

# 文本分类的代码实现

朴素贝叶斯、TF-IDF、中文分词

小胖



# 目录

## ONE 朴素贝叶斯模型

伯努利模型与多项式模型

## TWO Pipeline

模型的联结

## THREE 中文分词

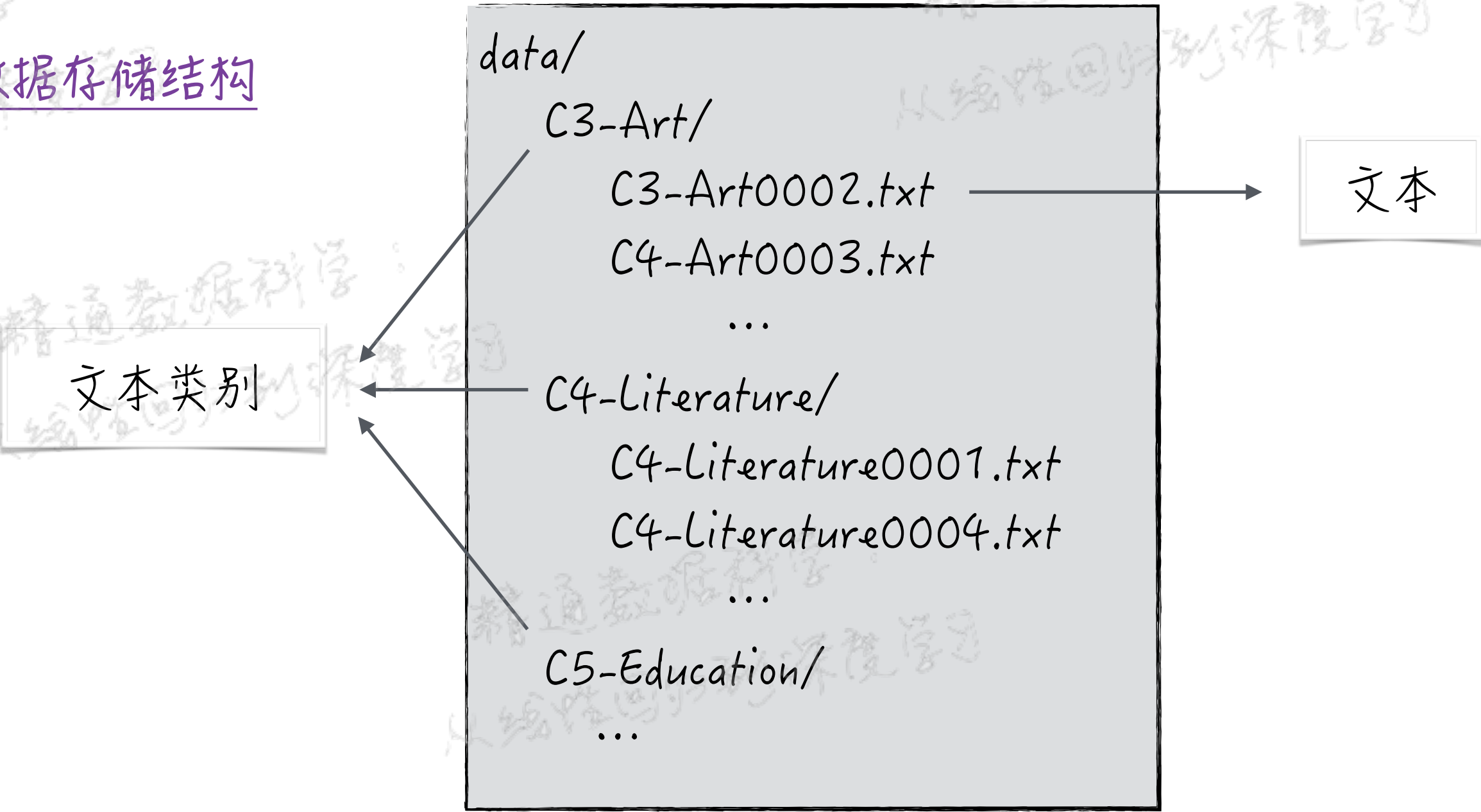
更精确地特征提取

# 朴素贝叶斯模型

模型数据

模型数据来源于复旦大学计算机信息与技术系国际数据库中心自然语言处理小组（李荣陆提供）

模型数据存储结构



# 朴素贝叶斯模型

伯努利模型与多项式模型

将数据分为训练集和测试集

利用字典对文本进行特征提取

使用伯努利模型对文本进行分类

使用多项式模型对文本进行分类

scikit-learn里的CountVectorizer

scikit-learn里的BernoulliNB

scikit-learn里的MultinomialNB

# 目录

## ONE 朴素贝叶斯模型

伯努利模型与多项式模型

## TWO Pipeline

模型的联结

## THREE 中文分词

更精确地特征提取



# Pipeline

模型的联结

模型的联结主义



训练数据

fit

transform

fit

transform

fit



Pipeline

# Pipeline

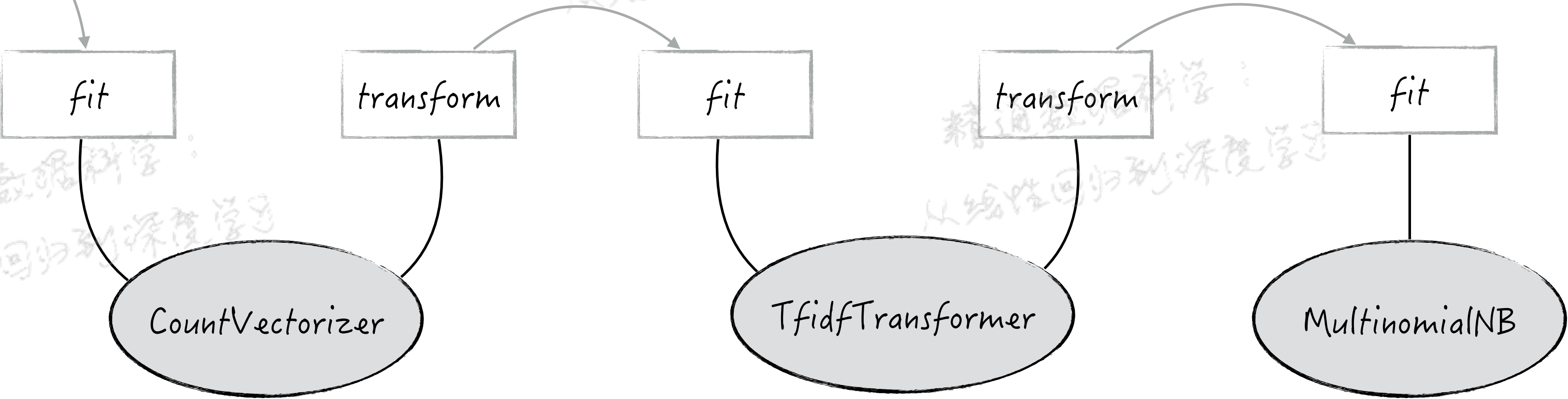
TF-IDF + 多项式模型



TF-IDF + 多项式模型



Pipeline



# 目录

## ONE 朴素贝叶斯模型

伯努利模型与多项式模型

## TWO Pipeline

模型的联结

## THREE 中文分词

更精确地特征提取



# 中文分词

jieba分词

使用 `pip install jieba` 安装jieba分词

不分词 + 伯努利

	precision	recall	f1-score	support
C11-Space	0.57	0.84	0.67	207
C19-Computer	0.89	0.88	0.89	389
C3-Art	0.77	0.71	0.74	222
C39-Sports	0.79	0.62	0.69	381
micro avg	0.76	0.76	0.76	1199
macro avg	0.75	0.76	0.75	1199
weighted avg	0.78	0.76	0.76	1199

分词 + 伯努利

	precision	recall	f1-score	support
C11-Space	0.92	0.82	0.87	207
C19-Computer	0.83	0.99	0.90	389
C3-Art	0.97	0.72	0.83	222
C39-Sports	0.86	0.86	0.86	381
micro avg	0.87	0.87	0.87	1199
macro avg	0.89	0.85	0.86	1199
weighted avg	0.88	0.87	0.87	1199

更好地特征提取，模型效果提升明显

特征提取是建模成功的关键

# THANK YOU

精通数据挖掘科学：  
从线性回归到深度学习