大数据导论 作业一

220110519 邢瑞龙

程序编写说明：

- 系统：Windows11
- 爬虫框架：Scrapy
- 爬虫网站：笔趣阁，豆瓣
- 爬虫额外措施：
  - 增大下载时间间隔：1s模拟人类行为反爬
  - User-agent：采用fake_useragent库随机生成
- 数据：300w中文(data.txt为笔趣阁爬取小说的原数据，processed_data.txt为去除停用词并分词后的数据) top250_movies.csv(豆瓣上爬取4000评论)
- 数据处理
  - 停用词词表来源：百度停用词
  - 分词：`jieba`库