爬虫 day6 代理

同一ip单位时间访问次数过多，被封ip，需要借助代理伪装ip进行爬取

urllib, requests, selenium都可设置代理

代理池

提前筛选，保留可用的代理放在代理池中备用

存储模块、获取模块、检测模块、接口模块

存储模块：
存储抓取下来的代理，不重复，使用redis sorted set，有序集合

获取模块：
定时抓取各代理网站代理

检测模块：
检测代理是否有效并设置分数，检测链接设为目标网站链接

接口模块：
可提供代理接口

分数规则：
分数100为可用，定时循环检测每个代理使用情况，一旦检测到可用代理设置为100， 不可用时分数减1，分数减为0后移除

新获取的代理分数为10， 如测试可行，分数设为100，不可行分数减1，分数减为0后代理移除

http://127.0.0.1:5555

http://127.0.0.1:5555/random

付费代理

级别	套餐	描述
第一梯队	讯代理独享代理、阿布云代理经典版、蘑菇代理、芝麻 HTTP 代理、讯代理优质代理	可用率 99% 以上
第二梯队	阿布云代理动态版、讯代理混播代理、云代理、站大爷短效优质代理、全网代理动态版、阿布云代理专业版	可用率 99% 以下，90% 以上
第三梯队	太阳 HTTP 代理、大象代理专业版、大象代理企业版	可用率 90% 以下，50% 以上
第四梯队	大象代理个人版、全网代理普通版、快代理	可用率 50% 以下，20% 以上
第五梯队	站大爷普通代理、西刺代理	可用率 20% 以下

蘑菇代理 - 企业级高品质HTTP代理IP技术平台|提供HTTP代理IP池租用与定制服务

http://piping.mogumiao.com/proxy/api/get_ip_bs?appKey=15914a3182764e548a82a29c171ccee3&count=20&expiryDate=0&format=1&newLine=2