贝叶斯垃圾邮件过滤系统

开发环境

Java JDK8 + Intellij + Maven

框架：Spring Boot + Thymeleaf

程序运行

Intellij 导入项目后，一般可以自动检测到 Maven 依赖管理和 Spring Boot 框架，能够自动 resolve 依赖，只是时间可能会有点长。太慢的话，可以把 maven 的仓库地址改成国内的。

运行 BayesApplication 项目

在 Intellij 中直接启动，将在本地根据传入的模型参数构建一个贝叶斯垃圾邮件分类器

```
bayes = new Bayes(256, "bayes_model.txt");
```

在模型训练过程中，会加载一些文件，需要一点时间。

最后是一个单例测试。

运行网页项目：SpamApplication

在 Intellij 中直接启动，Spring Boot 环境和 Tomcat 容器都会启动，同时会加载一些配置和 context，等到命令行出现类似的消息后，表示启动成功，可以在浏览器中访问 localhost:8080

```
Tomcat started on port(s): 8080 (http) with context path ''
Started SpamApplication in 44.963 seconds (JVM running for 46.494)
```

文件说明

- `data` 下面是模型相关的数据，以及停用词表（网上可以找到）和特征词表（程序生成）
- `src/main/java/com/example/spam` 中是 java 程序的主要代码
  - `configuration`：程序配置，如模型参数，模型路径配置

    - `controller`：前后端交互的数据传输
    - `jaba`：中文分词工具，直接复制的现成代码，因为这个没有打包好的 jar 包
    - `offline`：里面是贝叶斯分类算法的核心，有算法的实现过程，和使用算法预测的方法
    - `service`：后段服务层，为 `controller` 提供逻辑处理
    - `*Application`：贝叶斯本地分类器和网页应用的启动入口类
- `src/main/resources` 是一些资源文件
    - `static`：css，js 代码，网页样式和交互处理
    - `templates`：html 网页代码，网页内容
    - `application.properties`：Spring Boot 项目的配置文件
    - `dict.txt` 和 `prob_emit.txt`：中文分词用到的字典和支持文件
    - `logback-spring.xml`：日志配置文件
- `bayes_model.txt`：训练好的贝叶斯模型文件，里面保存了特征词在某类邮件中出现的对数概率