运行之前请注意以下事项：

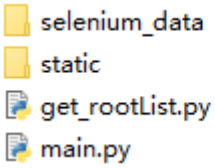
- 检查chrome.exe的路径是否在系统环境变量中，若不在请加入，然后重启计算机
- 下载chrome的webdriver，请注意版本和浏览器适配
- main.py 中需要自定义参数：

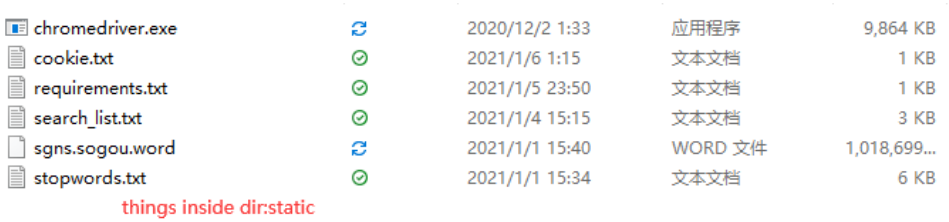
```
# global vars && parameters here
BILIBILI = 1
WEIBO = 2
ZHIHU = 4
WEIXIN = 8
DOUBAN = 16
XIMA = 32

WAIT_TIME = 2
LONG_WAIT_TIME = 15
threshold = 0.85  # when similarity >= threshold, 2 urls are considered relative.
'''
modify all these absolute paths to the according paths on your machine.
'''
stopwds_path = r'C:\Users\Neo\PycharmProjects\SQA_lab2\static'
stopwds_name = r'stopwords.txt'
# stopwords list
wordvec_path = r'C:\Users\Neo\PycharmProjects\SQA_lab2\static'
wordvec_name = r'sgns.sogou.word'
# pretrained word vector
# dimension of word vector default to 300
search_path = r'C:\Users\Neo\PycharmProjects\SQA_lab2\static'
search_name = r'search_list.txt'
chrome_path = r'C:\Users\Neo\PycharmProjects\SQA_lab2\static'
chrome_name = r'chromedriver.exe'
outputPath = r'C:\Users\Neo\PycharmProjects\SQA_lab2\static'
outputName = r'output.csv'
wechat_account = "***"
wechat_password = "***"
# 一个能够登陆入微信公众号平台的账号的邮箱和密码。不要用微信小程序的号!
zhihu_account = "***"
zhihu_password = "***"
# 一个能够登陆入知乎的账号和密码。 不是手机验证的免密登录，而是账号和密码!
weixin_url = 'https://mp.weixin.qq.com/'
cookie_path = r'C:\Users\Neo\PycharmProjects\SQA_lab2\static'
cookie_name = r'cookie.txt'
# f'store cookie from {weixin_url}'
search_list = read(os.path.join(search_path, search_name))
start = 0
# crawling from the start-th name
limit = 30
```

    - 在文件夹static中，有一些静态资源：
        - requirements.txt：pip install -r requirements.txt
        - search_list.txt: 从B站中随机抽取若干名up主组成的搜索根节点名称列表。具体地，在每一个up主分区中，在排名前100中随机抽10名。
        - sgns.sougou.word: 预训练词向量，用搜狗新闻的语料训练而成，300维。可以自己替换
        - stopwords.txt: 停顿词表，用于过滤无意义的字符
        - cookie是微信公众平台登录后返回的cookie，爬取链接时会用到
    - 在实际运行文件时，请将xxxx_name设置为对应资源文件的文件名，xxxx_path设置为对应资源文件的文件路径，或者将实际文件改成和截图中的一致

    - wechat_account, wechat_password: 自己的某个微信公众号的账号和密码，爬取微信公众平台时需要，请设置为一个可以登陆微信公众平台的账号和密码。
    - zhihu_account, zhihu_password: 自己的某个知乎账号的账号和密码，爬取知乎文章时会用到。
- 在正式运行前
    - pip install -r requirements.txt
    - 请先在包含main.py的目录下执行：

      mkdir selenium_data
    - 再进到当前目录下，执行：chrome.exe --remote-debugging-port=9222 --user-data-dir=./selenium_data
        - 这一步之前请保证chrome.exe已经被加入到系统环境变量
        - 这一步时目录应该长这个样子：
        - 

        - 

- get_rootList.py 为获取100+个b站up主的用户名，并且写到search_list.txt中
- 运行时，会先额外再弹出2个chrome窗口，但是没有请求页面，请等待，这是正在载入词向量

  第一次会要求知乎登陆，如果遇到了知乎登陆要输入验证码的情况，请手动点一下图片，或者重新
- 启动整个程序（我也不会输入验证码。。不过一般情况下不会要求验证码

  第一次会要求要微信客户端扫码，请在15秒内用对应的账号扫码登陆通过
- 由于爬取微信公众号太频繁，自己的微信公众号可能被tx封，表现是后续从的爬取无法再从微信公
- 众平台获取文章。所以一次爬取默认只爬30个用户名

  结果输出在output.csv中
-
- 本程序在python3.7.8 Win10 64bit 环境中测试通过

  评价：

- 获取根用户名时，爬取的是b站多个分区中，每个分区的排名前100的up主中的随机提挑选的10个的集合。
- 由于是机器学习模型，而且没有训练数据，因此encode句子的方法是wordAveraging，有时候可能会将不太相关的url认作相干，这一点在微信公众号平台尤为明显，因为推文往往字比较多，将一篇推文encode后句向量就非常接近平均值了。