

**项目开发报告**

**题目： 新闻数据情感分析 \_**

**课程名称： \_ 机器学习\_ \_\_**

**专业班级： \_ \_CS1908\_ \_ \_\_ \_**

**学 号： \_U201990067 \_ \_**

**姓 名： \_胡嘉慧\_ \_\_ \_**

**指导教师： \_\_黄 宏 \_ \_\_\_**

**报告日期： 2021/11/21\_\_ \_\_\_\_**

**计算机科学与技术学院**

# 项目开发报告

## 1.1项目目的

本项目的主要目的是使用基于在课堂学习的机器学习的知识，完成新闻数据的情感分析。通过清洗数据、把文本转化为词向量、预测文本的情感等步骤来解决新闻数据的情感分析。

## 1.2问题分析

要解决的问题是根据所给训练集和测试集来训练模型，得出对测试的新闻情感分析的结果，这是一个二分类问题，标签只有0和1，0代表不客观，1代表客观，训练模型后预测测试集的标签。首先需要把数据文本全部翻译成英语，之后要清洗数据，删除数据文本中的标点符号、网址、特殊符号、多余的空格等。第二步要文本转化为词向量，提取特征向量，训练模型。最后使用训练完成的模型来预测测试集的标签。

## 1.3设计与分析

### 1.3.1 数据分析

首先根据题目给出的训练集和测试集，它们的content都有不同的语言组成，所以首先要用google\_trans\_new把训练集和测试集的数据文本全部翻译成英语，然后才开始清洗数据文本的内容。

清洗数据文本的步骤如下:

1. 使用neattext库来删除文本中的邮箱、网址、标点符号、标签、多余的空格。
2. 观察执行步骤(1)后的数据文本，发现有网址没有删除，如图1.1和图1.2。查看未清洗数据可发现，因为网址前没有https://，故neattext检测不到，因此需要使用r'tmsnrtrs/[0-9a-zA-Z]\*'和r'pic.twitter.com/[0-9a-zA-Z]+'来删除。

