



به نام خدا

درس برنامه سازی پیشرفته

دانشکده مهندسی کامپیوتر

دانشگاه صنعتی شریف

ترم پاییز ۱۴۰۴-۱۴۰۵

استاد :

علی نجیمی

عنوان پروژه :

Web Scraping

Product Owner

آبین کریمیان

Product Manager

محمد حسین سورانی

فهرست مطالب

نکته قابل توجه

مشخصات پروژه

اهداف کلی

اهداف پروژه

شرح خلاصه‌ی پروژه

پروژه

۴	تحلیل نیازمندی‌ها
۷	طراحی و ساختار فنی
۸	تقسیم‌بندی نمره
۹	چک لیست تحویل نهایی



نکته قابل توجه

- این داک صرفا توضیحات مربوط به محتوای پروژه Web Scraping می‌باشد. توضیحات مربوط به فازهای پروژه، قوانین انجام پروژه و نحوه نمره‌دهی در **داک کلی راهنمای پروژه‌ها** نوشته شده. لذا قبل از مطالعه هر کدام از داک‌ها، داک کلی راهنمای پروژه‌ها را مطالعه نمایید.



مشخصات پروژه

• **عنوان :** استخراج کننده اطلاعات از وب سایت ها (Web Scraping)

• **تعداد اعضا :** ۳

• **هدف :** در حال حاضر بسیاری از شرکت ها برای استخراج داده های مفید از وب و تحلیل محتوای صفحات HTML از ابزارهای مختلفی استفاده می کنند. این پروژه با هدف آشنایی دانشجو با مفاهیم پایه ای Web Scraping و پیاده سازی یک ابزار ساده خط فرمان (CLI) طراحی شده است.

به طور کلی، این پروژه به عنوان یک تمرین بنیادین برای آشنایی با تحلیل داده، برنامه نویسی شی گرا، و طراحی نرم افزارهای کاربردی در صنعت در نظر گرفته شده است.



اهداف کلی

اهداف پروژه

- آشنایی با مفاهیم شیگرایی (Object-Oriented Programming) و تفکیک مناسب بخش‌های مختلف
- آشنایی و استفاده از design patterns و معماری سیستم‌های نرم افزاری مثل (Model View Control) است.
- با استفاده از JUnit، دانشجو یاد می‌گیرد برای اطمینان از صحت عملکرد کلاس‌های کلیدی مانند HtmlParser تست‌های مناسب بنویسد.
- آشنایی با سیستم‌های مدیریت نسخه مانند Git و کار در بستر یک مخزن GitHub از نکات مهم این پروژه است. تغییرات باید در بازه‌های کوتاه مدت commit شوند.
- آشنایی با مدیریت خطاهایی مانند FileNotFoundException و طراحی مکانیزم‌های امن برای کنترل ورودی‌های نامعتبر.

شرح و خلاصه پروژه

در این پروژه، برنامه‌ای خط فرمان طراحی می‌شود که مسیر یک فایل HTML را از کاربر می‌گیرد، محتوای آن را می‌خواند و تمام لینک‌های موجود در تگ‌های مختلف را استخراج کرده و نمایش می‌دهد. در صورت وارد شدن مسیر اشتباه یا فایل نامعتبر، خطای درستی مدیریت می‌شود. هدف کلی، آشنایی با شیگرایی، خواندن فایل‌ها و تحلیل محتوای HTML است.



پروژه

تحلیل نیازمندی ها

جدول ۱. نیازمندی های کاربردی (Functional Requirements)

کد	نیازمندی	تحلیل پیاده سازی
FR-01	دریافت فایل HTML	مسیر یک فایل HTML ذخیره شده روی دیسک را به عنوان ورودی از طریق خط فرمان دریافت کند.
FR-02	تحلیل فایل HTML	سیستم باید محتوای فایل HTML را بخواند و تحلیل کند (توسط <code>HtmlParser</code>).
FR-03	استخراج لینک ها از فایل	تمام لینک های موجود در تگ های <code><a></code> (ویژگی <code>href</code>) را از فایل استخراج کرده و در کنسول چاپ کند.
FR-04	دریافت URL	برنامه باید بتواند یک URL واقعی را به عنوان ورودی دریافت کند.
FR-05	واکشی محتوای URL	برنامه باید محتوای URL ورودی را از طریق اینترنت واکشی کند.
FR-06	استخراج داده های متنوع	استخراج انواع مختلف داده (مانند لینک ها، آدرس تصاویر <code></code> ، یا تیترهای <code><h1></code>) را داشته باشد.
FR-07	ذخیره نتایج در CSV	نتایج استخراج شده باید در یک فایل با فرمت CSV ذخیره شوند.
FR-08	دریافت لیست URL ها - سرور	سرور باید قادر باشد لیستی از URL ها را از کلاینت دریافت کند.
FR-09	اسکرپ همزمان	سرور باید URL های دریافتی را به صورت همزمان (Concurrent) اسکرپ کند.
FR-10	ارسال نتایج aggregated	نتایج جمع آوری شده از تمام URL ها باید به صورت یکپارچه به کلاینت بازگردانده شود.
FR-11	ذخیره نتایج سرور	[به عنوان جایگزین FR10] نتایج جمع آوری شده می تواند در یک فایل واحد روی سرور ذخیره شود.
FR-12	ارسال لیست URL - کلاینت	کلاینت باید بتواند به سرور متصل شده و لیستی از URL ها را برای آن ارسال کند.
FR-13	دریافت تاییدیه - کلاینت	کلاینت باید پس از اتمام پردازش توسط سرور، یک پیام تأیید یا نتایج aggregated را دریافت کند.



