《中文自然语言处理基础与实战》课程考查

项目开发文档

**一、项目概况**

1.题目：基于微博评论的情感分析

2.选题背景：

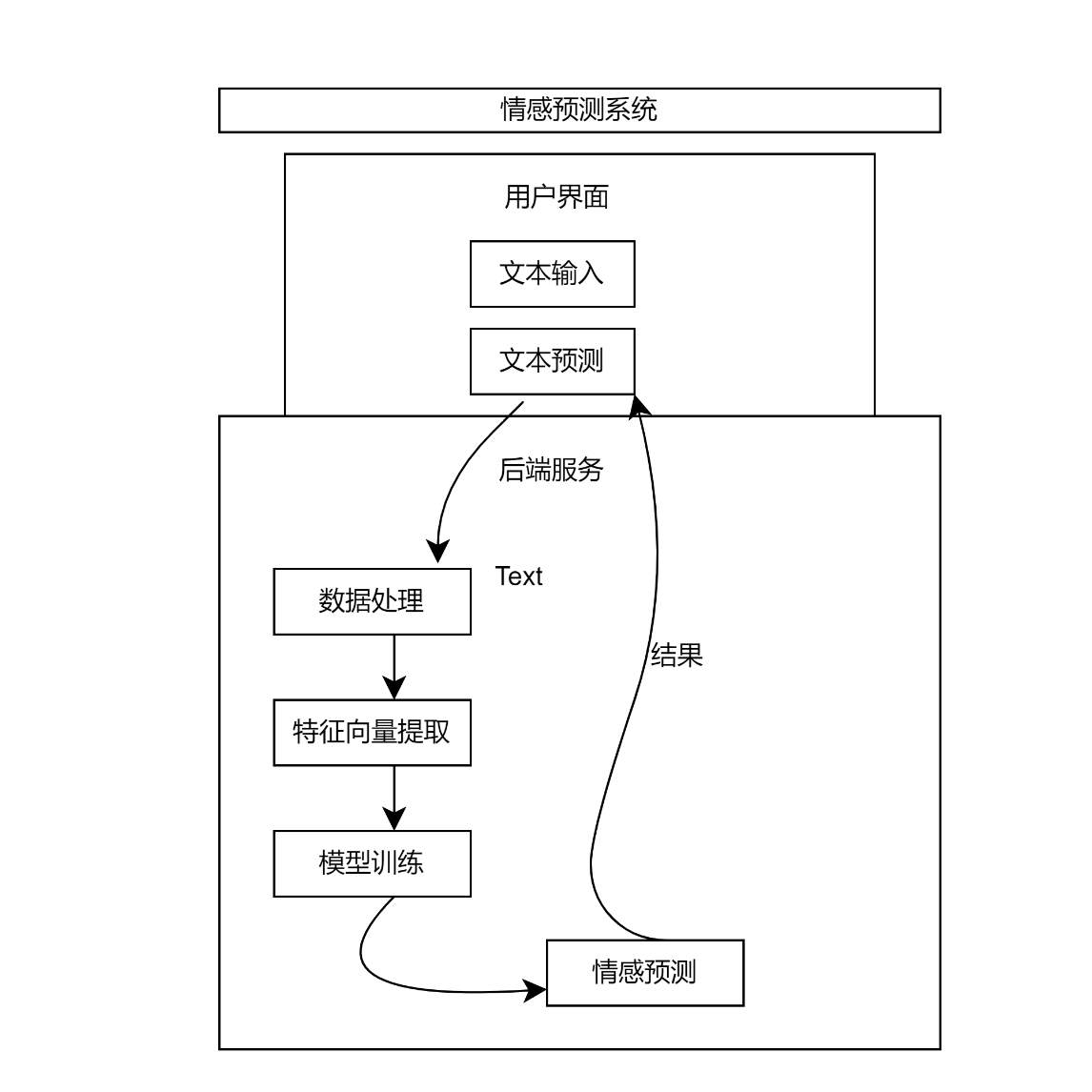
**1.社交媒体数据的丰富性与实时性**：随着互联网技术的飞速发展，微博作为中国最具影响力的社交媒体平台之一，每天产生海量的用户生成内容（UGC）。这些数据蕴含着丰富的社会舆情和情感倾向，微博评论中蕴含的情感信息对于了解公众对特定事件、产品或话题的态度和情绪倾向至关重要，对这些信息进行挖掘和分析，有助于政府、企业和研究机构提供科学的决策支持。

**2.情感分析在NLP中的重要性**：自然语言处理（NLP）是人工智能的一个重要分支，情感分析作为NLP领域的一个重要子领域，旨在从文本数据中自动提取和识别主观信息和情感倾向。情感分析在社交媒体、电子商务、新闻媒体等多个领域都有广泛的应用，对于企业和组织来说具有重要的价值。

**3.自然语言技术在情感分析中的应用**：传统的情感分析方法在处理大规模数据时效率较低，且准确性有待提升。近年来，随着自然语言技术的快速发展，如word2vec等模型在自然语言处理任务中表现出色，为微博情感分析提供了新的解决方案。word2vec能够学习词向量表示，将语义相似的词聚集在一起，在短文本情感分析中，可以利用word2vec将短文本中的词转换为向量表示，然后通过计算向量的相似度来分析文本的情感倾向。这种方法不仅提高了情感分析的效率，也提升了分析的准确性**。**

**二、需求分析**

1.**功能需求分析（功能图**）



**用户界面**：

**文本输入**：用户在Web页面的文本框中输入想要分析的文本。

**文本预测：**用户输入文本并提交表单，浏览器将表单数据通过POST请求发送到服务器，Flask后端接收到POST请求，调用/predict路由的predict函数，该函数提取文本，进行预处理，然后使用朴素贝叶斯模型进行情感预测。Flask将预测结果渲染并返回给前端，并展示在网页上。

**后端服务**：

**数据处理**：服务器接收到用户输入后，首先进行文本预处理。使用datachuli.py文件，涉及分词、去除标点和停用词等操作。

**特征提取**：预处理后的文本使用tf\_idf方法，由于我们使用的模型为NLTK的朴素贝叶斯分类器是属于基于特征集即通常是词袋模型的，而word2vec生成的是连续的向量表示，不能作为特征输入到NLTK的朴素贝叶斯分类器中，因为他们需要的输入格式不同，所以使用tf\_idf。

**模型训练**：使用train\_and\_save\_model()函数进行朴素贝叶斯分类器训练，train\_word2vec\_model函数进行Word2Vec模型。

**预测情感**：使用训练好的NLTK朴素贝叶斯模型对文本向量进行情感预测。使用 jieba库 对输入的句子进行分词。将分词结果转换为特征集，用于模型分类。

使用加载的模型对特征集进行分类，预测情感倾向，并返回结果。

2.非功能需求分析

**1.扩展性**

数据增长**：**

随着更多文本数据的积累，tf\_idf模型和朴素贝叶斯模型需要能够扩展以处理更大的数据集。这可能涉及到模型的定期再训练或使用更高效的数据存储解决方案。

功能扩展**：**

系统应设计为可以容易地添加新功能，如支持语言的情感分析、关键词提取功能与词性标注功能，或者集成更复杂的自然语言处理技术**。**

模块化设计**：**

代码应保持模块化，以便于单独更新或替换各个组件，如更换不同的机器学习模型或调整前端界面。

可视化界面多元：

不仅展示该次项目的情感分析界面，并且对数据集进行处理与展示——展示了数据集四种情感分布的饼图以及占比，通过tf\_idf特征提取展示了该数据集出现次数最多的前10个词的条形图，前100个词的词云图以及原数据集，便于更好了解原数据集的特征，还对分类模型的性能情况进行了混淆矩阵，精确率、召回率和F1分数条形图的展示，表现了模型的性能情况。

