle kept fun play. gla...	0.4404	0.8658
...	...	...	...	...	...	...
4880176	5.0	1		fun addictive exciting	0.0000	0.7579
4880177	1.0	2	worst game ever	worst game ever toxic people bad connection wo...	-0.6249	-0.7351
4880178	5.0	3	better!!!	fabulous game 10000 time better pocket edition...	0.5826	0.9666
4880179	5.0	1	everything need	awesome! upgraded coreldraw 8. worried load ol...	0.0000	0.4926
4880180	5.0	1	huge fan	omg fun keep playing	0.5574	0.6249

	rating	helpful_vote	title	text	title_to_polarity	text_to_polarity
				btw let's go gaming year		

4642375 rows × 6 columns

In [ ]:
```
df = pd.read_csv("software-with-polarity2-score.csv")
```

In [ ]:
```
df.loc[df["rating"] == 0, "rating"] = 1
```

In [ ]:
```
df.isna().any()
```

Out[ ]:
```
rating               False
helpful_vote         False
title                 True
text                  True
title_to_polarity    False
text_to_polarity     False
dtype: bool
```

In [ ]:
```
df.isna().any(axis=1).value_counts()
```

Out[ ]:
```
False    4514012
True      128363
Name: count, dtype: int64
```

In [ ]:
```
df.dropna(inplace=True)
df.isna().any()
```

Out[ ]:
```
rating               False
helpful_vote         False
title                False
text                 False
title_to_polarity    False
text_to_polarity     False
dtype: bool
```

In [ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 4514012 entries, 0 to 4642374
Data columns (total 6 columns):
 #   Column             Dtype  
---  ------             -----  
 0   rating             float64
 1   helpful_vote       int64  
 2   title              object 
 3   text               object 
 4   title_to_polarity  float64
 5   text_to_polarity   float64
dtypes: float64(3), int64(1), object(2)
memory usage: 241.1+ MB
```

In [ ]:
```
df.groupby("rating").agg({"helpful_vote" : ["min", "max", "mean"]})
```

Out[ ]:

	helpful_vote		
	min	max	mean
rating			
1.0	1	5817	6.257932
2.0	1	6729	5.339206
3.0	1	4661	4.982909
4.0	1	8185	5.185284
5.0	0	10268	6.181830

In [ ]:
```
df.groupby("rating").agg({"title_to_polarity" : ["min", "max", "mean"]})
```

Out[ ]:

	title_to_polarity		
	min	max	mean
rating			
1.0	-0.9979	0.9623	-0.055074
2.0	-0.9552	0.9442	0.051370
3.0	-0.9325	0.9628	0.158148
4.0	-0.9423	0.9968	0.307666
5.0	-0.9442	0.9981	0.351344

In [ ]:
```
df.groupby("rating").agg({"text_to_polarity" : ["min", "max", "mean"]})
```

Out[ ]:

	text_to_polarity		
	min	max	mean
rating			
1.0	-1.0000	0.9997	0.039715
2.0	-0.9959	0.9997	0.228855
3.0	-0.9995	0.9996	0.372423
4.0	-0.9944	1.0000	0.560669
5.0	-0.9989	1.0000	0.608485

Scaling

scale ข้อมูล helpful_vote ให้อยู่ในช่วงที่คอมพิวเตอร์ประมวลผลได้ง่าย

In [ ]:
```
from  sklearn.preprocessing import StandardScaler

