Univerzitet u Beogradu

Elektrotehnički fakultet



Analiza socijalnih mreža

Projektni zadatak

|  |  |
| --- | --- |
| **Kandidati:** | **Školska godina:** |
| Lazar Lazić 2020/3088  Nenad Babin 2020/3019 | 2020/2021. |

Beograd, februar 2021.

Sadržaj

[Sadržaj 2](#_Toc63769945)

[1. Čišćenje podataka 3](#_Toc63769946)

[2. Analiza mreže 4](#_Toc63769947)

1. Čišćenje podataka

Da bi analiza mreže bila ispravna, najpre je potrebno očistiti podatke koji se obrađuju. To podrazumeva utvrđivanje kvaliteta dostupnih podataka, zatim otkrivanje nedostataka i njihovo korigovanje ili uklanjanje.

Pre svega je provereno da li su podaci koji bi trebalo da budu jedinstveni (informacije o teniseru, informacije o meču, informacije o rejtingu), to i jesu. Pošto nisu svi bili jedinstveni, već su postojali duplikati, oni su uklonjeni. Zatim je ustanovljeno da nedostaju određeni podaci (imena, datumi rođenja i sl. pojedinih tenisera, neke informacije o mečevima). Kako ti podaci nisu bili potrebni pri analizi mreže, nisu bili korigovani.

Sekundarni skup podataka, dobijen čišćenjem primarnog skupa podataka, je ponovo proveren i utvrđeno je da zadovoljava kriterijume kvaliteta. Nakon toga, podaci su upisani u nove fajlove istog formata sa sufiksom \_cleaned koji su dalje korišćeni za analizu mreže.

1. Analiza mreže

Prilikom formiranja inicijalne mreže za analizu primećeno je da od skoro 55000 praćenih tenisera, manje od 500 njih igra po sezoni, stoga ne bi imalo smisla analizirati mrežu sa svim teniserima. Zato su prilikom analize mreža uzeti u obzir samo oni teniseri koji su odigrali bar jedan meč u odgovarajućem periodu.

Odgovori na pitanja postavljena u tekstu projektnog zadatka data su tabelarno, gde broj tabele odgovara broju pitanja. Odgovori koji nisu u formi tabele, kao i interpretacije dobijenih rezultata nalaze se u formi teksta između tabela.

Kako bismo odredili prosečan broj tenisera po svakom teniseru, potrebno je da izračunamo prosečan stepen svakog čvora u mreži. Rezultati koji se dobijaju se nalaze u tabeli 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **graf** | **2018.** | **2019.** | **2020.** | **agregirani** |
| **prosek** | 11.881 | 13.066 | 7.681 |  |

Tabela 1: Prosečan broj tenisera po svakom teniseru

Stepen čvora se može iskoristiti da vidimo i koji teniseri su igrali sa najvećim brojem različitih tenisera. Rezultati na kojima je predstavljeno top pet tenisera su dati u tabeli 2. U zagradama se nalaze konkretne vrednosti za svakog tenisera.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **r.b./graf** | **2018.** | **2019.** | **2020.** | **agregirani** |
| **1.** | Fognini (61) | Medvedev (60) | Rublev (43) |  |
| **2.** | Thiem (55) | Paire (51) | Aliassime (37) |  |
| **3.** | Tsitsipas (53) | Tsitsipas (51) | Tsitsipas (34) |  |
| **4.** | Zverev (51) | Zverev (51) | Djokovic (32) |  |
| **5.** | Gasquet (51) | Berrettini (50) | Khachanov (31) |  |

Tabela 2: Teniseri koji su se susretali sa najvećim brojem drugih tenisera

Dobijeni rezultati prikazani u tabelama 1 i 2 su očekivani, pošto je bilo manje mečeva u 2020. godini i samim tim su se teniseri međusobno manje susretali. Interesantno je da se u 2020. godini samo 50 tenisera susrelo sa 20 ili više različitih tenisera, dok je u 2018. i 2019. taj broj oko 100.

Sa usmerenog grafa kod koga su grane usmerene od igrača ka turnirima možemo da vidimo koji teniseri su učestvovali na najvećem broju različitih turnira. Između određenog igrača i nekog turnira može postojati isključivo jedna grana. Težina grane govori o broju mečeva koje je teniser odigrao na tom turniru. Primećeno je da su u skupu podataka svaki susret dveju država u Davis Cupu predstavljen kao zaseban turnir, pa su svi ti turniri agregirani u jedan. Rezultati su dati u okviru tabele 3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **r.b./graf** | **2018.** | **2019.** | **2020.** | **agregirani** |
| **1.** | Dzumhur (32) | Paire (31) | Aliassime (17) |  |
| **2.** | Zverev (30) | Fritz (30) | Bublik (17) |  |
| **3.** | Mannarino (30) | Sousa (29) | Mannarino (17) |  |
| **4.** | Basilashvili (29) | Millman (28) | Struff (16) |  |
| **5.** | Ramos (29) | Struff (28) | Millman (16) |  |

Tabela 3: Teniseri koji su se učestvovali na najvećem broju različitih turnira

2020. godine je veliki broj turnira otkazan, tako da je primetna razlika u broju turnira koji su igrači odigrali 2018. i 2019. godine u odnosu na 2020. godinu.

