编程作业说明：神经网络

**任务一：使用Logistic回归估计马疝病的死亡率**

**训练集**：包含于文件horseColicTraining.txt中，用于训练得到模型的最佳系数。训练集包含299个样本（299行），每个样本含有21个特征（前21列），这些特征包含医院检测马疝病的指标；最后1列为类别标签，表示病马的死亡情况；部分样本含有缺失值；

**测试集**：包含于文件horseColicTest.txt中，通过预测测试样本中病马的死亡情况，来评估训练模型的优劣。测试集包含69个样本，部分样本部分特征缺失；最后1列为类别标签，用于计算错误率。

**特征值缺失**：在该示例中，可以选择实数0来替换所有缺失值，理由是：1，在系数更新时不会影响系数的值（后面会介绍）；2，由于sigmoid(0)=0.5，对结果的预测不具有任何倾向性；3，该数据集中的特征值一般不为0，因此某种意义上也满足‘特殊值’这个要求。

**训练集中类别标签缺失**：在Logistic回归中，可将该样本直接丢弃。

所需提交材料：任务一需要编程实现Logistic回归完成对测试集的预测，并且计算得到精度，编写实验报告简述算法原理，实验结果等。

**任务二：使用神经网络完成新闻分类**

**（1）数据集讲解：**

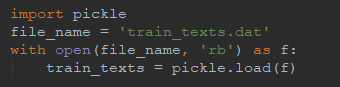
该数据集用于文本分类，包括大约20000个左右的新闻文档，均匀分为20个不同主题的新闻组集合，其中：

**训练集**：包括11314个新闻文档及其主题分类标签。训练数据在文件train目录下，训练新闻文档在train\_texts.dat文件中，训练新闻文档标签在train\_labels.txt文档中，编号为0~19，表示该文档分属的主题标号。

**测试集**：包括7532个新闻文档，标签并未给出。测试集文件在test目录下，测试集新闻文档在test\_texts.dat文件中。

**（2）实验思路：**

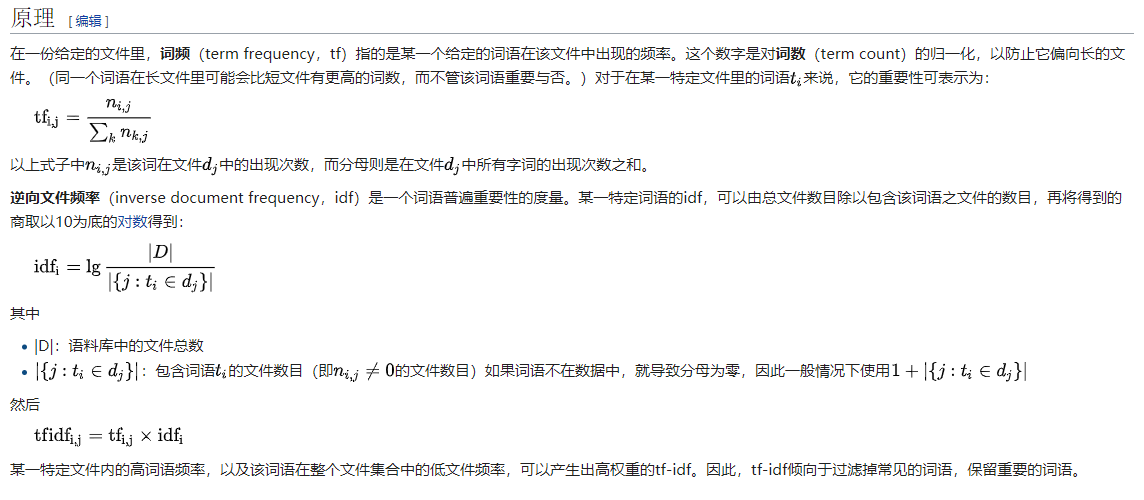
**数据集读取：**读取文件train\_texts.dat和test\_texts.dat方式如下，以train\_texts.dat为例，test\_texts.dat读取方式相同：



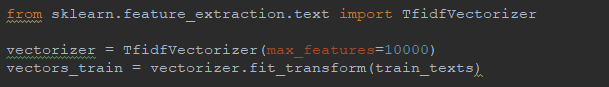
标签文件为正常txt文件，读取方式按照读取txt文件即可。

**文档特征提取：**

因为每篇新闻都是由英文字符表示而成，因此需要首先提取每篇文档的特征，把每篇文档抽取为特征向量，可以采用多种特征提取方式，这里给出建议为提取文档的TF-IDF特征，即词频（TF）-逆文本频率（IDF），TF-IDF简介如下：



由于这不是实验的重点，提取文档的TF-IDF特征可以通过sklearn. feature\_extraction.text中的TfidfVectorizer来完成，具体实现代码如下：



其中的train\_texts为之前从文件中读取的训练集新闻文档，max\_features=10000表示每个文档只提取10000维特征，该值设置过大会对提取的特征引入噪声，同时然后续计算变得更加复杂，设置过小则不能很好的保留每个文档的特征，这里建议初始实验设置为10000即可，后续完成实验后，可以自行调节。

（上述文档的TF-IDF特征提取仅仅为一种特征提取方式，其它特征提取方式也可以进行尝试）

**后续算法**：在完成每篇新闻文档的特征提取后，就可以构建神经网络模型进行训练，具体的网络模型结构，大家可以自行尝试。

**（3）实验要求**

预测测试集新闻文档分类结果，将预测结果与训练集标签相同的存储方式进行存储，存储为txt文件，每一行表示一个新闻的分类标号，将预测结果发送到邮箱zhangzizhuo@hust.edu.cn。需编写实验报告说明具体实验流程，模型结构等。

两次任务的实验报告编写在一份实验报告中。

（希望同学们得出理想的实验结果，学习到更多的知识！）