**2.稳定性**

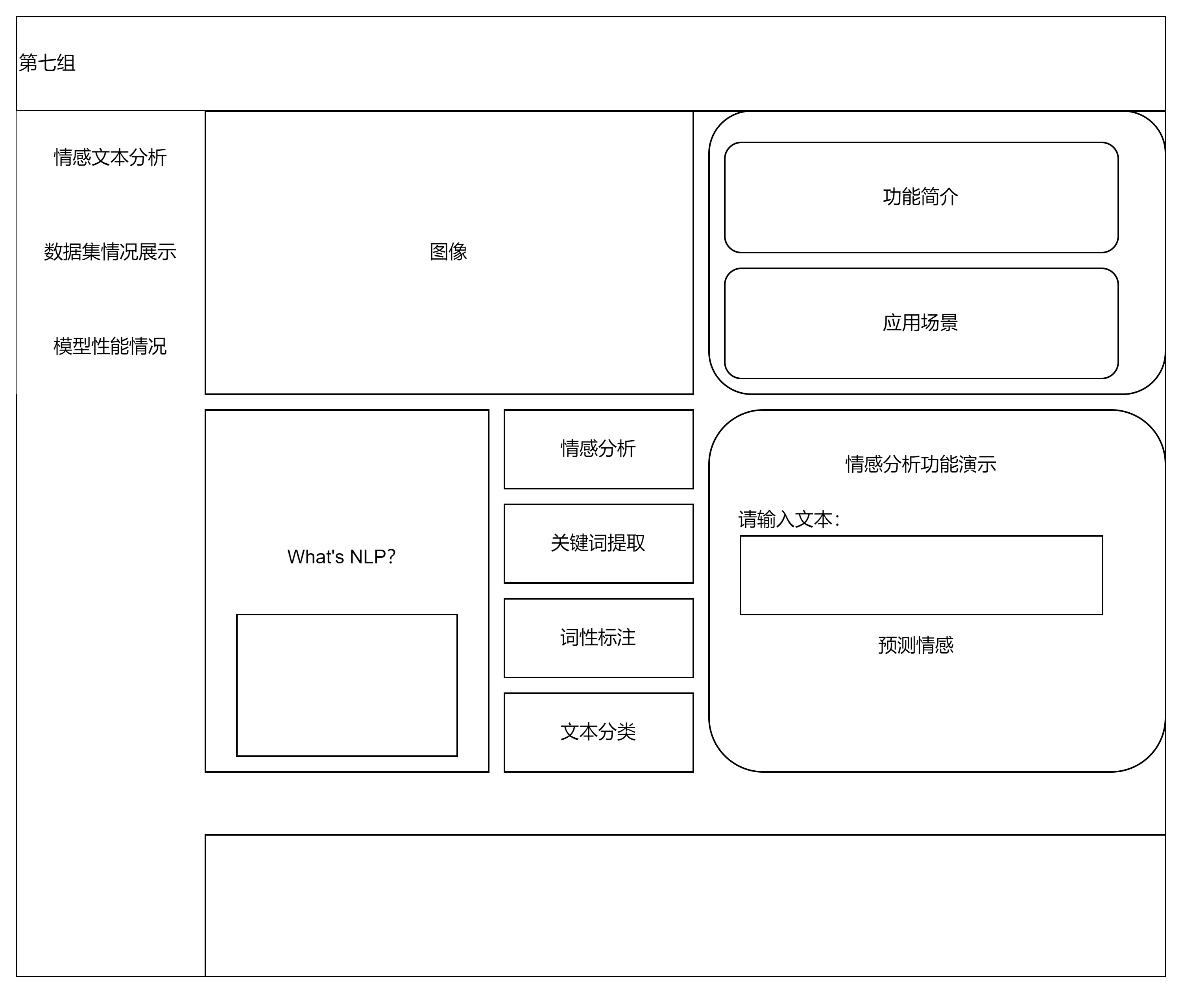
错误处理机制：系统应能够及时地处理错误情况，如输入无效数据、模型预测失败等，并给出清晰的用户反馈。

