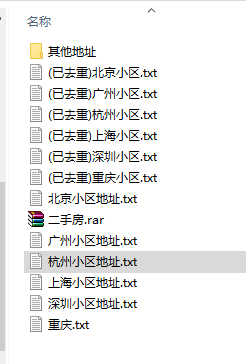
## 地址数据获取，入库

### 房天下数据获取清洗

房天下数据页面。



使用Python工具获取了包括北京市区，周边等小区名，地区，详细地址，单价数据，并进行去重操作。之后又获取了广州，杭州，上海，深圳，重庆等地的数据。



### 北京政务数据资源网数据获取，清洗

《<http://www.bjdata.gov.cn/>》

北京政务数据资源网数据包括，**信用服务，财税金融，经济建设，旅游住宿，交通服务，医疗健康，房屋住宅等各类数据**



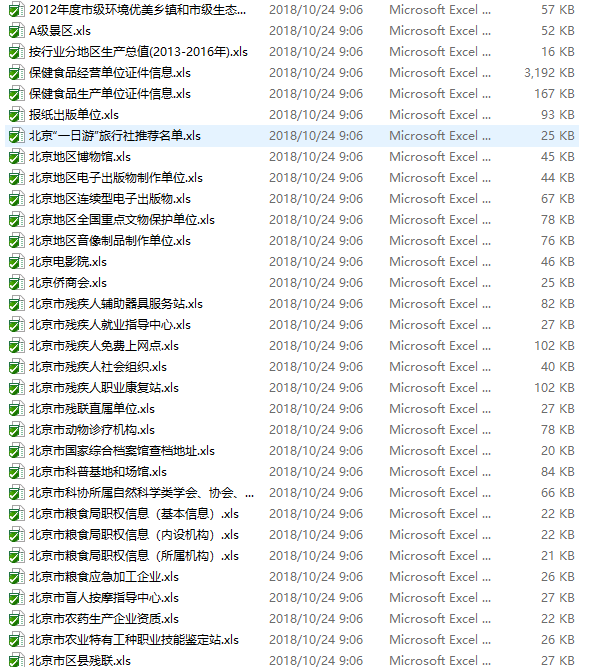
每类数据包括xls，csv，zip等不同数据类型的格式，



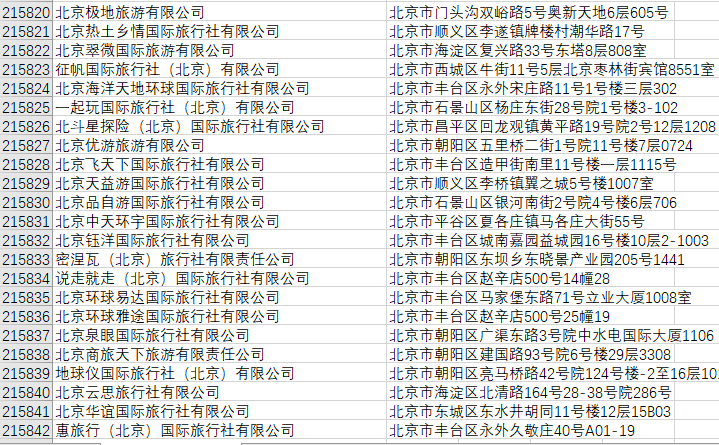
使用Python爬虫等工具对这些数据进行全部下载到本地

。

之后就是对数据进行清理，把含有地址数据的文件提取出来



并读取所有含有数据的文件，提取地址名称和地址信息到新的csv中，以便入库，最终可用数据为215842条



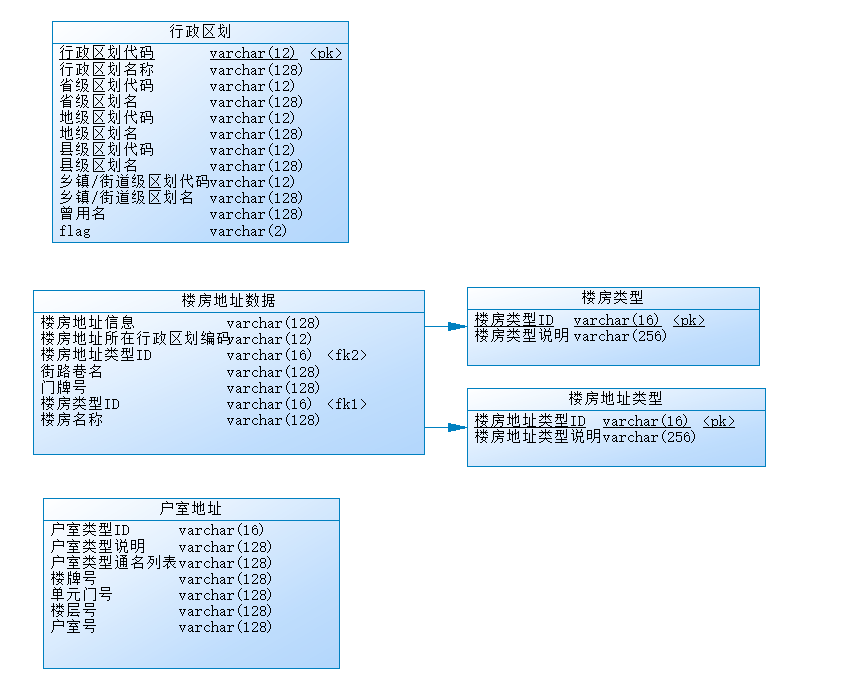
### 行政区划代码获取

<http://www.mca.gov.cn/article/sj/xzqh/2018/>

获取最新，详细到乡镇街道区划的地址

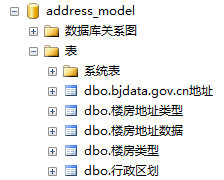
### 地址数据设计

详见\getdata\数据库设计《标准地址数据结构.docx》



### 地址数据入库

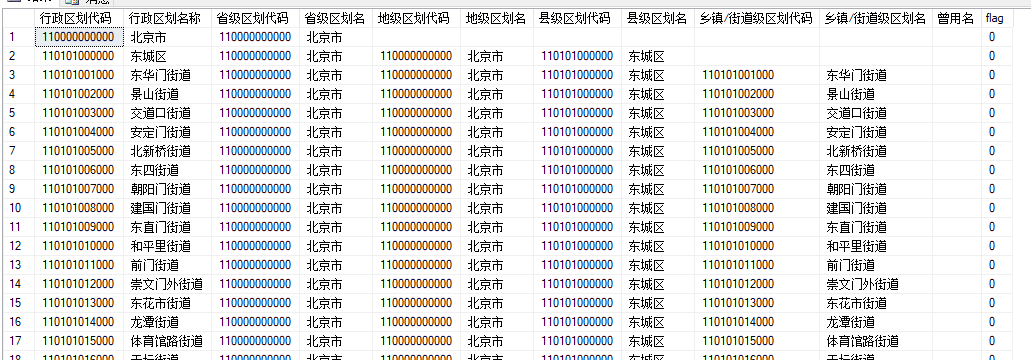
Address\_model库



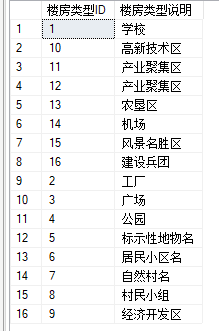
1. 楼房地址数据包括安居客数据，和房天下数据，其中需要使用jieba，hanlp等文字识别包识别小区行政区划代码，楼房类型id，街巷名，门牌号（可能为空），楼房类型id等，



1. 行政区划代码需识别省市县各级别，并分类入库，源码：补全行政区划代码库.py