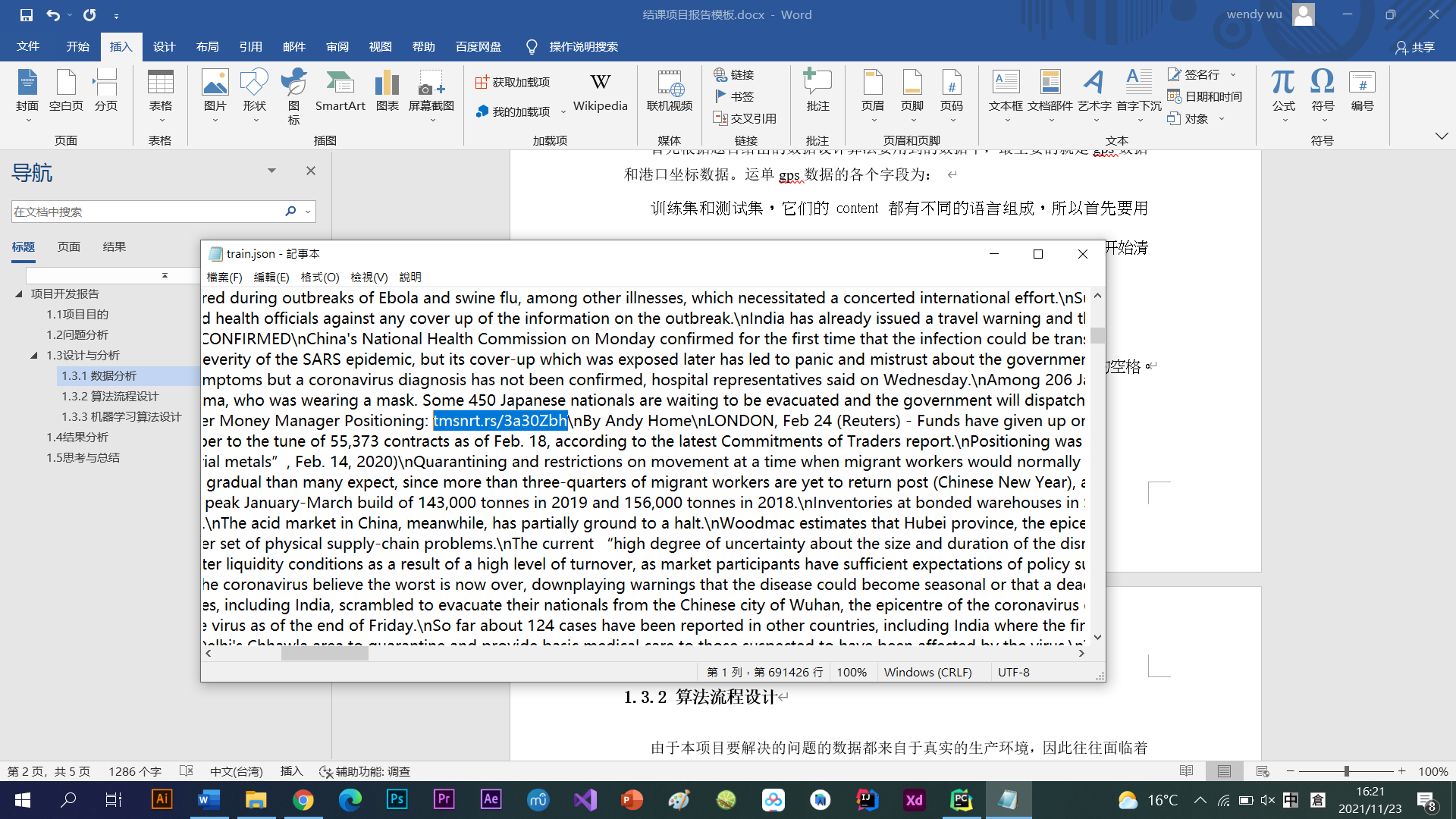


图1.1 原数据的网址1

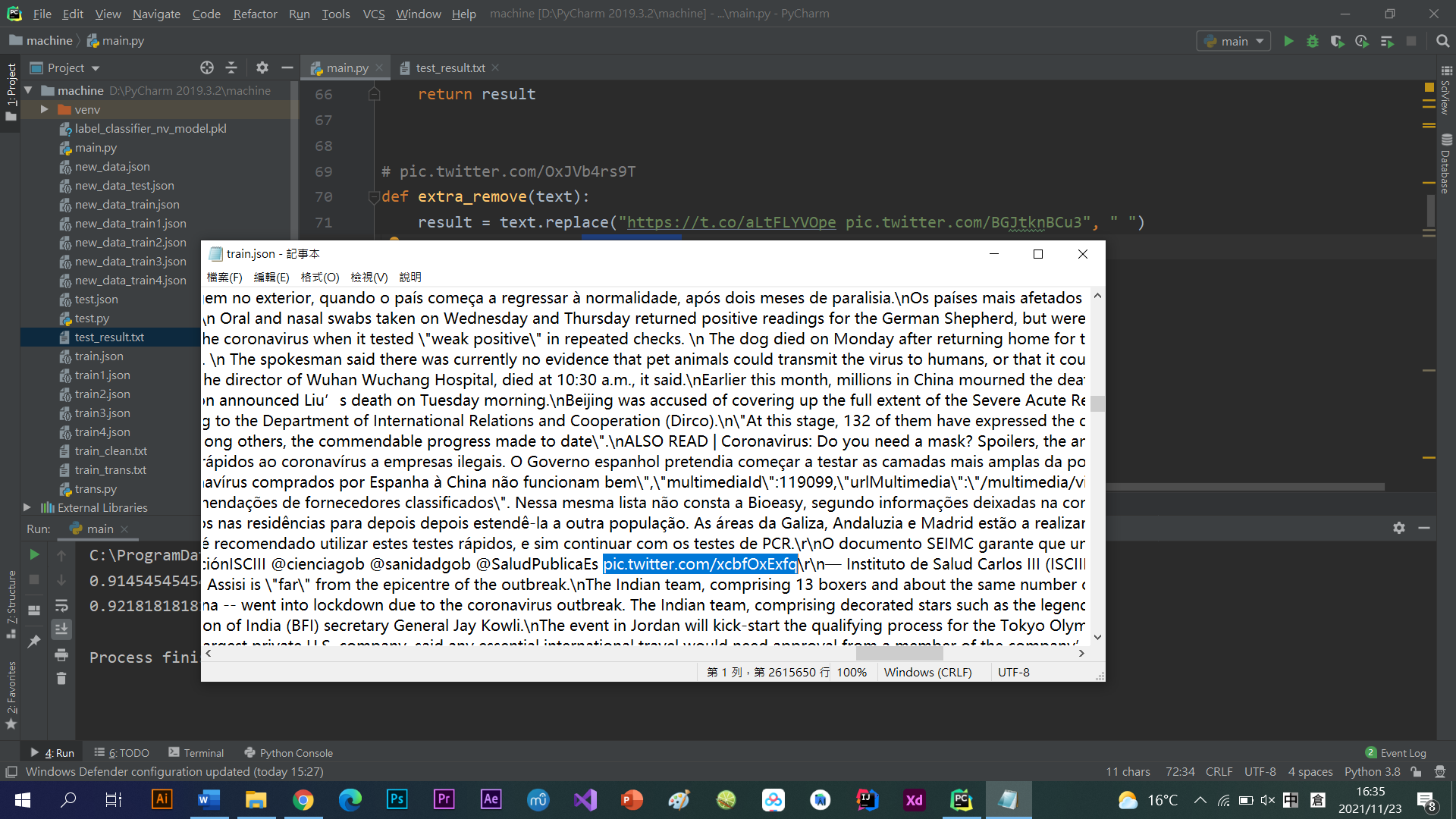


图1.2 原数据的网址2

1. 接着观察步骤(1)和(2)后的数据文本，发现有多个虽然单词不一样，但意思一样的单词，如下:

① Macau和Macao，中文意思都是澳门

②U.S.、USA和United States，中文意思都是美国

③ Cuvid-19和Covid-19

④ February和Feb

⑤ January和Jan The

⑥World Health Organization和WHO

以上的单词都可替成其中的一个单词。

1. 单词的单数和众数也是有一样的意思，所以应把所有的众数转换成单数，使用了textblob来完成。
2. 利用re库结合正则表达式删除括号以及括号中的内容，因为具有冗余的意思。
3. 使用textblob来进行每个content的情感分析，如果sentiment大于0，则表示positive，如果小于0，则表示negative，否则，则表示neutral。

