

Primena veštačke inteligencije za moderaciju tekstualnih poruka na internetu

Autor: Bojan Rađenović

Dvanaesta beogradska gimnazija i Regionalni centar za talente Beograd II, E-mail: bojan@radjenovic.dev

Mentor: Mateja Opačić, Regionalni centar za talente Beograd II

1. Uvod

U današnjem društvu, većina komunikacije se dešava preko interneta. Na primer, kada želimo da se dogovorimo da izađemo sa nekim ili čak da upoznamo ljude sa kojima imamo zajednička interesovanja. Međutim, primetio sam problem na javnim „chat sobama“. Problem na javnim mrežama za dopisivanje jeste negativno ponašanje i nepristojnost ljudi. Njihovo širenje negativnosti zahteva prisustvo osoba koje bi nadgledale „chat sobe“, ali to može da bude zahtevan i ponekad neuspešan proces. Lično sam imao iskustva sa nadgledanjem soba i javila mi se želja za nečim što bi moglo da olakša taj ceo proces. Ideja ovog projekta jeste da se olakša proces moderacije tako što će program automatski da analizira poruke i da sam nadgleda „chat sobe“ ili da poruke prosledi na dalju analizu.

2. Materijali i Metodologija

Zamisao ovog projekta je da se napravi program osnovan na veštačkoj inteligenciji. Ideja je bila da se, na platformi za dopisivanje Discord, napravi Discord Bot koji će moći da vrši automatsku analizu poruka. Discord Bot je član „chat sobe“ koji je sličan korisnicima. Koristi se za automatizaciju raznih radnji pomoću Discord-ovog javnog „API“-a. Postoje mnoge biblioteke za Python, koje obezbeđuju lakši pristup Discord-ovom API-u. Za analizu poruka, odlučio sam da koristim Machine Learning. Za treniranje modela sam koristio biblioteku TensorFlow, a za Discord Bot-a biblioteku disnake (1). Prikupljanje podataka sam automatizovao. Napravljen je Discord Bot, koji je, iz jedne javne „chat sobe“, automatski sačuvao poruke u MySQL bazu. Kada sam skupio oko 30.000 poruka, odlučio sam da ih obradim. To je uključivalo brisanje zarezova, tačaka i nekih reči, a i samu klasifikaciju. Takođe sam morao da obrišem neke poruke jer su se ponavljale ili bile na različitim jezicima. Na kraju sam završio sa bazom koja je sadržala oko 20.000 poruka. Zbog efikasnosti treniranja modela i samog projekta, koristio sam već postojeći model od Google-a „Natural Language Processing Model“ (2). Ovaj model od Google-a je napravljen pomoću raznih članaka iz vesti i predstavlja reči pomoću vektora. Koristio sam svoje podatke sa ovim modelom. Kada sam završio samo treniranje modela, krenuo sam da radim deo koji upotrebljava Discord „API“. Biblioteka disnake radi po principu događaja. Za svaku radnju u samoj „chat sobi“ postoji događaj (korisnik uđe u sobu, izađe, pošalje poruku...). Pomoću događaja za slanje poruke, napravio sam da Discord Bot prosleđuje poruku modelu na dalju proveru.

3. Rezultati i diskusija

Kada sam radio na samom modelu, morao sam više puta da menjam parametre (learning rate, slojeve...), a za svako treniranje je trebalo 10-15 minuta. Na kraju sam završio sa modelom koji je imao 80-90% tačnost. Kada sam završio deo sa Discord-om, odlučio sam da dodam Discord Bot-a u više „chat soba“. Međutim, dešavalo se da Discord Bot preduzme pogrešnu radnju i dešavalo se da remeti razgovore i zbog toga sam odlučio da automatsko nadgledanje bude kao sporedna opcija. Discord Bot podrazumevano obaveštava osoblje te sobe umesto da sam vrši radnje. Korisnici takođe imaju opciju da provere da li je njihova poruka klasifikovana kao pozitivna ili negativna, tako što će je poslati privatno samom Discord Bot-u.

4. Zaključak

Za projekat minimalne vrednosti, ideja je bila napraviti program koji, na jednoj platformi, olakšava nadgledanje „chat soba“. Program radi tačno kako je i zamišljen. Pomaže u nadgledanju tako što prosleđuje sumnjive poruke ili sam vrši moderaciju. U budućnosti, ideja je da se ovaj projekat nastavi sa razvojem. Ideja je da se ovaj model primeni na drugim platformama koje dozvoljavaju javne „chat sobe“ i da se doda podrška za više jezika.

