UNIVERZITET U BEOGRADU

ELEKTROTEHNIČKI FAKULTET

ANALIZA SOCIJALNIH MREŽA

Projektni zadatak

Verzija 1.0

Predmetni nastavnici:

dr Marko Mišić, docent

dr Jelica Protić, redovni profesor

Predmetni saradnik:

Predrag Obradović, asistent

Školska godina:

2021/2022.

Beograd, decembar 2021.

SADRŽAJ

1. UVOD

U okviru ovog dokumenta su data uputstva za izradu projektnog zadatka na predmetu Analiza socijalnih mreža (13M111ASM) u školskoj 2021/2022. godini. Studenti treba da pažljivo pročitaju ovo uputstvo pre izrade projektnog zadatka. Studenti projektni zadatak rade **samostalno** ili **u paru**.

2. CILJ

Cilj projektnog zadatka na predmetu Analiza socijalnih mreža je praktična primena stečenog teorijskog znanja iz predmeta na primeru jednog konkretnog istraživačkog problema. Kroz zadati istraživački problem, studenti treba da izvrše prikupljanje, obradu i preliminarnu analizu primarnog (sirovog) skupa podataka, izdvoje neophodne podatke i modeliraju problem mrežom odgovarajućeg tipa. Modeliranu mrežu treba da analiziraju alatima za obradu socijalnih mreža po izboru i izvrše vizuelizaciju mreže. Dobijene rezultate analize treba na odgovarajući način interpretirati u skladu sa postavljenim istraživačkim pitanjima i predstaviti u obliku izveštaja.

3. POSTAVLJENI PROBLEM

U okviru ove sekcije je dat predlog projektnog zadatka za tekuću školsku godinu. Studenti mogu predložiti predmetnom nastavniku drugu temu. U tom slučaju, poželjno je priložiti i deo skupa podataka koji bi se analizirao, kako bi student na adekvatan način u saradnji sa nastavnikom postavio ciljeve istraživanja i istraživačka pitanja.

3.1. Analiza potadaka sa društvene mreže *Reddit*

Redit (engl. *Reddit*) je veb stranica organizovana u vidu foruma. Baziran je na postavljanju, komentarisanju i ocenjivanju veb sadržaja. Korisnici mogu biti registrovani, ali nije neophodno za pretraživanje veb stranice. Korisnici su tipično anonimni, odnosno mogu izabrati proizvoljno korisničko ime. Korisnici na sajt postavljaju sadržaj u vidu teksta, linkova i fotografija, a drugi korisnici taj sadržaj mogu oceniti i postavljati komentare na njega. Sadržaj je podeljen u sabredite (engl. *subreddit*), koji su najčeće tematski organzovani i pokrivaju teme poput novosti, nauke,

tehnologije, video-igara, knjiga, filmova, muzike, hrane, fotografija, kao i raznovrsan, drugi sadržaj koji internet nudi. Nekada su i geografski organizovani, pa, recimo, postoji i *r/serbia* sa materijalima, novostima i sadržajima interesantnim korisnicima sa našeg podneblja.

Redit postoji od 2005. godine. Trenutno beleži oko 52 miliona dnevno aktivnih korisnika, dok na mesečnom nivou broj aktivnih korisnika dostiže preko 430 miliona. Prosečan mesečni broj poseta u 2021. godini (uključujući i neregistrovane korisnike) iznosi oko 1.6 milijardi. Procenjeno je da oko 25% odraslih osoba u Sjedinjenim Američkim Državama koristi Redit.

Sabrediti se smatraju aktivnim ako se na njima zabeleži barem 5 objava ili komentara dnevno. Trenutno postoji preko 100 000 aktivnih sabredita, od kojih najveći broje nekoliko desetina miliona članova.

Sistem glasanja je zasnovan na principu „za“ i „protiv“ (eng. *upvote* i *downvote*) putem kojeg objava dobija svoju ocenu i popularnost, što je broj glasova za veći, veća je i šansa da će objavu više korisnika pročitati. Korisnici unutar objava ostavljaju komentare i započinju razgovore sa drugim korisnicima.

