基于机器学习的垃圾邮件识别系统的设计与实现

开发、运行环境准备

> Python + Django + Scikit-learn

1. Python 3.8，需要安装下列库，一般直接 `pip install` + 包名即可
   - django 3.2
   - jieba 中文分词
   - numpy，joblib，scikit-learn，scipy...
   - 其他如果有报错说缺什么，就装什么
2. PyCharm 尽量用 Professional 版本，Community 也行
3. 网页 UI 需要的 css/js 都在项目文件中了，不需要额外下载

项目结构

1. `model` 中保存了贝叶斯分类模型参数，可以被加载调用
2. `NewsSummaryClassification` 这里面是网页项目运行、启动相关的配置，一般不需要修改
3. `polls` 是网页展示模块，主要关注
   - `views.py`：实现前后端数据交互，处理请求，训练模型或加载模型进行测试，并将结果返回给前端渲染
   - `urls.py`：定义了网页 URL Request 与处理函数的映射
4. `spam`：垃圾邮件分类算法的实现，包括数据、模型文件等
   - 带有 data 的一般都是数据文件，即垃圾邮件数据集
   - `model` 中保存了贝叶斯模型对应的参数
   - bayes 相关的是贝叶斯模型实现
   - dt 相关的是决策树模型实现

- svm 相关的是 SVM 分类模型实现
- 其他文件大多是三种算法模型文件，可以被加载用来测试

5. `SpamWeb`: 网页项目的配置
6. `static`：静态文件，包括 css 样式、js 脚本和图片等
7. `templates`：包括网页 HTML
8. `manage.py`：网页项目的运行入口
9. 其他文件一般是模板项目生成的，与项目关系不大

运行方式

1. 离线方式运行三种算法模型的训练和测试
   - 贝叶斯：运行 `spam/bayes.py`
   - 决策树：运行 `spam/dt_main.py`
   - SVM：运行 `spam/svm_sci.py`
2. 网页展示
   - PyCharm 专业版可以直接启动 Django 项目，或者命令行输入 `python manage.py runserver 8000`，等命令行出现

     ```
     April 26, 2021 - 16:08:40
     Django version 3.2, using settings 'SpamWeb.settings'
     Starting development server at http://127.0.0.1:8000/
     Quit the server with CONTROL-C.
     ```

   - 表示成功启动，如果有一些 warning 可以忽略
   - 浏览器打开访问 127.0.0.1:8000