Pronalaženje skrivenog znanja – Projektni zadatak za 2019. godinu

Projektni zadatak se sastoji iz pet celina na kojima se može ostvariti ukupno 60 poena. Zadaci se odnose na prikupljanje podataka, njihovu analizu, vizuelizaciju i implementaciju algoritama mašinskog učenja.

Zadatak 1: Prikupljanje podataka

Realizovati veb indekser (eng. *web crawler*) sa veb parserom (eng. *web scraper*), koji prikuplja podatke sa muzičkog sajta www.discogs.com. Baza sajta *Discogs* sadrži veliku količinu podataka o muzičkim albumima na celom svetu, kao i mnogo različitih detalja o tim albumima, kao i njihovim autorima. Baza obuhvata više od 12 miliona zapisa. Za potrebe ovog zadatka potrebno je da prikupite sve podatke iz Srbije - oko 12 hiljada zapisa, i podatke iz vremena bivše Jugoslavije (Yugoslavia) – oko 56 hiljada zapisa. Sve prikupljene podatke uneti u relacionu bazu podataka, koju treba da napravite u *MySQL* ili *PostgreSQL*.

Šta je veb indekser?

Cilj veb indeksera je da se poveže na određenu veb stranu i da preuzme njen sadržaj. Parsiranjem date strane možemo naći linkove, koji vode na neke druge strane, na koje veb-indekser ponovo može da uđe i da ponovi celu proceduru. Pored otkrivanja linkova, parser može da prepozna i druge sadržaje koje veb strana ima. U vašu bazu treba da prikupite informacije o albumima, autorima i drugim ulogama na albumu, imena žanrova, stilova, zemlje porekla.

Implementaciju veb-indeksera možete raditi u programskim jezicima: C, C++, C#, Java, Python, NodeJS ili PHP. Dozvoljeno je i korišćenje i prilagođavanje neke od postojećih implementacija otvorenog koda: crawler4j, Heritrix, Nutch, Scrapy, PHP-Crawler, itd.

Šta je veb parser?

Uloga veb parsera je da otkrije potreban sadržaj sa primljenih veb strana. Pri tome potrebno je odrediti značenje sadržaja kako bi se baza podataka popunjavala tačnim podacima. Najčešće tehnike koje se koriste pri implementaciji veb parsera su: HTML parser, DOM parser, tehnika regularnih izraza koji izdvajaju potreban sadržaj i tehnika prepoznavanja semantičkih anotacija. Za potrebe veb parsiranja takođe možete koristiti neku od postojećih implementacija (npr. biblioteka jsoup – parsira veb stranu kao stablo elemenata).

Kao rezultat zadatka 1 treba da prikažete realizovanu relacionu bazu podataka popunjenu traženim muzičkim zapisima i da priložite implementacije koje su korišćene za dohvatanje podataka. Podaci treba da budu preuzeti u konačnom vremenskom intervalu (~ 120-180 min).

Zadatak 2: Analiza podataka

Iz navedenih zapisa ubačenih u bazu podataka (iz zadatka 1), potrebno je uraditi sledeće:

a) izlistati koliko zapisa pripada svakom od žanrova
b) izlistati koliko zapisa pripada svakom od stilova
c) prikazati rang listu prvih 20 albuma koji imaju najveći broj izdatih verzija (više albuma može deliti jedno mesto na rang listi, pa konačan broj albuma na listi može biti i veći od 20)

d) prikazati prvih 100 osoba koje imaju:
   - najveći generalni rejting u pesmama (*Credits*)
   - najviše učešća kao vokal (*Vocals*)
   - najviše napisanih pesama – aranžman, reči teksta, muzika (*po kategorijama: Arranged by, Lyrics by, Music by*)

e) prikazati prvih 100 pesama koje se nalaze na najviše albuma, i osim broja albuma (COUNT), uz svaku pesmu napisati podatke o tim albumima (*Format*, *Country*, *Year*/*Relased*, *Genre*, *Style*)

f) sve grupe i pojedinačne izvođače analiziranih pesama koje imaju veb sajt (popunjeno polje *Sites*), u formatu: Naziv izvođača, Sajt

Kao rezultat zadatka 2 treba priložiti bazu podataka (zadatak 1), realizovane upite i generisane rezultate.

Zadatak 3: Vizuelizacija podataka

Iz navedenih zapisa ubačenih u bazu podataka (iz zadatka 1), potrebno je vizuelizovati sledeće podatke:

a) 6 najzastupljenijih žanrova i broj albuma koji su izdati u svakom od tih žanrova

b) Broj pesama prema trajanju (do 90 sekundi, 91-180, 181-240, 241-300, 301-360, 361 sekunda i više)

c) Broj albuma po dekadama (1950-1959, 1960-1969, 1970-1979, 1980-1989, 1990-1999, 2000-2009, 2010-2019)

d) Broj (i procentualni odnos) albuma čiji su naslovi napisani ćiriličnim pismom i latiničnim pismom

e) Broj (i procentualni odnos) albuma koji pripadaju:
   - samo jednom žanru,
   - tačno 2 žanra,
   - tačno 3 žanra,
   - 4 ili više žanrova.

Kao rezultat zadatka 3 treba priložiti bazu podataka (zadatak 1), realizovane upite i generisane rezultate u vidu grafikona (*charts*). Za grafikone možete koristiti bilo koji alat / implementaciju.

Zadatak 4: Implementacija algoritma k-Means

Realizovati malu aplikaciju koja iz navedenih zapisa primenom metoda K srednjih vrednosti (eng. *K-means*) nenadgledanog učenja, na osnovu ulaznih podataka, daje rezultat izvršavanja tog algoritma. Na osnovu ulaznog parametra K, koji može da postavi korisnik, grupisati sve ulazne podatke u K klastera. Svaki ulaz mora pripadati jednom klasteru. Podatke analizirati prema nekoliko različitih ulaza (žanrovi, stilovi, godine izdanja,...), a kroz aplikaciju korisnik treba da ima mogućnost odabira ulaza i odabira broja klastera.

Zadatak 5: Implementacija nekog drugog algoritma (po želji)

U istoj aplikaciji, primeniti još jedan algoritam (po želji) iz grupe nenadgledanog učenja.

Kao rezultat zadataka 4 i 5 treba priložiti programski kod realizovane aplikacije.