مستندات پروژه واکشی اطلاعات از اینستاگرام

InstaAppScrapAutomation

پروژه ای برای انجام اتوماسیون تحت برنامه اندرویدی اینستاگرام و واکشی اطلاعات پست های مد نظر با بررسی دو پست قبلی و بعدی و مقایسه آن با یک مقیاس عددی مشخص (که از طرف کاربر وارد می شود)

نحوه کارکرد و ساختار برنامه(اسکریپت)

در این پروژه با استفاده از زبان برنامه نویسی پایتون ، برنامه appium و selenium ، برنامه‌ای طراحی شده که عملیات مد نظر (استخراج اطلاعات از پست های مد نظر) را به شکل خودکار برروی برنامه اندرویدی اینستاگرام پیاده‌سازی کند. در اینجا دستور العمل ها و تعاریف به زبان پایتون نوشته شده که شامل خودکار سازی کار ها برروی برنامه است. این اسکریپت به برنامه‌ای واسط که appium باشد متصل میشود. اینجا ما از نسخه تحت خط فرمان آن استفاده می‌کنیم که به appium server نیز شهرت دارد. این برنامه دستورات را از زبان پایتون به جاوا تبدیل کرده و با استفاده از SDK و ADB اندروید ، آن‌ها را روی دستگاه اندرویدی که از طریق کابل به کامپیوتر متصل شده و حالت USB-Debug آن فعال است اجرا میکند. برای این کار اینستاگرام باید روی دستگاه هدف نصب شده باشد.

آموزش نصب

روش الف) مطمئن نیست و تا به حال فقط روی برخی سیستم‌ها جوابگو بوده.

بعد از منسوخ شدن نسخه GUI برنامه appium به دلایل امنیتی ، تنها نسخه موجود ،نسخه CLI یا خط فرمان آن است.

در مرحله اول **NodeJS** را دانلود و نصب کنید. در هنگام نصب تیک نصب برنامه‌های مورد نیاز را نیز بزنید. CMD ویندوز را باز کنید و دستور زیر را وارد کنید.

```
npm -g install appium
```

با این دستور برنامه رابط نصب می شود.

حالا باید با فراخوانی اسم برنامه رابط آنرا اجرا کنیم. نام برنامه را به عنوان دستور در CMD وارد کنید:

```
appium
```

برنامه باید بدون خطا اجرا شود و آماده به کار باشد.

>> فایل‌های ورودی را طبق نیاز خود ویرایش کنید. (در قسمت بعد از نصب توضیح داده خواهد شد.)

حالا موبایل اندرویدی با USB-Debug فعال را به کامپیوتر متصل کنید یا شبیه ساز اندرویدی خود را راه اندازی کنید. سپس بر روی فایل main.exe کلیک کنید تا برنامه اجرا شود.

روش ب) >> در صورتی که روش الف کار نکرد :

برنامه JDK را نصب کنید و محل نصب آن را یادداشت کنید.

در صورتی که NodeJS را نصب نکرده‌اید ، آنرا نصب کنید.
سپس از طریق کنترل پنل به آدس زیر بروید :
Control Panel -> System -> Advanced System Settings -> Environmental Variables -> Click New.

در پنجره باز شده ، فیلد اول را Java_Home و برای فیلد دوم آدرس نصبی که یادداشت کردید را وارد کنید. برای اطمینان از نصب درست برنامه در CMD دستور زیر را وارد کنید:
java -version
در صورتی که نصب درست انجام شده باشد ، باید ورژن جاوا نصب شده نمایش داده شود. مانند : java version 1.8.0_144. (عدد این متن بر اساس نسخه نصب شده متفاوت است)

حالا SDK را دانلود کنید. آنرا در محلی مانند C:\\SDK از حالت فشرده خارج کنید. یعنی در درایو C فولدری با نام SDK بسازید و فایل‌ها و فولدر های اصلی استخراج شده را در آن قرار دهید.
مثل مرحله قبل ، وارد کنترل پنل شده و به مسیر زیر بروید:
Control Panel -> System -> Advanced System Settings -> Environmental Variables -> Click New.
حالا در پنجره باز شده ، در فیلد اول Android_Home را وارد کنید و در فیلد دوم آدرس نصب SDK را وارد کنید.

