# Stock Sentiment Analysis

Author : 李攀 21180330

朱泽邕 21180429

李晨 65180205

Abstract ：

In this research ,we mainly focus on the stock sentiment analysis based on news ,involved techniques include :

Data cleaning , remove stop words, and 6 models of NLP :

1. Random Forest
2. Logic regression
3. Naive Bayes
4. Gredient Boosting
5. SGD
6. KNN

We use cleaned data to train those models and get their corresponding accuracy

Keywords :

NLP , Data Cleaning,Deep learning

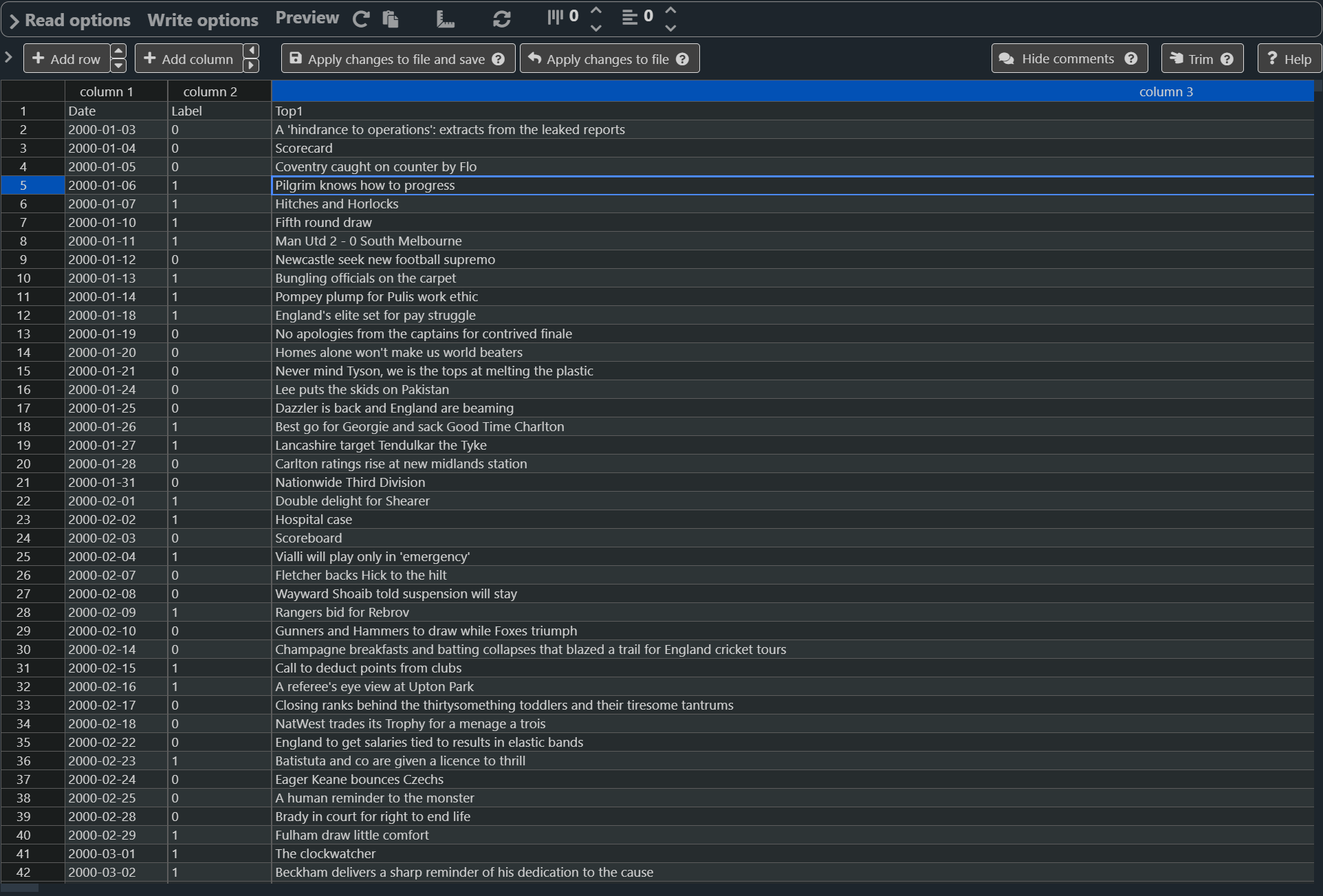
Data:

Our data are mainly gathered from Kaggle , and divide data into train set and test set based on date:

Train set : all news prior to 2015010 are categorized into train set

Test set: all news after 20141231 are categorized into train set

Table schema:



layout :

Date : Label (1 for positive , 0 for negative ) : Top1 - Top25 headline of that day

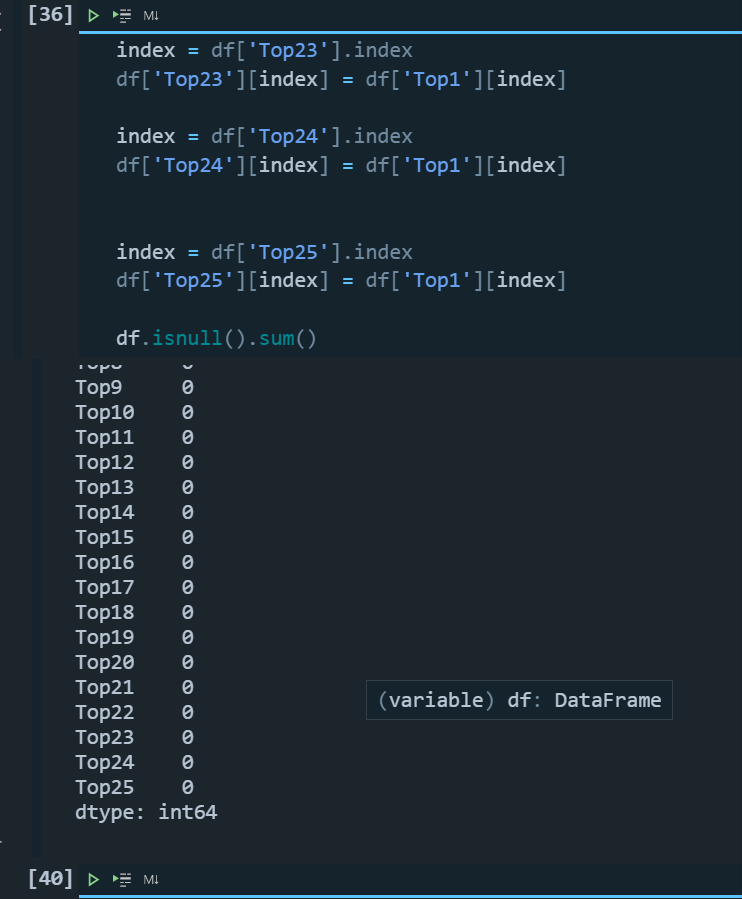
Process :