### 1.3.2 算法流程设计

由于本项目要解决的问题的数据文本不是统一的语言，而且有很多脏数据，所以整个算法的流程分为六步：读取文件，数据翻译，数据清洗，提取文本特征，训练模型和使用模型预测测试集。

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

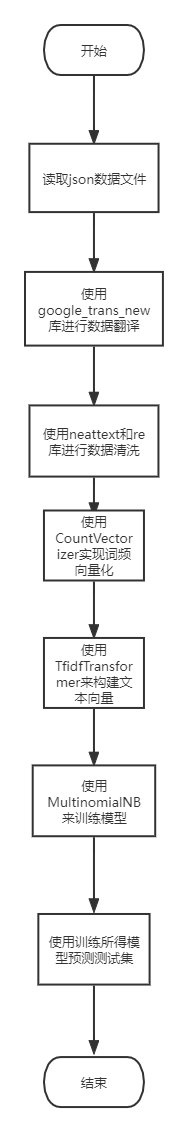


图1.3 算法流程图

### 1.3.3 机器学习算法设计

训练模型使用了LogisticRegression、MultinomialNB和SVM来完成，对比使用LogisticRegression和SVM训练出来的结果，MultinomialNB较优，而且因为贝叶斯分类器是基于贝叶斯理论的分类器，在NLP领域有着广泛的应用，如垃圾邮件检测，个人邮件排序，文本分类，色情内容检测等等，故最后使用MultinomialNB完成。MultinomialNB要计算先验概率，设为[0.1,0.3]为最优。

因为训练集中的label的值为1的数据有2000条，但label的值为0的数据只有200条，数据不平衡，所以在分割训练集和测试集的时候设置stratify=ylabels，就可以让测试集和训练集中的结果也保证这种分布，testsize设置为0.2，random设置为1000。

由于数据集中每一条数据都是很长的一个字符串，所以我們需要将该字符串转换成向量，使用CountVectorizer实现词频向量化，并且lowercase设置为true。因为就算大小写不一样，也是同样的单词。但是仅仅通过统计词频的方式来将文本转换成向量会出现一个问题：长的文章词语出现的次数会比短的文章要多，而实际上两篇文章可能谈论的都是同一个主题，所以我們要使用TfidfTransformer来构建文本向量。