Na profilu korisnika se koristi sistem karme. Korisnici karmu prikupljaju tako što njihove objave i komentare drugi korisnici ocenjuju. Što više glasova „za“ korisnici dobiju, to imaju više karme. Karma služi kao vid reputacije korisnika na Reditu.

3.2. Skup podataka za analizu

U okviru predloženog projektnog zadatka je potrebno analizirati objave (eng. *submssion*) i komentare (eng. *comment*) postavljene na sajtu Reddit u toku 2008. godine, koja je izabrana zbog relativno malog obima podataka i događaja koji su se u njoj desili, kao što je svetska ekonomska kriza. Primarni skup podataka (eng. *primary dataset*) za analizu dostupan je u vidu odgovarajućih fajlova u CSV formatu u arhivi koja je priložena uz tekst projektnog zadatka.

Originalni podaci su preuzeti sa repozitorijumu koji održava Džejson M. Baumgartner [1]. Sastoje se od kompresovanih JSON datoteka koje sadrže podatke na mesečnom nivou o objavama i komentarima postavljenim u toku 2008. godine. Preuzeti su 14.12.2021. godine. Kompresovane datoteke koje sadrže podatke o komentarima imaju nazive oblika *RC_yyyy_mm.bz2*, gde *yyyy* predstavlja godinu u opsegu od 2005. do 2021. a sa *mm* je označen redni broj meseca od 01 do 12. Datoteke *RS_yyyy_mm.bz2* sa istom nomenklaturom sadrže objave iz istog perioda. Primarni skup podataka je nastao dekompresijom i transformacijom originalnog skupa podataka iz JSON u CSV

format. Studenti po želji mogu koristiti i originalni skup podataka za izradu projektnog zadatka, ali treba imati u vidu da su originalni podaci značajno memorijski zahtevniji.

Primarni skup podataka sastoji se od većeg broja CSV fajlova koji sadrže podatke o objavama (u folderu *submissions_2018_asm*) i komentarima (u folderu *comments_2018_asm*) u okviru kojih su objave, odnosno komentari sortirani po datumu postavljanja. U skupu podataka se nalaze podaci o 2519853 objave i 7242871 komentaru u navedenom periodu. Uklonjene su neke od kolona koje za analizu nisu neophodne, a opis preostalih kolona na engleskom jeziku dat je u tabelama u nastavku i preuzet je iz rada koji opisuje *PushShift Reddit* skup podataka [2]. Između ostalog, uklonjene su kolone sa tekstualnim sadržajem objave, odnosno komentara, jer najviše doprinose memorijskom zauzeću, a za analize u okviru ovog projektnog zadatka nisu od koristi.

Field	**Description**
id	The submission's identifier, e.g., "5lcgjh" (String).
url	The URL that the submission is posting. This is the same with the permalink in cases where the submission is a self post. E.g., "https://www.reddit.com/r/AskReddit/
permalink	Relative URL of the permanent link that points to this specific submission, e.g., "/r/AskReddit/comments/5lcgj9/what did you think of the ending of rogue one/" (String).
author	The account name of the poster, e.g., "example username" (String).
created utc	UNIX timestamp referring to the time of the submission's creation, e.g., 1483228803 (Integer).
subreddit	Name of the subreddit that the submission is posted. Note that it excludes the prefix /r/. E.g., 'AskReddit' (String).
subreddit id	The identifier of the subreddit, e.g., "t5 2qh1i" (String).
num comments	The number of comments associated with this submission, e.g., 7 (Integer).
score	The score that the submission has accumulated. The score is the number of upvotes minus the number of downvotes. E.g., 5 (Integer). **NB:** Reddit fuzzes the real score to prevent spam bots.
over 18	Flag that indicates whether the submission is Not-Safe-For-Work, e.g., false (Boolean).
distinguished	Flag to determine whether the submission is posted by moderators or admins. "null" means not distinguished (String).
domain	The domain of the submission, e.g., self.AskReddit (String).
locked	Flag indicating whether the submission is currently closed to new comments, e.g., false (Boolean).
hide score	Flag indicating if the submission's score is hidden, e.g., false (Boolean).

