

بسم الله الرحمن الرحيم

راهنمای developer مازول web kit

فصل ۱

۱-۱ نگارش

در جدول زیر تاریخچه نگارش های مختلف ماژول آورده شده است.

تاریخ	نسخه	توضیحات
آذر ۹۷	۲,۰	نسخه اولیه document آماده شد.

۱-۲- توضیحات اولیه

ماژول web kit برای پیاده سازی ماژول های مبتنی browser فراهم شده است. این ماژول زبان Python و بر اساس ابزارهای موجود در binding کتابخانه QT (pyQt)، با کمک JavaScript نوشته شده است و برای استفاده از آن یک wrapper به زبان python نوشته شده است که شامل متدهایی برای استفاده از web kit می باشد. ورژن قبلی web kit نیز سازوکار مشابهی داشت با این تفاوت که web kit به زبان C++ توسعه داده شده بود و مستقیماً از Qt استفاده می شد. دلیل توسعه نسخه جدید، چالش های موجود در آن و امکانات حال حاضر web kit در ادامه آمده است.

۱-۳- دلیل پیاده سازی نسخه جدید

- ۱ - قدیمی شدن نسخه مرورگر استفاده شده در پیاده سازی قبلی
- ۲ - پشتیبانی نکردن از برخی سایت ها و همچنین باز نکردن سایت هایی که از captcha جدید گوگل استفاده می کردند مانند linkedin
- ۳ - امکانات بسیار زیاد و کاربردی نسخه جدید
- ۴ - عدم وجود نمونه پیاده سازی شده با نسخه جدید webkit chromium

۱-۴- مشکلات پیش آمده در زمان شروع پیاده سازی

- ۱- نبود نمونه برای گرفتن ایده و پیاده سازی
- ۲- مستندات برای زبان پایتون وجود نداشت و فقط مستندات C++ موجود بود، در حالت کلی مشکلی نداشت اما بعضی معادل سازی ها به دلیل تعاریف متفاوت دو زبان از شی گرایی و کار با object نیاز به نرمال سازی و variation & type converting داشت.
- ۳- برخی متدها مانند اجرای کدهای جاوا اسکریپت که بخش اصلی Hrobot را تشکیل می دهند مانند قبل همزمان نبودند و به صورت غیر همزمان طراحی شده بود که با طراحی EventLoop و ارسال signal حل شد.
- ۴- ارتباط بین پایتون و جاوا اسکریپت به صورت string برقرار بود و امکان دسترسی به object های یکدیگر را نداشتند که با پیاده سازی یک رابط برقرار شد اما وقت زیادی به دلیل وجود تعاریف متفاوت در دو زبان که باید مشترک می شد صرف شد.

۵- در برخی موارد باید رفتار کاربر مانند تایپ با کیبورد و موس شبیه سازی شود و نمی توان با جاوا اسکریپت کلیک یا متنی را جایگذاری کرد. برای اینکار متدهای ماوس و کیبورد فیزیکی طراحی شد.

۶- متدهای کار با NodeList های DOM باید پیاده سازی شود.

۷- متدهای کار با فرم، dropdown های انتخابی و دیگر موجودیت های یک صفحه وب باید پیاده سازی شود.

۸- مدیریت خطاهای جاوا اسکریپت.

۹- در روند پیاده سازی hrobot بر پایه نسخه جدید QtWebEngine به مشکلی برخورد کردیم که ظاهراً در Qt هنوز برطرف نشده، البته در نسخه های تست ظاهراً برطرف شده اما تا زمان ارائه رسمی قابل استفاده نمی باشد و باید راه حل موقتی برای این موضوع تا زمان ارائه رسمی نسخه بعدی پیدا کرد.

مشکل در مشخص کردن شروع لود یک صفحه یا اسکریپت و پایان آن است، لینک ها معمولی مشکلی ندارند اما لینک ها جاوا اسکریپتی یا anchor لینک ها فقط متد شروع لود را فراخوانی میکنند اما متد پایان لود را فراخوانی نمی کنند.

راه حل های ساده زیادی برای رفع این موضوع وجود دارد، اما همیشه و در همه حال قابل اعتماد نیستند، حل صحیح این مشکل به این دلیل اهمیت زیادی دارد، چون پیاده سازی به صورت غیر هم زمان باید انجام شود و اگر جایی با این مشکل برخورد کنیم و راه حل جواب ندهد، برنامه در حالت لودینگ باقی می ماند، مگر اینکه صفحه را دوباره لود یا به صفحه دیگری برویم.

از آنجایی که سیگنال loading به کاربر فرستاده نمی شود، کاربر باید به صورت polling به صفحه سر بزند و کنترل کند آیا element که انتظار دارد load شده است یا خیر. اگر فواصل بین polling ۰ در نظر گرفته شود کارایی سیستم پایین می آید، اگر بین این فواصل Wait گذاشته شود، از آنجایی که process کاربر و hrobot یکسان است و در دو thread مختلف اجرا می شوند wait کردن باعث می شود که تمام thread های فرایند sleep شوند و loading هم متوقف می شد، بنابراین هیچوقت این فرایند به پایان نمی رسد.

نهایتاً برای حل این مشکل thread مربوط به کاربر را درون یک QEventLoop مشغول نگه داشتیم که به صورت polling به thread مربوط به hrobot سر می زند. در صورت که load مربوط به element مورد نظر به پایان رسیده بود، از loop تعریف شده خارج می شود.

۱۰- برای نصب hrobot روی Fsystem مشکلی وجود داشت زیرا ماژول در thread متفاوتی نسبت به QApplication تعریف شده در process مربوطه اجرا می شود و نهایتاً Hrobot نیز در یک thread متفاوت نسبت به QApplication اجرا می گردد و باعث می شود خطای اجرا رخ دهد، زیرا Widget های گرافیکی QWebEngine باید در main thread اجرا گردد. برای حل این مساله ماژول هایی که از hrobot استفاده می کنند به صورت جداگانه در یک sub-process دیگر اجرا می شوند و نتایج خود را در آدرسی مشخص ذخیره می کنند و SystemApi نتایج را از روی فایل خوانده و به کاربر بازمی گرداند.

۵-۱- متدهای موجود در hrobot

۱- متد Save_as_pdf

این متد یک export با فرمت pdf از صفحه حال حاضر hrobot می گرد.

۲- متد Scroll_to_end

این متد صفحه حال حاضر hrobot را تا قسمت پایین scroll می کند.

۳- متد find_by_css

المنتهای موجود در صفحه را از طریق css selector بر می گرداند.

۴- متد find_by_xpath

المنتهای موجود در صفحه را از طریق xpath بر می گرداند.

۵- متد wait_until_css

تا زمان لود شدن المنت css مورد نظر، کاربر صبر می کند.

۶- متد wait_until_xpath

تا زمان لود شدن المنت xpath مورد نظر، کاربر صبر می کند.

۷- متد find_by_sibling

Sibling بعدی node ورودی را بر می گرداند.

۸- متد find_by_preceding

preceding بعدی node ورودی را بر می گرداند.

۹- متد find_element_has_attribute

المنتهایی که دارای یک attribute خاص هستند را بر می گرداند.

۱۰- متد find_by_attribute_value

المنتهایی که دارای attribute با یک value مشخص هستند.

۱۱- سایر متدها

المنتهای ذکر شده متدهای کلی هستند و سایر متدها مشتقاتی از این متدها می باشند.