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:

2020/2021.

Beograd, decembar 2020.

SADRŽAJ

1. UVOD

U okviru ovog dokumenta su data uputstva za izradu projektnog zadatka na predmetu Analiza socijalnih mreža (13M111ASM) u školskoj 2020/2021. 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.

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 stanja profesionalnog muškog tenisa u periodu od 2018. do 2020. godine

Tenis je sport sa reketima koji mogu da igraju dva igrača (singl) ili dva para igrača (dubl) jedan protiv drugog. Moderni tenis je počeo da se igra u Engleskoj u 19. veku i brzo se širio među višim staležima u engleskom govornom području. Danas je olimpijski sport.

Teniska sezona traje jedanaest meseci. Turniri su organizovani od strane sledećih organizacija: Međunarodne teniske federacije (eng. *International Tennis Federation* – ITF), Asocijacije teniskih profesionalaca (eng. *Association of Tennis Professionals* – ATP) i Ženske teniske asocijacije (eng.

Women's Tennis Association – WTA). Mogu da se igraju na nekoliko različitih podloga. Postoje tri glavne vrste podloga: zemljana, travnata i tvrda. Profesionalni teniseri i teniserke se rangiraju u zavisnosti od rezultata na ATP i WTA rang listama.

Tenis je individualni sport čiji mečevi su posebno pogodni za mrežno modelovanje. U velikom broju radova se mreže konstruišu tako što igrači predstavljaju čvorove, a karakteristike njihovih međusobnih mečeva formiraju grane [2]. Na tako dobijenim mrežama se ispituju standardne metrike i principi iz teorije mreža. Prednost profesionalnog tenisa kao sporta je duga tradicija vođenja statističkih i drugih informacija o igračima, mečevima, rezultatima, što ih čini vrlo pogodnim za analizu na različite načine. Za razliku od timskih sportova, broj mečeva je značajno veći, pa je samim tim i veća količina informacija dostupnih za analizu.

Igrači tenisa dolaze iz različitih delova sveta i teniskih škola, što utiče na njihove stilove igre i strategije koje koriste tokom mečeva. Takođe, u zavisnosti od faktora, kao štu su iskustvo, godine starost, raspored turnira, bodovi koje brane ili koje mogu da osvoje, podloga terena, geografska lokacija odigravanja turnira i sl., teniseri odlučuju na kojim će turnirima učestvovati. To dalje utiče na mogućnost susretanja pojedinih tenisera u okviru profesionalnih mečeva. Na mogućnost susretanja utiče i žreb turnira koji je, po pravilu, dirigovan na osnovu pozicije tenisera na ATP listi.

Tema projektnog zadatka u tekućoj školskoj godini je kvantitativna i kvalitativna analiza stanja profesionalnog muškog tenisa u singl konkurenciji primenom tehnika za analizu socijalnih i kompleksnih mreža i statističkih metoda. Poslednja dekada predstavlja period koji je obeležila dominacija profesionalnih tenisera poznatijih pod zajedničkim nazivom *Velika trojka* (eng. *Big Three*) [1]: Rodžera Federera, Rafaela Nadala i Novaka Đokovića. U proteklih deceniju i po od poslednjih 70 grend slemova i prvog koji je osvojio Rafael Nadal 2005. na Rolan Garosu, samo je devet trofeja pripalo drugim igračima. Ipak, pomenuti igrači su ušli u zrele igračke godine, a Rodžer Federer je na zalasku karijere. Za potrebe ove analize biće posmatran period od 2018. do 2020. godine. U teniskom svetu, 2018. i 2019. godina se mogu smatrati sasvim uspešnim godinama, dok je 2020. bila obeležena prekidima u okviru sezone zbog zdravstvene situacije usled pandemije virusa COVID-19, što je značajno uticalo na tok sezone i rangiranje na ATP listi.

3.2. Skup podataka za analizu

U okviru ovog projektnog zadatka je potrebno analizirati muške teniske mečeve u singl konkurenciji sa ATP turnirima u periodu od 2018. do 2020. godine. Podaci za analizu (primarni skup podataka) su dostupni u vidu odgovarajućih *Comma Separated Values* (CSV) datoteka u arhivi koja

je priložena uz tekst projektnog zadatka. U okviru arhive se nalaze i kratka uputstva sa opisom skupa podataka.

