[Naive Bayes Classifier](https://github.com/kk7nc/Text_Classification" \l "id33)

朴素贝叶斯文本分类已经在工业界和学术界使用了很长时间（由 Thomas Bayes 在 1701-1761 年间提出）。自1950年代以来，人们一直在研究这种技术用于文本和文档分类。朴素贝叶斯分类器（NBC）是一种广泛应用于信息检索的生成模型。 许多研究人员针对他们的应用解决并开发了这项技术。 我们从最基本的 NBC 版本开始，它是通过计算文档中的单词数量，使用词频（词袋）特征提取技术开发。

1. **from** sklearn.naive\_bayes **import** MultinomialNB
2. **from** sklearn.pipeline **import** Pipeline
3. **from** sklearn **import** metrics
4. **from** sklearn.feature\_extraction.text **import** CountVectorizer
5. **from** sklearn.feature\_extraction.text **import** TfidfTransformer
6. **from** sklearn.datasets **import** fetch\_20newsgroups
8. newsgroups\_train = fetch\_20newsgroups(subset='train')
9. newsgroups\_test = fetch\_20newsgroups(subset='test')
10. X\_train = newsgroups\_train.data
11. X\_test = newsgroups\_test.data
12. y\_train = newsgroups\_train.target
13. y\_test = newsgroups\_test.target
15. text\_clf = Pipeline([('vect', CountVectorizer()),
16. ('tfidf', TfidfTransformer()),
17. ('clf', MultinomialNB()),
18. ])
19. text\_clf.fit(X\_train, y\_train)
20. predicted = text\_clf.predict(X\_test)
21. **print**(metrics.classification\_report(y\_test, predicted))

训练集与测试集：



分类效果：