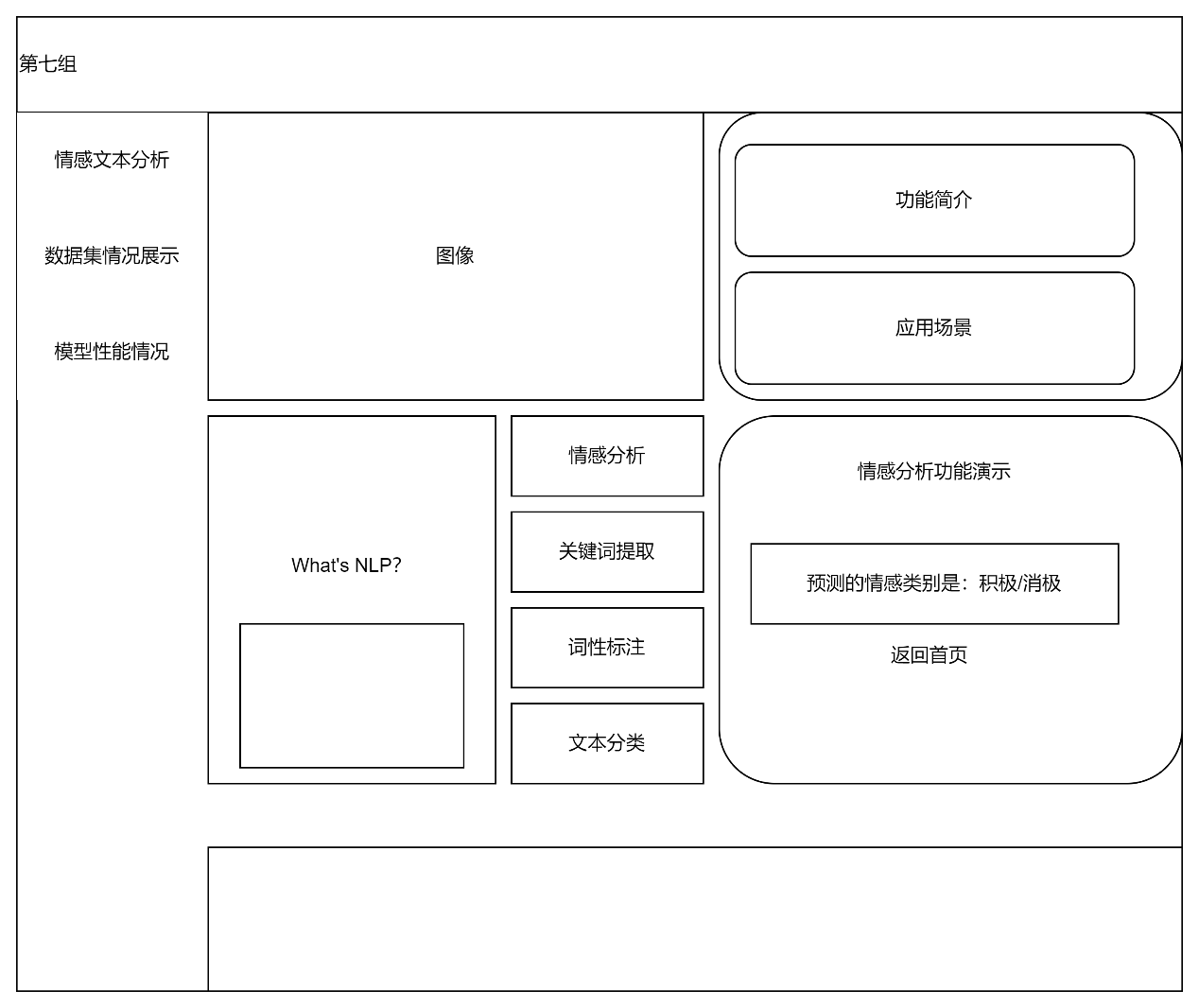
系统正常运行时间：系统应设计为高可用，确保在大部分时间内都能正常运行

模型训练及测试：定期进行模型训练和测试，确保模型准确度，为其提供持续改进的机会，增强模块间兼容性、提升整体稳定性。

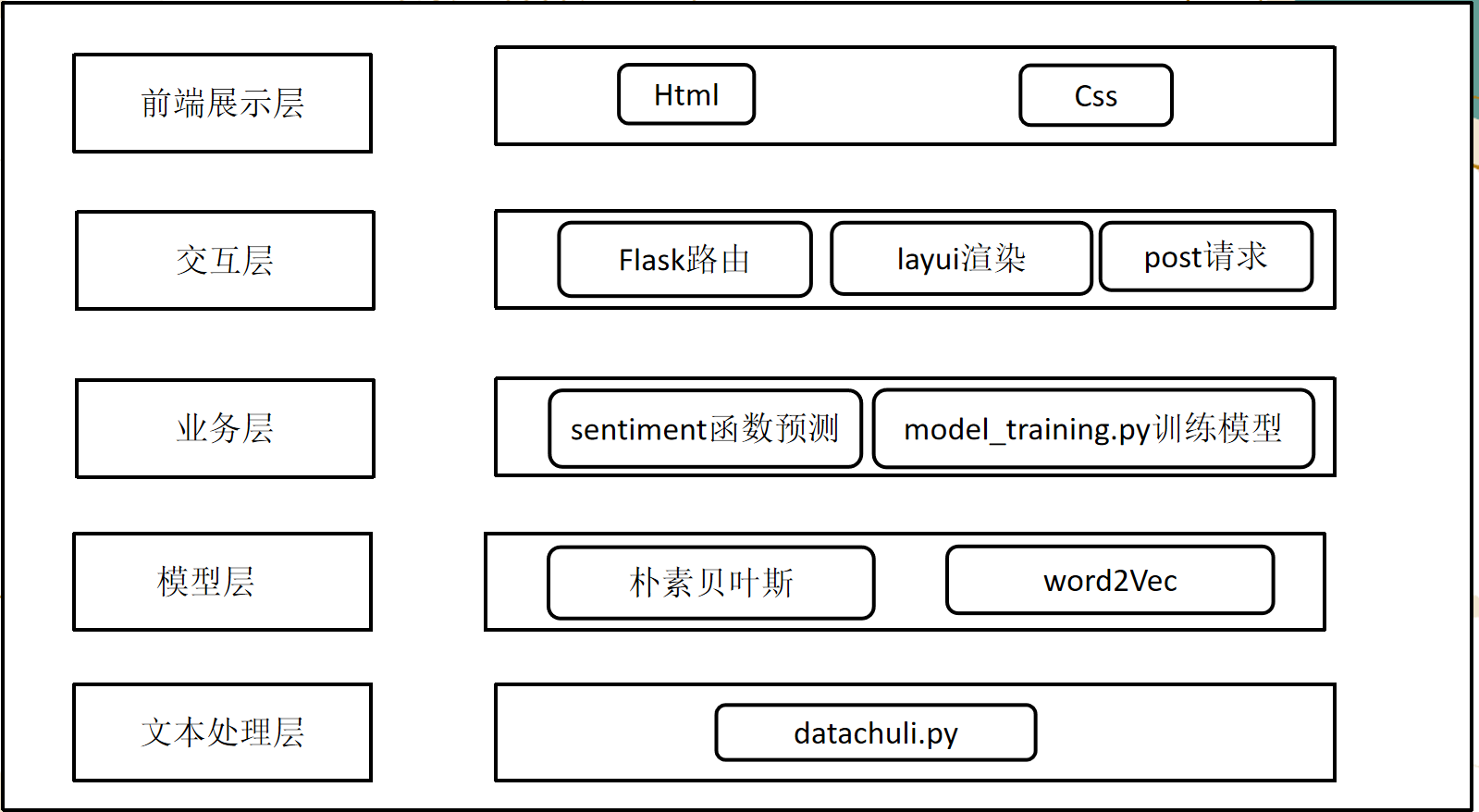
**三、项目设计**

1. 主界面：





2）项目技术框架



3）功能设计

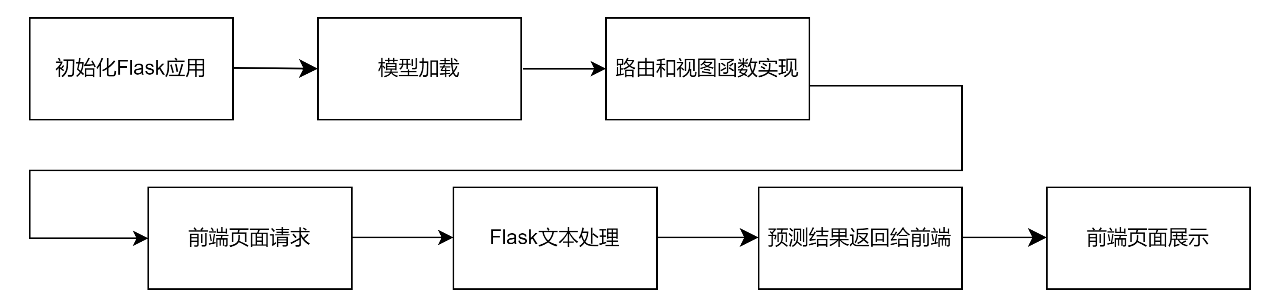
1. 网页功能：使用css+HTML实现网页界面文本输入与预测结果的展示。

流程图：



1. 后端功能：使用pycharm来编写代码。

流程图：



算法设计：使用py文件进行文本处理,Python编写的Flask进行网页框架的搭建。

**四、项目实现**

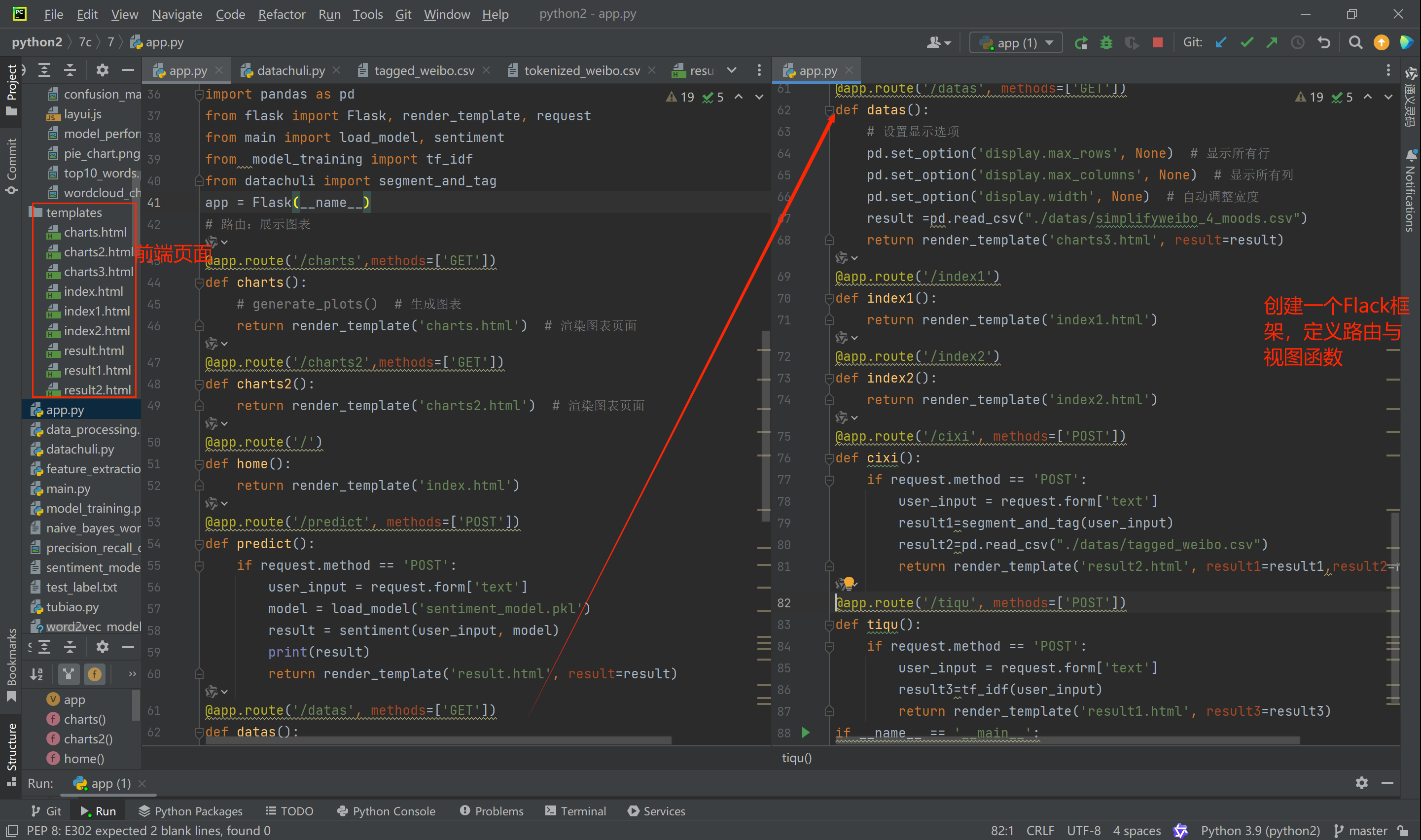
1. **界面实现**
2. 主界面



1. **核心功能实现**

