

**项目开发报告**

**题目：基于Bert和NaiveBayes的新闻情感分类模型**

**课程名称： \_\_\_机器学习 \_**

**专业班级： 计科1906 \_**

**学 号： U201914213\_\_\_\_\_**

**姓 名： 吴官骏\_\_ \_\_\_**

**指导教师： \_\_李\_玉\_华\_\_\_\_**

**报告日期： 2021.11.30\_\_\_\_\_\_**

**计算机科学与技术学院**

# 项目开发报告

## 1.1项目目的

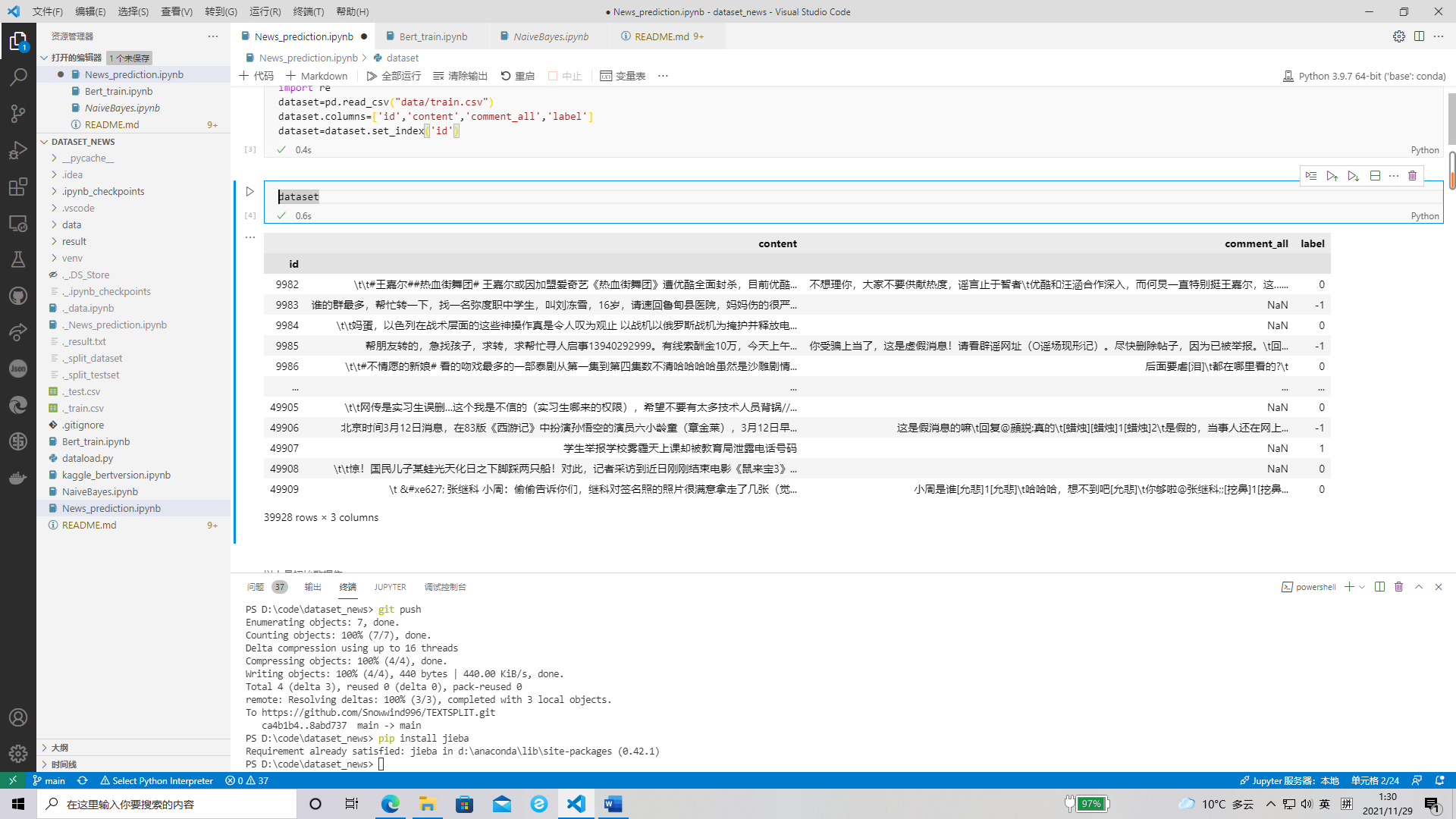
本项目是机器学习课程的课设，任务是实现一个情感分类的模型，给定40000条新闻的信息及其评论，通过实现一个新闻文本分类模型来对其的真假性进行判断。在这个项目中，我使用了两种算法，第一种是词嵌入模型与朴素贝叶斯算法，在测试集上达到了87.4%的准确率、查准率。第二种是基于预训练词向量的Bert[1]微调模型，预训练词向量来自于哈工大讯飞实验室的CN-WWM模型[2]，开发框架来自于复旦大学的fastNLP[3]。