最后便可以使用训练所得的模型去预测测试集的值了。

## 1.4结果分析

程序训练过程的运行截图如下图所示：

1. 翻译和清理数据集后的结果

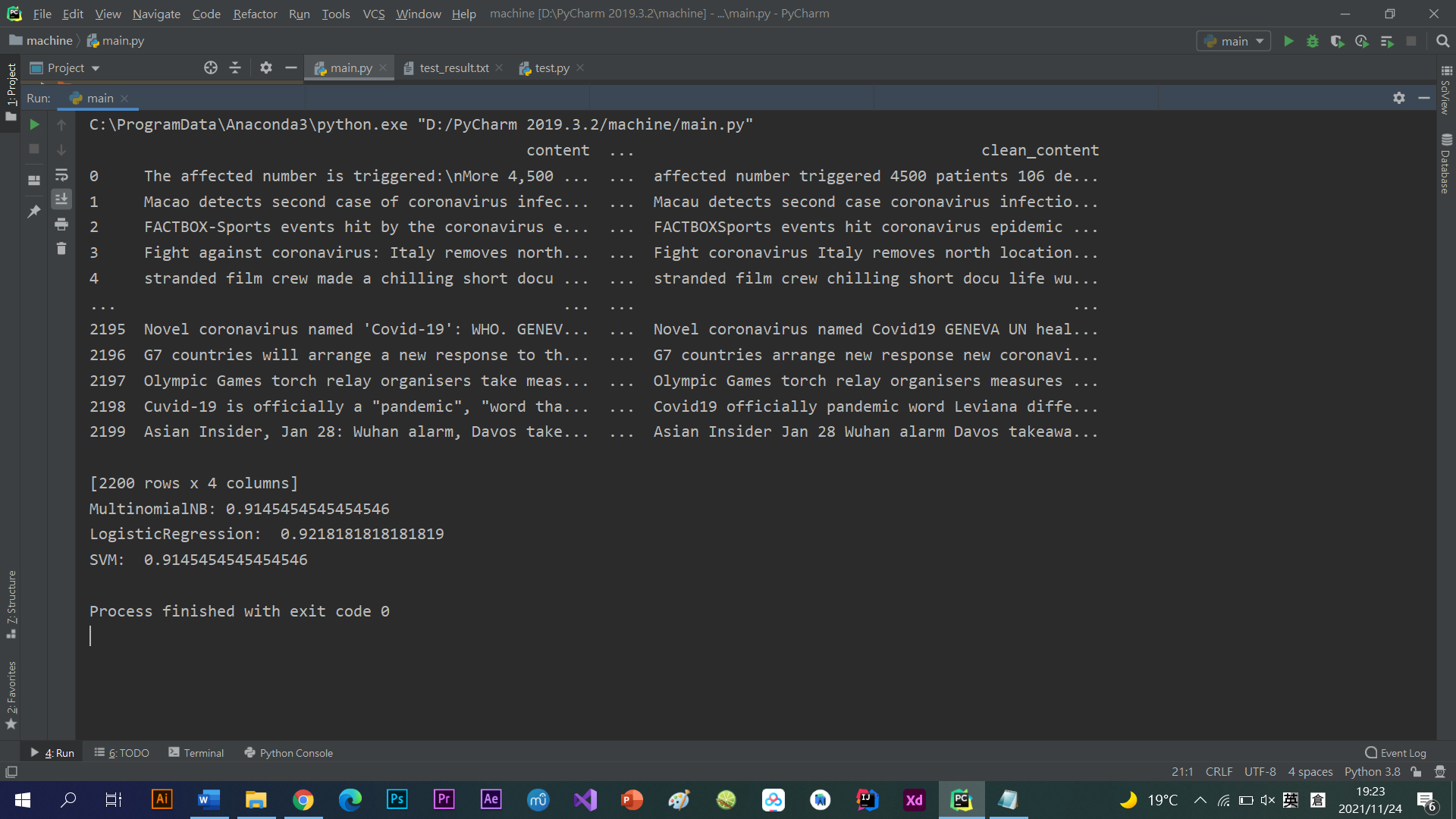


图1.4 清理前和清理后的数据文本对比

