**实验3 基于机器学习的文本分类**

**【实验目的】**

掌握基于机器学习实现文本分类的基本过程，掌握文本预处理、分词和分布式文本表示的具体实现，掌握基于朴素贝叶斯、逻辑回归和支持向量机的新闻文本分类。

**【实验要求】**

1.掌握文本的分布式表示：word2vec及Golve。

2.掌握朴素贝叶斯、逻辑回归、MLP和支持向量机的文本分类原理

3.编写代码实现基于朴素贝叶斯、逻辑回归、支持向量机和MLP的文本分类，依据精确率和召回率定量分析这三种分类算法。

<https://blog.csdn.net/m0_74053536/article/details/134367257>

<https://blog.csdn.net/weixin_47058355/article/details/134517331>

**【实验准备】**

1. **文本分类数据集的准备**

实验数据为14个新闻主题类别文件夹：财经、彩票、房产、股票、家具、教育、科技、社会、时尚、时政、体育、星座、游戏和娱乐

每个子文件夹下面都有若干个txt文档集，txt文本的内容是去掉新闻网页标签后的纯文本内容，包含数字、标点符号等信息。

停词表文件为：stop\_words.utf8

1. **掌握sklearn中相关函数的使用**

(1)数据集划分 sklearn.model\_selection中train\_test\_split

(2)文本表示的两种方法：word2vec和glove

(3)朴素贝叶斯 sklearn.naive\_bayes

(4)逻辑回归 sklearn.linear\_model

(5)分类结果评估 sklearn.metrics

**【实验设计】**

1. **文本数据读入并建立标签**

**定义函数 get\_data( )**

输出：新闻文本列表，标签列表

①读取新闻文本数据，获取新闻txt文本和对应的新闻类别（财经等）

②建立多分类标签

③将十四个文件夹内的新闻txt文本整合为一个新闻文本列表。列表元素就是一个新闻文本

④将十四个新闻类别的标签组合成邮件文本数据集标签。

1. **新闻文本数据集的预处理**

**每个邮件文本预处理**

**定义预处理函数** normalize\_corpus(corpus, labels)

输入：get\_data()输出的新闻文本列表corpus，标签列表labels

输出：预处理后的新闻列表，标签列表

**预处理函数完成的工作**

1. 分词

定义分词函数 tokenize\_text(text)

输入：text 新闻文本

调用jieba.cut分词，并去除前后空格

1. 去除停用词

定义去除停用词函数 remove\_stopwords(text)

输入：text 新闻文本

输出：去除停用词的邮件文本

1. **邮件数据集分类**

运用sklearn.model\_selection中train\_test\_split对邮件数据集进行训练集和测试集的分类

1. **邮件文本的特征表示**

本次实验word2vec和Glove，建立分布式词向量矩阵。

1. **定义函数Word2Vec\_extractor( )**

涉及到的第三方库：gensim

导入Word2Vec

from gensim.models.word2vec import Word2Vec，LineSentence

LineSentence()类主要是处理文本中文本转换成句子信息。

sentences = LineSentence(dataset\_path)

类里的参数解释

Word2Vec（）类里的参数解释

gensim.models.word2vec.Word2Vec(sentences=None, corpus\_file=None, vector\_size=100, alpha=0.025, window=5, min\_count=5, max\_vocab\_size=None, sample=0.001, seed=1, workers=3, min\_alpha=0.0001, sg=0, hs=0, negative=5, ns\_exponent=0.75, cbow\_mean=1, hashfxn=<built-in function hash>, epochs=5, null\_word=0, trim\_rule=None, sorted\_vocab=1, batch\_words=10000, compute\_loss=False, callbacks=(), comment=None, max\_final\_vocab=None, shrink\_windows=True)



**图1 Word2Vec参数介绍**

附：这里同学们可以修改word2vec的参数来得到不同的版本的词向量表示，然后可以拿不同版本的词向量表示来对比分析分类效果。

1. **定义函数Glove\_extractor( ),实现Glove的文本特征矩阵**

**这里Glove模块并没有封装到任何一个库里，比如说gensim是不能直接调用Glove的。这里有两个方案：**

**方案一：拿Glove的源代码，训练中文语料得到词向量矩阵**

**参考链接：**

[**https://blog.csdn.net/m0\_45176278/article/details/126686429**](https://blog.csdn.net/m0_45176278/article/details/126686429)

**方案二：直接去官网找中文语料库训练好的Glove训练出来的词向量矩阵，模块直接本地读取词向量矩阵。**

**参考链接：**

[**https://gitcode.com/Embedding/Chinese-Word-Vectors/overview?utm\_source=csdn\_github\_accelerator**](https://gitcode.com/Embedding/Chinese-Word-Vectors/overview?utm_source=csdn_github_accelerator)

**6. 分类器训练和测试**

1. 朴素贝叶斯分类
2. 逻辑回归分类
3. 支持向量机分类
4. MLP分类

**7.分类结果的量化分析**

分类结果的定量分析指标：准确率和召回率

①不同文本特征表示，相同分类器分类结果的对比

②相同文本表示，不同分类器之间的对比

**8.分类结果的定性分析**

分类结果的定性分析主要是**欠拟合、过拟合、泛化性等方向**，从文本到分类整个流程中哪些因素对分类结果产生积极影响，哪些因素产生消极影响。从自身经验上给出一些分类结果的提升建议。

**9. 实验总结**