Tabela 1: Opis kolona CSV fajlova sa podacima o objavama

Field	Description
id	The comment's identifier, e.g., "dbumnq8" (String).
author	The account name of the poster, e.g., "example username" (String).
link id	Identifier of the submission that this comment is in, e.g., "t3 5l954r" (String).
parent id	Identifier of the parent of this comment, might be the identifier of the submission if it is top-level comment or the identifier of another comment, e.g., "t1 dbu5bpp" (String).
created utc	UNIX timestamp that refers to the time of the submission's creation, e.g., 1483228803 (Integer).
subreddit	Name of the subreddit that the comment is posted. Note that it excludes the prefix /r/. E.g., 'AskRed-dit' (String).
subreddit id	The identifier of the subreddit where the comment is posted, e.g., "t5 2qh1i" (String).
score	The score of the comment. The score is the number of upvotes minus the number of downvotes. Note that Reddit fuzzes the real score to prevent spam bots. E.g., 5 (Integer).
distinguished	Flag to determine whether the comment is made by the moderators or admins. "null" means not distinguished (String).
gilded	The number of times this comment received Reddit gold, e.g., 0 (Integer).
controversiality	Number that indicates whether the comment is controversial, e.g., 0 (Integer).

Tabela 2: Opis kolona CSV fajlova sa podacima o komentarima

Na osnovu primarnog skupa podataka treba formirati sekundarni skup podataka (eng. *secondary dataset*) koji predstavlja prečišćenu verziju podataka za analizu. Prečišćavanje izvršiti prema potrebama zadatka i ciljevima istraživanja. Prilikom prečišćavanja se mogu izostaviti svi nepotrebni podaci.

3.3. Modelovanje mreže

Sekundarni skup podataka je potrebno iskoristiti za modelovanje odgovarajućih socijalnih ili drugih mreža. Prilikom modelovanja mreža implementirati odgovarajući tip mreže (usmerena, neusmerena, težinska i sl.) u skladu sa postavljenim istraživačkim pitanjima i ciljevima.

3.3.1. Mreže sabredita SNet, SNetF i SNetT

Potrebno je formirati tri mreže koje modeluju interakcije između sabredita na osnovu aktivnosti korisnika. U okviru mreža sabredita, sabrediti treba da predstavljaju čvorove mreže, a

vezu između dva čvora treba uspostaviti ukoliko postoje korisnici koji su bili aktivni na oba sabredita. Težinu grane odrediti agregacijom brojanjem ili putem modela zastarivanja veze (grane fomirane pre *n* meseci doprinose sa težinom α^n, gde je α realan broj između 0 i 1).

Mreža *SNet* (od engleskog *subreddit network*) se dobija na osnovu kompletnih podataka, sadrži sve sabredite i grane formirane na osnovu svih komentara i objava. Ovo je osnovna mreža koja će se koristiti za analizu sistema.

Prilikom konstrukcije mreže *SNet*, a nakon izvršene agregacije paralelnih grana, potrebno je izvršiti filtriranje grana male težine i na taj način konstruisati filtriranu mrežu sabredita *SNetF* (od engleskog *filtered subreddit network*). Za potrebe filtriranja odrediti i vizuelizovati raspodelu težine grane, pa odbaciti grane čija je težina ispod razumno izabranog praga *w_threshold*. Način izbora *w_threshold* neophodno je jasno dokumentovati.