1. Read in csv file and find missing values

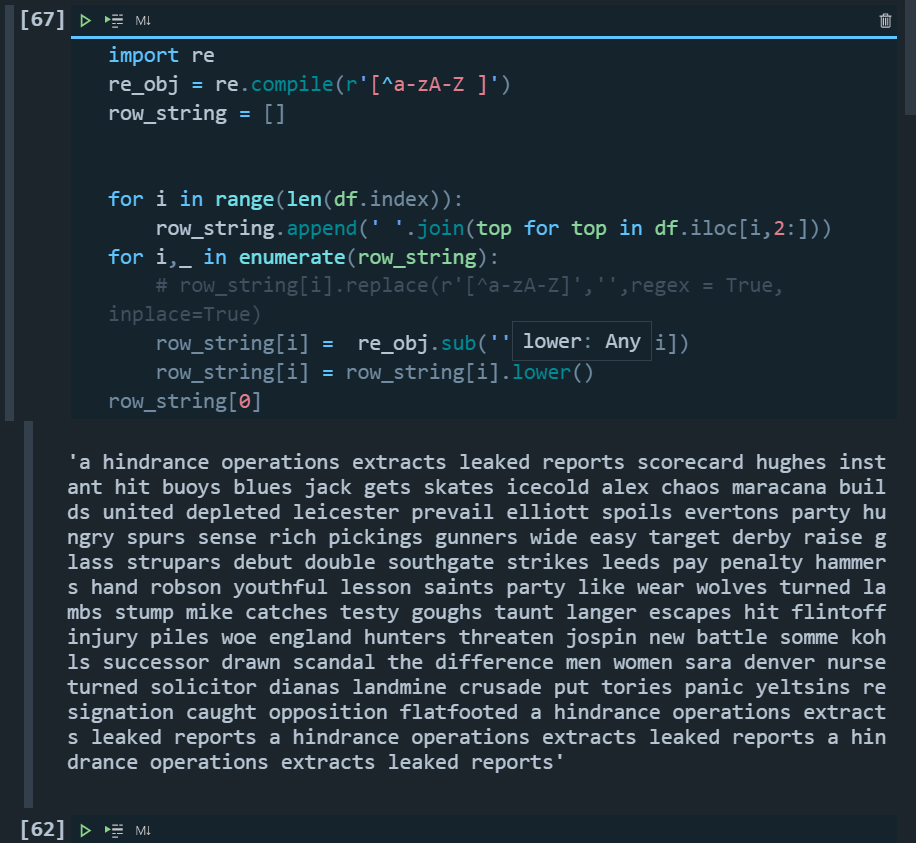


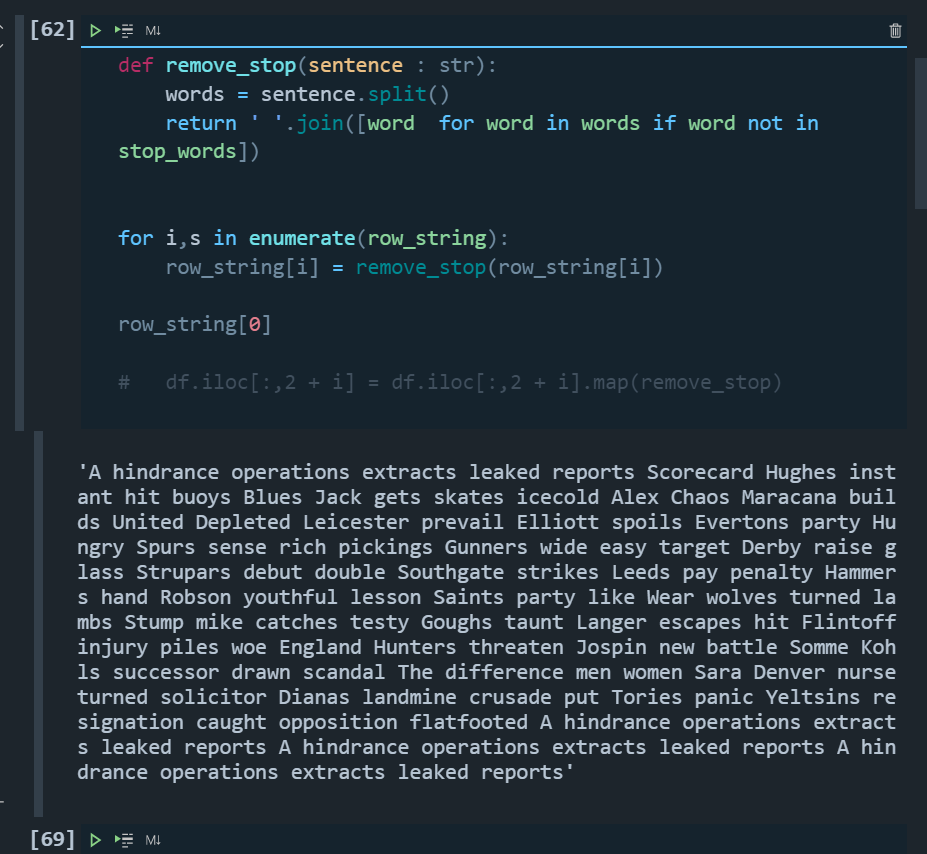
From the result,we inspect that top 23,top 24,top25 has missing values