## 1.2问题分析

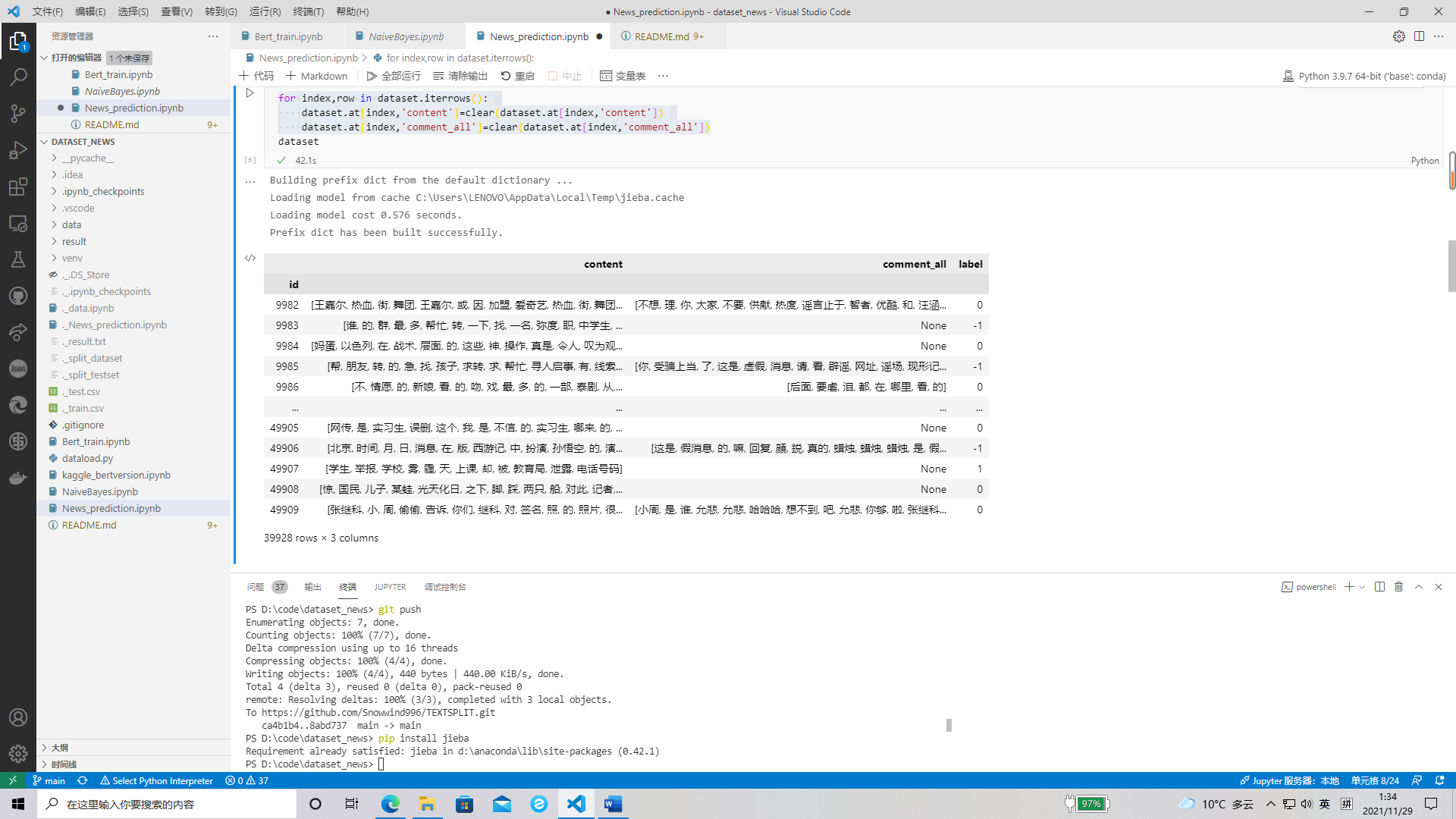
本问题是一个典型的三分类模型，输入为一系列的文本内容。在这里对输入内容做了简化，删去了新闻内容中的非汉字字符，仅仅留下了汉字字符，然后再用分词库对句子进行分词。在朴素贝叶斯实现中，通过TF-IDF词袋模型对每一个词语在各种样本类别的出现概率进行记录，在预测时用朴素贝叶斯算法计算每个句子属于哪一类的最大概率，确定分类。在Bert实现中，分词之后通过加载预训练的词向量模型对汉语文本进行词嵌入，得到合适的句向量编码，最后加载训练完成的Bert参数，在Bert模型的最后一层输出层加上全连接层作为微调，实现预测分类。