1. 使用textblob进行情感分析

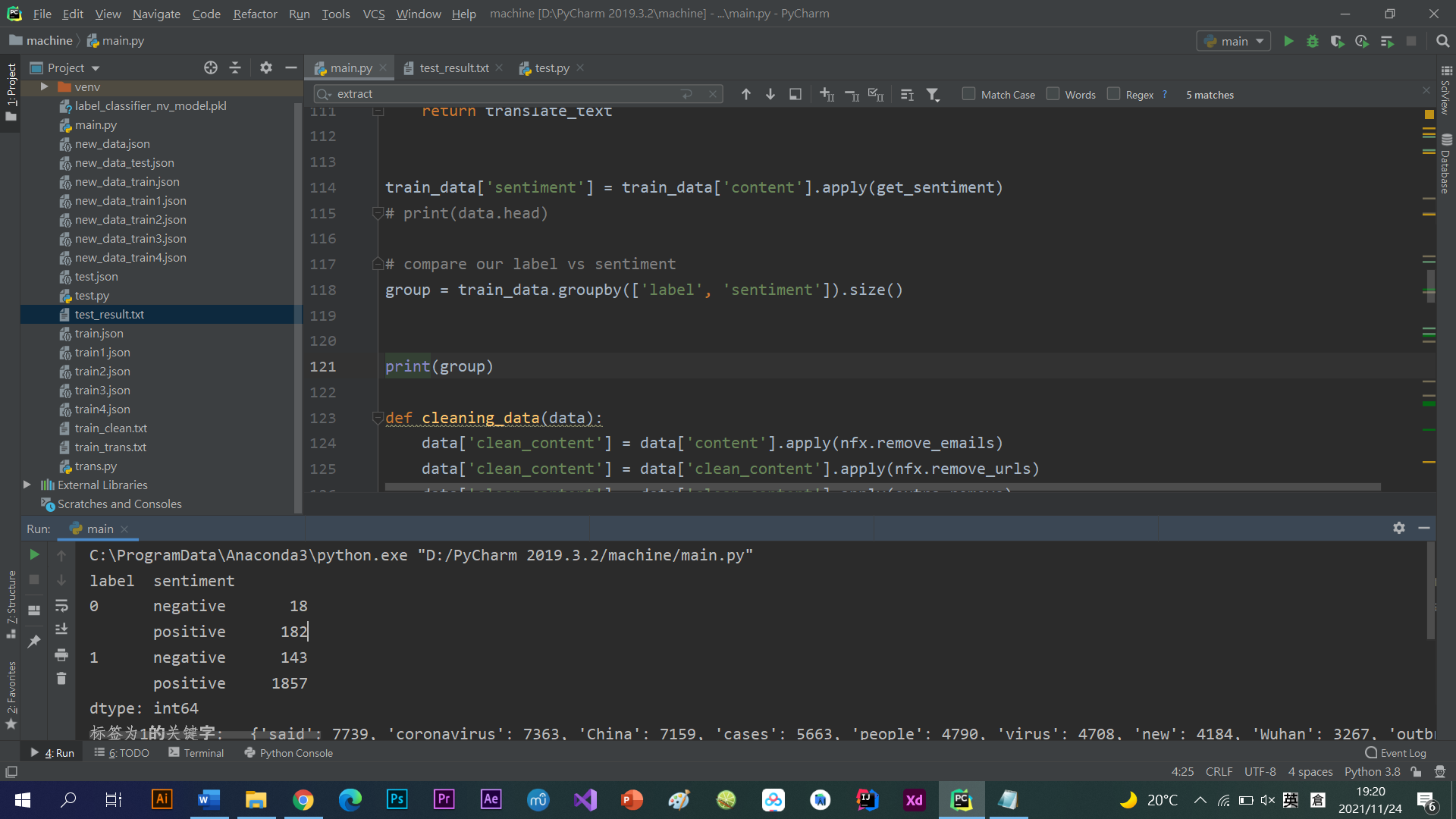


图1.5 文本的情感分析结果

1. 分别提取标签0和1的关键字

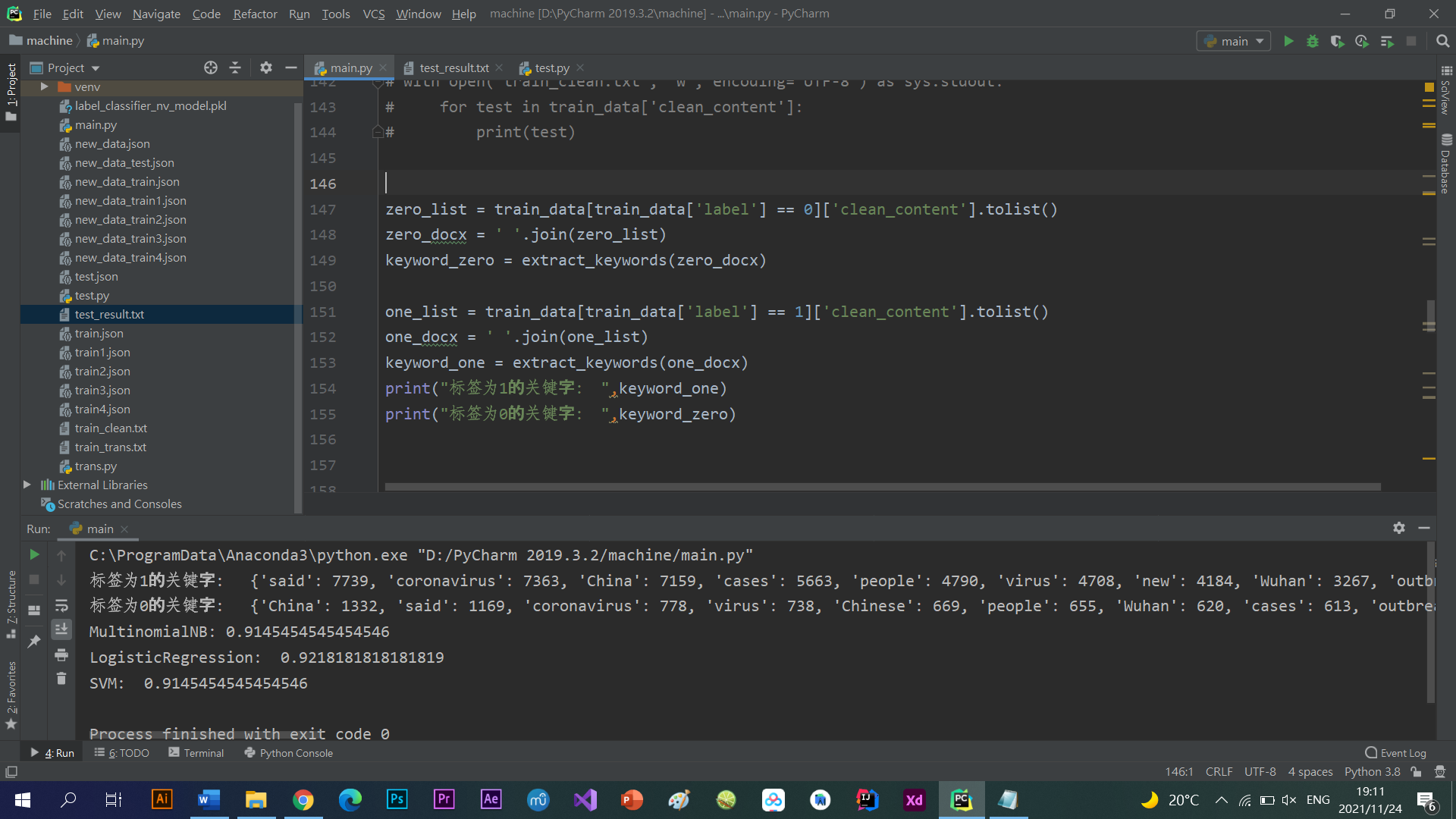