（1）**网页功能：**

**技术介绍**：创建Flack框架，使用HTML来构建用户输入界面和结果展示界面，使用CSS进行界面的样式设计，使其美观且用户友好



**3．后端功能实现**

**1）文本预处理功能：**

**（1）文本清洗**

**目标**：移除文本中无关的信息，如HTML标签、特殊字符、数字等。

**技术**：使用正则表达式（re模块）来识别和移除不需要的字符和模式。

（2）**分词**

**目标**：将文本分割成单独的词汇或短语，以便进行进一步分析。

**技术**：使用jieba库进行中文分词，这是中文文本预处理的关键步骤。

**（3）去除停用词**

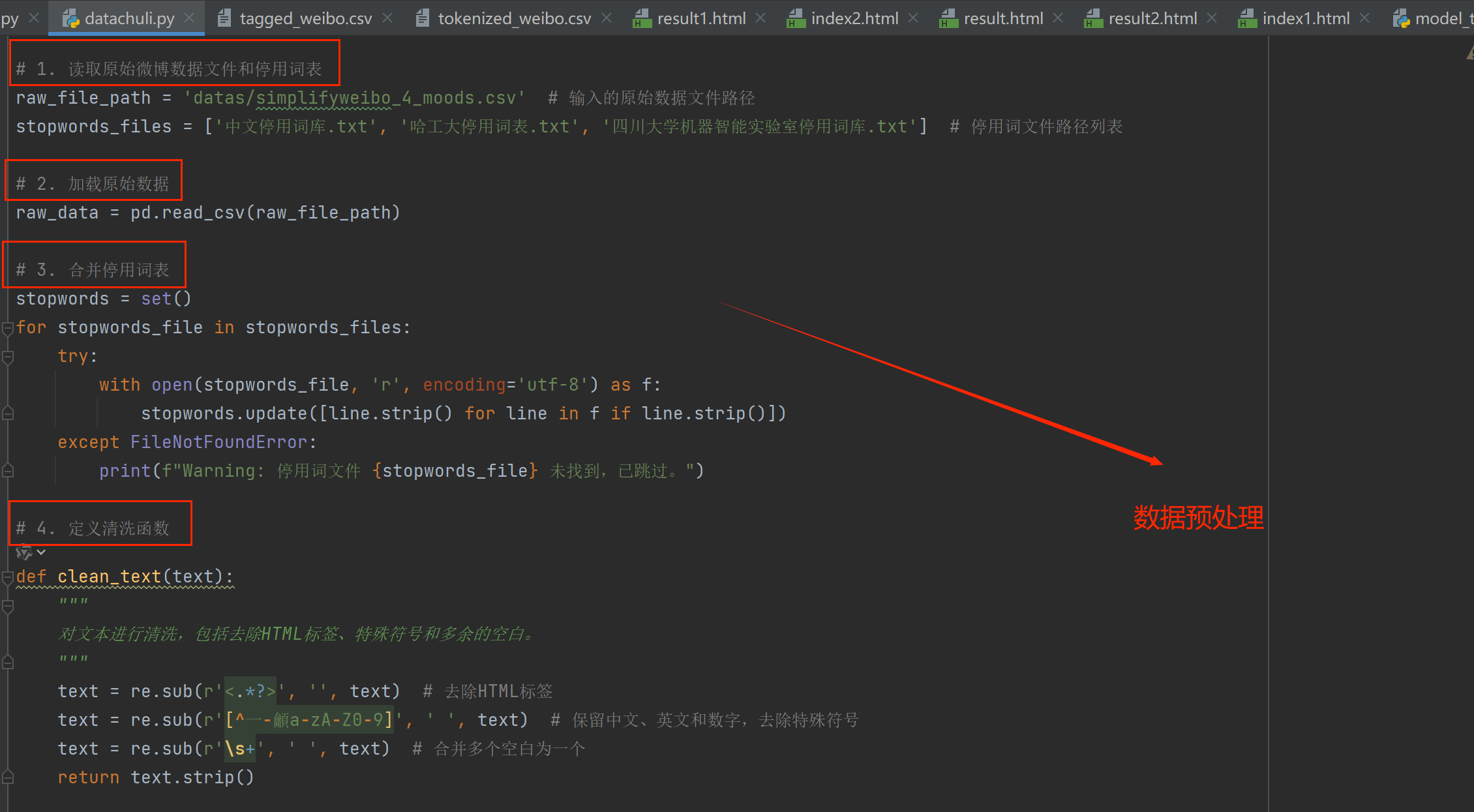
**目标**：移除文本中的常见词汇。

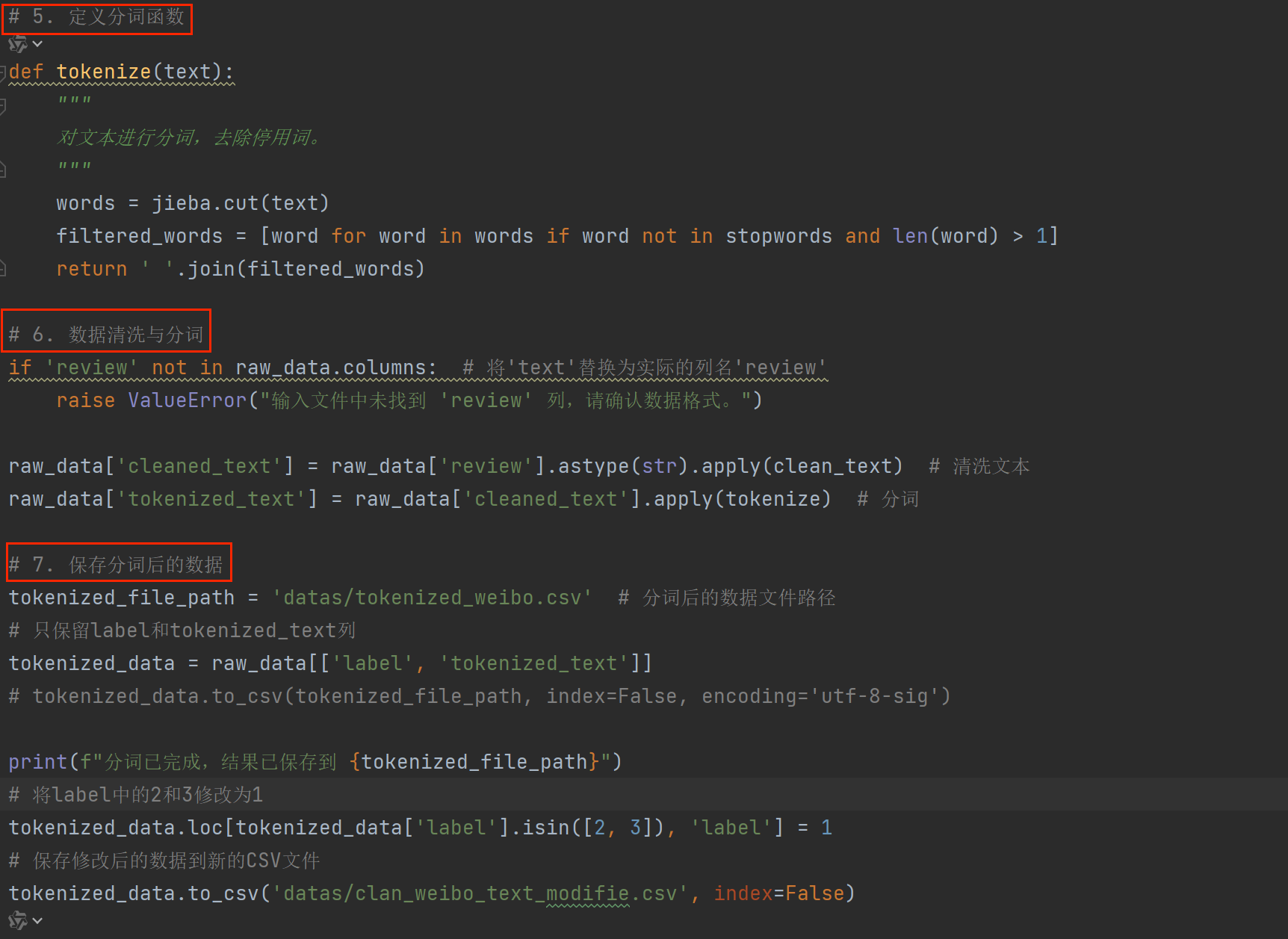
**技术**: 维护一个停用词集合（stopwords），其中包含了在文本处理中需要被去除的常见词汇，在分词中完成去除停用词。