## 1.3设计与分析

### 1.3.1 数据分析与预处理

首先通过Pandas库加载数据集，简要查看数据集的情况。如图所示

从中可以看出，文本形式主要为中文，但是还存在很多杂项，比如各种转义符\n\t以及标记#!等符号。这些符号在词嵌入的时候都十分的不方便，虽然某些符号也能表达一定的语义，但是为了简化处理，直接删掉也没关系。因此可以先用Python的正则表达式对数据进行预处理，再用Jieba库对其进行分词，得到结果如下图所示，发现所有content和comment\_all中的词语都已经被分开，而且基本保持了语义的连贯性，比如id为9982中的“王嘉尔”作为单独的人名就被分开，以及id为49909的“张继科”也作为人名被识别出来。虽然分词的粒度有点细，比如“热血街舞团”被分成了“热血”“街”“舞团”，但不影响整体语义的表达，而且要训练一个合理的分词器也要用更深的模型，比如BERT的分类代码。



因为本实验要实现两个模型：BERT和NaiveBayes，因此初步的预处理就到这里为止，接下来分开实现朴素贝叶斯的词嵌入和BERT的词嵌入。但在实现词嵌入之前，首先需要简单介绍一下BERT和朴素贝叶斯的核心思想与流程。

### 1.3.2 算法流程设计

由于本项目要解决的问题的数据都来自于真实的生产环境，因此往往面临着数据噪声大，难以处理，脏数据多，有许多缺失数据等特征，所以在设计算法的过程中对数据的预处理就显得尤其重要。同时由于比赛提供的数据量极大，全部的GPS记录有1亿多条，在本地的电脑上很难同时读取，因此在选取训练数据的过程中，也必须对完整的训练及进行初步筛选，不能使用全部数据进行训练，而只采用其中一小部分，基于上述考虑，整个算法的流程分为三步：数据筛选，数据清洗以及最终的学习算法。

我使用的编程环境是Python，主要用到的第三方库有pandas，numpy和xgboost。总体的算法流程图描述如下：

。。。。。。。

### 1.3.3 机器学习算法设计

具体的机器学习算法在设计过程中也充满了挑战性，主要的难题是，比赛给出的数据及数据类型十分独特。比如测试集中要求预测的是每条订单预计的到达目的地的时间，但是训练集中实际给出的却是每个订单中若干GPS点的数据，每条数据仅仅代表船在某一时刻的状态，而并不能反映这个订单在整个航行过程中的某个特征。因此在设计算法时，如何对数据进行组织就至关重要。此外，即使是不同的订单，数据也可能出现相同路由或者相同船只的情况，因此数据的组合方式十分复杂，为算法的设计带来了诸多困难。

