第19课词频统计

退出时必要的礼貌要有的吧；退出也要挣扎一下

爬虫及商业价值(请假搞钱)

新知识点

pip install jieba（结巴库，分词库）

看程序1：学习《从清华到MIT》中的词频统计

分词代码示例：

```
import jieba

cut = jieba.cut("我随便写个北京河图创意图片有限公司", cut_all=True)
print("全模式: " + "/ ".join(cut))  # 全模式

cut = jieba.cut("我随便写个北京河图创意图片有限公司", cut_all=False)
print("精确模式: " + "/ ".join(cut))  # 精确模式

cut = jieba.cut("我随便写个北京河图创意图片有限公司")  # 默认是cut_all=False，精确模式
print("默认模式: "+"/ ".join(cut))

cut = jieba.cut_for_search("我随便写个北京河图创意图片有限公司")  # 搜索引擎模式
print("搜索引擎模式: "+"/ ".join(cut))
```

结果输出：

```
全模式: 我/ 随便/ 写/ 个/ 北京/ 河图/ 创意/ 意图/ 图片/ 有限/ 有限公司/ 公司
精确模式: 我/ 随便/ 写个/ 北京/ 河图/ 创意/ 图片/ 有限公司
默认模式: 我/ 随便/ 写个/ 北京/ 河图/ 创意/ 图片/ 有限公司
搜索引擎模式: 我/ 随便/ 写个/ 北京/ 河图/ 创意/ 图片/ 有限/ 公司/ 有限公司
```

登录后复制

看程序2：数据呈现，学习使用第三方库实现柱状图呈现（生成网页文件）

作业任务

任务1：分别统计出《政府工作报告2021》《政府工作报告2022》中前20个两字高频词并画出柱状图，统计出“教育”的词频

任务2：分析出两年的报告中共同的前10高频词与不同的前5高频词，画出饼状图（使用不同的颜色块和颜色文字）

任务3：将《论语》按要求提纯，输出提纯后的文本文件

任务4：对网站www.icourse163.org/university/view/all.htm中大学、学院输出显示，并进行统计大学多少所，学院多少所和绘网页型柱状图