**（4）提取特征值**

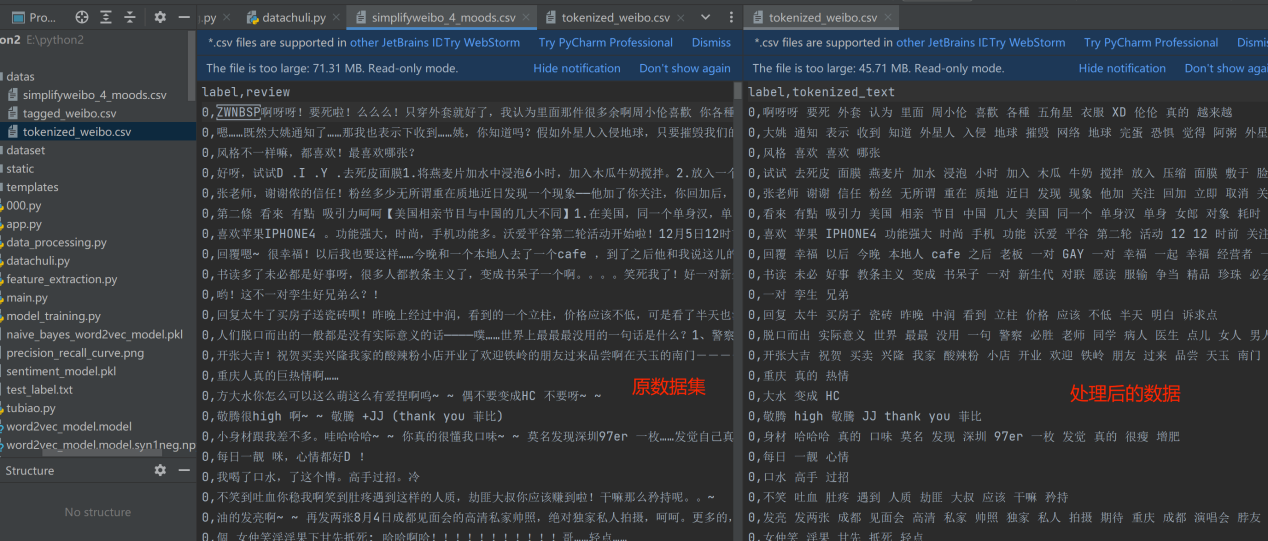
**目标**：目的是为了衡量词语在单个文档中的重要性，并有效抑制高频常见词（如“的”“是”等）的干扰，从而更准确地表征文档内容。

