简介

Pinterest有visual-search的功能，例如：

<https://www.pinterest.com/pin/172684966938649859/visual-search>

本爬虫可获取搜索结果中的左上角的原始图片、右上角的标签列表、右下角的搜索结果列表。

依赖安装

本爬虫依赖selenium。selenium是一个可以操控浏览器行为的工具。下面以Ubuntu下使用selenium操控firefox为例介绍安装步骤，其他平台请访问selenium官网查看具体的安装步骤。

1. 使用pip安装selenium: sudo pip install selenium
2. 安装Firefox
3. 下载<https://github.com/mozilla/geckodriver/releases/download/v0.14.0/geckodriver-v0.14.0-linux64.tar.gz>，并解压到/usr/bin/目录

配置说明

settings.py:

username/password: Pinterest用户名和密码

limit: 单个搜索最大的搜索结果数，默认为100

links: 被搜索的链接列表文件名，默认为”links.txt”

proxy: 如果网络状况不佳，可配置http代理，格式为ip:port

pageLoadtimeout: 页面的加载超时设置

elementLoadTimeout: 页面上使用ajax加载的元素的超时设置

searchInterval: 两次搜索之间的时间间隔，防止搜索过于频繁

scrollInterval: 操纵浏览器滚动条滚动到搜索结果最下方的时间间隔

scrollTimeout: 滚动到搜索结果最下方后，该时间范围内检测新的搜索结果

links.txt:

每行一个Pinterest图片的链接，例如：

https://www.pinterest.com/pin/172684966938649859/

抓取结果说明

每个链接的抓取结果保存在该链接对应的id的目录里，目录里有三个文件：

source.txt: 原始图片的图片链接

tags.txt: 搜索结果的标签列表

search-results.txt: 搜索结果的图片链接列表

使用说明

安装依赖并配置后运行python visual-search.py即可

抓取的是图片链接，要下载图片，可用wget下载，例如：

wget <https://s-media-cache-ak0.pinimg.com/474x/ae/83/62/ae8362291ab24cd2feb173354a6acd81.jpg>

注意事项

建议在性能较好的机器上运行

建议在网速较快的环境下运行

异常退出的可能原因：

1. 主页加载超时，登录失败
2. 搜索结果加载超时
3. 无法在搜索结果中获取被搜索图片
4. 无法在搜索结果中获取标签
5. 无法在搜索结果中获取搜索结果列表