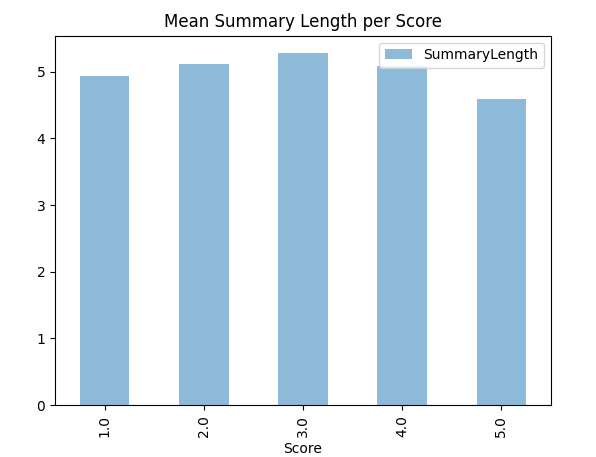
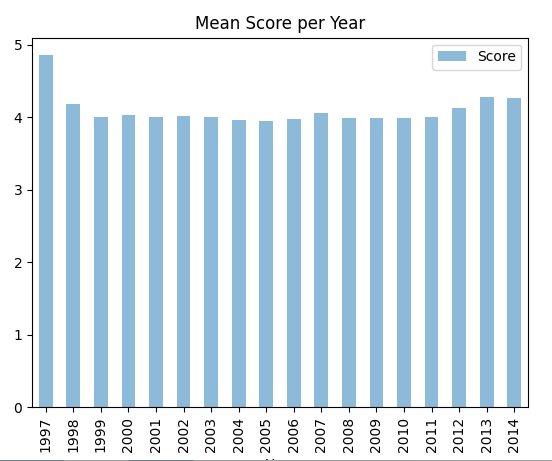
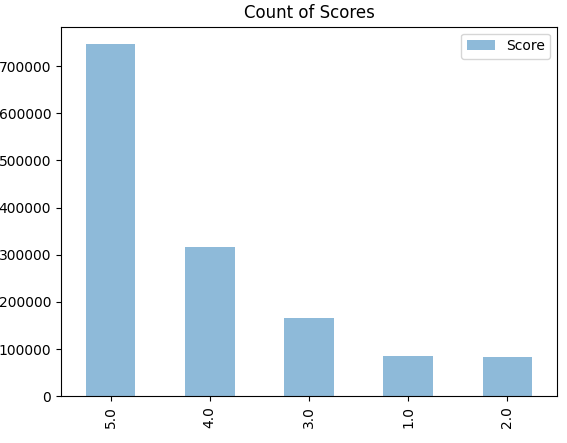
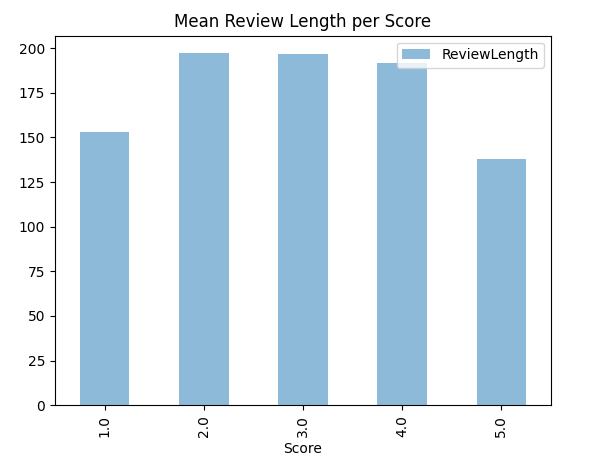
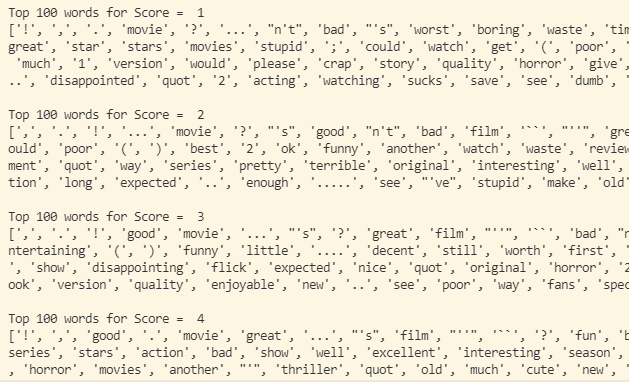
The code is in main.ipynb.

model.ipynb and nlp.ipynb is some attempts, not the final model.

# Data Analysis:

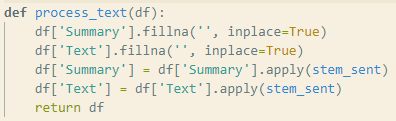
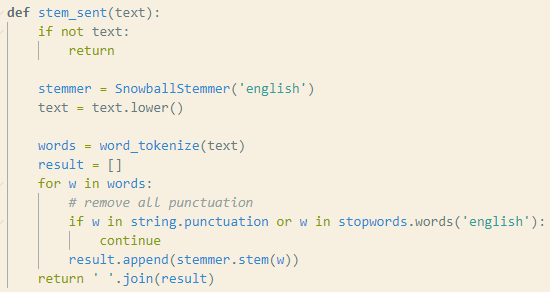
1. train.csv shape (1697533, 9). Could use more fold in cross validation.
2. Year affects the score. We should use Time as features.
3. Text is longer than summary, which has more information. Use NLP to extra features.
4. Text contains significant information, but needs to do extra preprocess. Like tokenize and stem.