Dobri kandidati za predstavnike profesionalnih tenisera su oni koji imaju visoku eigenvector centralnost, zato što su povezani sa drugim uticajnim čvorovima u mreži, pa bi trebalo da imaju dobar uvid u stanje tenisa. Takvi teniseri ne moraju nužno da imaju mnogo veza, ali oni imaju veze ka čvorovima koji su dobro povezani. Ukoliko bismo posmatrali više predstavnika koji imaju visoku eigenvector centralnost ne bi bilo značajnih razlika zato što su takvi čvorovi međusobno dobro povezani. Teniseri sa najvećim eigenvector centralnostima dati su u tabeli 4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **r.b./graf** | **2018.** | **2019.** | **2020.** | **agregirani** |
| **1.** | Thiem (0,147) | Medvedev (0,162) | Rublev (0,215) |  |
| **2.** | Zverev (0,146) | Berrettini (0,143) | Tsitsipas (0,182) |  |
| **3.** | Fognini (0,144) | Tsitsipas (0,141) | Aliassime (0,180) |  |
| **4.** | Gasquet (0,138) | Zverev (0,138) | Zverev (0,163) |  |
| **5.** | Del Potro (0,137) | Goffin (0,137) | Djokovic (0,159) |  |

Tabela 4: Teniseri sa najvećim vrednostima eigenvector centralnosti

Imajući u vidu da je Novak Đoković pokušao da osnuje svoju asocijaciju tenisera, interesantno je videti njegove eigenvector centralnosti. One iznose 0,128, 0,132 i 0,159 u 2018, 2019. i 2020. godini, respektivno. 2019. i 2020. godine ima eigenvector centralnost koja je bliska teniserima čije vrednosti prednjače.

Teniseri koji su se susretali sa najvećim brojem drugih tenisera su visoko rangirani, što ima smisla jer je na turnirima moguće izgubiti samo jednom, a pobediti mnogo puta. Stoga, da bi se teniser sreo sa velikim brojem drugih tenisera, mora da pobeđuje. Pobedama i napredovanjem na turnirima se ostvaruje visok rang.

U tabeli 6 su date države sa najvećim brojem aktivnih tenisera u proteklom periodu. Računati su teniseri koji su odigrali barem jedan meč u godini.

|  |  |  |  |
| --- | --- | --- | --- |
| **r.b./godina** | **2018.** | **2019.** | **2020.** |
| **1.** | SAD (34) | SAD (31) | SAD (28) |
| **2.** | Francuska (25) | Francuska (22) | Francuska (22) |
| **3.** | Španija (24) | Nemačka (22) | Italija (18) |
| **4.** | Nemačka (18) | Italija (21) | Nemačka (15) |
| **5.** | Italija (17) | Španija (18) | Australija (14) |

Tabela 6: Države sa najvećim brojem aktivnih tenisera u proteklom periodu

Promene koje se dešavaju u toku ove tri godine mogu biti usled odlazaka pojedinih igrača u penziju ili probijanja mladih tenisera na veliku scenu. Postoji i mogućnost da su neki igrači bili odsutni usled povrede, pa u toj godini nisu nastupali.

Kako bismo saznali koji su igrači bili najuspešniji igrači u proteklom periodu i iz kojih država dolaze, posmatrali smo rang listu tenisera na kraju svake godine. Pre svakog turnira se brišu osvojeni poeni prethodne godine na tom turniru, dok se na kraju turnira dodaju poeni teniserima u zavisnosti od njihovog plasmana na turniru i veličine turnira. U tabeli 7 su prikazani najbolje rangiranih pet igrača na kraju svake godine, države iz kojih dolaze i broj ATP poena koji su osvojili te godine.

|  |  |  |  |
| --- | --- | --- | --- |
| **r.b./godina** | **2018.** | **2019.** | **2020.** |
| **1.** | Djokovic (Srbija) – 9045 | Nadal (Španija) – 9985 | Djokovic (Srbija) – 12030 |
| **2.** | Nadal (Španija) – 7480 | Djokovic (Srbija) – 9145 | Nadal (Španija) – 9850 |
| **3.** | Federer (Švajcarska) - 6420 | Federer (Švajcarska) - 6590 | Thiem (Nemačka) - 9125 |
| **4.** | Zverev (Nemačka) - 6385 | Thiem (Nemačka) - 5825 | Medvedev (Rusija) - 8470 |
| **5.** | Del Potro (Argentina) - 5300 | Medvedev (Rusija) - 5705 | Federer (Švajcarska) - 6630 |

