Drugi projektni zadatak

Algoritmi i strukture podataka

Verzija 26.05.

EdgeRank algoritam

Implementirati sistem koja simulira EdgeRank algoritam. EdgeRank algoritam je razvila kompanija Facebook u cilju utvrđivanja koje objave (post-ovi) bi trebalo da se prikažu korisniku (u tzv. News feed-u).

Na konferenciji Facebook-a, 2010. godine, na osnovu faktora koji se uzimaju u obzir prilikom rangiranja objava, EdgeRank je opisan kao:

$$\sum_{edges\ e} u_e w_e d_e$$

Gde je:

u_e korisnikova sklonost

w_e popularnost, odnosno procena sadržaja

d_e vremenski baziran parametar raspada.

Korisnikova sklonost

Korisnikova sklonost se računa na osnovu akcija koje je korisnik prethodno preduzimao u kontaktu sa autorom sadržaja. Ovaj rezultat se faktoriše sa:

1) Jačinom akcije (interakcije). Pod akcijom se smatra reagovanje na sadržaj, komentarisanje i deljenje sadržaja, kao i prijateljstvo sa autorom sadržaja. Svaka interakcija ima određen stepen težine i važnosti, na primer, reagovanje ima manju važnost od deljenja sadržaja.
2) Vremenom koje je prošlo od akcije. Manje se uzimaju u obzir interakcije koje su davno izvršene od skorijih.

Na korisnikovu sklonost ka autoru sadržaja utiču, u manjoj meri, i sklonosti njegovih prijatelja kao i sklonosti prijatelja prijatelja.

Sklonost korisnika nije simetrična relacija. Korisnik A ne mora da ima istu sklonost ka korisniku B kao korisnik B ka korisniku A.

Procena sadržaja

Sadržaj koji je izazvao puno pažnje drugih ljudi (u vidu reakcija, komentarisanja itd.) je verovatno zanimljiviji za širi krug ljudi.

Vremenski baziran parametar raspada

Stare objave gube na relevantnosti. Novije objave se prikazuju sa većom verovatnoćom, a kako vreme prolazi, smanjuje se verovatnoća da će objava biti prikazana.

Pretraga

Korisnik može da pretražuje objave unoseći jednu ili više reči razdvojene razmakom. Na rangiranje rezultata pretrage, osim učestalosti pojavljivanja reči u objavi, utiče i releventnosti sadržaja koje se izvodi iz EdgeRank algoritma.

Pod rezultatom pretrage se smatra niz od maksimalno 10 objava.

Za maksimalnih 10 poena:

- Implementirati EdgeRank algoritam i odabir objava za ulogovanog korisnika u jednostavnijem obliku. Formirati proizvoljnu strukturu podataka za reprezentaciju interakcija. Uzeti da sva tri faktora (sklonost, popularnost, vremenska komponenta) jednako utiču na rangiranje rezultata. Uzeti da na korisnikovu sklonost utiče samo vrsta reakcije (bez vremenske komponente). Nije potrebno uzeti u obzir reakcije prijatelja niti prijatelja prijatelja.
- Implementirati rangiranje rezultata pretrage tako da na rang rezultata utiče broj pojavljivanja traženih reči u objavi uz korišćenje proizvoljnih struktura podataka.
- Ukoliko korisnik unosi upit sastavljen od više reči, rangiranje objave po svakoj pojedinačnoj reči utiče na sveukupno rangiranje određene objave. U ovom slučaju, ne treba insistirati na prisustvu svake od reči u rezultatima, ali bi trebalo bolje rangirati rezultate u kojima se pojavljuju sve reči.

Za više od 10 poena:

- Implementacija EdgeRank algoritma. (5 poena)
- Za organizovanje korisnika za EdgeRank koristiti graf. (2 poena)
- Kod rangiranja rezultata pretrage, osim broja pojavljivanja traženih reči u objavi, treba utiče i rezultat EdgeRank algoritma (2 poena)
- Za efikasnu pretragu reči u objavi koristiti strukturu podataka Trie. (3 poena)
- Serijalizacija: U cilju efikasnijeg izvršavanja operacija, omogućiti serijalizaciju formiranih struktura podataka kako se ne bi trošilo vreme za njihovo rekreiranje prilikom svakog pokretanja. (2 poena)

Za više od 24 poena:

- (2 poena) Autocomplete: Odabirom ove opcije, korisniku se nudi nekoliko popularnih završetaka zadatog upita, npr. ako korisnik unese dan* ponuđene opcije mogu biti Danette, Dane ili Daniel (takođe u case insensitive režimu).

