Sveučilište Jurja Dobrile u Puli

Tehnički fakultet u Puli


DANIEL VORIĆ

# AkcijoSC

Dokumentacija

JMBAG: 03030969417, redoviti student

Studijski smjer: Sveučilišni diplomski studij Računarstva

Predmet: Raspodijeljeni sustavi

Predmetni nastavnik: doc. dr. sc. Nikola Tanković

Pula, Rujan, 2025. godine

## Sadržaj

## 1. Opis projekta

AkcijoSC je distribuirani sustav za scraping proizvoda na akciji s više web trgovina poput linksa, chipoteke i instara. Korištenjem FastAPI-ja, Celery workera s Redisom kao posrednikom za poruke postiže se paralelno izvršavanje zadataka za prikupljanje podataka, ti podaci se pohranjuju u MongoDB bazu podataka. Za korištenje aplikacije koristi se jednostavno web sučelje.

## 2. Komponente

1. FastAPI aplikacija:
    - Endpointi:
        - POST /scrape/all - zakazuje scraping (Instar + Links + Chipoteka); vraća task ID-ove.
        - POST /results/merge - spaja završene rezultate (iz task id-eva); vraća stavke.
        - POST /results/save - sprema spojene stavke u Mongo (iz task id-eva); vraća broj spremljenih.
        - GET /database/list - vraća sve spremljene stavke.
        - DELETE /database/clear - briše sve iz kolekcije.
        - GET /database/ping - brza provjera konekcije na bazu.
2. Celery worker:
    - Definira Celery aplikaciju s Redisom kao brokerom I backendom.
    - Odrađuje dodijeljene zadatke asinkrono (Gevent pool za bolje performanse I asinkronost)
3. Scrapers:
    - Tri različita scraper modula za različite izvore podataka;
        - Scraper_instar – za webshop Instar
        - Scraper_chipoteka – za webshop Chipoteka
        - Scraper_links – za webshop Links

- Koristi BeautifulSoup za obradu HTML sadržaja I izvlačenje informacija o proizvodima

4. MongoDB:

- Pohrana scrapeanih podataka.
- Funkcije za dohvat i brisanje podataka preko FastAPI endpointa.
- JSON model:
  - - name (string) – naziv artikla
  - - price_new (string) – trenutna cijena
  - - price_old (string | null) – stara cijena
  - - discount_pct (float | null) – postotak popusta, ako se može izračunati
  - - source (string) – izvor: `instar`, `links`, `chipoteka`
  - - scraped_at (string) – timestamp u formatu `DD.MM.YYYY/HH:MM`
- - Naziv baze: baza_artikli
- - Kolekcija: artikli
  -

5. Docker:

- Kontejnerizacija

## 3. Karakteristike

1. Automatizirano prikupljanje akcijskih proizvoda s više izvora
2. Paralelizacija i skaliranje scraping zadataka (Celery + Redis)
3. Jednostavan prikaz i upravljanje (UI + REST API)
4. Centralna pohrana podataka (MongoDB) radi daljnje analize

## 4. Funkcionalnosti

1. Scraping podataka: dohvaćanje podataka s ciljanih web shopova, obrađivanje i pohranjivanje
2. Distribuirana obrada: Celery radnici raspoređuju zadatke na više workera i procesa za bržu obradu.
3. Baza podataka: MongoDB se koristi za sigurno skladištenje podataka, omogućujući lak pristup i upravljanje prikupljenim podacima.
4. Korisničko sučelje: Omogućuje jednostavno korištenje scrapera i prikaz podataka na localhostu.