1. 楼房类型为资料查找并手动录入。



# 数据脱敏

数据脱敏主要分为敏感数据识别和数据脱敏

## 敏感数据识别

1. 已知敏感数据包括，邮件，国内的手机号码，18和15位的身份证号码，省市县地址，中文姓名，日期，IPv4，ipv6，URL，组织机构代码，银行卡，邮编，IP地址，mac地址。
2. 准备工作：使用Python的fake工具，造出若干数据，一遍之后识别。脚本名为：fake\_data.py
3. 写Python脚本，进行识别，主要使用正则匹配进行简单的识别。
4. 重新整理脚本。全用正则匹配冗余代码过多，统一写到xml文件中进行读取，并对结果数据进行归类，识别百分比计算等。

具体思路可详见《finder文档.docx》

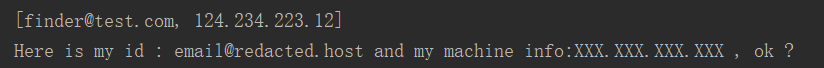
1. 因为Python脚本的复用性不好，改为调用jar包的方式。测试代码详见DetectMaskProgram（ps：jpype访问jar包经常报错，找不到方法，之后经过各种命令行等方式，发现，建立java项目时，建为maven项目，打出来的jar包基本都能访问。）

## 敏感数据加密

Mask的方法暂时只能替换为“\*\*”，暂时并不能提供保留格式脱敏，format-preservingencryption（FPE）**格式保留加密**还在学习中。。。

测试数据：Here is my id : finder@test.com and my machine info:124.234.223.12 , ok

输出敏感数据邮箱和mac地址，以及进行简单脱敏。



# 元数据获取

1. 元数据获取时，考虑使用sql获取各类元数据，但是经过尝试之后，获取数据库版本，数据库url，驱动版本等，受数据库版本，数据库类型等因素影响，复用性更是可怜，
2. 之后使用了pyodbc包，调用像SQLTables这样的函数，它相当于JDBC getTables函数，但是对于之前预想获取的元数据好多都获取不到，远远达不到jdbc的功能。
3. 之前有jpype调用jar包的经验，所以直接写一个连接各类数据库的java类，供Python来调用。