> 作者：苏婉芳
>
> 日期：2022年6月6日

本研究的代码共有以下几个部分，分别对应以下文件夹：

1. 华为应用市场爬虫与相关数据处理：huawei_appstore
2. 豌豆荚APK下载的URL爬取：apk_spider
3. APK下载和反编译、权限抽取：apk_analyse
4. API-权限映射关系处理与分析：clawer-api-permission_mapping
5. 应用权限推荐（支持度计算处理）/可视化界面：map_topic2perms
6. 部分结果数据：data
7. 数据库：db_LDA.db

1 huawei_appstore

read_results2excel.py：解析getMethodMap_new.py的运行结果写入excel

read_excel2db_sourceCodePerm.py：从excel读取结果写入数据库

1.1 huawei_appdescrip_spider

stopwords：中文停用词库源，包括1）4个常用的；2）迭代增加的

huawei_appdescrip_spider.py：华为应用市场爬取

low_freq_words_filter.py：低频错误词处理

word_cut.py：分词处理

1.2 cut_none_sence_topic_words

cut_none_sence_topic_words.py：迭代补充停用词

1.3 LDA

对华为应用市场描述进行多种LDA分类，及分类结果

含有各种LDA方式，包括直接分类（多种参数）、二级分类、原参数分类等

分别对应各个.py文件，查看代码即可知道具体方式

1.4 ltp_data

分词模型所在位置

2 apk_spider

apkurl_spider.py：从豌豆荚的各大类的“全部”爬取

apkurl_spider_subcate.py：从豌豆荚的各大类的各个子类爬取

3 apk_analyse

add_39_doc_permissions：补充2020年最新研究的manifest权限信息

download_apk.py：下载爬取的豌豆荚APK链接

extract_perm_xml_1/2.py：从APK解析权限数据

该部分需要在交我算运行近一周时间，本地解析速度过慢，容易出现memory error

4 clawer-api-permission_mapping

4.1 clawer_API29to32

补充新版Android的权限映射关系

extract_class_urls.py：抽取类文档链接

getMethodMap_new.py：解析文档提取API-权限映射关系

multi_process_classesPage：多线程处理getMethodMap_new.py的工作

add_exception：补充遗漏的类分析，类似getMethodMap_new.py

4.2 multi_process

多线程处理getMethodMap_new.py的工作

4.3 permissionList

所有结果的集合，excel形式

5 map_topic2perms

map_topic2perms.py：运行入口，内有处理论文的3.3与3.4节的函数，运行即可实现输入APP描述，得到推荐权限等结果

read_db2json.py：读取数据库内信息到json文件便于查询使用

UI文件夹：可视化工具，入口为main.py

6 data

安卓权限.xls：分析得到的官方文档manifest.permissions的所有数据

manifests_source_code：ASOP源码的部分文件

几个txt文件：获取的完整权限列表

apk：测试数据集（部分，一些已经在分析后删除）