Tabela 7: Najbolje rangirani teniseri u proteklom periodu

U tabelama 8-1, 8-2 i 8-3 se nalaze podaci o tome kako su u okviru skupa podataka okarakterisani igrači iz Srbije.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ime** | **prosečan rang** | **broj protivnika** | **broj mečeva** | **ruka** |
| Dusan Lajovic | 66,21 ± 71 | 42 | 48 | desna |
| Novak Djokovic | 9,23 ± 7 | 42 | 65 | desna |
| Laslo Djere | 92,77 ± 6 | 27 | 29 | desna |
| Filip Krajinovic | 44,52 ± 26 | 25 | 31 | desna |
| Viktor Troicki | 120,75 ± 59 | 19 | 19 | desna |
| Pedja Krstin | 203,77 ± 24 | 2 | 2 | desna |

Tabela 8-1: Srpski igrači u skupu podataka (2018)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ime** | **prosečan rang** | **broj protivnika** | **broj mečeva** | **ruka** |
| Novak Djokovic | 1,19 ± 0 | 47 | 69 | desna |
| Dusan Lajovic | 34,85 ± 7 | 44 | 48 | desna |
| Filip Krajinovic | 63,49 ± 23 | 41 | 49 | desna |
| Miomir Kecmanovic | 77,74 ± 27 | 36 | 42 | desna |
| Laslo Djere | 42,26 ± 19 | 35 | 43 | desna |
| Janko Tipsarevic | 302,98 ± 89 | 17 | 18 | desna |
| Viktor Troicki | 189,91 ± 39 | 8 | 8 | desna |
| Nikola Milojevic | 162,51 ± 15 | 2 | 2 | desna |
| Pedja Krstin | 214,3 ± 25 | 1 | 1 | desna |

Tabela 8-2: Srpski igrači u skupu podataka (2019)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ime** | **prosečan rang** | **broj protivnika** | **broj mečeva** | **ruka** |
| Novak Djokovic | 1,12 ± 0 | 32 | 46 | desna |
| Dusan Lajovic | 25,15 ± 2 | 27 | 30 | desna |
| Miomir Kecmanovic | 46,31 ± 6 | 26 | 29 | desna |
| Filip Krajinovic | 32,62 ± 4 | 24 | 28 | desna |
| Laslo Djere | 59,5 ± 15 | 19 | 22 | desna |
| Nikola Milojevic | 142,19 ± 6 | 4 | 4 | desna |
| Viktor Troicki | 186,54 ± 13 | 4 | 4 | desna |
| Danilo Petrovic | 152,35 ± 10 | 4 | 4 | desna |

Tabela 8-3: Srpski igrači u skupu podataka (2020)

Rađena je detekcija komuna u okviru grafova. Analizirali smo graf na kome su prikazani mečevi tokom 2018. godine.

Prvo je pokrenut Girvan-Newman algoritam koji je podelio mrežu na 203 komune, od kojih jedna ima preko 30% tenisera koji se nalaze na vrhu ATP liste, i mnogo malih komuna koje imaju desetak ili manje čvorova. Ova podela ima smisla jer dobri igrači češće i igraju međusobno, dok lošiji igrači igraju ređe i sa manjim brojem drugih tenisera. Koeficijent modularnosti ove podele je ~0,08 što se ne smatra dobrom podelom.

Zatim je pokrenut Louvain algoritam. Dobijeno je 27 komuna. Najveća komuna ima 58 tenisera i manje su razlike između veličina komuna. Najvećih 9 komuna su jako dobro međusobno povezane i one obuhvataju visoko rangirane tenisere, dok ostale obuhvataju manje uspešne tenisere. Interesantno je primetiti da su Nadal i Đoković u istoj, najvećoj komuni, dok je Federer u skoro duplo manjoj komuni. Koeficijent modularnosti je ~0,22 što ukazuje na bolju podelu. Na slici 9 se nalazi graf sa detektovanim komunama.

