

**项目开发报告**

**题目： 互联网虚假新闻检测**

**课程名称： 机器学习\_\_ \_**

**专业班级： CS1903 \_\_\_ \_**

**学 号： U201914995 \_**

**姓 名： 罗虞阳 \_\_ \_\_\_**

**指导教师： 李玉华\_\_ \_**

**报告日期： 2021.11.25 \_\_\_\_**

**计算机科学与技术学院**

# 项目开发报告

## 1.1项目目的

1. 使用机器学习相关知识设计算法完成互联网虚假新闻检测。

2. 加强使用python的熟练程度。

3. 加强自身工程项目开发以及调试能力，能够对常见错误进行修正。

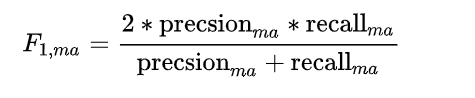
4. 熟悉python基础库，包括中文分词jieba库，机器学习sklearn库，数据处理pandas和numpy库，学习并调用合适库中的库函数。

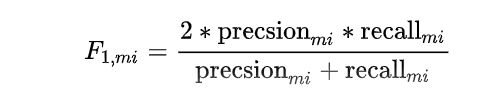
## 1.2问题分析

### 1.2.1 问题描述

使用机器学习相关知识完成互联网虚假新闻检测。为了完成本关任务，你需要掌握：1.如何清洗数据，2.如何将文本转化为词向量，3.如何进行虚假新闻的检测。

评价指标为F1 score，这是一个权衡Precision和Recall 的指标，他表示为这两个值的调和平均。 当任务为多分类任务时，precision和recall的计算方式就需要权衡每一类的precision\_i和recall\_i，Micro和Macro就是两种不同的权衡方式。

计算方法：将所有类别的Precision和Recall求平均，然后计算F1值作为macro-F1； 计算方法：先计算所有类别的总的Precision和Recall，然后计算出来的F1值即为micro-F1； macro-F1的计算公式为： 

micro-F1的计算公式为： 

### 1.2.2 求解思路

要想完成互联网虚假新闻检测，需要先分析数据集和测试集，并且对数据进行清洗。然后要对清洗好的文本进行转换，将其转换为词向量。最后就是对得到的每个句子的向量，挑选合适的机器学习模型进行学习和预测。然后将得到的预测结果转化为我们需要提交的数据格式，进行提交。

## 1.3设计与分析

### 1.3.1 数据分析

数据中有训练集和测试集，训练集包含两列，一列是给出的文本，还有一列是文本的分类，分类分为无需判断、虚假新闻、真实新闻三类，标签分别为0，-1，1。测试集只有一列，就是带判断文本。数据为文本，需要处理成机器学习能处理的向量模式。

### 1.3.2 算法流程设计

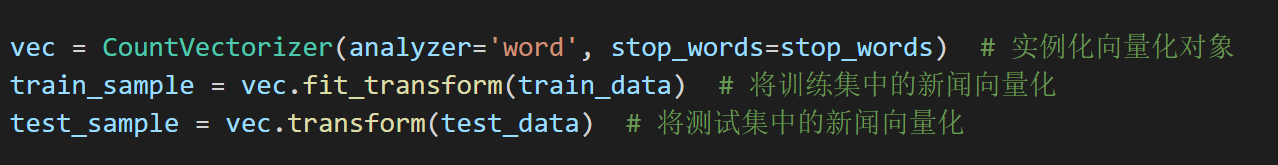
首先使用pandas的read\_csv函数读取训练集，然后将训练集中的空缺部分删去。接着使用jieba库中的cut函数进行对评论分词的操作，将分词后得到的结果保存在train\_data中，将标签保存在train\_label中。



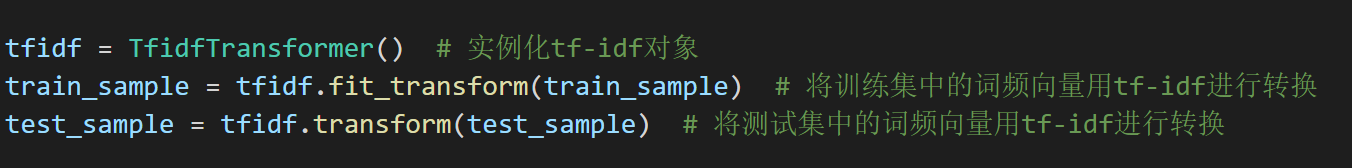
然后对训练集做同样的操作，只是没有对标签的处理。



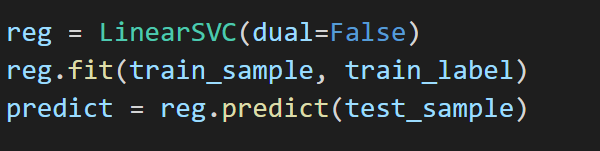
然后使用CountVectorizer构建词向量。对train\_data和test\_data中的数据使用停用词进行清洗，清洗掉对结果有影响但对人理解无影响的词，得到train\_sample和test\_sample。



