Programiranje 3: Ispit, Februar 2023

Prezime Ime: _____ **Broj indeksa:** _____

Uputstva

- Ispit traje **120** min.
- Projekat je potrebno kreirati na desktopu i dati mu naziv u formatu: **Prezime_Ime_Indeks** (izbegavati naša slova u imenu i prezimenu).
- Dozvoljeno je korišćenje samo onih materijala koji se nalaze u folderu **P3 Feb 2023** na desktop-u.
- **Nije dozvoljeno** korišćenje Interneta, niti korišćenje pomoćnih materijala u elektronskom, papirnom, ili bilo kom drugom obliku.
- Ovaj list sa tekstom zadatka mora da bude potpisan i ostavljen na tastaturi na kraju ispita.
- **Strogo je zabranjeno iznošenje, fotografisanje ili umnožavanje zadataka** na bilo koji način. Povreda ovog pravila se strogo kažnjava.
- Preporučuje se često snimanje (Save) onoga što je urađeno, kako u slučaju problema sa računarom, nestankom struje i sl. ne bi bilo izgubljeno ono što je uradjeno.
- Na kraju rada treba proveriti još jednom da li su sačuvani sve file-ovi u projektu. Delovi projekta koji nisu snimljeni neće biti preneti prilikom kopiranja zadataka za pregledanje i naknadne intervencije nisu moguće.
- Zadaci sa kompajlerskim greškama se ne pregledaju i automatski se ocenjuju sa 0 poena.

Zadatak

I'm only myself when I have a guitar in my hands.
George Harrison

Kreirati projekat (v. *Uputstva*) i u njemu dva foldera: *notebooks* i *data*. U folderu *notebooks* kreirati notebook *feb2023.ipynb* i u njemu raditi sve ostale stvari koje se traže u zadatku. Kreirati u projektu i folder *data* u koji treba postaviti dataset koji se kreira na kraju zadatka. Ovakva struktura projekta je obavezna i eliminatorna.

Svaka značajna celina u zadatku treba da bude u notebooku *feb2023.ipynb* u zasebnoj code ćeliji, a ispred te ćelije treba da stoji markdown ćelija sa kratkim objašnjenjem koda u ćeliji ispod nje. I ovo je obavezno i eliminatorno.

Potrebno je uraditi Web scraping sa sajta Louder Sound, *The 50 Best Guitarists Of All Time*, koji prikazuje najbolje gitariste svih vremena. Početna stranica je https://www.loudersound.com/features/the-50-greatest-guitarists-of-all-time. Za svakog gitaristu treba sa tog sajta sakupiti sledeće informacije: ime (u daljem tekstu *name*), rang (pozicija na listi, u daljem tekstu *rank*), ime autora članka o tom gitaristi (*author*), ime grupe (*band*) u kojoj je taj autor svirao (ako je dato; u protivnom, postaviti taj podatak na vrednost *None*) i link na YouTube video prikazan u članku o tom gitaristi (*link*). Sakupljene podatke treba zapisati u jedan *csv* file u folderu *data* i zatim prikazati sadržaj tog file-a na ekranu u vidu odgovarajućeg Pandas *DataFrame* objekta.

Koraci koji se ocenjuju u zadatku su[1]:

1. Kreirati i testirati funkciju *get_soup()* koja sa zadate stranice preuzima HTML sadržaj i na osnovu njega kreira *BeautifulSoup* objekat. Za preuzimanje HTML sadržaja koristiti paket po želji (*requests* ili *selenium*).
2. Kreirati i testirati funkciju *get_specific_page()* koja polazi od početne URL adrese sajta sa više stranica i rednog broja željene stranice i vraća URL željene stranice.

[1] Pored ovih koraka, dozvoljeno je napraviti i pomoćne korake radi lakšeg rešavanja zadatka, ali se oni ne ocenjuju.

Programiranje 3: Ispit, Februar 2023

Prezime Ime: _____ **Broj indeksa:** _____

3. Kreirati i testirati funkciju *get_next_soup()* koja vraća odgovarajući *BeautifulSoup* objekat sa zadate stranice sajta sa više stranica.
4. Kreirati i testirati funkciju *find_text_info()* koja polazi od odgovarajućeg *BeautifulSoup* objekta i vraća listu uređenih četvorki tekstualnih informacija za svakog autora u obliku (*name*, *rank*, *author*, *band*).
5. Kreirati i testirati funkciju *get_youtube_video()* koja polazi od odgovarajućeg *BeautifulSoup* objekta i rednog broja odgovarajućeg YouTube videa i vraća link na taj video[2].
6. Kreirati i testirati funkciju *get_yt_links()* koja polazi od odgovarajućeg *BeautifulSoup* objekta i vraća listu svih YouTube video linkova u tom objektu.
7. Kreirati i testirati generatorsku funkciju *crawl()* koja polazi od URL-a početne stranice sajta sa kojeg se vrši Web scraping i maksimalnog broja stranica pokojima treba uraditi Web crawling za prikupljanje traženih podataka i vraća odgovarajuće *BeautifulSoup* objekte.
8. Kreirati i testirati funkciju *get_article_info_list()* koja polazi od URL-a početne stranice sajta sa kojeg se vrši Web scraping i maksimalnog broja stranica pokojima treba uraditi Web crawling za prikupljanje traženih podataka i vraća listu uređenih petorki za svakog gitaristu *na celom tom sajtu* u obliku (*name*, *rank*, *author*, *band*).
9. Na osnovu liste formirane u prethodnom koraku formirati odgovarajući *csv* file u *data* folderu.
10. Prikazati na ekranu odgovarajući Pandas *DataFrame* objekat na osnovu formiranog *csv* file-a.

Način ocenjivanja

Deo zadatka	**Maksimalan broj poena za taj deo zadatka**
Funkcija **get_soup**	3
Funkcija **get_specific_page()**	3
Funkcija **get_next_soup()**	3
Funkcija **find_text_info()**	18
Funkcija **get_youtube_video()**	13
Funkcija **get_yt_links()**	8
Funkcija **crawl()**	3
Funkcija **get_article_info_list()**	13
Kreiranje **csv** file-a	5
Prikazivanje **csv** file-a	1

[2] Voditi računa o tome da su linkovi na YouTube video clip-ove oblika npr. https://www.youtube.com/watch?v=Lm8xCYZjB-M, a ne https://www.youtube.com/embedLm8xCYZjB-M.