图1.6 标签的主要关键字

1. 查看训练集中的特征词及其索引

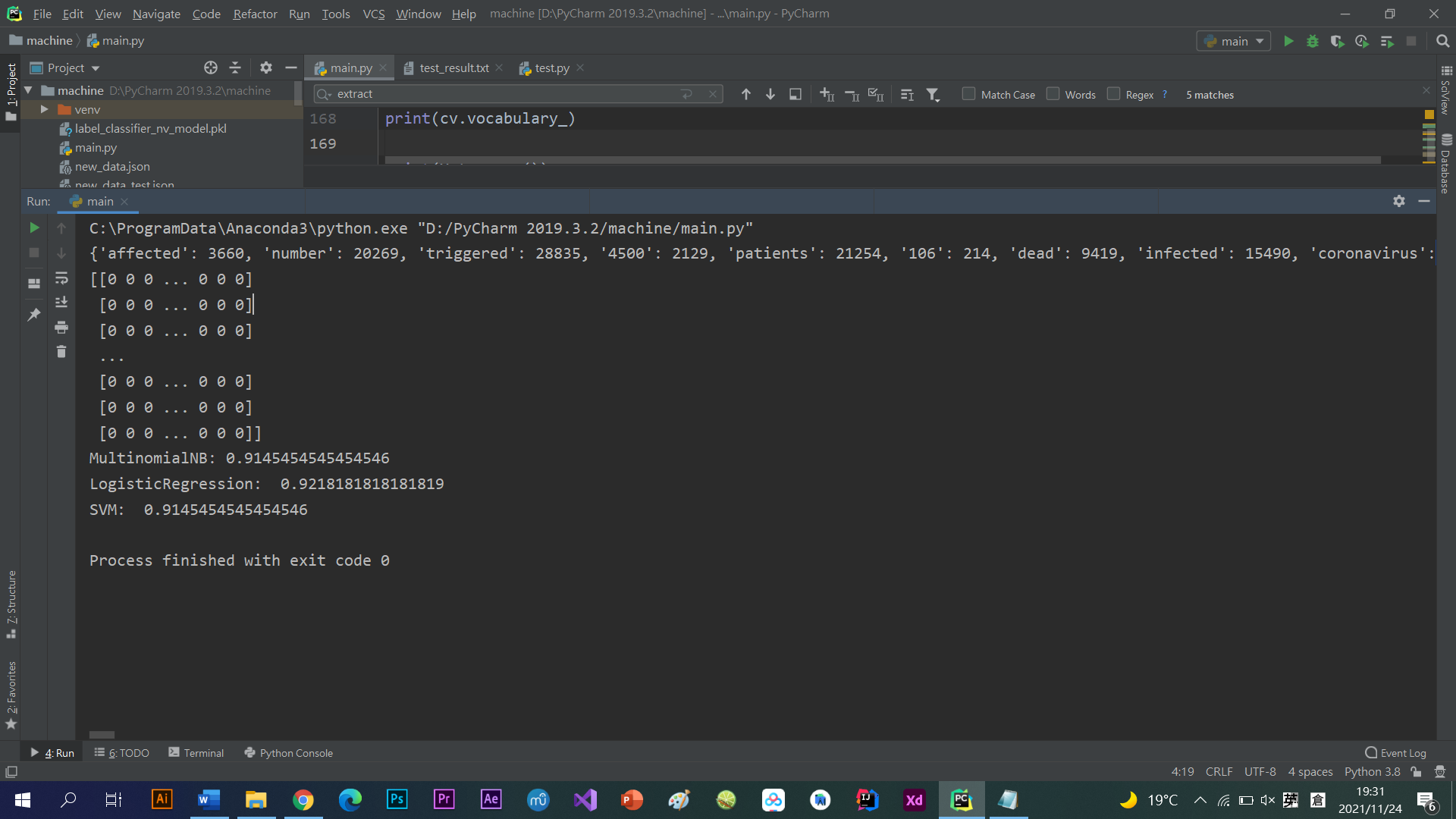


图1.7 特征词及其索引

1. 训练三种不一样的模型及其正确率

