

Prikupljanje i priprema podataka

Prikupljanje i parsiranje podataka provedeno je kroz nekoliko osnovnih koraka.

- Pristup resursima
- Traženje svih lokacija o dostupnim resursima
- Spremanje podataka o komentarima u direktorije po datotekama s imenima piva
- Parsiranje podataka
- Kreiranje struktura
- Baza podatak

Svi moduli potrebni za navedene akcije implementirani su u pythonu korištenjem standardnih paketa i alata besplatno dostupnih na internetu. Više o paketima i čemu služe te o strukturi direktorija i pozicijama navedenih skripti unutar aplikacija možete pronaći u aplikaciji pod sustavnom wiki.

Pristup resursima i spremanje

Sve kritike koje čine resurse aplikacije preuzete su sa Beeradvocate internet servisa. Beeradvocate pruža pregled kroz tri osnovne kategorije za koje su kreirani parseri koji prikupljaju bitne informacije o korisnicima i pivama koje se navode u dotičnim kategorijama. Konkretno radi se o kategorijama : **najboljih 250, najgorih 100 i najkomentiranih 250**. Skripta **beeradvocate.py** prikuplja podatke tako da za svaku kategoriju, od navedenih, isparsira ime, stil i mjesto podrijetla te njihove web adrese. Nakon izdvajanja navedenih podataka, skripta s presloškom za izdvajanje komentara provjerava stranice svakog piva te izdvaja podatke o korisnicima i njihovim ocjenjenim komentarima. Konačno kreiraju se direktoriji za svaku od navedenih kategorija gdje se spremaju svi komentari za pivo iz kategorije u txt datoteku pod nazivom koje odgovara imenu piva. Bez smanjenja općenitosti možemo kao id koristiti imena piva jer se radi o jakim svjetskim brendovima i sva imena su zaštićena zakonom o autorskim pravima. Ovo saznanje u kasnijim fazama uvelike olakšava analizu i filtriranje podataka.

Parsiranje i strukture

Prilikom spremanja podataka u datoteke posvećena je pažnja u kojem formatu i po kojem pravilu su spremjeni komentri. Strukturiranje podataka uvelike je olakšalo parsiranje koje je provedeno kroz nekoliko jednostavnih regularnih izraza. Kako je aplikacija zamišljena kao portabilno web rješenje, ali istodobno treba pružati najbrži mogući odziv u slučaju potrebe za potpunim pokretanjem sustava, parsiranje se može vidjeti u nekoliko različitih formata. Formati su namijenjeni za web rješenje i za shell-script pokretanje. Nakon parsiranja podaci su spremjeni u pogodne strukture za daljnju analizu i integraciju s bazom podataka. Parser sa regularnim izrazima nalazi se u **BAClass.py**. Za spremanje kreiranih struktura koristi se python modul pickle.

Spremanje u bazu

Za potrebe aplikacije kreirana je baza podataka. Baza sadrži jednostavne tablice u kojima su spremljene sve pive koje se pojavljuju u resursima koje će biti dostupni korisnicima aplikacije. Tako naprimjer tablica `podaci_beer` sadrži podatke o pivama i njihovim općenitim ocjenama. Sve detalje oko strukture baze možete pronaći u skripti **`models.py`** koja je smještena u direktoriju `podaci`. Skripta **`storing_data.py`** brine se o pravilnom prebacivanju podataka iz struktura u bazu.

U sklopu dokumentacije opisan je proces obnove podataka i skripte koje koriste iste metode ali na način da se prilagode bazi podataka. Detaljnije o obnovi u cjelini o obnovi podataka.