Универзитет у Београду

Електротехнички факултет



Извештај за пројектни задатак

из Анализе социјалних мрежа

|  |  |
| --- | --- |
|  | Кандидати:  Ивана Васић ГГГГ/ББББ |
|  | Данило Лалић 2021/3260 |

Београд, 2022.

# Увод

У овом раду ће бити изложени резултати анализе друштвене мреже *Reddit* за 2008. годину. Почетни скуп података се састоји од *CSV* датотека и подељен је у два скупа: први представља објаве, а други коментаре на платформи у наведеном периоду. За сваки скуп је дато дванаест датотека које представљају податке за сваки месец.

За анализу су коришћени програмски језик *Python*, неколико библиотека, од који се издвајају *NetworkX* i *Pandas*, и *Gephi* програм за визуелизацију, анализу и истраживање графова и мрежа /\* Још алата? \*/. Библиотека *Pandas* је искоришћена за статистичку анализу. За моделовање мрежа се користи библиотека *NetworkX*. Иста је искоришћена и за одређене анализе моделованих мрежа. Софтвер *Gephi* је употребљен за визуелизацију и анализу мрежа. /\* Касније може и детаљнији опис шта је за шта коришћено. \*/

У наставку ће бити изложен процес припреме и анализе података. У /\* број \*/ поглављу дат је опис пречишћавања иницијалног скупа податак. У /\* број \*/ поглављу дати су резултати статистичке анализе. У /\* број \*/ поглављу описано је на који су креиране мреже. Такође је објашњени и разлози доношења одређених архитектуралних одлука везаних за мрежу. У /\* број \*/ поглављу наведени су резултати и запажања анализе основних карактеристика мрежа и анализе мера централности истих, респективно. Затим је у /\* број \*/ поглављу обрађена детекција комуна, а у /\* број \*/ имамо поређење *SNet* и *SNetT* мрежа.

# Филтрирање и припрема података

Најпре се подаци учитавају у две *DataFrame* структуре, једна садржи коментаре, а друга објаве. Из поља претходно наведе структуре се може видети да број редова одговара оном наведеном у опису података, али колона има за један више. Разлог за неједнак број колона је тај што је у иницијалном скупу података додата неименована колона која представља индекс реда. Ти индекси се за сваки нови месец/датотеку ресетују на 0. Овде је одлучено да буде искоришћена већ дата индекс колона која долази уз *DataFrame* тип, па је према томе колона са већ датим индексима одбачена из обе структуре.

Након тога је извршена провера јединствености и присутности (потребно је да вредност не буде *null, NaN*…) идентификатора. Утврђено је да оба скупа имају јединствене идентификаторе, али је један идентификатор у скупу коментара имао вредност *NaN*. Идентификатор наредног реда је „nao“, а претходни је „nam“, што сугерише да су идентификатори додељени по лексикографски. На основу овога се може закључити да је идентификатор са вредношћу „nan“ погрешно протумачен ка *NaN* вредност, па је неопходно било експлицитно доделити вредност идентификатору овог коментара.

Провера вредности свих колона је показала да колона „*distinguished*“ у скупу коментара садржи нерегуларне вредности (*null, NaN*…), што је у складу са описом скупа података, јер *null* вредност за ову колону означава да коментар нису поставили модератор или админ. За скуп објава колоне „*distinguished*“ и „*domain*“ имају нерегуларне вредности. За колону „*distinguished*“ имамо исту ситуацију као и код претходног скупа. Разлог нерегуларних вредности за колону „*domain*“ није познат, али у даљем раду она и не представља значајну колону па ју је могуће једноставно избацити. /\* Да ли је заиста тако када је реч о домену? \*/

/\* делетед \*/

# Статистичка анализа података

Током статистичке анализе података утврђено је да постоје сабредити са истим идентификаторима, али различитим именима. Реч је о сабредитима са именима „*\_Descary*“ и „*Descary*“ са идентификатором „t5\_2qj0s“, и о „*\_\_Journalism*“ и „*Journalism*“ са идентификатором „t5\_2qhyl“. Овде се уводи претпоставка да је дошло до грешке приликом прикупљања података, тако да су сва имена са цртицама претворена у њихове парњаке без цртица.

Постоји 5032 различита сабретита у 2008 години. Једни од најважнији сабредита по броју коментара су „*subreddit.com*“, „*politics*“, „*programming*“… Најзначајнији сабредити по броју корисника су „*subreddit.com*“, „*politics*“, „*business*“… Табеларни приказ је дат у табели 3.1.

|  |  |  |  |
| --- | --- | --- | --- |
| По коментарима | | По корисницима | |
| Име сабредита | **Број коментара** | **Име сабредита** | **Број корисника** |
| *reddit.com* | 1768764 | *reddit.com* | 138153 |
| *politics* | 1059618 | *politics* | 15250 |
| *programming* | 421137 | *business* | 13009 |
| *pics* | 383239 | *technology* | 13002 |
| *science* | 340915 | *entertainment* | 12624 |

Табела 3.1 Најзначајнији сабредити по броју коментара и по броју корисника.