。。。。。。。

## 1.4结果分析

程序训练过程的运行截图如下图所示：

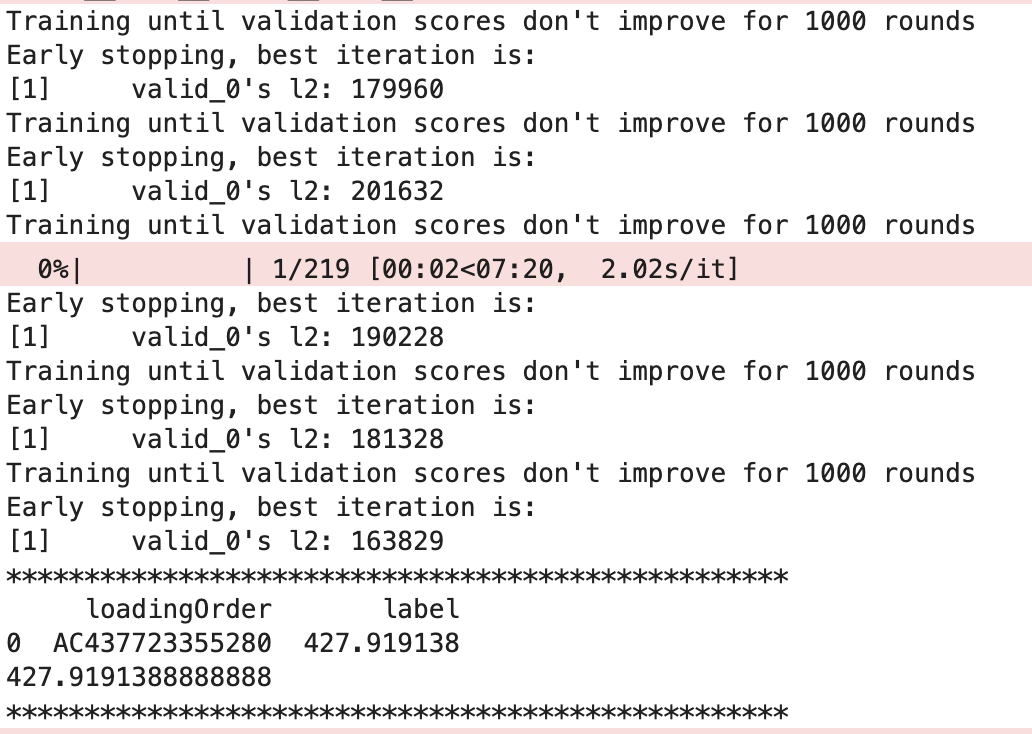


图1.3 程序运行截图

## 1.5思考与总结

本次项目我选择了自选题目的方式，主要是出于效率上的考虑。因为我现在正好在参加这个比赛，题目也与机器学习十分契合，因此就将其作为了机器学习课程的结课项目。同时，机器学习课程中学到的内容，也帮助我在算法的选择上提供了经验和判断。而且这个项目的数据量极大，并且由于是真实数据，数据本身的分布也极其复杂，因此总体来说难度较大，具有一定的挑战性。总的来说，这个项目也让我学到了很多，后续我也会继续优化算法，争取取得更好的效果。

## 1.6 相关文章

[1] https://arxiv.org/pdf/1810.04805.pdf

[2] [ymcui/Chinese-BERT-wwm: Pre-Training with Whole Word Masking for Chinese BERT（中文BERT-wwm系列模型） (github.com)](https://github.com/ymcui/Chinese-BERT-wwm)

[3][fastnlp/fastNLP: fastNLP: A Modularized and Extensible NLP Framework. Currently still in incubation. (github.com)](https://github.com/fastnlp/fastNLP)