Sveučilište Jurja Dobrile u Puli
Fakultet informatike

Enkodiranje Tokena na Raspodijeljenom Blockchain Sustavu za Unapređenje Sigurnosti Neuronskih Mreža - Dokumentacija Projekta

Krstačić Rafael

Mentor: doc. dr. sc. Nikola Tanković

Pula, rujan 2023.

1. Opis projekta

Repozitoriji projekta sa potrebnim datotekama se nalazi na (finalna verzija: Blockchain.ipynb): https://github.com/rkrstacic/distributed-tokenizer

Repozitoriji korišten za dohvat podataka sa Reddit-a: https://github.com/rkrstacic/RedditScraping

Reddit podaci preuzeti sa (samo jedan snapshot je korišten): https://files.pushshift.io/reddit/comments/

Ideja iza ovog projekta je pokazati da je moguće napraviti komunikaciju zajedničkog vokabulara u raspodjeljenom sustavu koji bi koristio blockchain tehnologiju za sigurnu komunikaciju. To bi se postiglo na način da svi korisnici imaju neki svoj privatni vokabular koji nije uključen u standardni, zajednički vokabular te međusobnim peer-to-peer glasanjem zajedno dolaze do konsenzusa da se korisnici raspodjele u klastere u kojem će korisnik jednog klastera biti članom onog klastera čiji drugi korisnici tog klastera imaju najsličniji vokabular.

Projekt je podijeljen u dvije cjeline:

1. Prikupljanje, priprema i analiza podataka
2. Implementacija i izvršavanje simulacije klastera

Prikupili su se podaci iz jednog snapshot-a Reddit-a (snapshot obuhvaća komentare jednog mjeseca) u kojem se uzelo 2 subreddit-a sa najviše komentara te su se uzeli najaktivnijih 100 korisnika iz oba subreddit-a.

Tokenizer je za konkatenaciju svih komentara pronašao mali broj nepoznatih tokena pa se zbog tog razloga uzelo 500 najčešćih tokena i napravio novi vokabular za tokenizer koji će imati više nepoznatih tokena.

Analizom tih tokena se dobila matrica sličnosti tokena svih korisnika koja govori da postoji veća sličnost korisnika iz istog subreddit-a u odnosu na druge korisnike na temelju “matching score-a”. Matching score je mjera sličnosti dvaju korisnika koja se računa na način da se podijeli broj zajedničkih tokena prvog i drugog korisnika sa

brojem tokena prvog korisnika. Može se promijeniti da se dijeli sa brojem svih nepoznatih tokena. Ova matrica je temelj za hipotezu da se korisnici mogu samostalno grupirati glasanjem svojim nepoznatim tokenima. Nepoznati tokeni jednog korisnika se može nazvati privatnim vokabularom pošto ga jedino taj korisnik poznaje. Analizom nepoznatih tokena također se ustanovilo da odabir subreddit-a nije najbolji pošto jedan subreddit ima veći broj nepoznatih tokena. Ukoliko se nasumično odaberu tokeni, veća je šansa da se izaberu tokeni koji pripadaju jednom subredditu nego drugom, što nije pogodno za simulaciju.

Svi nepoznati tokeni su nasumično podijeljeni u klastere. Korisnik sebe smatra članom onog klastera u kojem je najviše tokena iz njegovog privatnog vokabulara. Korak simulacije je modeliran na sljedeći način:

1. Nasumično se odabere jedan korisnik
2. Korisnik bira nekoliko tokena koji nisu u njegovom klasteru
3. Korisnik pita sve korisnike slažu li se da se svi predloženi tokeni prebace u njegov klaster
4. Svi ostali korisnici glasaju na sljedeći način
    a. Ako se poveća njihov najveći matching score sa svim klasterima, pozitivno glasaj. To znači da je korisniku u koristi da se tokeni prebace
    b. Ako se ne promijeni njihov najveći matching score sa svim klasterima, neutralno glasaj. To znači da je korisniku svejedno
    c. Ako se smanji njihov najveći matching score sa svim klasterima, negativno glasaj. To znači da korisniku nije u koristi da se tokeni prebace
5. Ukoliko je pozitivan broj glasova svih korisnika, predloženi tokeni se prebacuju u klaster u kojem se nalazi odabrani korisnik

Za potrebe brze simulacije korisnici se interpretiraju kao n-bitni broj (u slučaju ovog projekta to je 8502-bitni broj, broj sa 8502 znamenke u binarnom zapisu). Ova reprezentacija korisnika omogućuje brzu usporedbu korisnika, prebacivanje tokena i ostale operacije pošto se koriste bitwise operacije koje su vrlo brze u usporedbi sa reprezentacijom tokena kao niz stringova. Na primjer, korisnik mora prebrojati zajedničke tokene u svakom klasteru da bi znao u kojem pripada i to može napraviti na način da napravi “bitwise or” nad sobom i svakim klasterom te prebroji koliko

jedinica se nalazi u tim binarnim zapisima i uzme klaster sa najvećim brojem jedinica u tom zapisu.

Rezultati simulacije nisu očekivani, pošto se dešavaju jedan od dva scenarija:

1. Korisnici nakon određene iteracije stagniraju u svojim klasterima
2. Svi se korisnice prebace u jedan klaster

Kako nastaviti ovaj projekt?

- Namjestiti okruženje u kojem su svi tokeni već odvojeni u 2 klastera sa 100% purity i pokrenuti simulaciju (klasteri moraju imat sposobnost zadržati svoje korisnike i ne se spojiti u jedan veliki klaster)
- Pronaći načine kako dobiti više hot-tokena (tokeni koji se pojavljuju 4 puta više u jednom subredditu u odnosu na drugi i barem 10 korisnika jednog subreddita ih ima)
    - Odabrati bolje subreddita sa specifičnim temama i manje generičnog vokabulara
    - Eliminirati ekstremne tokene (previše poojavljivanja)
    - Promijeniti broj tokena koji se nalazi u malom vokabularu tokenizatora (trenutno ih je 500)
    - Probati sa drugom vrstom tokenizatora
- Proces lemmatizacije vjerojatno nije konzistentan i gube se neki tokeni
- Mjenjanje hiperparametara simulacije, promjena tehnika glasanja, promjena broja klastera…
- Za unapređenje brzine simulacije implementirati bitwise operacije što je više moguće (recimo prebacivanje batch-a tokena se radi sa petljom gdje se prebacuje token po token, umjesto da se svi tokeni odjednom prebace tako da se reprezentiraju svi tokeni kao jedan broj koji se oduzme iz jednog klastera i doda u drugi)