Podaci su preuzeti sa repozitorijumu koji održava Džef Sakman [3] i čine primarni skup podataka za analizu (eng. *primary dataset*). Podaci su preuzeti 22.12.2020. godine. Primarni skup podataka se sastoji od datoteka koje sadrže podatke o mečevima u singl konkurenciji u periodu od 2018. do 2020. godine, kao i datoteke *atp_players.csv* koja sadrži spisak svih igrača ikada rangiranih na ATP listi, *atp_rankings_10s.csv* koja sadrži podatke o rangiranju aktivnih tenisera na ATP listi u periodu od 2010. do 2019. godine i *atp_rankings_current.csv* koja sadrži podatke o trenutnom rangiranju aktivnih tenisera na ATP listi, tokom 2020. godine. Datoteke koje sadrže podatke o mečevima imaju nazive oblika *atp_matches_yyyy.csv*, gde *yyyy* predstavlja godinu u opsegu od 2018. do 2020.

U skupu podataka se nalaze podaci o 7117 mečeva u navedenom periodu. Podaci se sastoje od kolona čija su objašnjenja data u prilogu [2], [3]:

- Informacije o turniru
  - *tourney_id* – identifikator turnira gde prva četiri karaktera označavaju godinu održavanja.
  - *tourney_name* – naziv turnira.
  - *surface* – tip podloge koji može biti: *Hard* – tvrda, *Clay* – šljaka, *Carpet* – tepih ili *Grass* – trava.
  - *draw_size* – veličina žreba.
  - *tourney_level* – kategorija turnira koja može biti: *O* – olimpijske igre, *G* – grend slem, *PM* – obavezni premijer, *P* – premijer, *F* – završno prvenstvo, *W* – elitni kup, *D* – Fed kup, *I* – internacionalni, *C* – čelendžer, 100, 80, 75, 60, 50, 25, 15, 10 – ITF turniri sa odgvarajućim nagradnim fondom u hiljadama dolara.
  - *tourney_date* – datum turnira u formatu *GGGGMMDD*.
  - *best_of* – broj setova potrebnih za pobedu.
- Informacije o igračima su date za pobednika (prefiks *winner*) i gubitnika (prefiks *loser*)
  - *winner_id/loser_id* – identifikator igrača.
  - *winner_seed/loser_seed* – redni broj nosioca ako je igrač nosilac u žrebu ili nema vrednost.

- *winner_entry/loser_entry* – način ulaska u turnir. Ako je igrač kvalifikovan na osnovu svog ranga nema vrednost. Neke od mogućih vrednosti su: *WC* (eng. Wild Card), *SR* (eng. *Special Ranking*), *SE* (eng. *Special Exempt*), *Q* (eng. *Qualifier*), *LL* (eng. *Lucky Loser*), *JE* (eng. *Junior Exempt*), *A/Alt* (eng. *Alternate*).
- *winner_name/loser_name* – ime igrača.
- *winner_hand/loser_hand* – dominantna ruka igrača. Može imati vrednost *R* za desnu ruku, *L* za levu ili *U* ako je informacija nepoznata.
- *winner_ht/loser_ht* – visina igrača u centimetrima.
- *winner_ioc/loser_ioc* – troslovna oznaka države koju igrač predstavlja.
- *winner_age/loser_age* – godine starosti igrača.
- *winner_rank/loser_rank* – trenutni rang igrača.
- *winner_rank_points/loser_rank_points* – broj poena na rang listi.

- Informacije o meču
  - *score* – rezultat.
  - *round* – kolo turnira kojem pripada meč. Može imati jednu od sledećih vrednosti: *F, SF, QF, R16, R32, R64, R128, Q3, Q2, Q1, RR*.
  - *minutes* – dužina meča u minutima.
  - *match_num* – identifikator meča na nivou godine.
- Performanse igrača su takođe date za oba igrača. One vezane za pobednika imaju prefiks *w*, a one vezane za gubitnika prefiks *l*.
  - *w_ace/l_ace* – broj as udaraca.
  - *w_df/l_df* – broj duplih servis grešaka.
  - *w_svpt/l_svpt* – broj poena na servis igrača.
  - *w_1stIn/l_1stIn* – broj ubačenih prvih servisa.
  - *w_1stWon/l_1stWon* – broj osvojenih poena na prvi servis.
  - *w_2ndWon/l_2ndWon* – broj osvojenih poena na drugi servis.
  - *w_SvGms/l_SvGms* – broj servis gemova igrača.
  - *w_bpSaved/l_bpSaved* – broj brejk lopti koje je igrač spasio.

