Distribuirani web scraper

Nikola Nöthig

Raspodjeljeni sustavi

Sveučilište Jurja Dobrile u Puli

21.6.2024

Opis Projekta:

Ovaj projekt implementira raspodijeljeni web scraping sustav za prikupljanje podataka sa 3 web stranice koristeći FastAPI, Celery s Redisom kao posrednikom za poruke, i MongoDB za pohranu podataka. Sustav omogućava asinkrono i paralelno izvršavanje scraping zadataka, poboljšavajući performanse i efikasnost obrade.

Komponente:

1. **FastAPI aplikacija**:
    - **Endpointi**:
        - / : Početna stranica koja koristi index.html.
        - /scrape_all : Pokreće scraping zadatake za sve scrape funkcije, upravlja asinkronim izvršavanjem i pohranjuje rezultate u MongoDB.
    - **Background zadaci**:
        - Asinkrono izvršavanje i praćenje Celery zadataka.
        - Pohrana rezultata u MongoDB.
2. **Celery worker**:
    - Definira Celery aplikaciju s Redisom kao brokerom i backendom.
    - Zadaci dodijeljeni različitim redovima na osnovu scrape funkcija.
3. **Scrapers**:
    - Tri različita scraper modula (scraper1, scraper2, scraper3) za različite izvore podataka.
    - Koristi BeautifulSoup za obradu HTML sadržaja i izvlačenje informacija o proizvodima.
4. **MongoDB**:
    - Pohrana podataka prikupljenih scrapingom.
    - Funkcije za dohvat i brisanje podataka preko FastAPI endpointa.

Funkcionalnosti:

- **Scraping podataka**: Sustav periodično i automatski dohvaća podatke s ciljanih web stranica, obrađuje ih i pohranjuje.
- **Distribuirana obrada**: Celery radnici raspoređuju zadatke na više workera ili procesa za bržu obradu.
- **Baza podataka**: MongoDB se koristi za fleksibilno i sigurno skladištenje podataka, omogućavajući lako pristupanje i upravljanje prikupljenim informacijama.
- **Korisničko sučelje**: Služi za prikaz informacija i podataka na localhostu.

- **Praćenje i logiranje**: Integrirano logiranje pruža uvid u status scraping zadataka i omogućava otklanjanje grešaka.

Implementacija:

- **Asinkroni HTTP zahtjevi** uz pomoć httpx omogućavaju dohvaćanje, spremanje i brisanje podataka.
- **Celery taskovi** se definiraju i pozivaju za distribuiranu obradu svakog scraper modula, optimizirajući upotrebu resursa i vremensku efikasnost.

Ovaj projekt omogućuje efikasno i skalabilno prikupljanje podataka iz više izvora, koristeći modernu arhitekturu i alate za razvoj backend servisa.