zmv = StandardScaler()
```

```
zmv.fit(df[["helpful_vote"]])
df["helpful_vote_scaled"] = zmv.transform(df[["helpful_vote"]])
df
```

Out[ ]:

	rating	helpful_vote	title	text	title_to_polarity	text_to_polarity
0	5.0	1	lot fun	love playing tapped fun watch town grow earnin...	0.5106	0.9403
1	5.0	1	light dark	love flashlight app! really illuminates dark, ...	0.0000	0.9159
2	4.0	1	fun game	one favorite game	0.5106	0.4588
3	4.0	1	good kid	cute game. good kid are. love nik wallenda!	0.4404	0.8858
4	4.0	1	good game	made think , variety puzzle kept fun play. gla...	0.4404	0.8658
...	...	...	...	...	...	...
4642369	5.0	3	amazing game little flaw	really fun game	0.5859	0.5563
4642371	1.0	2	worst game ever	worst game ever toxic people bad connection wo...	-0.6249	-0.7351
4642372	5.0	3	better!!!	fabulous game 10000 time better pocket edition...	0.5826	0.9666
4642373	5.0	1	everything need	awesome! upgraded coreldraw 8. worried load ol...	0.0000	0.4926
4642374	5.0	1	huge fan	omg fun keep playing	0.5574	0.6249

	rating	helpful_vote	title	text	title_to_polarity	text_to_polarity
				btw let's go gaming year		

4514012 rows × 7 columns

In [ ]:
```
df.groupby("rating").agg({"helpful_vote_scaled" : ["min", "max", "mean"]})
```

Out[ ]:

	helpful_vote_scaled		
	min	max	mean
rating			
1.0	-0.140756	167.927599	0.011185
2.0	-0.140756	194.282197	-0.015364
3.0	-0.140756	134.521991	-0.025660
4.0	-0.140756	236.357081	-0.019812
5.0	-0.169654	296.550751	0.008986

Feature Crossing

- ในข้อมูลนี้มีอยู่ทั้งหมด 3 feature ได้แก่ tilte_to_polarity, text_to_polarity, helpful_vote_scaled
- และ feature helpful_vote_scaled นำมาใช้งานเดี่ยวๆ ได้ยาก
- ดังนั้นในส่วนนี้จะพยายามหาวิธีรวมแต่ละ Feature เข้าด้วยกันให้มีประสิทธิภาพมากที่สุด

In [ ]:
```
df["title_text_polarity"] = (df["title_to_polarity"] + df["text_to_polarity"]) /
df["title_text_helpful_polarity_diff"] = (df["title_text_polarity"] - df["helpfu
df.head()
```

Out[ ]:

	rating	helpful_vote	title	text	title_to_polarity	text_to_polarity	helpful_vote
0	5.0	1	lot fun	love playing tapped fun watch town grow earnin...	0.5106	0.9403	-0
1	5.0	1	light dark	love flashlight app! really illuminates dark, ...	0.0000	0.9159	-0
2	4.0	1	fun game	one favorite game	0.5106	0.4588	-0
3	4.0	1	good kid	cute game. good kid are. love nik wallenda!	0.4404	0.8858	-0
4	4.0	1	good game	made think , variety puzzle kept fun play. gla...	0.4404	0.8658	-0

In [ ]:
```
df.groupby("rating").agg({"title_text_polarity" : ["min", "max", "mean"]})
```

Out[ ]:

	title_text_polarity		
	min	max	mean
rating			
1.0	-0.96280	0.96230	-0.007679
2.0	-0.92225	0.95460	0.140112
3.0	-0.89715	0.97975	0.265286
4.0	-0.92215	0.97725	0.434167
5.0	-0.94360	0.99850	0.479915

In [ ]:
```
df.groupby("rating").agg({"title_text_helpful_polarity_diff" : ["min", "max", "m
```

Out[ ]:

	title_text_helpful_polarity_diff		
	min	max	mean
rating			
1.0	-3.076746e+06	55929.168293	-5.135206
2.0	-4.477428e+04	6441.053733	0.934268
3.0	-4.477428e+04	7202.720543	1.764857
4.0	-4.477428e+04	2965.951342	1.294527
5.0	-4.477428e+04	6441.053733	1.379026

In [ ]:
```
zmv = StandardScaler()
zmv.fit(df[["title_text_helpful_polarity_diff"]])
df["title_text_helpful_polarity_diff_scaled"] = zmv.transform(df[["title_text_he
df.groupby("rating").agg({"title_text_helpful_polarity_diff_scaled" : ["min", "m
```

Out[ ]:

	title_text_helpful_polarity_diff_scaled		
	min	max	mean
rating			
1.0	-2119.232160	38.523102	-0.003890
2.0	-30.840432	4.436181	0.000291
3.0	-30.840432	4.960809	0.000863
4.0	-30.840432	2.042565	0.000539
5.0	-30.840432	4.436181	0.000597

Engineering Good Features วิเคราะห์ความสำคัญของคุณลักษณะ