**技术**：TF-IDF技术通过结合两个统计量来度量词语的重要性：词频（TF）和逆文档频率（IDF）。









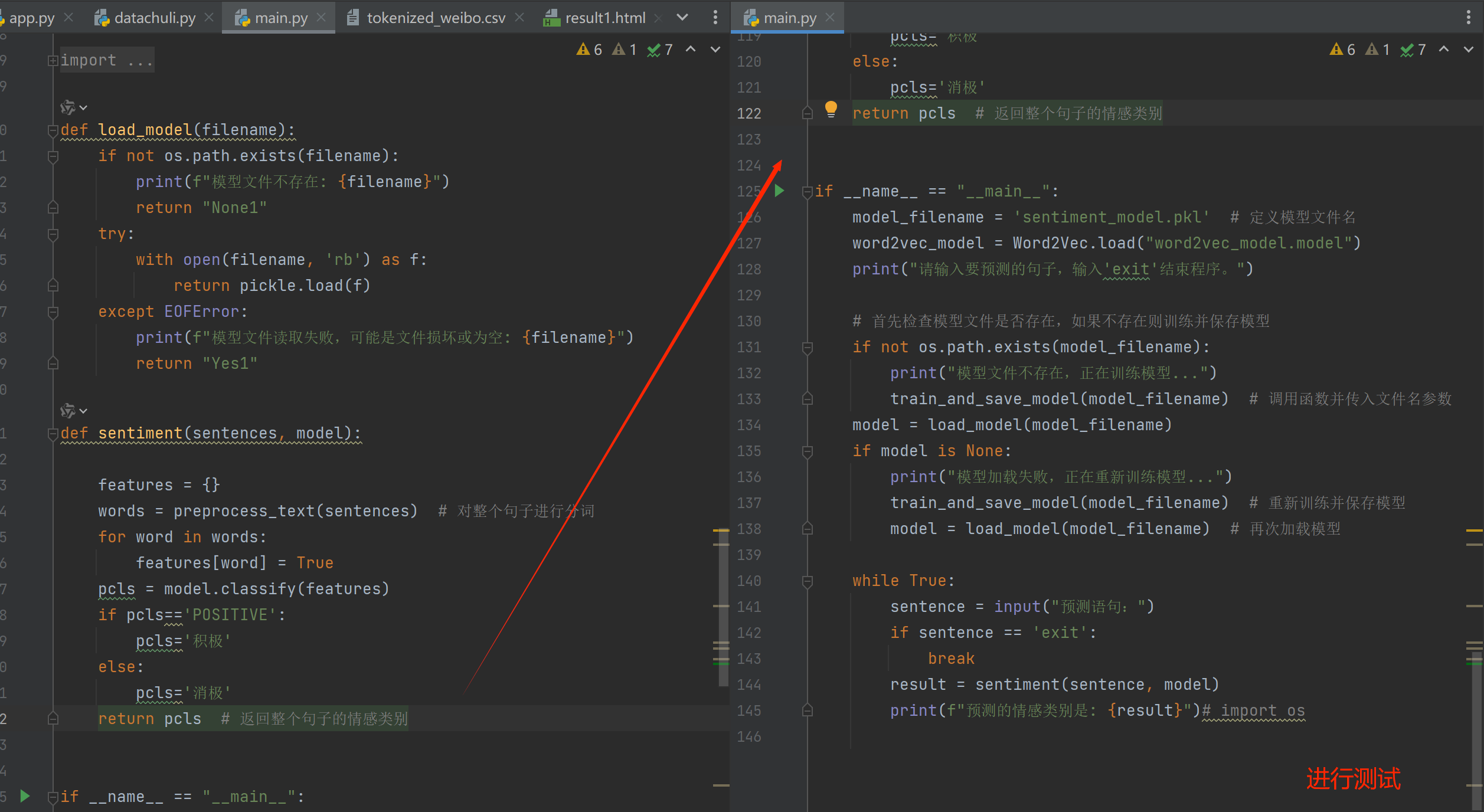
1. **情感预测功能**

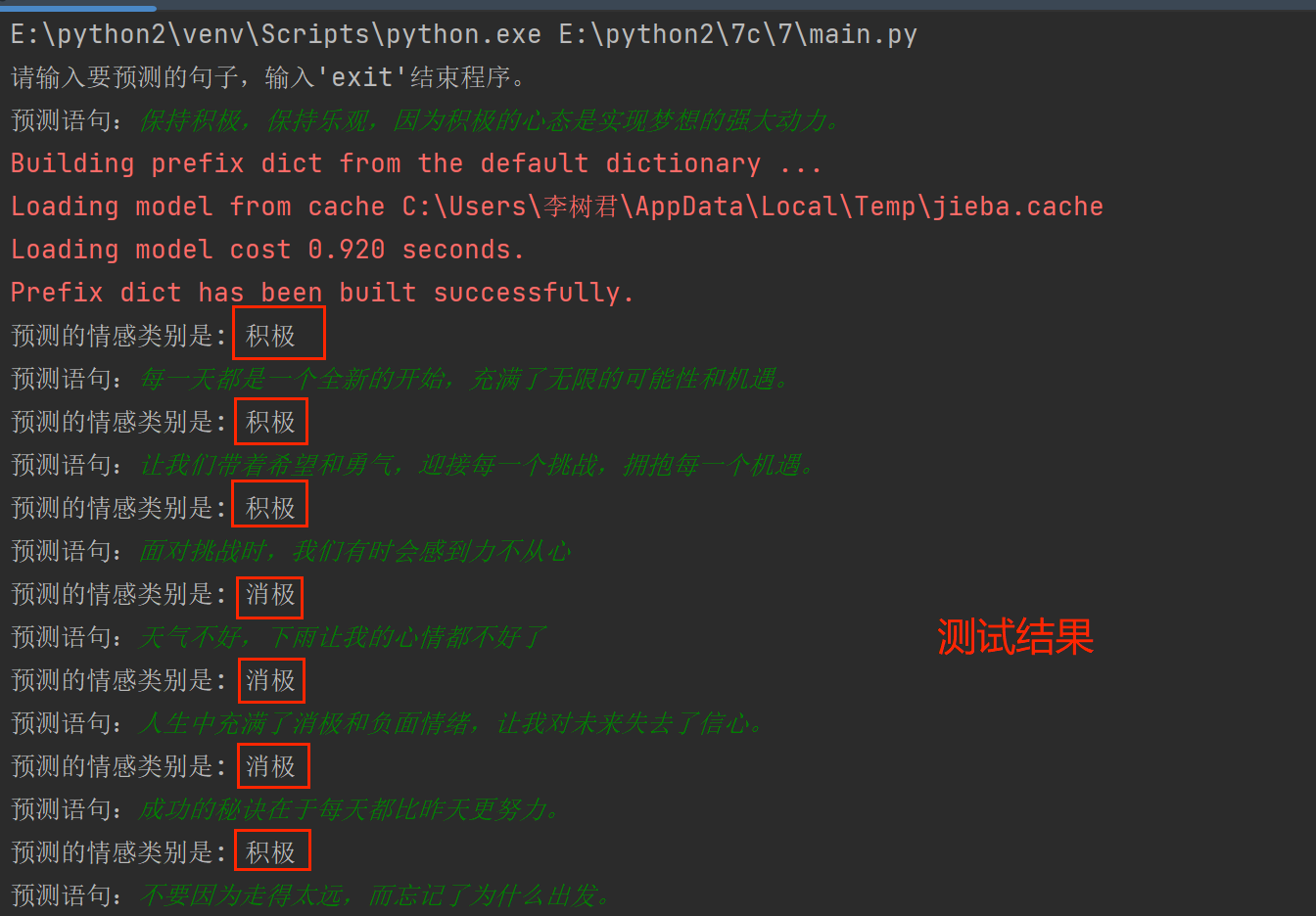
**技术介绍：**使用NLTK 的朴素贝叶斯分类器进行情感分类



**五、项目测试**

1、集成测试

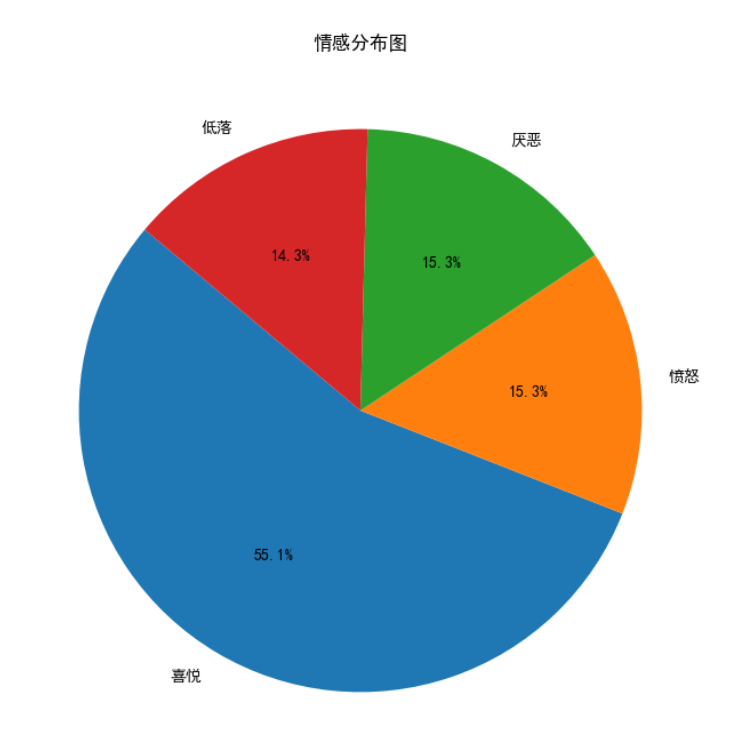




2.单元测试

数据集情况——展示四个情感类别在数据集的占比，以及出现次数最多的前10个词

1. 饼图展示了基于微博数据集的情感分布情况。图中将情感分为四个类别，并以百分比的形式表示每个类别在数据集中的占比。喜悦：占比最大，为55.1%。这表明在所分析的微博数据集中，表达喜悦情感的内容占据了超过一半的比例，显示出大多数用户在微博上分享的内容倾向于积极和正面的情感。愤怒：占比为15.3%。这表示有相当一部分微博内容表达了愤怒或不满的情感。厌恶：同样占比为15.3%，与愤怒的情感占比相同，说明厌恶情感在微博上也有一定的表达。低落：占比为14.3%，这是四个情感类别中占比最小的，但仍然有相当一部分内容表达了低落或悲伤的情感



（2）条形图展示了在某个数据集中出现次数最多的前10个词汇及其出现的次数。这些词汇可能代表了数据集中讨论的主要话题或常用词汇。以下是各个词汇及其出现次数的具体信息：回复：出现次数最多，为26834次。