Просечан број активних корисника по сабредиту је 129. Корисници са највећим бројем објава и највећим бројем коментар су дати у табели 3.2. Преглед корисника који су активни на највећем броју сабредита се може видети у табели 3.3.

|  |  |  |  |
| --- | --- | --- | --- |
| По објавама | | По коментарима | |
| Корисничко име | **Број објава** | **Корисничко име** | **Број коментара** |
| gst | 18870 | NoMoreNicksLeft | 13480 |
| qgyh2 | 12238 | malcontent | 12159 |
| rmuser | 9822 | matts2 | 11672 |
| twolf1 | 8597 | mexicodoug | 9169 |
| IAmperfectlyCalm | 8308 | 7oby | 9161 |

Табела 3.2 Корисници са највећим бројем објава и коментара.

|  |  |
| --- | --- |
| Корисничко име | Број сабредита на којима је активан |
| Escafane | 214 |
| MrKlaatu | 201 |
| scientologist2 | 177 |
| krugerlive | 164 |
| omfgninja | 157 |

Табела 3.3 Корисници који су активни на највећем броју сабредита.

/\* Пето питање. \*/

Идентификатори првих пет објава које имају највећи број коментара су: /\* Навести идентификаторе \*/. /\* Написати шта представља њихов садржај. \*/

# Моделовање мрежа

Мрежа *SNet* има за циљ представљање утицаја и ширења информација између сабредита путем корисника. Чворови су сабредити док грана између два сабредита постоји ако постоји корисник који је био активан на оба сабредита; тежина грана представља број коринсика који су били активни на оба сабредита.

# Основна карактеризација мрежа

## SNet мрежа

Густина мреже је 0,012. Дијаметар мреже је 5, а просечне дистанце пута су 2,09856. Мрежа садржи 1486 повезаних компонената. Највећа од њих садржи 3525 чворова и представља гигантску компоненту. Глобални коефицијент кластеризације је 0.267, а просечни је 0.907. Расподела локалног коефицијента кластеризације је дата на слици 5.1. Поредећи наведене вредности кластеризације са просечним и глобалним коефицијентом кластеризације *Erdos-Renyi* мреже, који су оба 0,0124, можемо закључити да је кластерисање у *SNet* изражено. С обзиром да мрежа има мали дијаметар са такође малом просечном дужином путева и изражено кластерисање, закључујемо да мрежа одражава особине малог света. Коефицијент асортативности за нетежински степен чворова је -0,436, а за тежински -0,032. У првом случају мрежа је дисасортативна, а у случају тежинског степена је неасортативна. Дистрибуција чворова по степену прати *power law* расподелу.

Chart, line chart

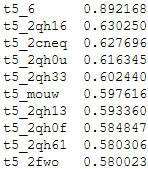
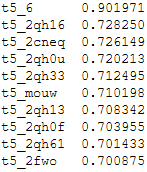
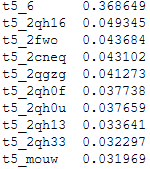
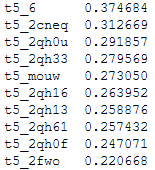
Description automatically generated

Слика 5.1 Расподела локалног коефицијента кластеризације код *SNet* мреже.

# Мере централности

## SNet мрежа

Све мере централности су вршене над доминантном компонентом која има 3525 чворова. Остале компоненте су занемарене зато што садрже 3 или мање чвора. На слици 6.1 је дат преглед мера централности. Висок степен централности по сопственом вектору указује на то да информације које су дељене на тим сабредитима су преовладавале на целој мрежи.

Слика 6.1 Првих десет истакнутих чворова и вредности мера централности по степену, блискости, релационој централности и централности по сопственом вектору за *SNet* мрежу, респективно.

Text

Description automatically generated Text

Description automatically generated Text

Description automatically generated Text

Description automatically generated Text

Description automatically generated

Слика 6.2 Вредности Kацове централности када је за све чворове.

Text

Description automatically generated Text

Description automatically generated Text, chat or text message

Description automatically generated Text

Description automatically generated Text

Description automatically generated

Слика 6.3 Вредности Kацове централности када је 0 за сaбредит *reddit.com*, а 1 за све остале.

За рачунанје Кацове централност су одабране вредности за из скупа вредности: 0,000005, 0,000004, 0,000003, 0,000002 и 0,000001. На слици 6.2 се може видети првих десет чворова рангираних по Кацоовј централности када је за све чворове, док се на слици 6.3 види поредак када је , за сабредит *reddit.com*.

# Детекција комуна