Mreža *SNetT* (od engleskog *targeted subreddit network*) se sastoji od podataka o ciljanoj grupi sabredita čija je tematika bliska temi ekonomske krize. Treba je konstruisati kao indukovani podgraf *SNet* mreže čiji su čvorovi sabrediti u skupu: {*reddit.com, pics, worldnews, programming, business, politics, obama, science, technology, WTF, AskReddit, netsec, philosophy, videos, offbeat, funny, entertainment, linux, geek, gaming, comics, gadgets, nsfw, news, environment, atheism, canada, math, Economics, scifi, bestof, cogsci, joel, Health, guns, photography, software, history, ideas*}. Skup sabredita je formiran na osnovu pretrage ključnih reči u objavama i komentarima tokom 2008. godine u vezi sa svetskom ekonomskom krizom prikupljenih iz različitih izvora [3][4][5][6].

3.3.2. Mreža korisnika UserNet

Mreža korisnika treba da modeluje interakcije između korisnika platforme *Reddit*. Neophodno je voditi evidenciju o tome ko je komentarisao čiju objavu ili komentar, pri čemu se komentarisanje na objavu i na komentar mogu, ali ne moraju smatrati istim tipom interakcije. Agregacija se može izvršiti slično kao u slučaju mreža sabredita.

3.4. Istraživačka pitanja i ciljevi

Prilikom obrade primarnog i sekundarnog skupa podataka pogodno je kao smernice koristiti prethodno definisana istraživačka pitanja. U okviru ove sekcije je postavljen jedan broj takvih pitanja, a studenti treba da, nakon što odgovore na ova pitanja, na osnovu analize problema i samih

podataka definišu dodatna pitanja ili specijalizuju navedena čime mogu bliže usmeriti samu analizu. Odgovore na pitanja treba dati u formi specificiranoj u poglavlju 4.

Pitanja su grupisana u kategorije po tematici koju obrađuju i tehnikama analize koje se u njima sprovode. Svako od pitanja se odnosi na svaku od konstruisanih mreže, ukoliko je za nju smisleno i primenljivo. Studentima se predlaže da prvo sve analize sprovedu za jednu od mreža, pa zatim pređu na sledeću i da svoje odgovore na taj način strukturiraju i izlože u izveštaju o projektu.

Da bi se odgovorilo na postavljena pitanja, potrebno je primeniti odgovarajuće mere i metode za analizu mreže ili statističke metode. Mreže bi trebalo karakterisati kako kroz osnovna svojstva mreže, tako i kroz složenije mere centralnosti i metode za detekciju komuna. Mere i metode izabrati prema adekvatnosti spram postavljenog problema. Tamo gde se očekuje odgovor u obliku neke vrste rangiranja, navesti listu od 5 do 10 najrelevantnijih rezultata.

3.4.1. Statistička obrada podataka [5 poena]

1) Koliko postoji različitih sabredita koji se pojavljuju u posmatranom periodu? Koji su najvažniji po broju korisnika, a koji po broju komentara?
2) Kakav je prosečan broj zabeleženih korisnika aktivnih u posmatranom periodu po sabreditu? Korisnik se smatra aktivnim na sabreditu ako je zabeležen barem jedan komentar ili objava tog korisnika.
3) Ko su korisnici sa najvećim brojem objava, a ko korisnici sa najvećim brojem komentara?
4) Koji korisnici su aktivni na najvećem broju sabredita? Na koliko su sabredita aktivni?
5) Kako su korelisani brojevi objava i brojevi komentara korisnika? Odrediti Pirsonov koeficijent korelacije i izvršiti vizuelzaciju.
6) Koje objave poseduju najveći broj komentara i na kojim su sabreditima postavljene? Prikazati podatke o tim objavama, uključujući to na kojem su sabreditu postavljene i šta im je sadržaj (ako je polje objave "*over 18*" postavljeno na *false*).

3.4.2. Osnovna karakterizacija modelovanih mreža [10 poena]

7) Kolika je gustina mreže?
8) Kolike su prosečne distance u okviru mreže i dijametar mreže?
9) U kojoj meri je mreža povezana i centralizovana? Navesti broj i veličine povezanih komponenata i proceniti da li postoji gigantska komponenta.