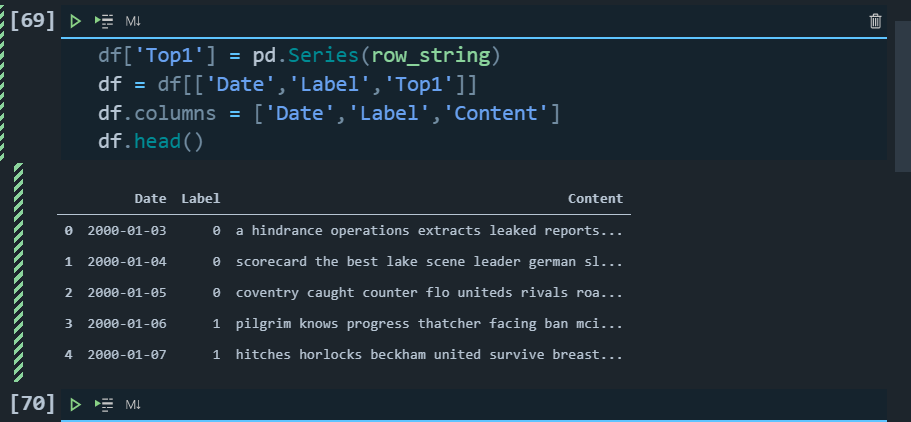
1. Replace missing values



All missing values are replaced with the content of Top1

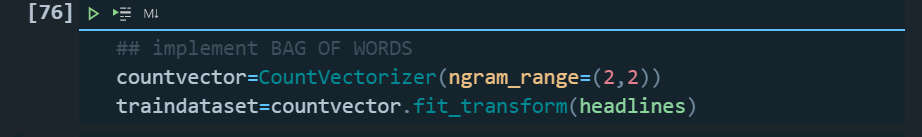
1. Remove punctuation with regex and convert all words into lower case
2. 
3. Remove stop words from a pre collected set of words.



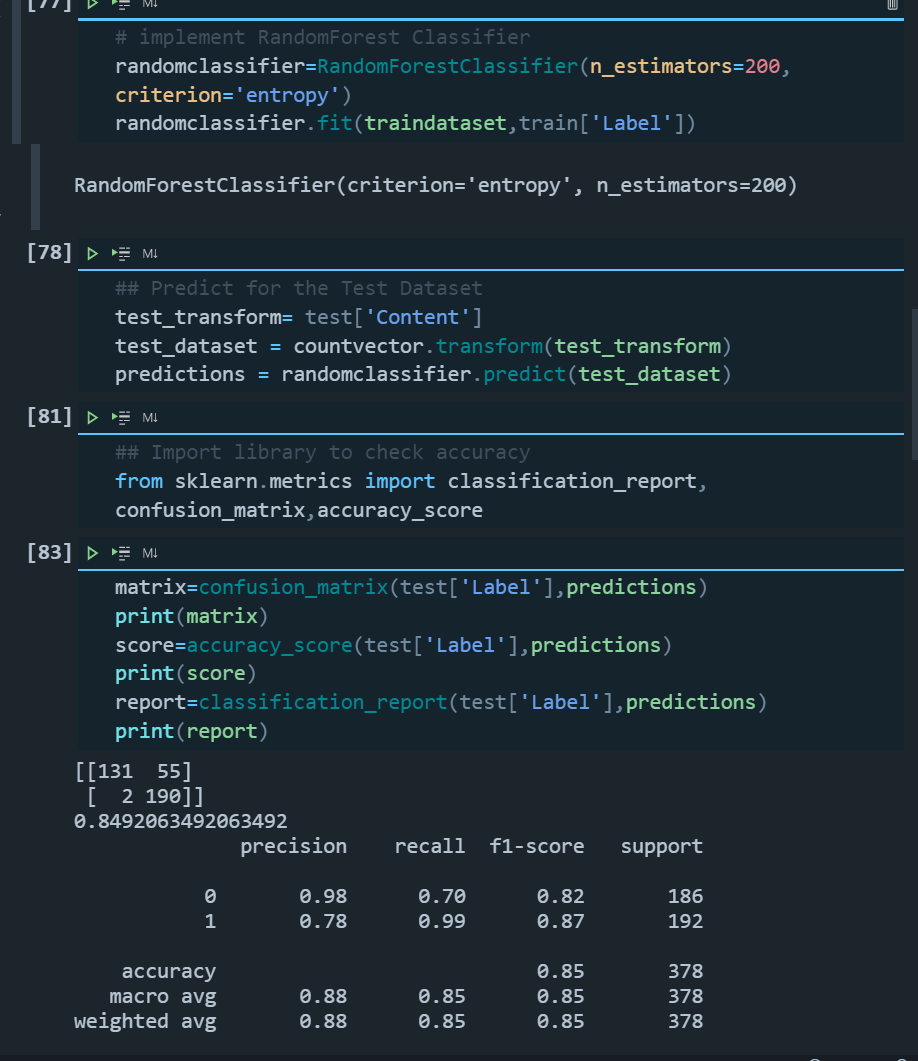
1. Replace data frame content with a single Series of processed string.