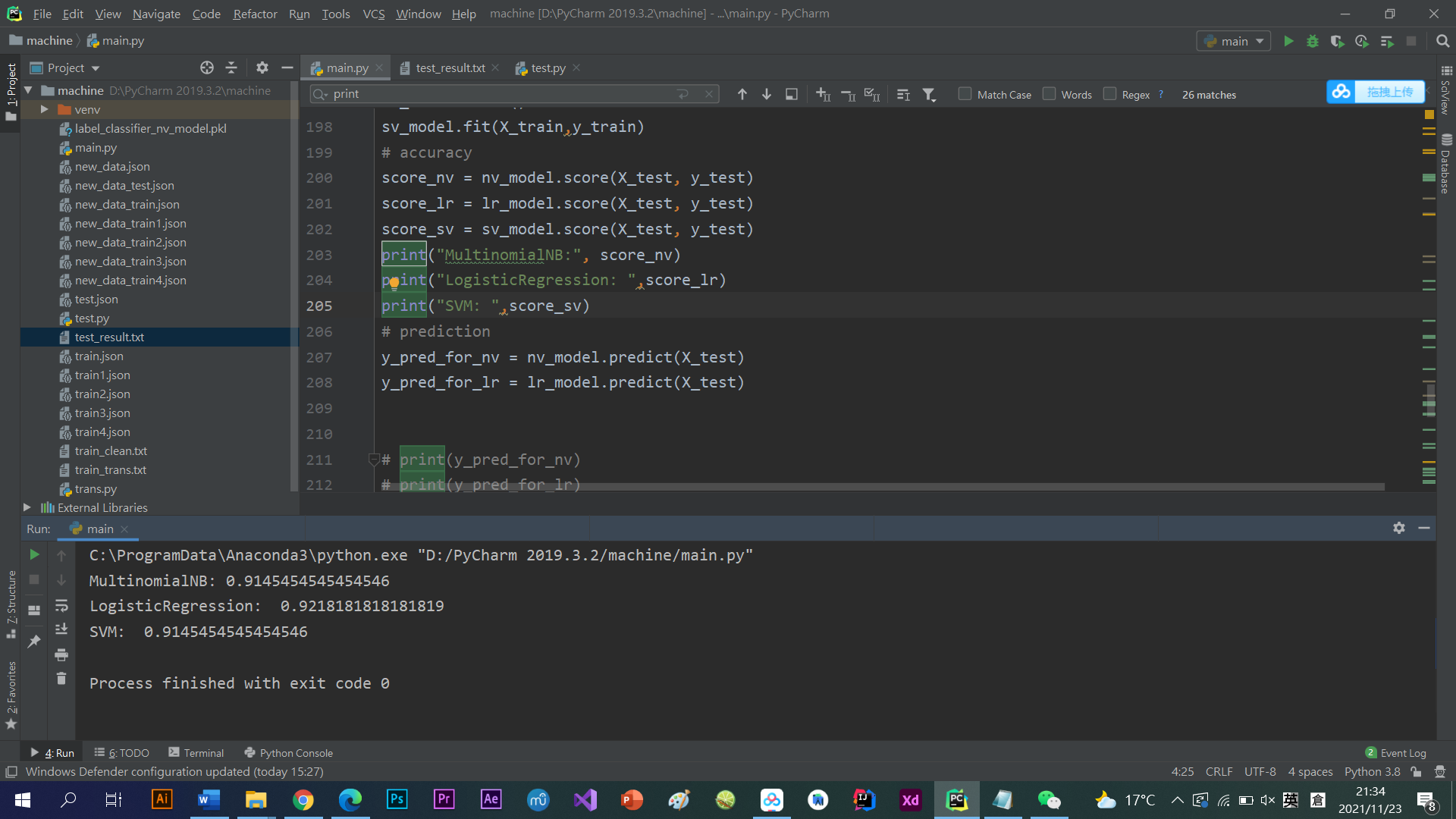


图1.8 三种模型的得分

1. 最优的预测结果

翻译文本和清洗数据后在头歌上的运行结果如图1.9。

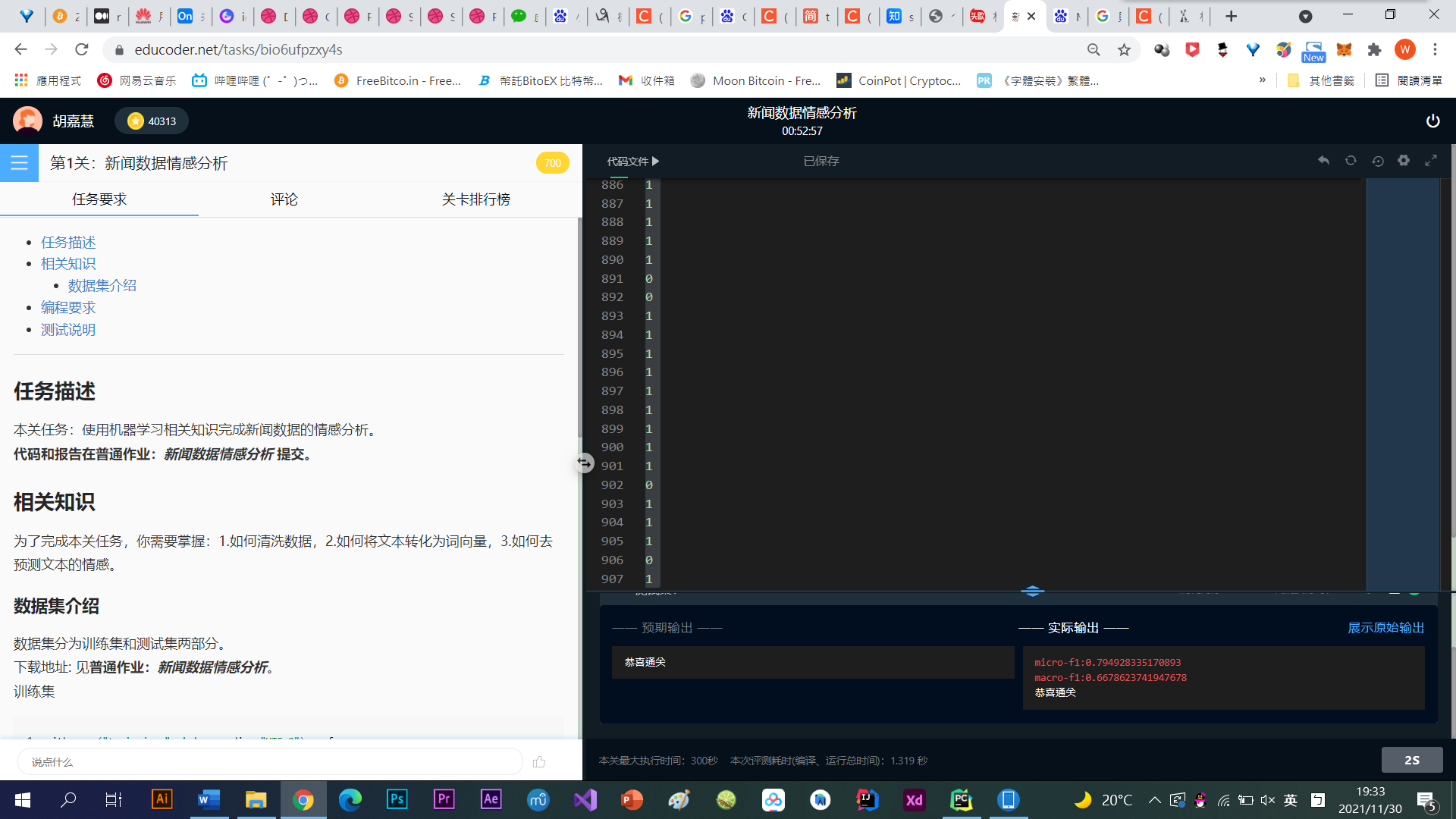


图1.9 头歌运行截图(1)