10) Koliki je prosečni, a koliki globalni koeficijent klasterizacije mreže? Kakva je raspodela lokalnog koeficijenta klasterizacije njenih čvorova? Da li je klasterisanje izraženo ili ne? Odgovor dati upoređivanjem sa slučajno generisanom *Erdos-Renyi* mrežom istih dimenzija.
11) Na osnovu odgovora na pitanja 8 i 10, proceniti da li mreža iskazuje osobine malog sveta.
12) Izvršiti asortativnu analizu po stepenu čvora i dati odgovor da li je izraženo asortativno mešanje. U slučaju da je mreža usmerena, analizu izvršiti i po ulaznom i po izlaznom stepenu čvora. Priložiti i vizuelizaciju.
13) Da li mreža ispoljava fenomen kluba bogatih (eng. *rich club phenomenon*)?
14) Kakva je distribucija čvorova po stepenu i da li prati *power law* raspodelu?
15) Odrediti najvažnije habove i autoritete u mreži. Kako su oni raspoređeni i ugrađeni u mrežu, da li su na periferiji ili u jezgru mreže?

3.4.3. Analiza mera centralnosti [10 poena]

16) Sprovesti analize centralnosti po stepenu, bliskosti i relacionoj centralnosti. Dati pregled najvažnijih aktera po svakoj od njih.
17) Ko su najvažniji akteri po centralnosti po sopstvenom vektoru? Šta nam to govori o njima?
18) Rangirati čvorove po Kacovoj centralnosti (eng. *Katz centrality*) sa varijacijom parametara. Pri računanju Kacove centralnosti, eksperimentisati sa dodeljivanjem drugačije vrednosti parametra β za sabredit koji se u priloženim CSV fajlovima identifikuje verdnošću kolone *subreddit* jednakom *reddit.com*. Dati pregled najvažnijih aktera u slučaju da je β isto za sve sabredite i u slučaju da je β navedenog sabredita značajno veće.
19) Na osnovu prethodna tri pitanja predložiti i konstruisati heuristiku (kompozitnu meru centralnosti) za pronalaženje najvažnijih aktera i pronaći ih. Obratiti pažnju na tip mreže koji se analizira (usmerena ili neusmerena) i, shodno tome, prilagoditi koliko različite mrežne metrike utiču na heuristiku.

3.4.4. Detekcija komuna [10 poena]

20) Ako veličina mreže dozvoljava, spektralnom analizom ili analizom dendrograma proceniti potencijalne kandidate za broj komuna u mreži.

21) Sprovesti klasterisanje Luvenskom metodom (maksimizacijom modularnosti) u alatu Gephi za tri različite vrednosti parametra rezolucije. Konstruisati vizuelizacije i diskutovati izbor parametra rezolucije na dobijeno klasterisanje (broj i veličina klastera).

22) Koje zajednice (komune) se mogu uočiti prilikom analize mreže? Da li postoji neko objašnjenje za detektovane komune?

23) Ko su akteri koji se mogu okarakterisati kao ključni brokeri (mostovi) u mreži? Šta ih čini brokerima?

3.4.5. Poređenje SNet i SNetT mreža [5 poena]

24) Uporediti karakteristike *SNet* i *SNetT* mreža. Komentarisati potencijalne razlike i proceniti da li su sabrediti iz *SNetT* aktivniji i bolje povezani od ostatka mreže.

25) Kako su raspoređeni čvorovi iz *SNetT* u okviru *SNet* mreže? Da li pripadaju jezgru ili periferiji ili su mešovito raspoređeni?

3.5. Preporučene metode i alati

Za analizu modelirane socijalne mreže se preporučuje korišćenje programskih jezika Python (NetworkX biblioteka) i R (*sna* i *igraph* paketi) ili softverskih alata Gephi, UCINET, ili Pajek. Obrada primarnog skupa podataka se može obaviti pomoću MS Excel alata ili pisanjem odgovarajućih skripti u programskom jeziku po izboru. Ukoliko nije moguće drugačije, razrešavanje eventualnih dvosmislenosti u primarnom skupu podataka izvršiti ručno.