เนื่องจากข้อมูลทั้งหมดมีขนาดใหญ่การจะทดสอบความสำคัญของแต่ละ Feature จะใช้เวลานานดังนั้นผู้เขียนจึงจะทำการ แบ่งข้อมูลเพียงบางส่วนมาทดสอบเท่านั้น (แบ่งแบบ Startified Sampling)

In [ ]:
```
from sklearn.model_selection import train_test_split
dont_used,df_sample = train_test_split(df, test_size=0.2, stratify=df["rating"])
```

In [ ]:
```
df_sample.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 902803 entries, 1126772 to 4198578
Data columns (total 10 columns):
 #   Column                                  Non-Null Count   Dtype
---  ------                                  --------------   -----
 0   rating                                  902803 non-null  float64
 1   helpful_vote                            902803 non-null  int64
 2   title                                   902803 non-null  object
 3   text                                    902803 non-null  object
 4   title_to_polarity                       902803 non-null  float64
 5   text_to_polarity                        902803 non-null  float64
 6   helpful_vote_scaled                     902803 non-null  float64
 7   title_text_polarity                     902803 non-null  float64
 8   title_text_helpful_polarity_diff        902803 non-null  float64
 9   title_text_helpful_polarity_diff_scaled  902803 non-null  float64
dtypes: float64(7), int64(1), object(2)
memory usage: 75.8+ MB
```

In [ ]:
```
df_sample["rating"].value_counts().sort_values(ascending=False)
```

Out[ ]:
```
rating
5.0    501452
4.0    161125
1.0    119702
3.0     77808
2.0     42716
Name: count, dtype: int64
```

In [ ]:
```
train, test = train_test_split(df_sample, test_size=0.3, random_state=1234)
len(train), len(test)
```

Out[ ]:
```
(631962, 270841)
```

In [ ]:
```
train
```

Out[ ]:

	rating	helpful_vote	title	text	title_to_polarity	text_to_polarit
4264572	5.0	1	awesome.	played original xbox. far exactly same. awesome.	0.6249	0.831
4457866	5.0	1	uno	game awesome,i used play child,it's great know...	0.0000	0.893
3654763	3.0	1	fair	game better added island b.s. contests. miss o...	0.3182	0.318
2555095	5.0	1	fantastic!!! easiest app use ever!!!	answer type getting information ! fantastic sp...	0.8209	0.795
490797	5.0	1	love variety	great app love pagan music - offer many channe...	0.6369	0.924
...	...	...	...	...	...	
1749690	5.0	20	great game	fun playing ramsay dash cool load fun blast. l...	0.6249	0.950
503775	4.0	1	okay	read lot review	0.2263	0.000
667205	4.0	3	fun	first one ever used lot sun it.	0.5106	0.000
3799243	5.0	1	fair	get coin pod. kindle cant get many coin hate g...	0.3182	-0.571
58714	1.0	1	make difference	hear difference set 40% recommended. try highe...	0.0000	0.421

631962 rows × 10 columns

In [ ]:

```
from sklearn.inspection import permutation_importance
from sklearn.linear_model import LinearRegression

features = ["title_to_polarity", "text_to_polarity", "title_text_polarity", "tit
price = "rating"

x_train, y_train = train[features], train[price]
x_test, y_test = test[features], test[price]
```

```
model = LinearRegression()
model.fit(x_train, y_train)

importances = permutation_importance(model, x_test, y_test)
importances= pd.Series(importances.importances_mean, index=features)
importances.sort_values(ascending=False)
importances
```

Out[ ]:
```
title_to_polarity                          9.871998e+17
text_to_polarity                           1.496341e+18
title_text_polarity                        3.418531e+18
title_text_helpful_polarity_diff_scaled    3.517533e-06
dtype: float64
```

In [ ]:
```
df.to_csv("software-3.csv")
```

In [ ]:
```

```