For now,all the preparation work has been done,next feed these data into models and train them,

8: convert string into bag of words

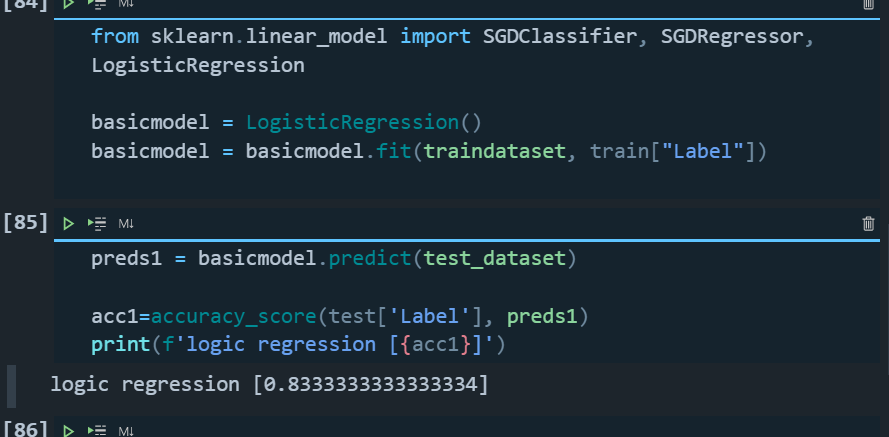


1. Random Forest model:



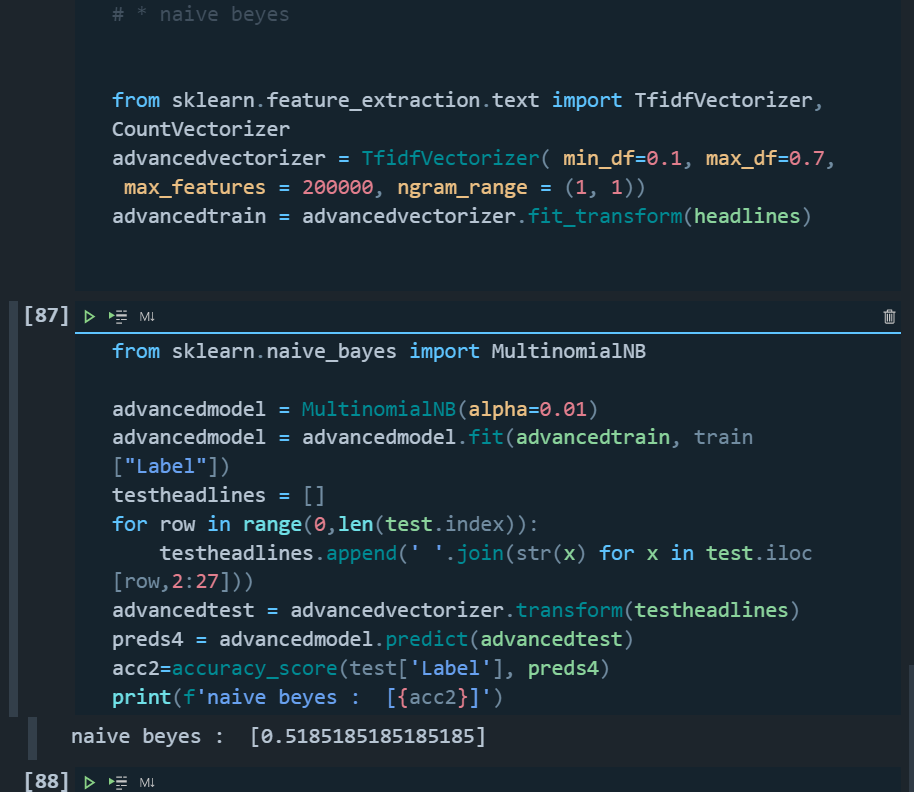
Accuracy : 0.849

10 : logic regression :