今天：出现了22900次。

中国：出现了18899次。

喜欢：出现了18823次。

知道：出现了16715次。

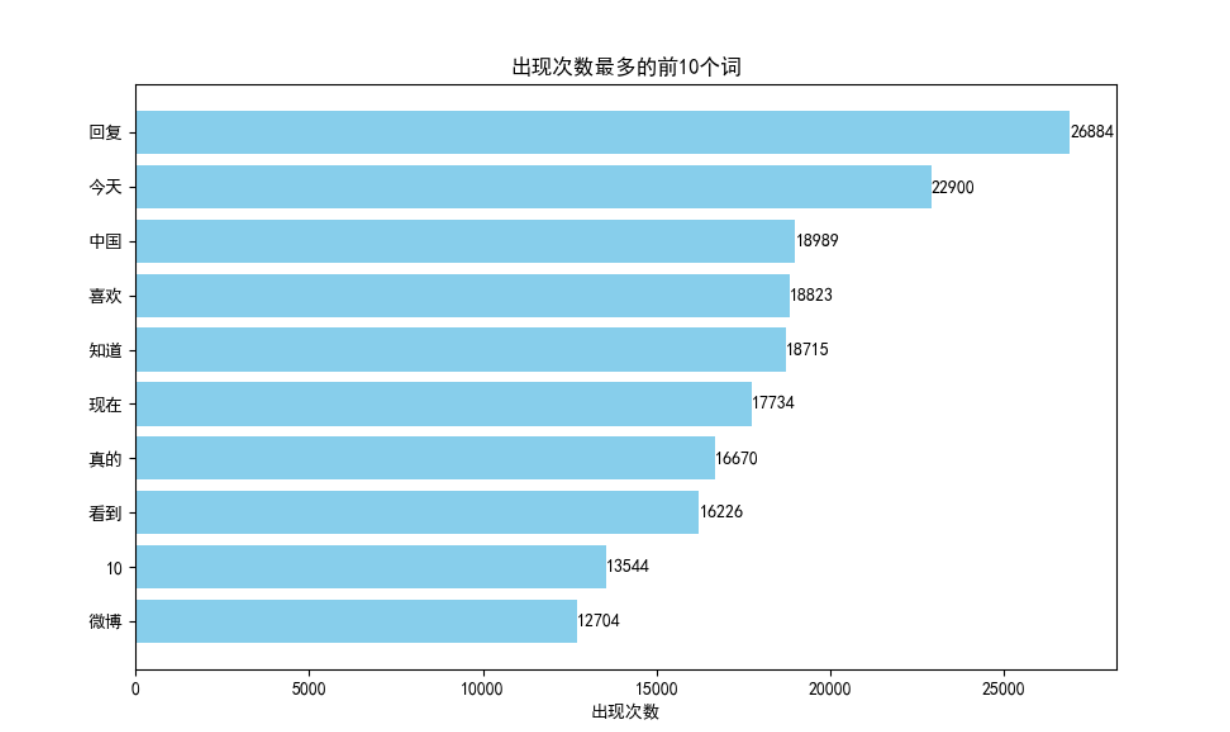
现在：出现了17734次。

真的：出现了16670次。

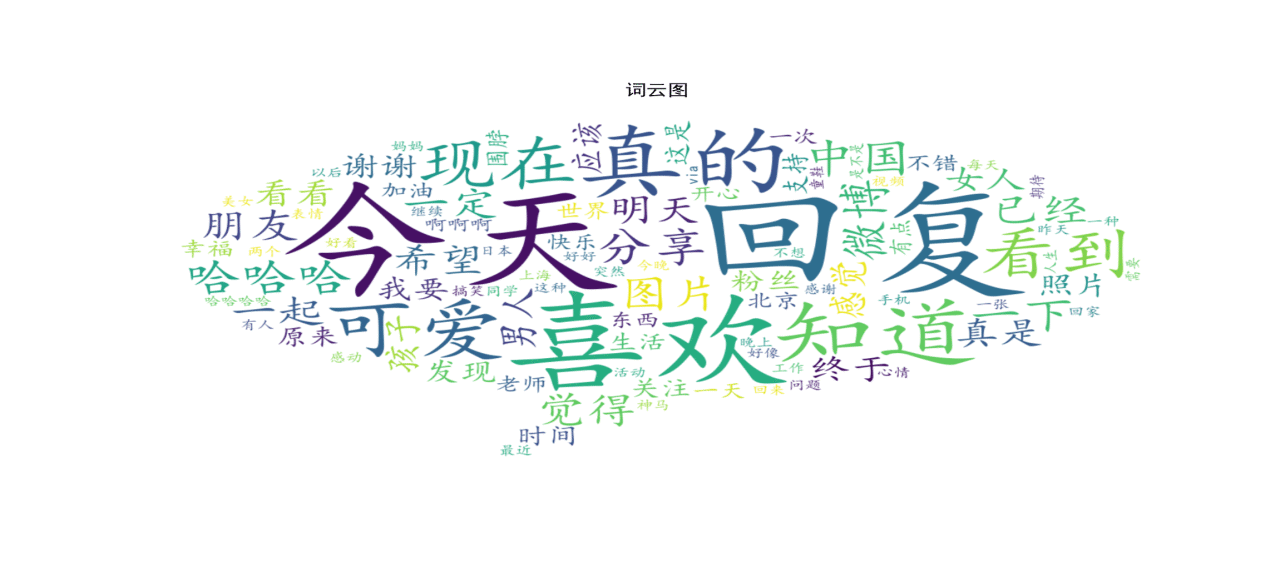
看到：出现了16226次。

怕：出现了13544次。

没有：出现了12704次。这些词汇的出现频率可以为分析数据集的主题、情绪倾向或用户行为提供线索。例如，如果“回复”和“今天”出现频率很高，可能表明数据集中有很多日常交流和即时互动。而“中国”和“喜欢”的高频率可能表明数据集中有很多关于中国话题的正面讨论。通过这些高频词汇，可以进一步挖掘数据集中的潜在模式和趋势。



（3）这张词云图展示了微博数据集中出现频率较高的词汇。词云图中，词汇的大小通常与其出现频率成正比，即词汇出现得越频繁，其在图中显示得越大。词云图是一种直观展示文本数据中词汇分布的方式，可以帮助快速识别数据集中的关键词和主题。通过分析这些词汇，可以对微博用户的兴趣、情感倾向和日常话题有一个大致的了解。