– *w_bpFaced/l_bpFaced*– broj brejk lopti sa kojima se igrač suočio.

U skupu podataka se nalaze podaci o 54975 igrača ikada rangiranih na ATP listi, odnosno igrača koji su u nekom trenutku za vreme posmatranog vremenskog perioda imali barem 1 poen na ATP listi. Podaci se sastoje od kolona čija su objašnjenja data u prilogu [3]:

- *player_id* – identifikator igrača
- *first_name* – ime igrača.
- *last_name* – prezime igrača.
- *hand* – dominantna ruka igrača. Može imati vrednost *R* za desnu ruku, *L* za levu ili *U* ako je informacija nepoznata.
- *birth_date* – datum turnira u formatu *GGGGMMDD.*
- *country_code* – troslovna oznaka države koju igrač predstavlja.

U skupu podataka se nalaze podaci o rangiranju igrača na ATP listi u navedenom periodu. Podaci se sastoje od kolona čija su objašnjenja data u prilogu [3]:

- *ranking_date* – datum rangiranja u formatu *GGGGMMDD.*
- *rank* – rang na listi.
- *player_id – identifikator igrača.*
- *points* – broj poena na ATP listi za navedeni datum.

Ako više igrača poseduju isti broj poena i, na primer, zauzimaju rangove od 1800 do 1850, treba obratiti pažnju da se za rang svih tih igrača nekada uzima vrednost 1800, u zavisnosti od drugih parametara, odnosno, može postojati više igrača sa istim rangom na ATP listi u istom trenutku.

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 mreža. Potrebno je modelovati mrežu svih tenisera na osnovu njihovih međusobnih susreta za svaku godinu ponaosob i za sve posmatrane godine zajedno, kao i *ego* mreže igrača *Velike trojke*: Rodžera Federera, Rafaela Nadala i Novaka Đokovića. Prilikom modelovanja mreže implementirati

odgovarajući tip mreže (usmerena, neusmerena, težinska i sl.) u skladu sa postavljenim istraživačkim pitanjima i ciljevima. Primarna mreža za analizu treba da bude mreža svih tenisera, a preostale treba iskoristiti u funkciji odgovaranja na određena istraživačka pitanja. Po potrebi se mogu napraviti i analizirati i druge mreže na osnovu zadatog skupa podataka.

U okviru mreže tenisera, igrači treba da predstavljaju čvorove mreže, a vezu između dva čvora treba uspostaviti ukoliko su se dva tenisera susrela na nekom turniru. Mrežu tenisera modelovati za svaku godinu ponaosob, kao i agregirnu mrežu za sve godine.

Potrebno je na adekvatan način modelovati informacije o broju međusobnih susreta i međusobnom odnosu broja pobeda i poraza. Na primer, u okviru pojedinačnih mreža po posmatranim godinama informacija o broju međusobnih susreta se može modelovati težinom grane, dok se u agregiranoj mreži za sve tri godine informacije o susretima po godinama prate kao posebni atributi grana, a težina grane modeluje ukupan broj susreta. Razmisliti i o tome kako bi se altiernativno sumirale te tri težine i da li je logično da sve tri budu ravnopravne.

U tom smislu, za definisanje težine grane u literaturi postoje različiti pristupi: broj odigranih mečeva (za neusmerene mreže), broj pobeda ili poraza, procenat broja pobeda ili poreza od ukupnog broja međusobnih susreta, broj poena na rang listi koje donosi pobeda, kao i nekoliko pristupa koji se odnose na poene odigrane u toku meča (za usmerene mreže). Tip mreže izabrati prema potrebama analize koja se sprovodi.