5. Reference

1. Python biblioteka koja služi za interakciju sa Discord API-em (<https://github.com/DisnakeDev/dsnake>)
2. Google-ov model koji predstavlja reči vektorom (<https://tfhub.dev/google/nnlm-en-dim50/2>)

ПРИМЕНА ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ ЗА МОДЕРАЦИЈУ ТЕКСТУАЛНИХ ПОРУКА НА ИНТЕРНЕТУ

APPLICATION OF ARTIFICIAL INTELLIGENCE ON ONLINE INSTANT MESSAGING PLATFORMS FOR THE PURPOSE OF MODERATION

Аутор:

БОЈАН РАЂЕНОВИЋ

*2. разред, Дванаеста београдска гимназија
Регионални центар за таленте Београд II*

Ментор:

МАТЕЈА ОПАЧИЋ

Регионални центар за таленте Београд II

РЕЗИМЕ: Вештачка интелигенција (AI) има потенцијал да промени платформе за дописивање омогућавајући ефикаснију модерацију и сигурније корисничко искуство. Једна од примарних примена вештачке интелигенције у платформама за дописивање је за сврху модерације, као што је идентификација и уклањање непожељног садржаја и спречавање узнемиравања. Пораст друштвених мрежа такође је довео до повећања штетног и токсичног садржаја који се дели на интернету[1]. Коришћење вештачке интелигенције за модерацију на платформама за дописивање може помоћи у спречавању таквог садржаја да се брзо шири и досеже велику публику. Додатно, истраживања су показала да већина људи проверава своје телефоне у првих неколико минута након буђења ујутро[2]. Стога, коришћење вештачке интелигенције за модерацију на друштвеним мрежама може помоћи у спречавању штетног и токсичног садржаја да буде међу првима са којима људи долазе у контакт сваког дана.

КЉУЧНЕ РЕЧИ: вештачка интелигенција, друштвене мреже, негативност

ABSTRACT: Artificial intelligence (AI) has the potential to revolutionize online instant messaging platforms by enabling more effective moderation and safer user experiences. One of the primary applications of AI in messaging platforms is for moderation purposes, such as identifying and removing inappropriate content, detecting spam, and preventing harassment. The rise of social media and messaging platforms has also led to an increase in harmful and toxic content that is shared online[1]. Using AI for moderation on messaging platforms can help prevent such content from spreading rapidly and reaching a large audience. Additionally, research has found that most people check their phones within the first few minutes of waking up in the morning[2]. Therefore, using AI for moderation on messaging platforms can help prevent harmful and toxic content from being among the first things people encounter each day.

KEYWORDS: artificial intelligence, social media, online negativity

УВОД

У данашње време, модерација садржаја на интернету постаје све значајнија тема због наглог раста броја корисника и прилива различитих садржаја. Многе друштвене мреже, форуми и друге сервисе на интернету користе модераторе како би се осигурали сигурност и заштиту корисника од непожељних или опасних садржаја. Међутим, учестале грешке у модерацији које се дешавају, указују на потребу за новим и напредним алатима који могу да помогну у овом процесу. У том смислу, идеја примене вештачке интелигенције за модерацију текстуалних порука на интернету, постаје све популарнија.[3]

Као модератор, суочио сам се са великим изазовом да пратим све поруке које се шаљу на платформи, посебно у ситуацијама када је прилив порука јако велик. Иако се трудим да будем што пажљивији, понекад се дешава да не приметим неки непожељни садржај, што би могло имати штетне последице по кориснике. Често сам размишљао о томе како би се овај процес могао олакшати и како бих могао бити сигурнији да ниједан непожељни садржај неће проћи испод мог радара.

С обзиром на то да је вештачка интелигенција у последње време напредовала много, јавила се жеља да истражим како би се ова технологија могла применити за модерацију текстуалних порука на интернету. Уверен сам да би коришћење вештачке интелигенције за ову сврху могло да олакша процес модерације и уједно смањи вероватноћу да пропустим неку поруку. Такође, примена ове технологије би могла да буде корисна у ситуацијама када има превише порука за ручну модерацију, што би могло да доведе до повећања ефикасности целокупног процеса.