جدول ۲. نیازمندی‌های غیر کاربردی (Non-Functional Requirements)

نیازمندی	کد
تحلیل پیاده سازی	
برنامه در فاز اول باید یک برنامه ساده خط فرمان (CLI) باشد.	رابط کاربری NFR-01
تمرز اصلی باید برروی پیاده سازی صحیح اصول شی‌عگرایی باشد.	مفاهیم پیاده سازی NFR-02
ورودی‌های نامعتبر مانند مسیر فایل اشتباه باید به شکل مناسبی مدیریت شوند.	مدیریت خط NFR-03
تست‌های جامع (JUnit) باید برای منطق استخراج لینک (HtmlParser) نوشته شود.	قابلیت تست NFR-04
برنامه باید به یک اپلیکیشن ماژولار تبدیل شود.	معماری ماژولار NFR-05
الگوی Strategy برای تعریف استراتژی‌های مختلف استخراج داده (لینک، تصویر، تیتر) استفاده شود.	الگوی Strategy NFR-06
کد باید بازنویسی (Refactor) شود تا با معماری جدید انطباق داشته و خوانایی بیشتری داشته باشد	کیفیت کد NFR-07
برنامه باید به یک مدل کلاینت-سرور تبدیل شود.	معماری کلاینت-سرور NFR-08
برنامه باید مقیاس پذیر شود (تبدیل به ابزار قدرتمندتر).	مقیاس پذیری NFR-09
یک Thread Pool (مانند ExecutorService) برای مدیریت همزمان فرآیند اسکرپ کردن استفاده کند.	همزمانی NFR-10
باید از ایجاد بی‌رویه نخ‌ها جلوگیری شده و منابع سیستم بهینه مدیریت شوند.	مدیریت منابع NFR-11
نتایج حاصل از هر نخ باید به صورت ایمن (Thread-safe) در یک مجموعه داده مشترک جمع‌آوری شوند.	ایمنی نخ NFR-12
ارتیکل شبکه باید با استفاده از Socket Programming و ServerSocket پیاده سازی شود.	ارتیکل شبکه NFR-12



طراحی و ساختار فنی

جدول ۳. استفاده از مفاهیم OOP

نمره	توضیح	مفهوم
الزامي	طراحی کلاس‌های جداگانه برای پخش‌های مختلف سیستم (ورودی CLI، پردازش HTML، استخراج داده و نمایش خروجی)	کلاس‌ها و شی‌گرایی
%۳	ویژگی‌ها و متدهای مربوط به تحلیل HTML، ذخیره لینک‌ها یا داده‌ها درون کلاس‌های مرتبط نگهداری شده و از طریق getter/setter کنترل می‌شوند.	Encapsulation
%۲	تعريف یک کلاس پایه برای پردازشگرهای عمومی و ارث بری از آن برای تحلیل تگ‌های خاص مثل <a>، یا سایر تگ‌ها.	Inheritance
%۱.۵	استفاده از متدهای override شده برای رفتارهای مختلف تحلیل تگ‌ها با نمایش داده در خروجی‌های متفاوت (مثلًاً کنسول یا فایل).	Polymorphism
%۱.۵	کلاس‌های سطح بالاتر (مثلًاً کنترل‌کننده) شامل نمونه‌هایی از کلاس‌های تحلیلگر و رابط خط فرمان هستند تا تعامل بین اجزا برقرار شود.	Composition
%۲	تعريف رابط‌های عمومی برای مازول‌هایی که وظایف خاصی دارند (مثل OutputHandler یا Parser) بدون اشاره به جزئیات پیاده‌سازی.	Abstraction

توجه: الزامي بودن نمره «کلاس‌ها و شی‌گرایی» به این معنی است که دریافت نمره نهایی ملزم به قبول شدن در این بخش است. به عنوان مثال، اگر نمره این بخش ۵٪ شود، باید بخش «کلاس‌ها و شی‌گرایی» را انجام داده باشد تا این ۵٪ نمره را کسب کنید.

جدول ۴. نمودارهای مورد نیاز

نمره	شرح	نمودار
%۵	نمایش نقش‌ها و عملکرد کلی سیستم	Use Case Diagram
%۵	نمایش ساختار کلاس‌ها و روابط آن‌ها	Class Diagram
%۵	نمایش مراحل اجرای برنامه از ورودی تا خروجی	Activity Diagram



تقسیم‌بندی نمره

درصد	شرح	بخش
%۷۲	هر نیازمندی کاربردی %۴.۵ و هر نیازمندی غیرکاربردی %۱.۵ (کسب %۴۰ برای نمره کامل این بخش کافیست)	پیاده‌سازی نیازمندی‌ها
%۱۰	مطابق جدول ۳	اصول OOP و ساختار فنی
%۱۵	مطابق جدول ۴	طراحی نمودارها
%۱۵	نوشتن سناریو تست، تست واحد، اجرای بدون خطای	تست و اعتبارسنجی
%۲۰	ارائه در جلسه‌ی نهایی کلاس، توضیح عملکرد، نوشتن مستند Word یا PDF مناسب	مستندسازی و ارائه

توجه: طراحی گرافیکی و UI جزو مباحث اصلی درس نیست. با این حال، برای افرادی که قصد پیاده‌سازی پروژه گرافیکی دارند نمره امتیازی در نظر گفته می‌شود. به این صورت که ۱۰٪ از نمره پروژه (۰.۸٪ از ۲۰٪) بطور اضافی و جدا از نمره‌های امتیازی دیگر محاسبه می‌شود.



چک لیست تحویل نهایی

● نمودارها

● کد برنامه با توضیحات

● فایل اجرایی

● گزارش تست و خروجی

● ارائه نهایی