实例化tf-idf对象，对训练集和测试集中的词频向量使用tf-idf进行转换，对特征进行提取。这里使用tf-idf是因为通过网上查阅资料了解到tf-idf是词频-逆向文件频率，用于信息检索和文本挖掘的加权技术，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度，如果某个词在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此类词或者短语具有很好的类别区分能力，适合用于分类。



最后使用sklearn库中的LinearSVC函数对训练集进行学习和对测试集的预测。



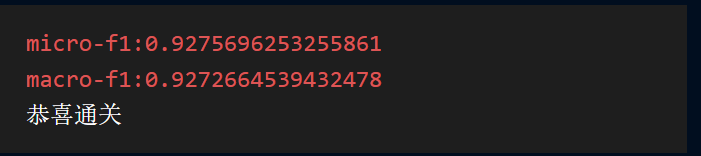
### 1.3.3 机器学习算法设计

此次实验样本数目较少，并且是多分类问题，所以使用SVM支持向量机来解决。基本思想是寻找一个超平面来对样本进行分割，把样本中的正例和反例通过超平面分开，其原则是使得正例和反例的间隔最大。

LinearSVC算法线性支持向量分类，类似于SVC，但是其使用的核函数是”linear“，所以它具有更大的灵活性在选择处罚和损失函数时，而且可以适应更大的数据集，他支持密集和稀疏的输入是通过一对一的方式解决的

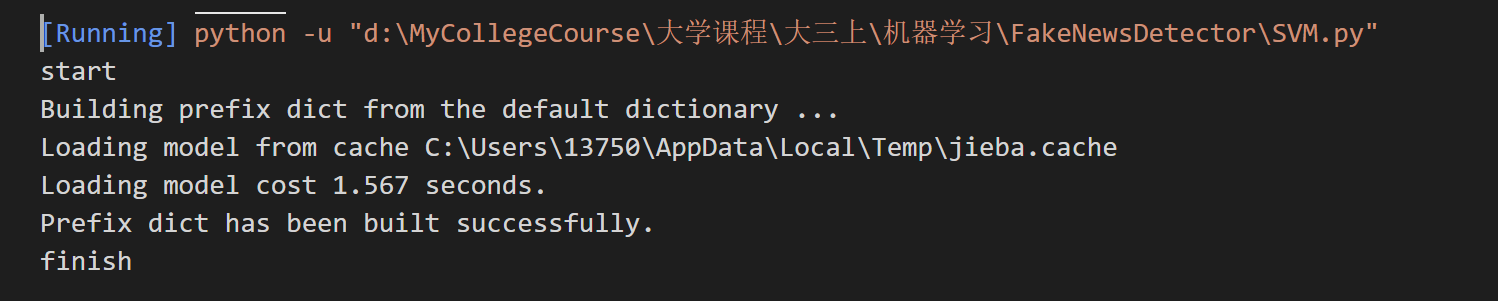
在 LinearSVC 中没有 kernel 这个参数，限制我们只能使用线性核函数。由于 LinearSVC 对线性分类做了优化，对于数据量大的线性可分问题，使用 LinearSVC 的效率要高于 SVC。

经过测试，LinearSCV()的准确率为：

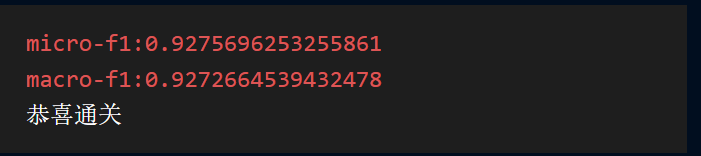


## 1.4结果分析

程序训练过程的运行截图如下图所示：



提交结果如图：



这次提交的结果是很好的，之前还有使用过朴素贝叶斯模型和神经网络，效果没有使用LinearSVC模型效果好。所以随后提交的就是LinearSVC模型。

## 1.5思考与总结

通过这次的机器学习课程结课作业，我学会了如何对数据集进行清洗；同时我也学习到了如何处理中文文本，包括对停用词的使用，还有就是使用tf-idf进行特征提取，以及使用jieba库进行分词并构建词向量；还有就是熟悉了如何使用sklearn.SVM库中的LinearSVC模型来处理词向量并且预测出结果。这次的实验难度并不是非常大，更多的是让我们熟悉一整个处理流程，通过自己一整个流程走下来还是学习到了许多。