У циљу истраживања примене вештачке интелигенције за модерацију текстуалних порука на интернету, одлучио сам да направим робота (у даљем тексту „Bot”) на Discord платформи који ће вршити аутоматску модерацију помоћу машинског учења. Discord јесте бесплатан програм за комуникацију на интернету који омогућава корисницима да направе своје сервере (chat собе), да деле мултимедијалан садржај и да играју видео игре са другим корисницима.[4] Discord Bot јесте програм који се интегрише на Discord платформи и омогућавају аутоматизовану комуникацију са корисницима.

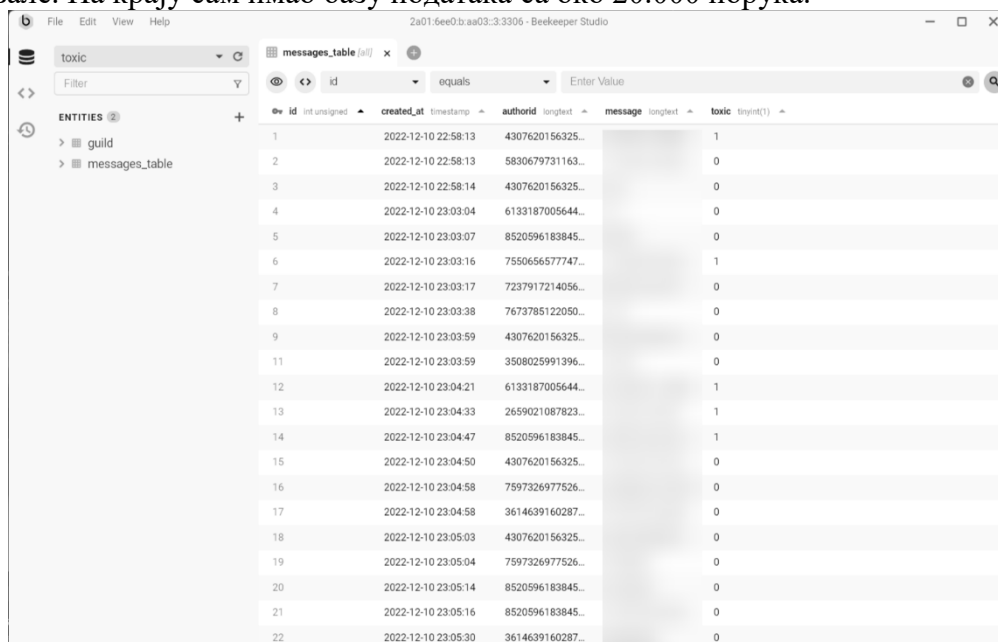
МАТЕРИЈАЛ И МЕТОДИКА РАДА

Методологија овог рада се састојала из више делова. Делови израде су били следећи: Самостално прикупљање података за базу података, затим анализа и класификација тих података, креирање модела за анализу порука, и на крају сам Discord Bot.

Прикупљање и обрада података за базу података

С обзиром да би ручно прикупљање података за овакву базу података било изазовно, одлучио сам да то аутоматизујем због ефикасности израде пројекта. У ту сврху, направљен је Discord Bot који ће, из јавних chat соба, прикупљати поруке и смештати их у базу података. За израду овог помоћног Discord Bot-а употребио сам Python библиотеке „disnake“[5], „aiomysql“[6], као и сам MySQL. „Disnake“ јесте open-source библиотека која олакшава интеракције са Discord API-ем, док је „aiomysql“ библиотека која служи за приступање MySQL бази у Python-у. Када особа у chat соби пошаље поруку, Discord Bot аутоматски прослеђује њу у базу.

Када сам скупио око 30.000 порука, одлучио сам да их обрадим. За овај процес, користио сам Beekeeper Studio[7]. Beekeeper Studio јесте open-source SQL едитор који је једноставан за коришћење. Обрађивање порука је укључивало брисање интерпункцијске знакове, а и самих порука уколико су на различитим језицима или уколико су се понављале. На крају сам имао базу података са око 20.000 порука.