1. 数据集情况展示



模型性能情况——检验模型的精确率、召回率和F1分数条形图，查看模型性能

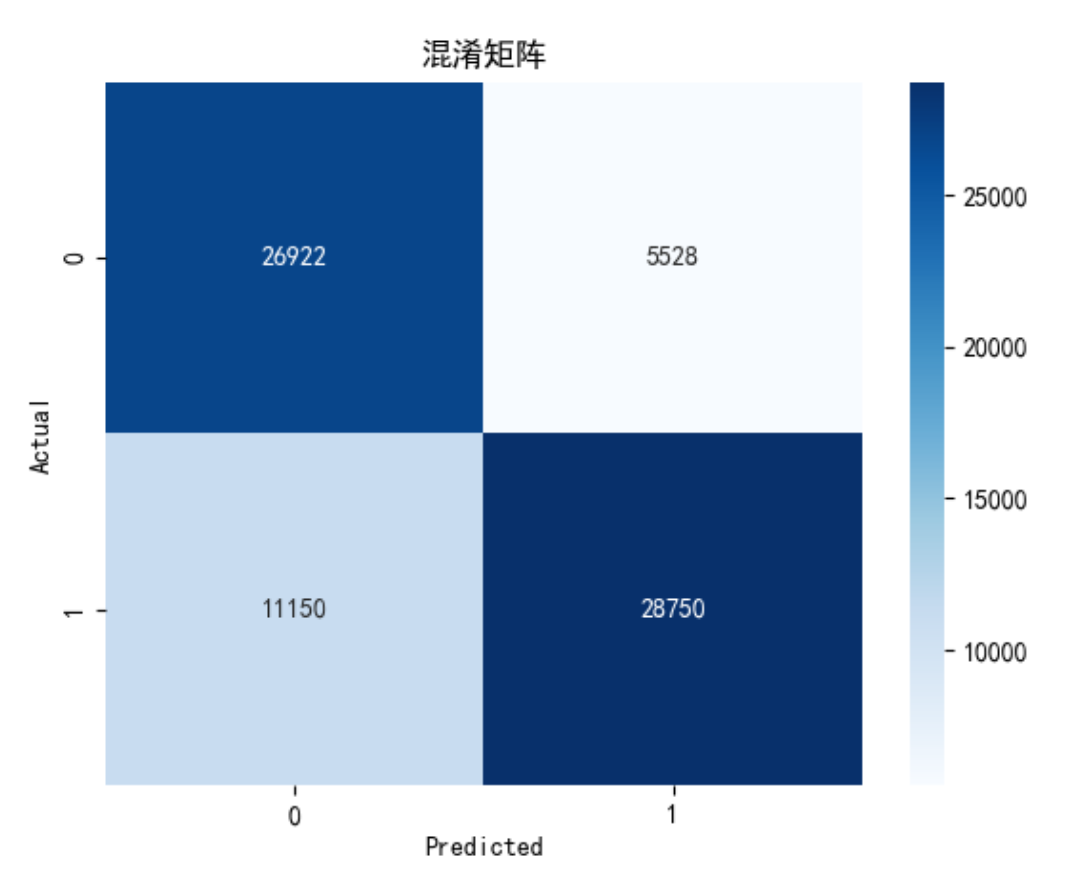
（1）混淆矩阵显示了模型预测结果与实际结果的对比：

True Negatives (TN): 左上角，值为 26922。模型正确预测为0（负类）的数量。

False Positives (FP): 右上角，值为 5528。模型错误预测为1（正类）但实际为0（负类）的数量。

False Negatives (FN): 左下角，值为 11150。模型错误预测为0（负类）但实际为1（正类）的数量。

True Positives (TP): 右下角，值为 28750。模型正确预测为1（正类）的数量。

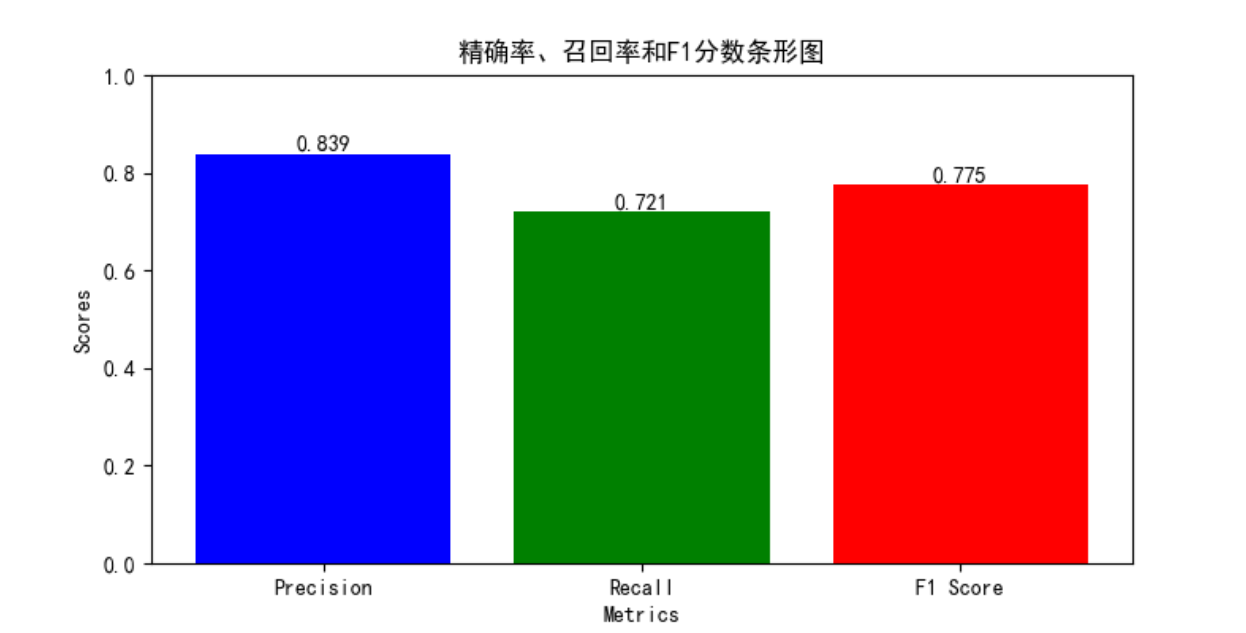


（2）精确率、召回率和F1分数条形图

Precision（精确率）: 蓝色条形，值为 0.839。精确率是指在所有被预测为正类的样本中，实际为正类的比例。计算公式为 TP / (TP + FP)。

Recall（召回率）: 绿色条形，值为 0.721。召回率是指在所有实际为正类的样本中，被正确预测为正类的比例。计算公式为 TP / (TP + FN)。

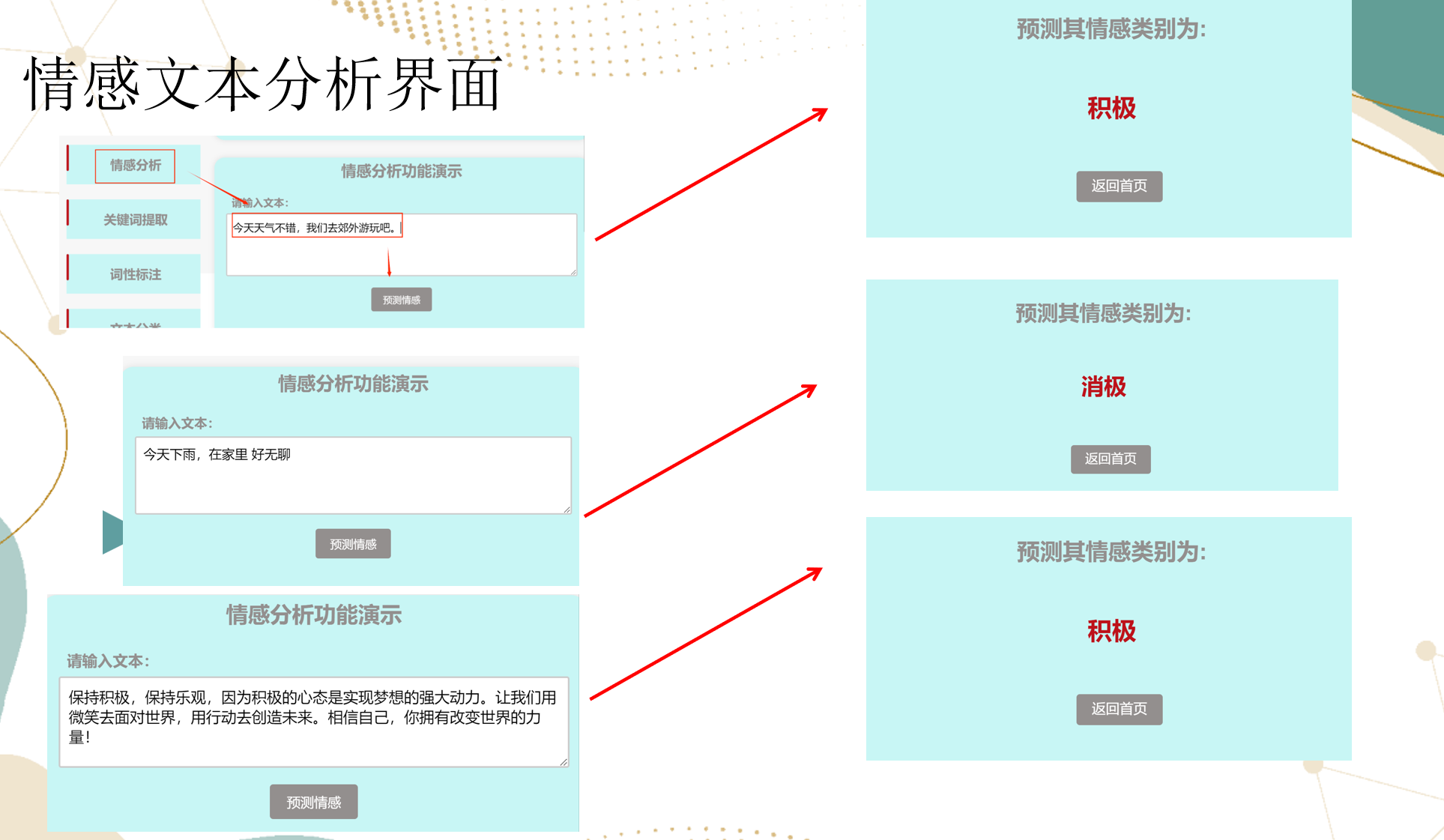
F1 Score: 红色条形，值为 0.775。F1分数是精确率和召回率的调和平均数，用于综合考虑精确率和召回率。计算公式为 2 \* (Precision \* Recall) / (Precision + Recall)。



**3.确认测试**

网页输入文本情感分析测试





**情感分析测试**、**关键词提取功能测试与词性标准功能测试**