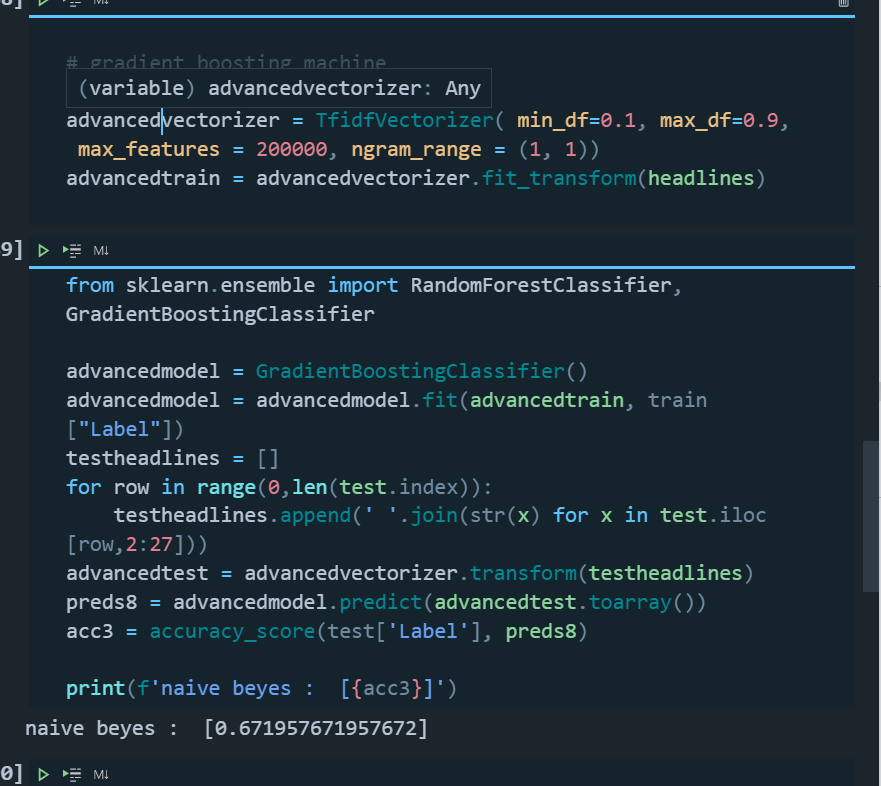


Accuracy : 0.8333

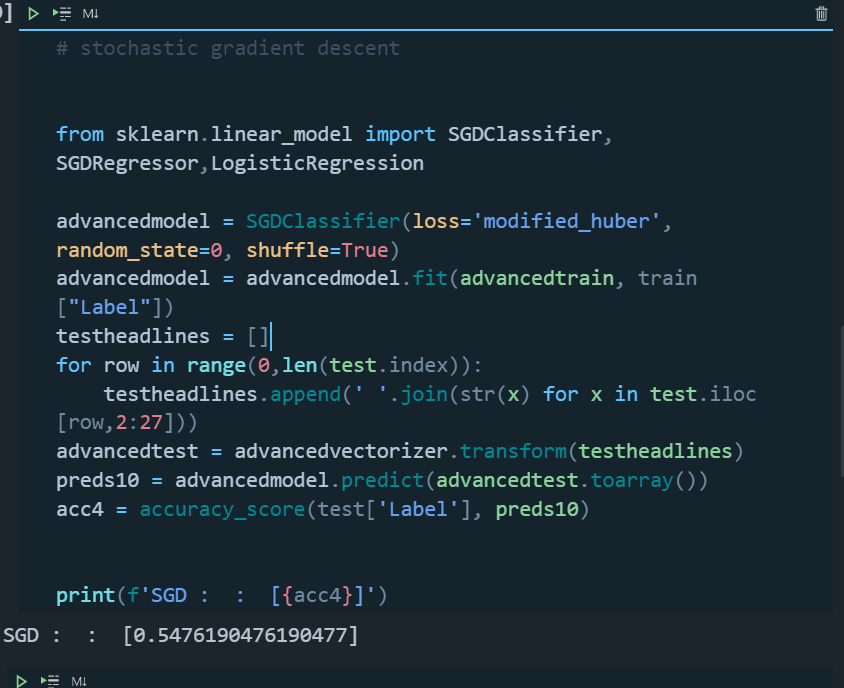
11 : Naive Bayes



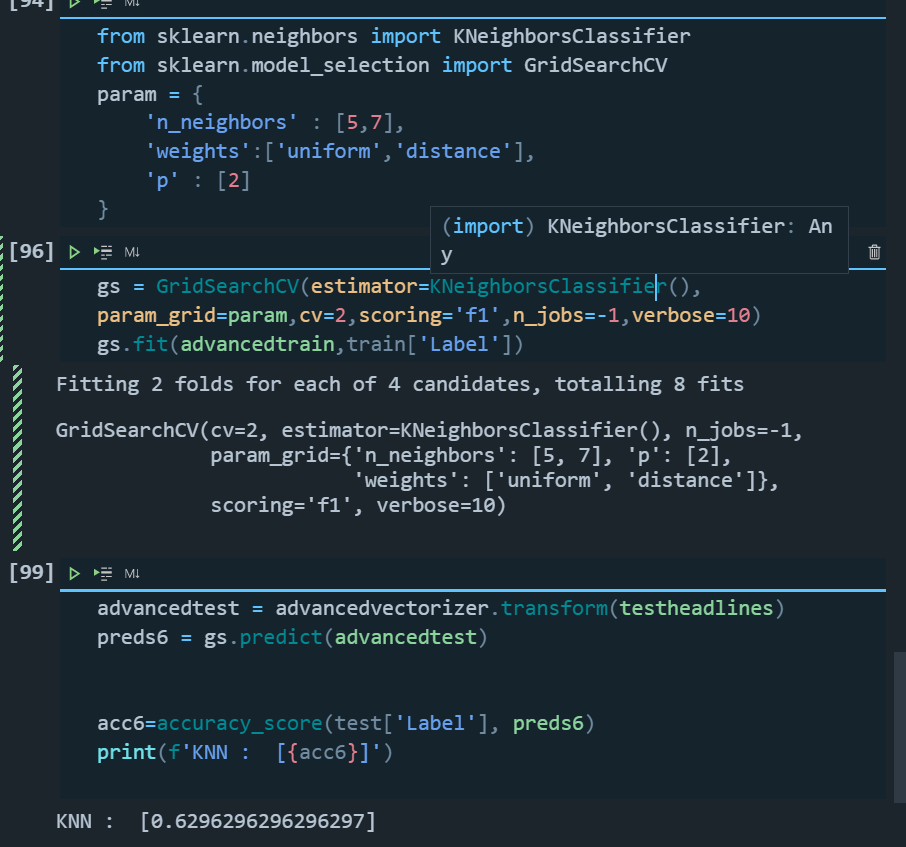
12 : Gradient Boosting Machine



13: Stochastic Gradient descent



14 : KNN



In summary :

Model Accuracy

Random Forest 0.849

logic regression 0.833

Naive Bayes 0.518

Gradient Boosting Machine 0.672

Stochastic Gradient descent 0.547

KNN 0.629

Conclusion : From the above results ,we could infer that Random Forest achieves the best accuracy,thought it is a simple algorithm,and Naive Bayes gain the worst result .

This is not the case that the more complicated an algorithm is ,the better it is , the results really depends on the data set and read-world situations

Bibliography ：

1. [Neuro–Linguistic Programming (goodtherapy.org)](https://www.goodtherapy.org/learn-about-therapy/types/neuro-linguistic-programming" \l ":~:text=Neuro-linguistic programming (NLP) is a psychological approach that,through experience to specific outcomes.)
2. [Natural language processing - Wikipedia](https://en.wikipedia.org/wiki/Natural_language_processing)
3. [Stock Price Movement Based On News Headline - Analytics Vidhya](https://www.analyticsvidhya.com/blog/2021/05/stock-price-movement-based-on-news-headline/)
4. [Stock-Prediction from News — A Naive Approach | by Andreas Stöckl | Medium](https://andreasstckl.medium.com/stock-prediction-from-news-a-naive-approach-c9b03cd4b8d9)
5. [Top 5 Pre-Trained NLP Language Models (daffodilsw.com)](https://insights.daffodilsw.com/blog/top-5-nlp-language-models)
6. [Learn Python - Free Interactive Python Tutorial](https://www.learnpython.org/)