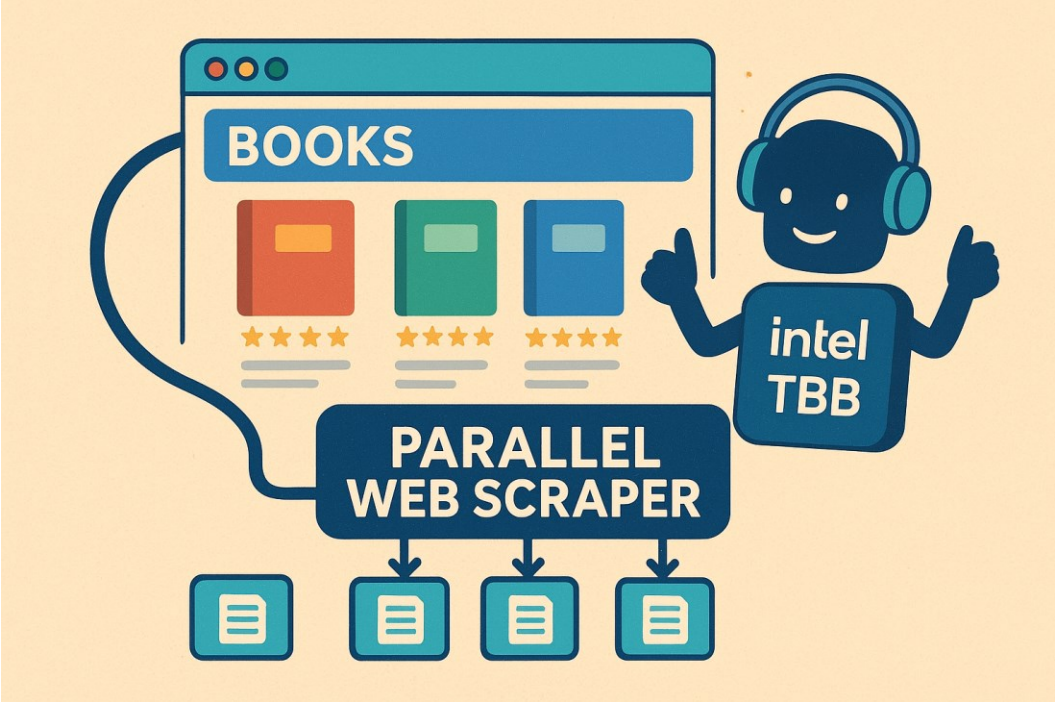
Realizacija paralelnog web sakupljača podataka upotrebom Intel TBB okvira



1. Pregled projekta

Cilj projekta je implementacija **paralelnog web sakupljača podataka** (engl. *web scraper*) koji efikasno analizira sadržaj web stranica koristeći **Intel Threading Building Blocks (TBB)**. Program će koristiti TBB model paralelizma zasnovan na zadacima i konkurentne strukture podataka kako bi se optimizovalo izvođenje na savremenim višejezgranim procesorima.

Sistem treba da podržava:

- Istovremeno preuzimanje prethodno indeksiranih web stranica sa sajta: **https://books.toscrape.com/index.html**
    - Indeksiranje nije potrebno uraditi automatski (moguće je ručno zadati različite indekse sa zadatog sajta)

    - **VAŽNO!** Korišćenje drugih web sajtova koji svojom politkom ne podržavaju sakupljanje podataka je zabranjeno.
- Paralelnu analizu HTML dokumenata.
- Bezbedno skladištenje (thread-safe) obrađenih podataka (URL-ova, metapodataka i teksta).
- Skalabilno izvršavanje sa podesivim nivoima paralelizma.

2. Programski zahtevi

Funkcionalni zahtevi

1. **Unos i inicijalizacija URL-ova**
    - Prihvatanje liste početnih URL-ova.
2. **Preuzimanje stranica**
    - Istovremeno preuzimanje stranica putem HTTP/HTTPS zahteva.
    - Podrška za timeout i retry logiku.
3. **Analiza sadržaja**
    - Paralelno izdvajanje podataka sa stranica.
4. **Skladištenje podataka**
    - Evidencija posećenih URL-ova u konkurentnoj strukturi podataka.
    - Čuvanje teksta i metapodataka u konkurentnim kontejnerima.
5. **Rezulati izvršavanja programa**
    - **Broj preuzetih stranica, jedinstvenih URL-ova.**
    - **Statistiku kao što je propusnost (analizirane stranice u jedinici vremena).**
    - **Pet rezultata analize sadržaja stranica**
        1. Broj knjiga koje su ocenjene sa 5 zvezdica
        2. Prosečnu cenu knjige
        3. Proizvoljno po vašem izboru

4. Proizvoljno po vašem izboru
5. Proizvoljno po vašem izboru

Rezultate ispisati u izlaznu datoteku.

Nefunkcionalni zahtevi

- **Paralelizam**: Efikasno zakazivanje zadataka koristeći TBB task groups i paralelne algoritme iz TBB okvira.
- **Skladištenje**: Upotreba konkurentnih kontejnera
- **Skalabilnost**: Skaliranje u skladu sa brojem CPU jezgara i analiza različitog ponašanja u zavisnosti od konfiguracije paralelnog izvršavanja.
- **Robusnost**: Otpornost na greške u mreži.

Dodatni poeni:

Napomena: Dodatnim poenima nije moguće preći maksimalni broj bodova koje projekat nosi, već nadoknaditi ukoliko neke od podrazumevanih bodova tokom odbrane izgubite

Za 2 dodatna poena napraviti paralelno autoamtsko indeksiranje sajta radi pronalaženja stranica koje program treba da analizira.

Za 3 dodatna poena upotrebiti parallel_pipeline iz TBB okvira za lančanje faza projekta, npr:

- Faza 1: Preuzimanje URL-ova.
- Faza 2: Analiza.
- Faza 3: Čuvanja rezultata analize.