حالا npm -g install appium را در CMD وارد کنید. دستور زیر را در مرحله بعد وارد کنید:
appium
برنامه رابط باید بدون ارور اجرا شود.

نحوه استفاده

آماده سازی دستگاه اندرویدی
(می توانید از شبیه ساز استفاده کنید و این مرحله را رد کنید)

بعد از نصب و اجرای appium ، دستگاه اندرویدی خود را بردارید ، developer mode آنرا فعال کنید (معمولاً با تپ برروی شماره سریال یا سری ساخت به تعداد ۷ یا چند بار امکان‌پذیر است). سپس وارد بخش developer mode یا developer options در تنظیمات دستگاه خود شوید. در لیست پیش رو USB-Debugging را پیدا کنید و آنرا فعال کنید.
دستگاه را با کابل USB به کامپیوتر (یا لپ تاپ) متصل کنید. در صورتی که اینستاگرام روی آن نصب نشده آنرا نصب کنید.

* توجه داشته باشید طی کار اسکریپت یک برنامه که از سمت appium (برنامه رابط که نمونه آن را روی ویندوز نصب کردید) روی موبایل شما خودکار نصب می شود. می‌توانید آنرا هر زمان پاک کنید و مشکلی پیش نخواهد آمد.

آماده سازی شبیه ساز اندروید

شبیه ساز مد نظر خود را نصب کنید. به عنوان مثال BlueStacks . تنها مورد مهم این هست که تنظیمات مربوط به ADBیا USB-Debug را در شبیه ساز فعال کنید. برای نسخه ۵ بلو استکس ، در تنظیمات و بخش advanced قرار دارد. آنرا فعال کنید.

وارد کردن و تنظیم ورودی های اسکریپت

در پوشه اصلی اسکریپت تعدادی فایل متن txt قرار دارد که ورودی ها از این طریق به برنامه منتقل می شود.
یوزرنیم های مد نظر برای بررسی را در فایل usernames.txt در هر خط جداگانه بدون علامت @ وارد کنید. برای مثال bijan53c.

در userpass.txt ، یوزرنیم و پسورد اکانت خود را وارد کنید ، برای دانلود پست اینستاگرام و یا لاگین نیاز خواهد شد. به دو نقطه بین یوزرنیم و پسورد دقت کنید که حذف نشود ، در غیر این صورت برنامه کار نخواهد کرد یا حداقل دانلود پست به مشکل خواهد خورد.

در Xrate.txt هم عدد مد نظر برای تعیین مقدار مقایسه لایک ها وارد کنید. فقط عدد وارد کنید ، در غیر این صورت برنامه با مشکل مواجه خواهد شد و کار نخواهد کرد.

نحوه اجرای برنامه:

فایل main.exe را اجرا کنید.
یا می‌توانید با نصب کتابخانه‌های مورد نیاز (خطوط اول فایل‌های اسکریپت) فایل main.py را اجرا کنید ، ولی پیشنهاد میکنم که فایل exe را اجرا کنید تا نیاز به نصب و کانفیگ کتابخانه‌های مختلف نداشته باشید.

در صورت بسته شدن ناگهانی برنامه اسکریپت (که معمولاً به معنی بروز ارور است) برنامه main.exe را از طریق CMD اجرا کنید. برای این کار ، فرض بر اینکه فولدر اسکریپت را برروی دسکتاپ قرار داده اید. Cmd را باز کنید. وارد دسکتاپ شوید:

cd Desktop

وارد فولدر برنامه شوید

cd InstaAuto-2023-05

برنامه را اجرا کنید

main.exe

اگر اروری مانند ارور زیر دریافت کردید
main.exe no such a file or directory
ممکن است لازم باشد تا این دستور را مجدد وارد کنید (ممکن است طی فشرده‌سازی ۲ پوشه ایجاد شده باشد)
cd InstaAuto-2023-05
سپس دوباره
main.exe

با این کار در صورتی که برنامه دچار ارور شود ، شما نوع ارور را خواهید دید.

برداشت اطلاعات استخراج شده

بعد از اتمام کار ، اطلاعات و فایل‌های دانلود شده به ازای هر ID در پوشه IDs و به شکل ID_ ذخیره شده. مثل bijan53c_. در کنار فایل‌های دانلودی ، یک فایل txt با اطلاعات مد نظر و لینک پست نیز ذخیره شده که در صورت عدم موفقیت در دانلود اطلاعات ذخیره شده باشند.