U okviru ego mreža, potrebno je za svakog od članova *Velike trojke* modelovati zasebnu mrežu, koja uključuje sve direktne veze *ego* čvora sa njegovim *alter*-ima, ali i međusobne veze *alter*-a. Tip mreže izabrati prema potrebama analize.

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.

1) Koliki je prosečan broj tenisera (saigrača) po svakom teniseru?
2) Koji teniseri su se susretali sa najvećim brojem drugih tenisera?
3) Koji teniseri su učestvovali na najvećem broju (različitih) turnira?

4) Koji teniseri su dobri kandidati za predstavnike profesionalnih tenisera? Da li i šta se menja ukoliko se umesto jednog bira skup od nekoliko predstavnika? Dobar predstavnik ili predstavnici bi trebalo da budu u kontaktu sa što većim skupom tenisera.
5) Kako su rangirani na ATP listi teniseri koji su se susretali sa najvećim brojem drugih tenisera?
6) Iz kojih zemalja dolazi najveći broj aktivnih igrača u proteklom periodu?
7) Iz kojih zemalja dolaze najuspešniji igrači u smislu osvojenih poena na ATP listi?
8) Kako su u okviru skupa podataka okarakterisani igrači iz Srbije?
9) Koje zajednice (komune) se mogu uočiti prilikom analize mreže? Da li postoji neko objašnjenje za detektovane komune?
10) Sprovesti analizu klasterisanja i asortativnu analizu, pa uporediti grupisanje na osnovu zemlje iz koje igrači dolaze, na osnovu broja mečeva koje igraju i na osnovu rejtinga.
11) U kojoj meri teniseri imaju tendenciju da se susreću sa istim teniserima? Da li na to utiče njihovo rangiranje na ATP listi?
12) Koji teniseri predstavljaju jezgro mreže?
13) Ko su teniseri koji povezuju različite grupe u okviru mreže?
14) Kolika je gustina svake od modelovanih mreža?
15) U kojoj meri su mreže povezane i centralizovane?
16) Koje su prosečne distance, a koliki dijametar u okviru modelovanih mreža?
17) Kakva je distribucija čvorova po stepenu i da li prati neku zakonomernost? Kako je stepen čvora korelisan sa rejtingom tenisera?
18) Da li u mreži postoje habovi i koji su?
19) Da li mreža tenisera iskazuje osobine malog sveta?
20) Kakve su karakteristike *ego* mreža članova *Velike trojke*? U kojoj meri se te karakteristike razlikuju?
21) Kakva je pozicija *ego* čvora u svakoj od *ego* mreža? Kako su oni strukturno ugrađeni u mrežu?
22) Kako su posmatrane ego mreže ugrađene u mrežu tenisera?
23) Analizirati mrežu dobijenu unifikacijom ego mreža članova *Velike trojke*. Koji procenat čvorova mreže svih tenisera učestvuje u njoj? Sprovesti klasterisanje ovako dobijene mreže

na tri klastera i na osnovu dobijenih rezultata dati interpretaciju kriterijuma pripadnosti klasterima.

24) Kakva je distribucija broja tenisera u odnosu na broj mečeva koji su odigrali?

25) Kakva je distribucija broja turnira u odnosu na podlogu i godinu održavanja?

26) Kakva je distribucija broja mečeva u odnosu na podlogu i godinu održavanja?

Da bi se odgovorilo na postavljena pitanja, potrebno je primeniti odgovarajuće mere i metode za analizu mreže ili statističke metode. Mrežu 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.

3.5. 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. Poeni sa jednom odbranjenog projektnog zadatka važe jednu školsku godinu. Postoji mogućnost da se dobro urađeni projektni zadaci prošire u završni, master rad.

LITERATURA

[1] *Big Three (tennis)*, https://en.wikipedia.org/wiki/Big_Three_(tennis), pristupano: 22.12.2020.

[2] M. Kostić, *Primena tehnika mašinskog učenja i analize socijalnih mreža u predviđanju ishoda teniskih mečeva*, master rad, Univerzitet u Beogradu – Elektrotehnički fakultet, septembar 2020.

[3] J. Sackmann, *Repozitorijum tennis_atp*, dostupno na: https://github.com/JeffSackmann, pristupano: 22.12.2020.