Diagram

Description automatically generated

Slika 9: Detektovane komune (2018)

Rađena je asortativna analiza, pa je upoređeno grupisanje na osnovu zemlje iz koje igrači dolaze, na osnovu broja mečeva koje igraju i na osnovu rejtinga. U tabeli 10 su date konkretna vrednosti.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **vrsta / graf** | **2018.** | **2019.** | **2020.** | **agregirani** |
| **država** | 0.018 | 0.005 | 0.007 |  |
| **broj mečeva** | 0.229 | 0.188 | 0.236 |  |
| **rejting** | 0.517 | 0.285 | 0.408 |  |

Tabela 10: Asortativnost tenisera prema njihovoj državi, broju odigranih mečeva i rejtingu

Iz priloženih koeficijenata asortativnosti možemo zaključiti nekoliko stvari. Koeficijent asortativnosti za tenisere iz istih zemalja je pozitivan, ali izuzetno mali što ukazuje da postoji slaba tendencija da međusobno igraju teniseri iz istih zemalja. Koeficijent asortativnosti za tenisere koji imaju slični rejting je izrazito pozitivan, što ukazuje na to da se uglavnom na turnirima sreću teniseri koji imaju sličan rang. Izuzetak je 2019. godina kada je koeficijent značajno niži. Koeficijent asortativnosti za tenisere koji imaju sličan broj odigranih mečeva je pozitivan, međutim ne izrazito jer se u ranim fazama takmičenja sreću teniseri koji mnogo igraju sa onim koji malo igraju (odnosno koji bivaju izbačeni od strane boljih tenisera).

U tabeli 11 se nalaze srednje vrednosti težine grana grafova. Na taj način smo želeli da vidimo koliku tendenciju imaju teniseri da se susreću sa istim teniserima.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **graf** | **2018.** | **2019.** | **2020.** | **agregirano** |
| **prosečna težina grane** | 1.161 | 1.169 | 1.092 |  |

Tabela 11: Prosečne težine grana neusmerenog grafa

Pošto je prosečna težina grana blizu 1 to znači da teniseri nemaju tendenciju da se susreću sa istim teniserima. Međutim, oni koji se susreću više puta su teniseri koji su bolje rangirani, tako da se može reći da rangiranje utiče na ponovne susrete.

Na slici 12 se nalazi jezgro mreže iz 2018. godine. Uz svaki čvor je predstavljen i prosečan rejting tokom godine. Sa slike se može videti da jezgro mreže predstavljaju visoko rangirani teniseri, okvirno prvih 100 igrača.

Chart

Description automatically generated with low confidence

Slika 12: Jezgro mreže (2018)

Teniseri koji povezuju različite grupe se mogu dobiti nakon računanja relacione centralnosti. U tabeli 13 se nalaze igrači sa najvećim relacionim centralnostima.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **r.b./graf** | **2018.** | **2019.** | **2020.** | **agregirani** |
| **1.** | Cuevas (0,07) | Cuevas (0,056) | Djokovic (0,033) |  |
| **2.** | Jaziri (0,044) | Sousa (0,039) | Giron (0,033) |  |
| **3.** | Fognini (0,034) | Berankis (0,032) | Tsitsipas (0,031) |  |
| **4.** | Copil (0,031) | Thiem (0,029) | Ruud (0,029) |  |
| **5.** | Vesely (0,031) | Pella (0,029) | Rublev (0,028) |  |

Tabela 13: Relacione centralnosti

Izračunate su i gustine modelovanih mreža. U tabeli 14 se nalaze izračunate gustine. Možemo videti da postoji oko 2-3% mogućih veza, što ukazuje na to da su mreže retke.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **graf** | **2018.** | **2019.** | **2020.** | **agregirani** |
| **gustina** | 0.028 | 0.036 | 0.022 |  |

Tabela 14: Gustine mreža

Za svaki graf je izdvojen broj slabo povezanih komponenti, kao i centralnosti na nivou mreže. Centralnosti su računate kao procenat u odnosu na odgovarajuće centralnosti zvezdaste mreže sa istim brojem čvorova. Vrednosti se nalaze u tabeli 15.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **graf** | **2018.** | **2019.** | **2020.** | **agregirani** |
| **Broj povezanih komponenti** | 18 | 11 | 25 |  |
| **Centralnost po stepenu** | 0.118 | 0.130 | 0.103 |  |
| **Centralnost po bliskosti** | 0.268 | 0.260 | 0.292 |  |
| **Relaciona centralnost** | 0.066 | 0.051 | 0.028 |  |

Tabela 15: Povezanost i centralnost na nivou mreže

U 2020. godini je primećen veći broj povezanih komponenti, što može biti posledica virusa COVID-19 i nemogućnosti tenisera da idu na određena mesta (turnire). U svim mrežama je uočena jedna ogromna komponenta sa preko 80% čvorova, i mnogo malih komponenti sa 10 ili manje čvorova.

Sračunate su i prosečne distance i dijametri mreža. Treba primetiti da su sve mreže sastavljene iz više povezanih komponenti pa su ove vrednosti zapravo prosečne vrednosti prosečnih distanci povezanih komponenata i najveće vrednosti dijametara u svim povezanim komponentama.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **graf** | **2018.** | **2019.** | **2020.** | **agregirani** |
| **Prosečne distance** | 3.137 | 3.046 | 3.185 |  |
| **Dijametar** | 11 | 9 | 9 |  |

Tabela 16: Prosečne distance i dijametri mreža