مستند های مربوط به سوس کد

برنامه اصلی به شکل exe است ، ولی تمام فایل‌های سورس کد با پسوند .py هستند.
این برنامه از دو اسکریپت main.py و InstaAutoRowScroll.py تشکیل شده. اسکریپت اول یوزر نیم های لیست ورودی را میخواند ، به ازای هر کدام یک فولدر در زیر مجموعه IDs میسازد و برای هر کدام از آن‌ها عملگر اصلی که در InstaAutoRowScroll.py را فرا میخواند. همانطور که گفتیم ، فایل دوم شامل عملگر اصلی است.
عملگر اصلی ، یوزرنیم و پسورد پیج برای لاگین در اینستاگرام را میخواند ، عدد نسبت مشخص شده را میخواند ،در صورتی که برنامه اینستاگرام لاگین نشده باشد ، این کار را انجام می دهد، وارد بخش جستجو می‌شود ، یوزرنیم های مد نظر را می‌خواند و آن‌ها را یکی یکی در اینستاگرام جستجو می‌کند. وارد پیج می‌شود ، پست های پیج را یکی یکی بررسی میکند و اطلاعات مورد نیاز ، شامل محتوای متنی(شامل کپشن نمیشود ، بلکه شامل نوع پست و تعداد لایک میشود) ،مقدار عددی لایک ، تاریخ انتشار و لینک پست را استخراج می کند. به ازای هر پست تعداد لایک های آنرا با دو پست قبلی و بعدی و با در نظر گرفتن نسب متغیر بررسی میکند. در صورتی که لایک های پست اصلی بیشتر از موارد مذکور بود ، پست دانلود می شود. در صورت عدم موفقیت در دانلود اطلاعات استخراج شده مذکور در فایل متنی ذخیره می شود.

توضیحات کد main.py
در خط ۳ قابلیت تعامل با سیستم عامل برای ساخت فولدر اضافه می‌شود
در خط ۵ ، عملگر (function) "scrapper" از اسکریپت InstaAutoRowScroll.py وارد می شود.
در خط ۸ و ۹ ، یوزرنیم ها از فایل مروبطه از هر خط آن به شکل جداگانه خوانده می شوند.

در خط ۱۱ به بعد هم با حلقه ، ابتدا فاصله های اضافی حذف می‌شوند ، سپس فولدر ساخته می‌شود و در نهایت عملگر فراخوانی می‌شود تا کار انجام شود.

توضیحات کد InstaAutoRowScroll.py
بعد از وارد کردن کتابخانه‌های مربوطه در خط ۲۱ ، یوزرنیم و پسورد پیج اصلی خوانده می شود. برای ورود به برنامه ، مشاهد سایر پیج ها و ...

در خط ۳۰ ، متغیر نسبت لایک ها خوانده می شود.

در خط ۳۶ و خطوط زیر مجموعه آن ،عملگری تعریف می‌شود تا قابلیت‌های appium برای آن ارسال می‌شود تا appium بتواند با مشخصات داده شده با دستگاه اندرویدی تعامل کند. آدرس اصلی سرور appium مورد استفاده localhost:4723 میباشد.
در خط ۵۶ فراخوانی می شود.

در خط ۶۱ حداکثر زمان تایم اوت (صبر) برای برنامه تعریف می شود. در صورتی که طی این زمان اتصال برقرار نشود عدم موفقیت و ارور اعلام می شود.

در خط ۶۵ و تمامی زیر مجموعه های آن ، سعی می‌شود در صورت نیاز به برنامه اندرویدی اینستاگرام ، عمل لاگین انجام شود. در این بخش المنت های مربوطه همچون فیلد یوزر نیم و پسورد و ... شناسایی و مورد استفاده قرار میگیرند.

در خط ۱۲۳ ، برای برنامه تعریف می‌شود تا با استفاده از متغیر تایم اوت جداگانه در خط ۱۱۴ ، دکمه اکسپلور اینستاگرام پیدا شود. و در خطوط بعدی تپ (کلیک) شود.

