روزبه پورجعفریان (۴۰۰۱۰۶۲۷۹)
مبانی علم داده
استاد درس: دکتر خلج

تاریخ دریافت: ۱۴۰۳/۱۰/۲۱
تاریخ ارسال: ۱۷/ ۱۴۰۳/۱۱
فاز صفر پروژه

Scraping برای جمع‌آوری مقالات علمی

۱. مقدمه

در این فاز، هدف جمع‌آوری مقالات علمی مرتبط با پنج موضوع زیر از Semantic Scholar بود:

- Foundation Models
- Generative Models
- LLM (Large Language Models)
- VLM (Vision-Language Models)
- Diffusion Models

از وب کراولینگ (Web Crawling) با Selenium برای جمع‌آوری خودکار مقالات استفاده شد. داده‌های استخراج‌شده شامل:

- عنوان مقاله (Title)
- نویسندگان (Authors)
- سال انتشار (Year)
- تعداد استنادات (Citations)
- چکیده (Abstract)
- لینک مقاله (URL)

تمام این اطلاعات در فایل‌های CSV ذخیره شدند تا در مراحل بعدی پروژه استفاده شوند.

---

۲. روش اجرا

۲.۱ ابزارها و تکنولوژی‌ها

Selenium WebDriver: برای تعامل با صفحات وب
Chrome WebDriver: برای اجرای خودکار مرورگر
Pandas: برای پردازش و ذخیره داده‌ها در قالبCSV
WebDriver Manager: برای مدیریت خودکار درایور کروم

۲.۲ مراحل اجرا

۱ـ جستجوی موضوعات:

- ورود به سایت Semantic Scholar
- یافتن فیلد جستجو و وارد کردن نام موضوع
- ارسال درخواست جست‌وجو

۲ـ جمع‌آوری داده‌ها:

- دریافت مقالات نمایش داده‌شده
- استخراج عنوان، نویسندگان، سال، استنادات، چکیده و لینک مقاله
- فیلتر کردن مقالات قدیمی

۳ـ بارگیری مقالات بیشتر:

- استفاده از اسکرول اتوماتیک برای نمایش مقالات بیشتر
- کلیک روی دکمه "Next Page" در صورت موجود بودن
- انجام این کار تا زمانی که ۲۰۰۰ مقاله برای هر موضوع جمع‌آوری شود

۴ـ ذخیره داده‌ها:

- ذخیره مقالات در فایل‌های CSV جداگانه برای هر موضوع
- ساخت پوشه‌ای مخصوص ذخیره داده‌ها

---

۳. چالش‌ها و راهکارها

۳.۱ چالش‌های فنی

Semantic Scholar تعداد محدودی مقاله را نمایش می‌دهد
راهکار :استفاده از اسکرول بی‌نهایت و کلیک روی "Next Page"

محدودیت در دسترسی به نتایج زیاد
راهکار: افزایش تعداد اسکرول‌ها و تأخیر تصادفی بین درخواست‌ها

عدم نمایش مقالات برای برخی موضوعات
راهکار: تغییر عبارات جستجو به نام‌های مرتبط‌تر و دقیق‌تر

۴ـ نتایج و خروجی‌ها

فایل‌های CSV ذخیره‌شده:

نام فایل	مقالات جمع‌آوری‌شده	موضوع
Foundation_Models.csv	+۲۰۰۰	Foundation Models
Generative_Models.csv	+۲۰۰۰	Generative Models
LLM.csv	+۸۰۰	LLM
VLM.csv	+۵۰۰	VLM
Diffusion_Models.csv	+۶۰۰	Diffusion Models

۵ـ نتیجه‌گیری و اقدامات بعدی

- این فاز به طور کلی موفق بود و مقالات مورد نیاز از Semantic Scholar استخراج شدند.
- روش Web crawling کارآمد بود اما برخی موضوعات نیاز به جستجوی پیشرفته‌تر دارند.
- داده‌های جمع‌آوری‌شده برای فازهای بعدی پردازش و استفاده خواهند شد.