## 5. Postavljanje I pokretanje

### Opcija A: Pokretanje uz Docker

Iz root projekta pokrenite:

```
docker compose build --no-cache
docker compose up -d
docker compose ps
```

Otvorite:

- UI: http://localhost:8000/ui za GUI sučelje
- Docs: http://localhost:8000/docs za Swagger UI

### Opcija B: Pokretanje lokalno (bez Dockera)

1. Kreirajte i aktivirajte virtualno okruženje, pa instalirajte `requirements.txt`:

```
py -3.11 -m venv venv
venv\\Scripts\\Activate.ps1
pip install -r requirements.txt
```

2. Postavite varijable okruženja (PowerShell):

```
$env:REDIS_URL = "redis://localhost:6379/0"
$env:MONGODB_URI = "mongodb+srv://<user>:<pass>@<cluster>/?retryWrites=true&w=majority" (za vaš db)
```

3. Pokrenite Redis:

- Instalirajte Redis lokalno i pokrenite ga na portu 6379.
- Navigirajte se putem Powershella u instalation folder i upišite komandu redis-server.
- Iz drugog Powershella isprobajte funkcionalnost sa: redis-cli ping

3. Pokrenite Redis:

- Instalirajte Redis lokalno i pokrenite ga na portu 6379.
- Navigirajte se putem Powershella u instalation folder i upišite komandu redis-server.
- Iz drugog Powershella isprobajte funkcionalnost sa: redis-cli ping

4. Pokrenite Celery workere (u odvojenim terminalima):

!Važno! Ako u postavkama nije dozvoljeno izvršavanje vanjskih skripti (i ne želite staviti unrestrited) potrebno je upisati ovu naredbu u svaki terminal. Za bolje performanse povečati broj nakon -c. (concurrency level)

```
Set-ExecutionPolicy -ExecutionPolicy RemoteSigned -Scope Process
```

```
venv\\Scripts\\Activate.ps1
python -m celery -A scraperi.celery_app:app worker -Q links_queue -P gevent -c 100 -Ofair -n links@%h -l INFO
```

```
venv\\Scripts\\Activate.ps1
python -m celery -A scraperi.celery_app:app worker -Q chipoteka_queue -P gevent -c 100 -Ofair -n chipoteka@%h -l INFO
```

```
venv\\Scripts\\Activate.ps1
python -m celery -A scraperi.celery_app:app worker -Q instar_queue -P gevent -c 100 -Ofair -n instar@%h -l INFO
```

4. Pokrenite Celery workere (u odvojenim terminalima):

!Važno! Ako u postavkama nije dozvoljeno izvršavanje vanjskih skripti (i ne želite staviti unrestrited) potrebno je upisati ovu naredbu u svaki terminal. Za bolje performanse povečati broj nakon -c. (concurrency level)

```
Set-ExecutionPolicy -ExecutionPolicy RemoteSigned -Scope Process
```

```
venv\\Scripts\\Activate.ps1
python -m celery -A scraperi.celery_app:app worker -Q links_queue -P gevent -c 100 -Ofair -n links@%h -l INFO
```

```
venv\\Scripts\\Activate.ps1
python -m celery -A scraperi.celery_app:app worker -Q chipoteka_queue -P gevent -c 100 -Ofair -n chipoteka@%h -l INFO
```

```
venv\\Scripts\\Activate.ps1
python -m celery -A scraperi.celery_app:app worker -Q instar_queue -P gevent -c 100 -Ofair -n instar@%h -l INFO
```

5. Pokrenite API:

```
venv\\Scripts\\Activate.ps1
uvicorn scraperi.main:app --host 0.0.0.0 --port 8000
```

6. Otvorite UI i koristite aplikaciju:

- UI: http://localhost:8000/ui
- Docs: http://localhost:8000/docs za Swagger UI

## 6. Korištenje

1. U UI-ju kliknite “Start Scrape” za pokretanje zadataka.
2. Pričekajte da merge prikaže pronađene stavke i da nema više pending zadataka.
3. Kliknite “Save to DB” za spremanje rezultata u MongoDB.
4. Kliknite Load Table (from merge)
5. “Load from DB” učitava spremljene stavke.

## 7. Zaključak

AkcijoSC demonstrira distribuirani scraping s paralelizacijom i trajnom pohranom. Uz Docker, pokretanje je jednostavno, a lokalno pokretanje omogućuje fleksibilno debugiranje. Dokument pokriva ključne informacije o aplikaciji te korake za instalaciju i korištenje.