## 基于Bayes 的垃圾邮件分类器的设计与实现

1. **开发环境：**
   1. Windows 10 64 位系统 / macOS
   2. Python 3.6.x (64 位)，需要安装下列依赖库，如果安装中提示缺少什么，根据提示安装即可

可以直接 pip install xxx，附上开发时用的版本

* + 1. jieba – 0.39
    2. Django – 2.1.3
    3. mysqlclient – 1.3.14
    4. xmnlp – 0.1.8
  1. PyCharm Professional 版本不要太旧就行，可直接创建、导入 Django 项目，免去许多配置
  2. MySQL 5.x 或者 8.x 的新版本，设置好中文编码 utf8mb4

为保证正确运行，请尽量使用相同的开发环境。 涉及到python 的路径设置，请在 PyCharm中调试时，选择正确的运行路径。注意：涉及到中文文本的处理，确保所有的代码文件是 UTF-8 编码！

1. **项目结构：（基于 Django 开发的Web界面）**
2. mail，管理整个 Web 的前后端交互、请求、渲染等，大多数的业务逻辑在 views.py 中完成，而与数据库相关的接口在 mysqldb.py 中实现。
3. bayes，其中的 bayesclass.py 是贝叶斯朴素分类器的实现
4. SpamMail 文件夹，Django 项目的设置
5. Static 文件夹，网页端使用到的 css 样式和一些 js 代码，用于渲染网页
6. Templates 文件夹，html 网页代码，通过 Django 注入数据，作为前端 UI
7. Logs 文件夹，保存整个项目运行过程中的日志信息
8. 各种 txt 文件，有 data 的是数据文件，bayes.txt 和 bayes.pickle 是贝叶斯模型的参数，stop.txt 是中文常见停用词。
9. **主要代码文件说明**

本项目是 Django 框架开发的，也可以抛开网页 UI 直接运行代码，命令行输出。

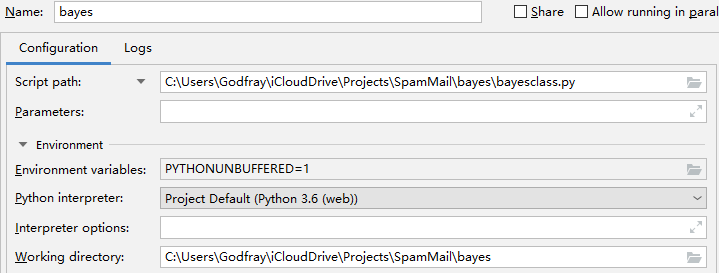
* 1. Bayes 垃圾邮件分类器的实现，主要在 bayes 文件夹下
* bayesclass.py，利用 data.txt 中的标注数据，提取了中文的特征词，训练了一个基于4700多个特征词的朴素贝叶斯分类模型。具体的操作可以看最后的一部分注释，模型的实现过程中也有对应注释。
  1. Django 框架，网页 UI
* 主要是 SpamMail/mail 下的 views.py，里面提供了各种请求的处理和前端渲染。如用户注册登陆、查看收件箱、发送邮件、删除邮件、标记垃圾邮件、个人关键词过滤的设置。
* mysqldb.py 中是和数据库相关的操作。**首先需要按照演示视频中的数据，创建好对应的数据库 mail，首先建好user表，存储用户信息，再建 userlog 这个全局的邮件内容表。用户相关的数据表会自动生成和更新。**在 get\_new\_mails 这个函数中，调用了模糊匹配的算法，会首先修正文本中的错别字后，首先对比用户自定义的过滤词典，一旦匹配即认为是垃圾邮件。否则会使用贝叶斯模型分析是否为垃圾邮件。

1. **程序运行方式**

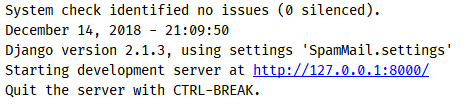
可以通过 Pycharm 导入打开，支持命令行运行和 Web UI 运行。为了提高演示的效率，对实际演示的数据集进行了缩减，完整结果请根据注释中 main 函数的地方或者是 views.py 的 index 函数中的对应地方修改。

*注意：由于 python 环境和项目设置的路径不同，需要根据实际情况修改。直接运行可能会出错。有的时候需要重新选择一下 Python interpreter 就没有问题了。*

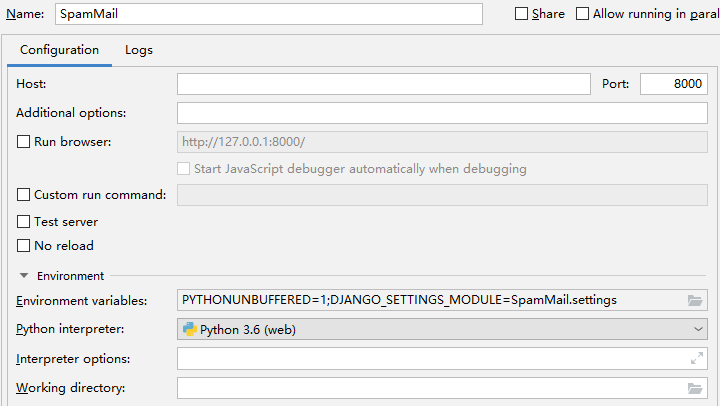
1. 在Pycharm 中的设置主要是 Script path 和 Working directory，如果直接运行 bayesclass.py 的话，保持如下配置：



1. Web UI 演示运行，命令行中，在第一层 SpamMail 文件夹下，使用 “python manage.py runserver 8000” 运行，当有如下的提示后，可以访问 <http://127.0.0.1:8000> ，看到项目的 UI 界面，下面的操作请直接在网页上操作即可。



推荐在 Pycharm 中调试与运行，应该导入项目后，保持默认配置即可：



1. **关于模糊搜索**

由于 windows 环境无法安装pycorrector，选择 xmnlp 这个并不算好的替代工具。没有自己实现的原因是，中文的文本纠错、模糊匹配，非常非常复杂，具体的可以参考pycorrector 的说明文档：<https://github.com/shibing624/pycorrector>，可以作为论文的重要参考。