

# 实战演练1

## 学习目的

将之前作业中的词频统计功能应用到实际案例中

## 自学任务

2019年1月10日，微信官方发出一篇文章《张小龙4小时“拖堂”演讲，全面总结微信8年（官方完整版）》，有人立即使用 MS Word 统计进行分析，写出了一篇10万+的文章《张小龙的四小时演讲没有用哪些词？》。我们用 Python 来代替 MS Word 做这件事。

### 1. 通过网络请求获得网页内容，使用分词工具对中文字符串进行分词，统计词频，得出结果，发送给指定的邮箱

1. 阅读参考材料3
2. 通过 conda 安装第三方 yagmail（邮箱库）、requests（网络请求库）和 pyquery（xml 文档解析库）
3. 阅读 yagmail 和 requests 的 README 文档
4. 创建作业文件 `d11_training1.py`
5. 使用 requests 请求 参考资料4 微信公众号文章的链接，获取返回结果（response）
6. 将 response.text 用 pyquery 把微信公众号正文提取出来

# 提取微信公众号正文伪代码示例

```
from pyquery import PyQuery
document = PyQuery(response.text)
content = document('#js_content').text()
```

7. 使用 stats\_word 中的 stats\_text 对提取结果进行分析和词频统计处理（返回前 100 个词的统计结果），将结果转换成 str 类型，
8. 通过 yagmail 登录自己的邮箱，将统计结果发送到 [pythoncamp@163.com](mailto:pythoncamp@163.com)（标题格式：班级编号+Github用户名，示例 19100101 zhangsan）

### 注意事项

1. 大部分邮箱（如 qq 和 163）需要在网页设置里开启 SMTP，有要求用授权码登录的则密码就填授权码；Gmail 国内网络无法发邮件
2. 不要在文件内填写自己的邮箱和密码，会被公开的。用下面的方式实现邮箱和密码的输入

```
import getpass
sender = input('输入发件人邮箱:')
password = getpass.getpass('输入发件人邮箱密码(可复制粘贴):')
recipients = input('输入收件人邮箱:')
```

#TODO: 在下面实现自己的分词统计和发送邮件的操作

## 2. 提交作业

1. 将本地仓库关于本次作业的变更提交为一个 commit
2. 通过 Github 桌面客户端将本地电脑的变更推送到自己账户下的作业仓库
3. 回到 Github 自己账户下的作业仓库页面，向远程公用作业仓库的 master 分支发起 Pull Request，在提交的 Pull Request 的标题（title）中填写自己的班级编号+作业编号，如示例：19100101 day1，并在评论（comment）中 @自己的教练 提醒他检查作业

## 参考资料

1. [Yet Another GMAIL/SMTP client](#)
2. [Requests: HTTP for Humans™](#)
3. 正则表达式操作，[英文](#)，[中文](#)
4. [张小龙4小时“拖堂”演讲，全面总结微信8年（官方完整版）](#)
5. [张小龙的四小时演讲没有用哪些词？](#)