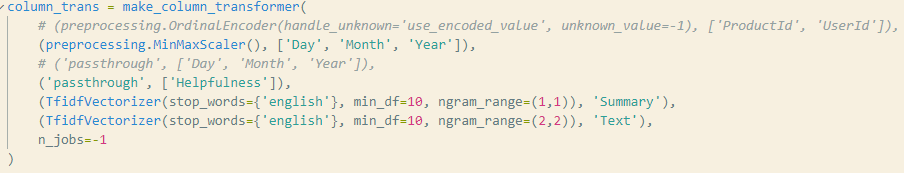
# Preprocess:

For Text and Summary, use SnowballStemmer to stem words, save to csv file.

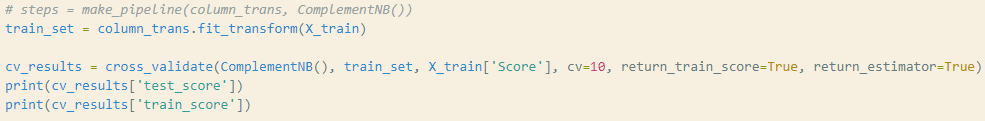
# Feature extraction:

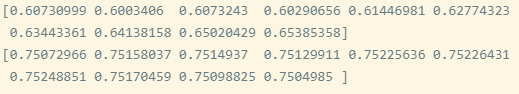
1. Use TF-IDF to embed words. Use N-gram (N=2) in Text, because Text has more words than Summary.
2. For Time, extra the Year, Month, Day, and scale to [0, 1].
3. Calculate Helpfulness = HelpfulnessNumerator / HelpfulnessDenominator.
4. Ignore other features.



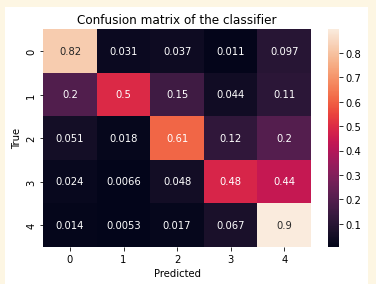
# Model:

Use Naïve Bayes model, cross validation to get the best model.

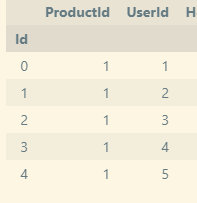
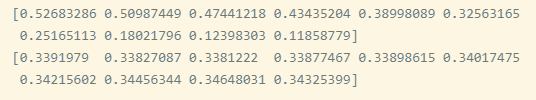


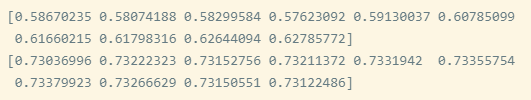


Confuse Matrix on all train set.

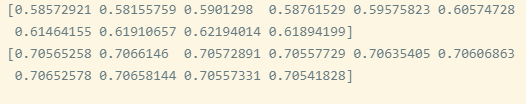


# Decision tried:

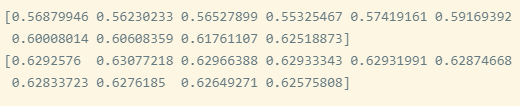
1. One-hot to encode Text and Summary. But TF-IDF better than it.
2. Use ProductId and UserId as input features. Tried to use OrdinalEncoder to label every user and product. But the accuracy is really lowIt totally destroyed the original model 
3. Only use Text column as input. The result is actually good, which means Text contains a large part of information.



1. Not normalize the year, month, day. The result is worse than use it.

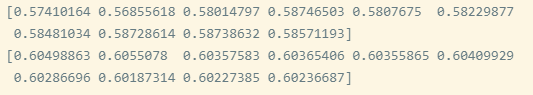


1. Not use N-gram in Text. The result is worse than use it.

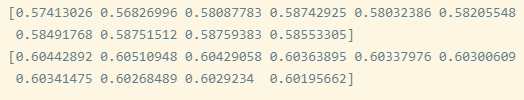


1. Use SVM model. But it’s not as good as Naïve Bayes.

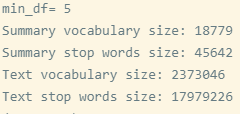
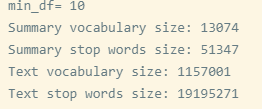
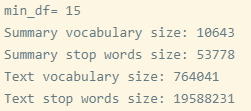
SGDClassifier:



LinearSVC:



1. Use different min\_df in TF-IDF. When min\_df=5, it’s overfitting. When min\_df=15, the result has no improvement. Besides min\_df=10 reduces the vocabulary size to a reasonable range.



1. Try to use SVD after TF-IDF. It needs too long time and large memory to work. (Memory Error)