- (2 poena) Pretraga fraza. Fraza je tekst za pretragu koji se navodi pod navodnicima. U rezultatima se prikazuju (uz rangiranje) objave u kojima se navedeni delovi fraze pojavljuju uzastopno, u istom redosledu.

U cilju podrške traženim operacijama, obezbediti konzolni meni. Korisnik može da se uloguje, posle čega mu se nude opcije da pregleda objave kao i da započne pretragu. Posebne vrste pretrage ne treba da budu dodatne opcije u meniju već korisnik samim načinom unosa sugeriše koja vrsta pretrage treba da se izvrši (uvođenjem navodnika ili zvezdice, navođenjem više reči...).

Dodatna objašnjenja

Set objava koje se prikazuju korisniku se izračunava dinamički, u trenutku kada se korisnik uloguje.

Vrste interakcija: reakcija, komentar, deljenje
Vrste reakcije: likes, loves, wows, hahas, sads, angrys, special
Podržati postojanje samo osobe kao aktere sistema (bez grupa, stranica itd.).

Set podataka

Za lakši razvoj, dati su fajlovi sa test podacima.
Svaki fajl u prvom redu ima objašnjenje o sadržaju svake od kolona.

statuses.csv – sadrži podatke o objavama. Svaka objava sadrži identifikator objave (status_id), tekst objave (status_message), naziv veze (link_name), tip objave (status_type, može biti slika, link itd.), veza (status_link), datum i vreme objavljivanja (status_published), autor (author), broj reakcija (num_reactions), broj komentara (num_comments), broj deljenja (num_shares), broj sviđanja (num_likes), broj reakcija “voli” (num_loves), broj “wow” reakcija (num_wows), broj “haha” reakcija (num_hahas), broj tužnih reakcija (num_sads), broj besnih reakcija (num_angrys), broj posebnih reakcija (num_special).

comments.csv – sadrži podatke o komentarima na objave. Svaki komentar sadrži identifikator komentara (comment_id), identifikator statusa kome pripada (status_id), identifikator roditeljskog komentara (parent_id, u slučaju da se komentariše komentar), poruka komentara (comment_message), autor (comment_author), datum i vreme objavljivanja komentara (comment_published), broj reakcija (num_reactions), broj sviđanja (num_likes), broj reakcija “voli” (num_loves), broj “wow” reakcija (num_wows), broj “haha” reakcija (num_hahas), broj tužnih reakcija (num_sads), broj besnih reakcija (num_angrys), broj posebnih reakcija (num_special).
*** EdgeRank algoritam ne treba da uzme u obzir reakcije na komentare

friends.csv – sadrži podatke o prijateljstvima korisnika. Svaka linija fajla sadrži ime korisnika (person) ,broj prijatelja (number_of_friends), spisak prijatelja (friends).

reactions.csv – sadrži podatke o reakcijama korisnika. Svaka reakcija sadrži identifikator statusa na koji se odnosi (status_id), vrstu reakcije (type_of_reaction, od dostupnih vrsta), korisnika koji je reagovao (reactor), datum i vreme reagovanja (reacted).

shares.csv – sadrži podatke o deljenju objava korisnika. Svako deljenje sadrži identifikator statusa na koji se odnosi (status_id), korisnika koji je podelio (sharer), datum i vreme deljenja (shared).

Pored test podataka, dat je fajl ***parse_files.py*** koji bi trebalo da olakša parsiranje sadržaja fajlova ***comments.csv*** i ***statuses.csv***. Upotreba funkcija iz ovog fajla nije obavezna.

Druga unapređenja funkcionalnosti koje predmetni profesor ili asistent odobre mogu doneti do 3 dodatna poena.

Opšte informacije o slanju i upload-u zadatka:

- Zadatak nosi 25 poena.
- Smestiti sve fajlove zadatka u **folder** pod nazivom projekat2_sv_XX_YYYY gde se umesto XX_YYYY navodi broj indeksa - broj upisa i godina upisa (primer: projekat2_sv_02_2022)
- Ubaciti fajl u **zip** arhivu i nazvati je isto kao i zadatak (projekat2_sv_XX_YYYY.zip)
- Uploadovati zip arhivu kao assignment na enastavu.
- Ukoliko bude problema sa uploadom, možete u predviđenom roku poslati zip na email adresu predmetnog asistenta.

Rok za slanje arhive je 25.06.