در خط ۱۳۰ تا ۱۳۸ ، ابتدا ،فیلد سرچ کلیک می‌شود ، @ به یوزرنیم وارد می‌شود تا نتیجه دقیق‌تری بدست بیاید. در ۱۳۶ ، یوزرنیم با علامت @ وارد فیلد می شود.

در خط ۱۳۹ ، کلید enter فشرده می‌شود تا سرچ انجام شود.

در خطوط ۱۴۴ تا ۱۵۲ ، ابتدا نام یوزرنیم پیدا شده برداشته می‌شود ، سپس با یوزر نیم خوانده شده (و سرچ شده) مقایسه می‌شود تا تأیید شود که پیج درست پیدا شده. و در بخش if – else این مورد تطبیق داده می شود.

در خط ۱۵۴ ، اسم اصلی که کاربر برای پیج خود انتخاب کرده پیدا شده و ذخیره می شود. داشتن این نام برای شناسایی پست ها مورد استفاده قرار خواهد گرفت.

در خط ۱۶۱ تا ۱۶۹ ، عملگر اسکرول تعریف می‌شود تا در موارد مورد نیاز (همچون دیدن تمام پست های پیج از بخش اصلی خود پیج (بخش جدولی شکل)) استفاده شود. (در واقع به معنی شبیه سازی تاچ و کشیدن است تا صفحه به سمت پایین حرکت کند.) این عملگر مشخصات مقدار حرکت را در ۴ متغیر دریافت می کند.

در خط ۱۷۶ ، لیستی تعریف می‌شود تا محتوای متنی پست ها را که قبلاً گفتیم ذخیره کند.

در خط بعدی با استفاده از نام پیج که استخراج کردیم و ترکیب آن به رشته مشخص شده متغیری میسازیم تا به ما کمک کند المنت های نگهدارنده پست ها را شناسایی کنیم.

در خط‌های ۱۷۹ و ۱۸۰ ، متغیر هایی برای شناسایی اتمام عملیات تعریف می شود. ابتدا متغیر اول برای شناسایی تعداد اسکرول بی نتیجه (از نظر شناسایی پست جدید) است که بدین معنی است که به آخر پیج رسیده اید.
و در خط بعد تعداد پست های باز شده و استخراج شده ذخیره می شود. (که اگر به ۱۰۰۰ برسد به اتمام می رسد.)

در خطوط ۱۸۱ و ۱۸۲ ، لیست هایی که اطلاعات استخراج شده را ذخیره می‌کنند تعریف شده اند. مورد اول برای ذخیره پارامتر ها (لایک ، محتوای متن ، لینک ، تاریخ و ...) و مورد دوم صرفاً لایک ها.

در قسمت بعد یک حلقه کلی while تعریف شده که پروسه اصلی را تا جایی که پست وجود دارد ادامه می دهد.

بعد از آن یک حلقه for است که ۳ بار به ازای هر ردیف از پست ها تکرار می شود. به همین دلیل که هر ردیف ۳ پست را شامل می شود.

از ۱۸۸ تا ۱۹۴ ، المنت های پست ها شناسایی می شوند.

از ۱۹۶ تا ۲۰۲ ، نوع پست شناسایی می‌شود که IGTV است یا خیر. این امر برای استخراج اطلاعات مهم است چون المنت نگهدارنده اطلاعات انواع مختلف پست ها متفاوت است. در ۲۰۲ هم پست کلیک می شود.

نکته اینجا است که در ۱۹۲ ، در صورتی که پست قبلاً کلیک نشده باشد و اطلاعات آن ذخیره نشده باشد ، کلیک خواهد شد.

با try -except از خط ۲۰۶ ، برنامه سعی میکند تا نوعی از پست چند تصویری را پیدا کند یا اینکه انواع دیگر (به دلیل متفاوت بودن ساختار المنت) ، سپس متغیر محتوای آن‌ها پیدا می‌شود. این متغیر شامل تعداد لایک هاست. پس از خطوط ۲۱۱ تا ۲۱۵ با تقسیم‌بندی رشته کلمات مقدار لایک ها را استخراج می کند.