	id	int unsigned	created_at	timestamp	authorid	longtext	message	longtext	toxic	tinyint(1)
1			2022-12-10 22:58:13		4307620156325...				1	
2			2022-12-10 22:58:13		5830679731163...				0	
3			2022-12-10 22:58:14		4307620156325...				0	
4			2022-12-10 23:03:04		6133187005644...				0	
5			2022-12-10 23:03:07		8520596183845...				0	
6			2022-12-10 23:03:16		7550656577747...				1	
7			2022-12-10 23:03:17		7237917214056...				0	
8			2022-12-10 23:03:38		7673785122050...				0	
9			2022-12-10 23:03:59		4307620156325...				0	
11			2022-12-10 23:03:59		3508025991396...				0	
12			2022-12-10 23:04:21		6133187005644...				1	
13			2022-12-10 23:04:33		2659021087823...				1	
14			2022-12-10 23:04:47		8520596183845...				1	
15			2022-12-10 23:04:50		4307620156325...				0	
16			2022-12-10 23:04:58		7597326977526...				0	
17			2022-12-10 23:04:58		3614639160287...				0	
18			2022-12-10 23:05:03		4307620156325...				0	
19			2022-12-10 23:05:04		7597326977526...				0	
20			2022-12-10 23:05:14		8520596183845...				0	
21			2022-12-10 23:05:16		8520596183845...				0	
22			2022-12-10 23:05:30		3614639160287...				0	

СЛИКА 1. Снимак екрана Beekeeper Studio-а који илуструје поља MySQL табеле са порукама.
FIGURE 1. Screenshot of Beekeeper Studio which illustrates fields of the MySQL table with messages.

Тренирање модела вештачке интелигенције

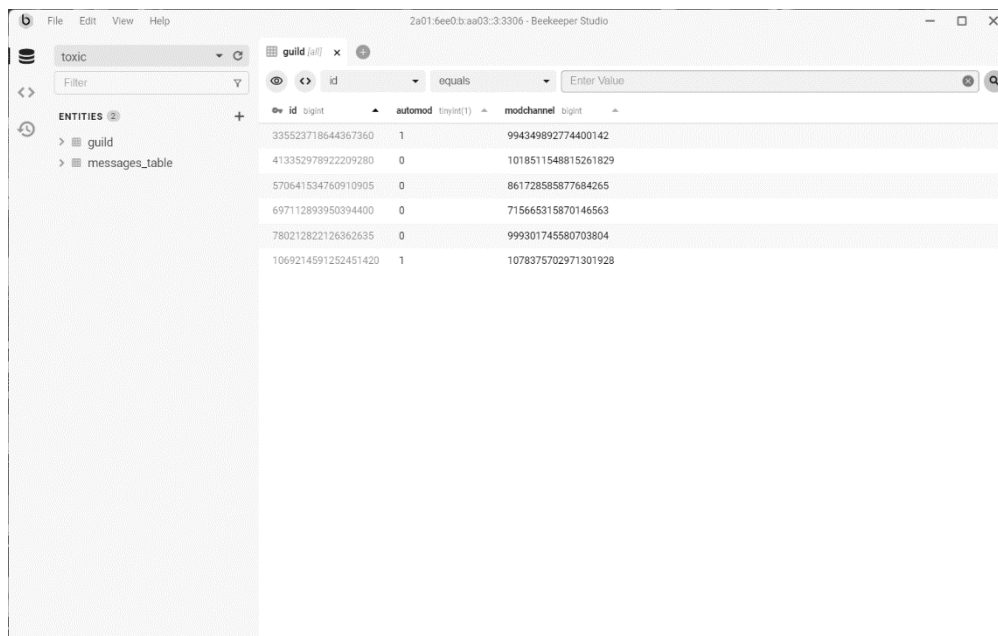
Након прикупљања података и уређивања података, следи тренирање модела вештачке интелигенције. За креирање модела, одлучио сам да користим библиотеку „TensorFlow“. Изабрао сам њу због тога што је open-source, има много ресурса и видео лекција, а и због чињенице да подржава пренос учења. Пренос учења представља приступ машинског учења у којем се претходно научено знање из једног проблема примењује на решавање другог проблема.

Због ефикасности израде овог пројекта, одлучио сам да користим модел „nnlm-en-dim50“[8]. Овај модел се може користити као основа за решавање проблема обраде текста, већ је обучен на огромној количини текста (на енглеском језику) како би научио репрезентацију речи и реченица у енглеском језику.

Discord Bot

Као што је већ напоменуто, за приступ Discord-овом API-у користио сам библиотеку „disnake“[5]. Библиотека „disnake“ ради по принципу догађаја. За сваку радњу у chat соби, постоји догађај (корисник уђе у собу, изађе из собе, пошаље поруку...). Помоћу догађаја за слање порука, направио сам да Discord Bot прослеђује поруку на даљу проверу.