Vizuelizacija mreže se može obaviti korišćenjem alata Gephi, NodeXL ili kroz podršku u okviru programskih jezika Python (*matplotlib, graphviz* i *graph-tool* biblioteke) i R (*igraph* paket).

4. REZULTATI

Projektni zadatak se predaje u vidu pisanog izveštaja koji sadrži rezultate sprovedene analize i pisana objašnjenja uočenih fenomena. Uz izveštaj se dostavljaju i odgovarajuće dopunske datoteke, kao što su tabele sa rezultatima analize, izvorni programski kod skripti ili programa korišćenih u analizi, datoteke koje sadrže produkovane vizuelizacije i sl. Potpuno odsustvo dopunskih datoteke koje predstavljaju rezultate rada može povući umanjenje broja poena na projektnom zadatku. Za pisanje izveštaja se može koristiti šablon koji se nalazi u odgovarajućoj sekciji na sajtu predmeta. Preporučeni obim izveštaja je do 10 stranica teksta.

5. PREDAJA, ODBRANA I VREDNOVANJE

Projektni zadatak se predaje elektronskim putem najkasnije do termina ispita u odgovarajućem ispitnom roku na način kako to bude specificirao predmetni nastavnik. Na odbranu je potrebno doneti štampanu verziju izveštaja. Po pravilu, projektni zadatak se brani pred predmetnim nastavnikom ili saradnikom u ispitnom roku u kome student želi da polaže ispit. Ukoliko student želi da brani zadatak u nekom drugom terminu, treba o tome da blagovremeno obavesti predmetnog nastavnika, radi eventualnog dogovora. Ukoliko se projektni zadatak radi u paru, studenti zajedno brane projektni zadatak.

Projektni zadatak nosi 40 poena. Raspodela poena po tematskim oblastima je prikazana u sekciji 3.4. Studenti ne moraju realizovati sve zahteve u okviru projekta pre odbrane. Ne postoji minimalan broj poena koji je potrebno osvojiti na projektnom zadatku da bi se položio ispit.

Poeni sa jednom odbranjenog projektnog zadatka važe jednu školsku godinu. Postoji mogućnost da se dobro urađeni projektni zadaci prošire u završni, master rad. Upit u vezi sa takvom mogućnošću studenti mogu uputiti predmetnom nastavniku ili saradniku.

LITERATURA

[1] Baumgartner, J.M., *Reddit PushShift dumps*, dostupno na: https://files.pushshift.io/reddit/, pristupano: 14.12.2021.

[2] Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). *The Pushshift Reddit Dataset*. Proceedings of the International AAAI Conference on Web and Social Media, 14(1), 830-839. dostupno na https://ojs.aaai.org/index.php/ICWSM/article/view/7347, pristupano: 23.12.2021.

[3] Aigner, E., Aistleitner, M., Gloetzl, F., & Kapeller, J., *The Focus of Academic Economics: Before and after the Crisis* (May 22, 2018). Institute for New Economic Thinking Working Paper Series No. 75, Available at SSRN: https://ssrn.com/abstract=3228774 or http://dx.doi.org/10.2139/ssrn.3228774, pristupano: 24.12.2021.

[4] *Glossary of 2008 financial crisis*, https://www.france24.com/en/20180904-glossary-2008-financial-crisis, pristupano: 24.12.2021.

[5] *Financial crisis glossary - Economic and monetary affairs*, https://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+IM-PRESS+20100414FCS72750+0+DOC+XML+V0//EN, pristupano: 24.12.2021.

[6] *Financial crisis of 2007–2008*, https://en.wikipedia.org/wiki/Financial_crisis_of_2007%E2%80%932008, pristupano: 24.12.2021.