در خط ۲۱۸ ابتدا متغیری برای تشخیص اینکه تاریخ پست استخراج شده یا خیر تعریف شده و بر اساس آن زیر مجموعه های آن ، شامل حلقه while تعریف شده که با متن مشخص شده برای المنت مد نظر جستجو میکند و متغیر m ، بخش مد نظر از متن کلی استخراج می شود.
و با except در ادامه آن ، به این شکل تعریف شده که برنامه سعی میکند تا متن تاریخ را پیدا کند (در صورت طولانی بودن پست ، که المنت پیدا نشود) اسکرول انجام می‌شود و صفحه به پایین هدایت می‌شود تا المنت دیده شود.

در خط ۲۳۴ به متغیر مربوطه تعداد پست اسکرپ شده اضافه می شود. در ۲۳۹ بر اساس پست که اگر IGTV باشد ، لینک از طریق دکمه گزینه ها کپی می‌شود و در ۲۴۶ متن کپی شده از اندروید به اسکریپت منتقل می شود. در ادامه اصلاحات لازم برای حذف متن های بلا استفاده و دیکد رشته از حالت بایت به رشته انجام می شود.

در ۲۵۸ نیز ، همین اتفاق برای سایر پست ها با این تفاوت که از طریق دکمه share post اینستاگرام برای کپی لینک استفاده می‌شود چون نوع پست ، باعث ایجاد این تفاوت می

شود. در ۲۷۲ از دکه برگشت گوشی استفاده می‌شود تا بخش باز شده مخصوص share بسته شود.

در ۲۸۰ متغیر کامل را تعریف می‌کنیم که تمام پارامتر های مهم پست را ذخیره می‌کند که برای فایل‌های متنی txt استفاده می‌شود و مقایسه پست ها.

در ۲۸۵ ، لیست های پارامتر ها و لایک ها ، پست مد نظر را اضافه می‌کنند تا ذخیره به ترتیب ایندکس لیست ذخیره شوند.

در ۲۹۴ ، از دکمه برگشت خود اینستاگرام استفاده می شود. این مورد برای برگشت از پست به بخش اصلی پیج است.

در خط ۳۰۳ تعریف می‌کنیم که اگر تعداد اسکرول ها بیشتر از مقدار نوشته شده باشد ، بدین معنی است که به پایان پست های یک پیج رسیده ایم. در ۳۰۷ هم تعریف می‌شود در غیر این صورت باز اسکرول انجام شود. درواقع برنامه ۳ بار سعی می‌کند تا اسکرول کند ، اگر طی این تایم و تلاش‌ها پست جدیدی که در لیست پست ها نباشد ، پیدا نشود ، برنامه نوعی تایم اوت اعلام می‌کند بر این مبنا که پست های پیج به اتمام رسیده.

در خط ۳۱۲ هم تعریف شده اگر تعداد پست ها به ۱۰۰۰ رسید ، کافی است و اگر حتی پست های بیشتری هم داخل پیج باشد اسکرپ نشود. یعنی فقط ۱۰۰۰ پست جدید تر پیج.

عملگر خط ۳۱۷ ، اطلاعات فایل txt را فقط ذخیره می کند.
عملگر ۳۲۳ ، اطلاعات ذخیره شده را به مسیر مشخص انتقال می دهد.

از خط ۳۳۵ ، با تمام خط‌های زیر مجموعه آن ، تعریف می‌کنیم که با حلقه for به ازای هر آیتم (ایندکس) داخل لیست (به ازای هر پست) ، تعداد لایک های آن را با ۲ ایندکس قبلی و ۲ ایندکس بعدی مقایسه می‌کند (تفاوت مقدار لایک ها) و در صورتی که این تفاوت از مقدار نسبت بیشتر بود ، آن پست انتخاب می‌شود (خط ۳۲۵ و زیر مجموعه های آن) و ذخیره می شود.

سپس در خط ۳۶۷ با instaloader یک لاگین انجام می‌شود ، سپس از طریق لینک بدست آمده دانلود پست انجام می شود. در خط ۳۷۷ نیز نام فایل به شکل مد نظر (شامل لایک ، تاریخ و ...) ذخیره می شود.
در بهش except نیز به این شکل تعریف شده که در صورت خطا در اینستالودر و عدم دانلود ، اطلاعات به شکل txt ذخیره شود تا حداقل اسکرپ فایده خود را حفظ کند اطلاعاتی مثل لینک در دست باشد.

بعد از اتمام ، مجدداً این پروسه برای یوزرنیم بعدی تکرار می شود.