У зависности од конфигурације Discord Bot-а за chat собу, он може да врши више радњи. Ова конфигурација се чува на бази сваке chat собе у MySQL бази.



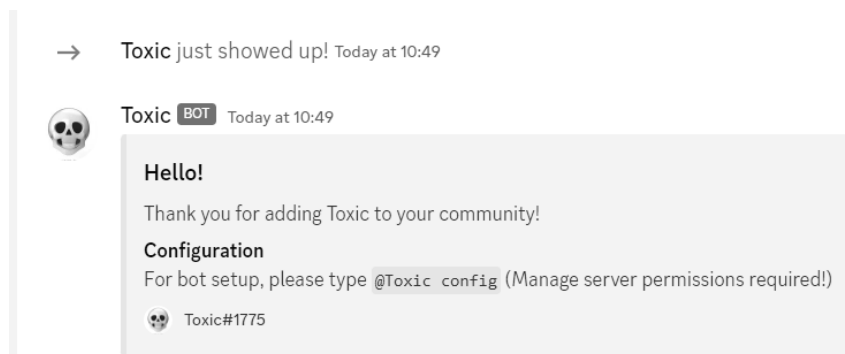
id	bigint	automod	tinyint(1)	modchannel	bigint
335523718644367360	1			994349892774400142	
413352978922209280	0			1018511548815261829	
570641534760910905	0			861728585877684265	
697112893950394400	0			715665315870146563	
780212822126362635	0			999301745580703804	
1069214591252451420	1			1078375702971301928	

СЛИКА 2. Снимак екрана Beekeeper Studio-а који показује сачувана подешавања у MySQL бази.

FIGURE 2. Screenshot of Beekeeper Studio which shows saved guild settings in the MySQL database.

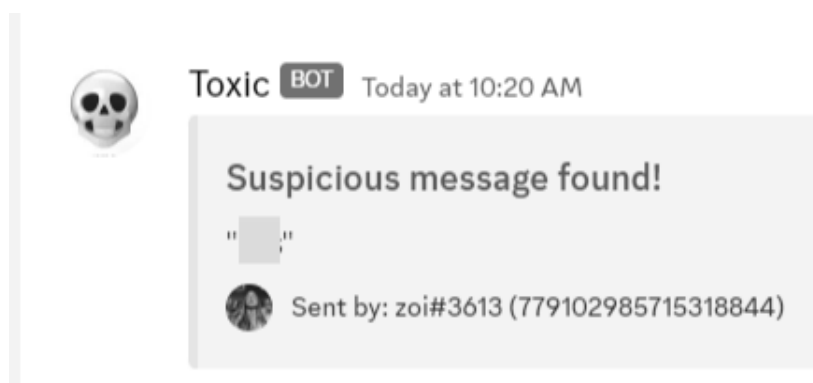
Када се Discord Bot придружи chat соби, он пошаље поруку добродошлице у којој објашњава како се он може конфигурисати.

Discord Bot се може конфигурисати на више начина: да само обавештава модераторе о сумњивим порукама или да аутоматски врши модерацију.



СЛИКА 3. Порука коју Discord Bot шаље у chat собу када се придружи њој.

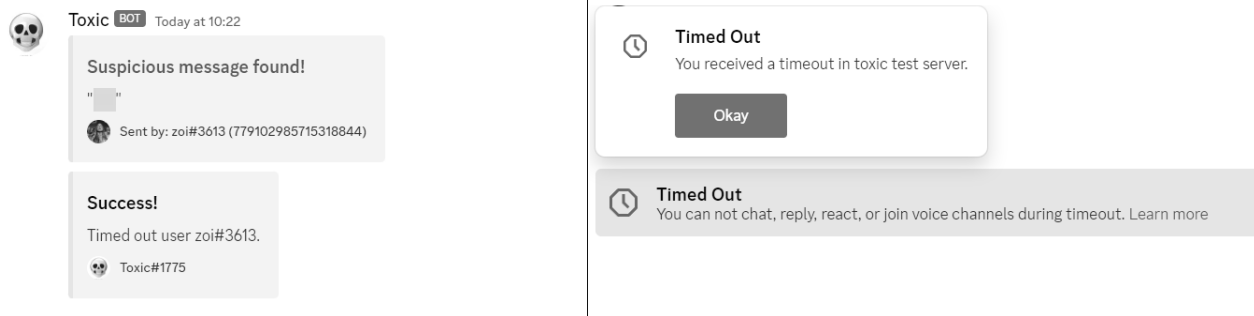
FIGURE 3. Message that the Discord Bot sends when it joins a chat room.