不进行文本翻译文本和清洗数据在头歌上的运行结果如图1.10。

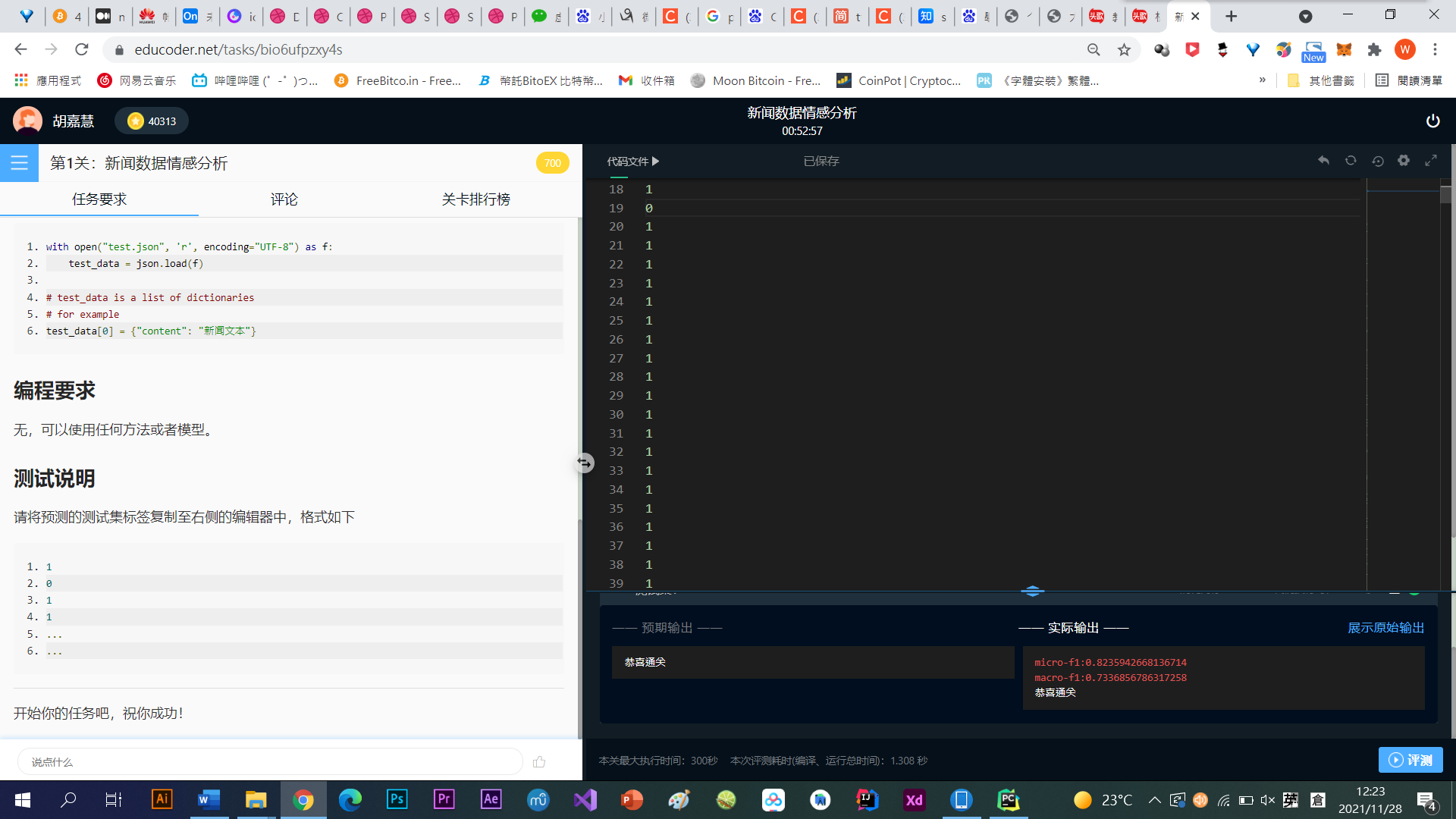


图1.10 头歌运行截图(2)

由图1.9和图1.10可得知不对文本进行翻译和清洗训练出的结果更优。

## 1.5思考与总结

本次项目使用了机器学习的方法对新闻数据进行了情感分析，通过这次项目，我学会了最基本的处理自然语言的方法，运用了很多不一样的方法来清洗数据文本，以达到最佳的效果，并实践应用了多个模型，最后比较模型的预测结果，从中选出最优的模型作为最终的结果。

通过机器学习这门课程，我学习到了很多知识，无论是理论上的，还是实践上的。虽然平时课上的内容都是偏理论的，而且课件都是英文的，理解上难免有些困难，但通过老师的讲解还是能够帮助我去深入了解到机器学习的基本知识，这是我第一次接触机器学习，靠着在课上所学的理论知识来查阅训练模型的实操方法，这个过程虽然繁锁，但也很有趣，最重要的是我在这个过程中受益匪浅。