به نام خدا

مباحثی در علوم کامپیوتر
تمرین سوم: خزش داده‌ها و جمع‌آوری دادگان
۷ دی ۱۳۹۹

استاد درس: دکتر اکبری
دانشگاه صنعتی امیرکبیر
دانشکده ریاضی و علوم کامپیوتر

Data scraping، که به عنوان Web scraping نیز شناخته می‌شود، فرآیند وارد کردن اطلاعات از وب‌سایت‌ها و برنامه‌ها به یک فایل csv یا پوشه‌ای در کامپیوتر شخصی شما است. این یکی از کارآمدترین راه‌ها برای دریافت داده از وب و در برخی موارد فرستادن آن به وب سایت دیگری است. کاربردهای معمول خزش[1] داده ها عبارتند از:

- تحقیق برای محتوای وب و هوش تجاری
- قیمت‌گذاری برای سایت‌های سفر رزرو و سایت‌های مقایسه قیمت
- یافتن سرنخ‌های فروش و انجام تحقیقات در یک بازار با crawl منابع داده عمومی (به عنوان مثال Yell و Twitter)
- ارسال اطلاعات محصول از یک سایت تجارت الکترونیکی به یک فروشنده آنلاین دیگر (به عنوان مثال قسمت Shopping در Google)
- ساخت داده‌گان‌های مناسب برای الگوریتم‌های یادگیری ماشین و یادگیری عمیق

۱ شرح پروژه

در این پروژه هدف استخراج تعدادی عکس از سایت Houzz و جمع‌آوری دادگانی است که می‌تواند برای انجام کارهای مختلف از جمله تشخیص اشیاء یا دسته‌بندی عکس‌ها مفید باشد. ابتدا سایت lewis john را بررسی کنید و توضیح دهید چرا سایت Houzz گزینه‌ی بهتری برای خزش است. در این بررسی راجع‌به مسیر robots.txt هم تحقیق کنید و این فایل را برای هر دو سایت بررسی کنید. برای خزش مسیرهای مربوط به میزها، صندلی‌ها، تخت‌ها و مبل‌ها را از سایت پیدا کرده و به‌عنوان ورودی به خزنده بدهید. لازم است از هر صفحه موارد زیر را جدا کنید:

- اسم محصول
- دو عکس اول هر محصول (در صورت وجود)
- در بخش توضیحات محصول، تگ This Product Has Been Described As

برای این‌کار می‌توانید از کتابخانه‌های مختلف مانند selenium و scrapy استفاده کنید. توضیح دهید تفاوت استفاده از این دو کتابخانه در چیست؟

[1] Crawl

۱.۱ امتیازی

- خودتان با JavaScript یا jquery و ابزارهای مشابه از میان آخرین وبلاگ‌های به‌روزشده در بلاگفا، از هر وبلاگ ۱۰ مطلب اخیر را جمع‌آوری کنید.
- برخی سایت‌ها با ردیابی ip و مشخصات مرورگر، با محدودکردن دسترسی‌های با فرکانس بالا از یک ip و یا مرورگر، خزش داده‌ها را مشکل می‌کند. کد خزنده را طوری بنویسید که در هر n درخواست متوالی، ip، مشخصات مرورگر و ویژگی‌هایی که با JavaScript قابل دریافت و ردیابی است را عوض کند.

۲ معیارهای تصحیح و ارزیابی

- فرمت داده‌های جمع‌آوری‌شده در نهایت منظم باشد و قابلیت استفاده مجدد را داشته باشد. (برای مثال در فرمت json یا csv ولی تمیز!)
- مستندسازی برای دیتاست جدیدتان و نوشتن توضیحاتی برای نفرات بعدی‌ای که از دیتاست استفاده خواهند کرد. (برای مثال ویژگی‌هایی مانند این که چند نمونه در دیتاست موجود است، هر نمونه به چه فرمت است و برای چه تسک‌هایی می‌توان از این دیتاست استفاده کرد.)
- خوانایی کد مورد استفاده برای عملیات خزش
- نوشتن کد مربوط به خزش در قالب یک کلاس پایتون که قابلیت استفاده‌ی دوباره از آن برای سایت‌های دیگر هم تا حدودی فراهم باشد. (یعنی کد نوشته‌شده فقط مخصوص یک وبسایت نباشد و قابلیت تعمیم‌پذیری داشته باشد.)
- قابل‌اعتماد و منعطف بودن کد خزنده، به طوری که در صورت قطع اینترنت، قطع برق و یا مشکلات مشابه، کل داده‌هایی که تا به حال خزش شده‌اند از دست نرود.

۳ ارسال پاسخ

مهلت ارسال پاسخ: ۱۵ دی ۱۳۹۹
مهلت ارسال با تاخیر (۱۰درصد کسر نمره به ازای هر روز): ۱۷ دی ۱۳۹۹

ارتباط با ما

ایمیل:

یاسمن امی
yassi.ommi@gmail.com

آرمان ملک‌زاده
malekzadeh@ieee.org

موفق باشید!