# 1．业务挑战

随着大数据时代的来临，人们面临着越来越多的数据，其中有很多对他们并没有用，比如说邮箱里的垃圾邮件，我们应该怎么对这部分数据实现过滤？

# 2．价值体现

通过机器学习算法使计算机学会分辨出邮件是否为垃圾邮件，通过对来往邮件进行检测，实现自动过滤的功能。

# 3.解决方法概要

通过有监督学习，找到垃圾邮件与其特征的对应关系，通过tfidf算法，得到特征向量，利用逻辑回归算法，再结合是否为垃圾邮件的标签值，学习出对垃圾邮件的辨别能力。

# 3. Demo 链接，用户名和密码

demo地址IP：<https://9.30.109.52:9443>

在IBM公司内网的情况下，用chrom浏览器打开，用以下用户登录进入

username:tuser01

password:c6deshop

在Meetup工程下打开notebook: EmailSpam

**Note:firefox打开网页的话，可能会出错。如果不能看，也可以通过以下github:**

<https://github.ibm.com/ds-ml/bank-churn/blob/master/Churn_analysis%252Btraining.ipynb>

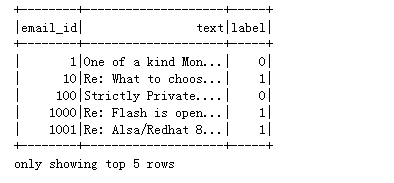
# 4. 数据文件介绍

# Parquet文件

文件名截图如下:

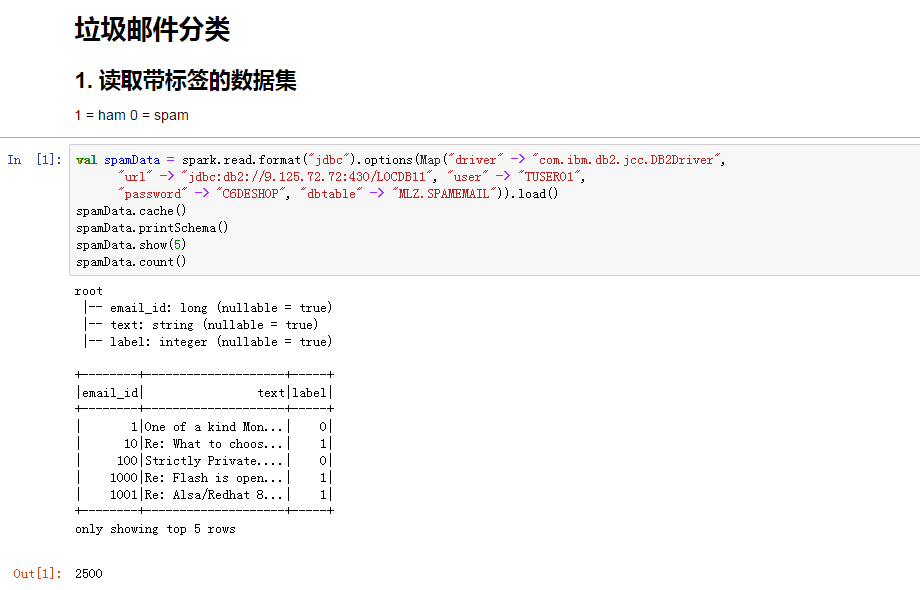


文件内容分为三列，邮件id号，邮件内容，是否为垃圾邮件的标签

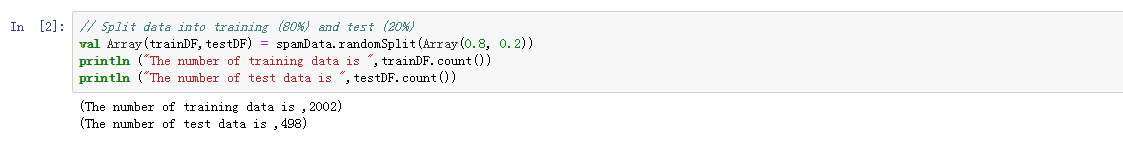


# 5. 代码详细步骤

1> 从DB2中读取数据，显示出前五行，并计算共有多少行数据



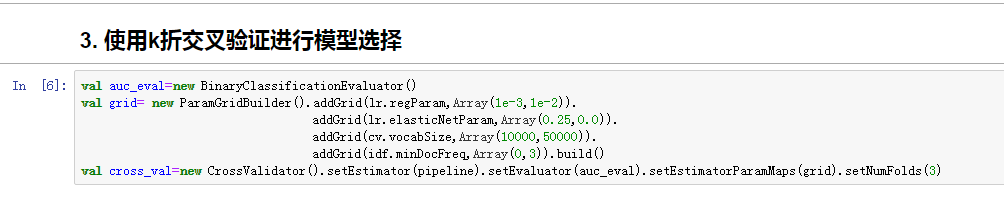
2>将数据切分为两部分，80%用于训练算法，20%用于测试



3>添加所需要的函数库，构造tfidf算法和算法流水线pipeline,将tokenizer、countvectorizer和idf串联起来。



3>构建交叉验证算法，使算法的LR算法的正则化参数regparam范围为0.001，0.01两个数之间变化，使LR算法的elasticNetparam参数在0.25和0之间变化，使countvectorizer算法的vocabsize的参数在10000,50000之间变化，使idf算法的最小文档频率mindocFreq参数在0和3之间变化。





5.把预测结果的label列强制转换为double类型，并将数据格从dataframe类型转换为RDD类型，使用MLlib库中的BinaryClassificationMetrics方法对结果结果进行统计，然后构建dataframe用于画roc曲线。

添加brunel包，画出roc曲线

65

6> 对所有的数据进行预测，并将预测结果与原有的标签做对比。