СЛИКА 4. Пример поруке коју Discord Bot шаље у chat собу са модераторима када је искључена аутоматска модерација.

FIGURE 4. Example message which illustrates the message that the Discord Bot sends to the chat room with moderators when automatic moderation is disabled.

У случају да chat соба има укључену аутоматску модерацију, корисник који пошаље сумњиву поруку ће бити „timeout“-ован. Timeout представља једну од опција модерације на платформи Discord. Корисник који је „timeout“-ован не може да шаље поруке, да „реагује“ на поруке и да улази у собе за причање.



СЛИКА 5. Пример поруке коју Discord Bot шаље у chat собу са модераторима када је укључена аутоматска модерација и шта корисник који је послао сумњиву поруку види.

FIGURE 5. Example message which illustrates the message that the Discord Bot sends to the chat room with moderators when automatic moderation is enabled and what the user who had sent the suspicious message sees.

РЕЗУЛТАТИ И ДИСКУСИЈА

Када сам радио на самом моделу, морао сам много пута да мењам параметре, као што су learning rate и слојеви, како бих што боље прилагодио модел. Сваки пут када сам тренирао модел, морао сам чекати 20-30 минута за тренирање. Успео сам да креирам модел који је имао тачност између 80-90%.

Када сам завршио део са Discord-ом, одлучио сам да додам Discord Bot-a у више chat соба како бих га тестирао у реалном свету. Међутим, суочио сам се са проблемом да Discord Bot понекад предузме погрешну и ремети разговоре. Како бих спречио такве ситуације, одлучио сам да имплементирам аутоматско надгледање као споредну опцију, не као главну.

Осим тога, корисници који имају додатну опцију - могу проверити да ли је њихова порука класификована као позитивна или негативна тако што ће је послати приватно самом Discord Bot-у. На тај начин, корисници ће добити повратну информацију о томе како је њихова порука оцењена, што ће им помоћи да побољшају своје писање и изражавање у будућности.

ЗАКЉУЧАК

За пројекат минималне вредности, идеја је била направити програм који ће олакшати надгледање chat соба на једној платформи. Програм ради тачно како је и замишљен. Програм аутоматски прослеђује сумњиве поруке и врши модерацију како би се спречило било какво непримерено понашање. То помаже администраторима и модераторима да имају бољи преглед ситуације и да одрже chat собу сигурном и пријатном за све кориснике.

Међутим, ово је само почетак. Планирам да наставим са развојем овог пројекта у будућности. Идеја је да овај програм не само да се примени на друге платформе које дозвољавају јавне chat собе, већ да се и прошири на више језика. На тај начин, желим да омогућим коришћење овог модела у што већем броју заједница, како би се побољшало корисничко искуство и омогућила сигурна и позитивна околина за све.

ЛИТЕРАТУРА

1. Web документ: Anti-Defamation League: *Online Hate and Harassment Report: The American Experience*. Преузето 2. децембра 2022. са сајта <https://www.adl.org/resources/report/online-hate-and-harassment-report-american-experience-2020>
2. Kliestik, T., Scott, J., Musa, H., and Suler, P. (2020): “*Addictive Smartphone Behavior, Anxiety Symptom Severity, and Depressive Stress*,”
doi:10.22381/AM1920204
3. Gollatz, Kirsten, Felix Beer, and Christian Katzenbach. (2018): “*The turn to artificial intelligence in governing communication online*.” doi:10.31235/osf.io/vwpcz
4. Web документ: Discord Inc.: *What is Discord?*, Преузето 15. децембра 2022. са сајта <https://discord.com/safety/360044149331-what-is-discord>
5. <https://github.com/DisnakeDev/disnake> - „GitHub” страница disnake библиотеке
Аутор: DisnakeDev
6. <https://github.com/aio-libs/aiomysql> - „GitHub” страница aiomysql библиотеке
Аутор: aio-libs
7. <https://github.com/beekeeper-studio/beekeeper-studio> - „GitHub” страница
Beekeeper Studio Аутор: beekeeper-studio
8. <https://tfhub.dev/google/nnlm-en-dim50/2> - „TensorFlow Hub” страница nnlm-en-dim50 модела Аутор: Google