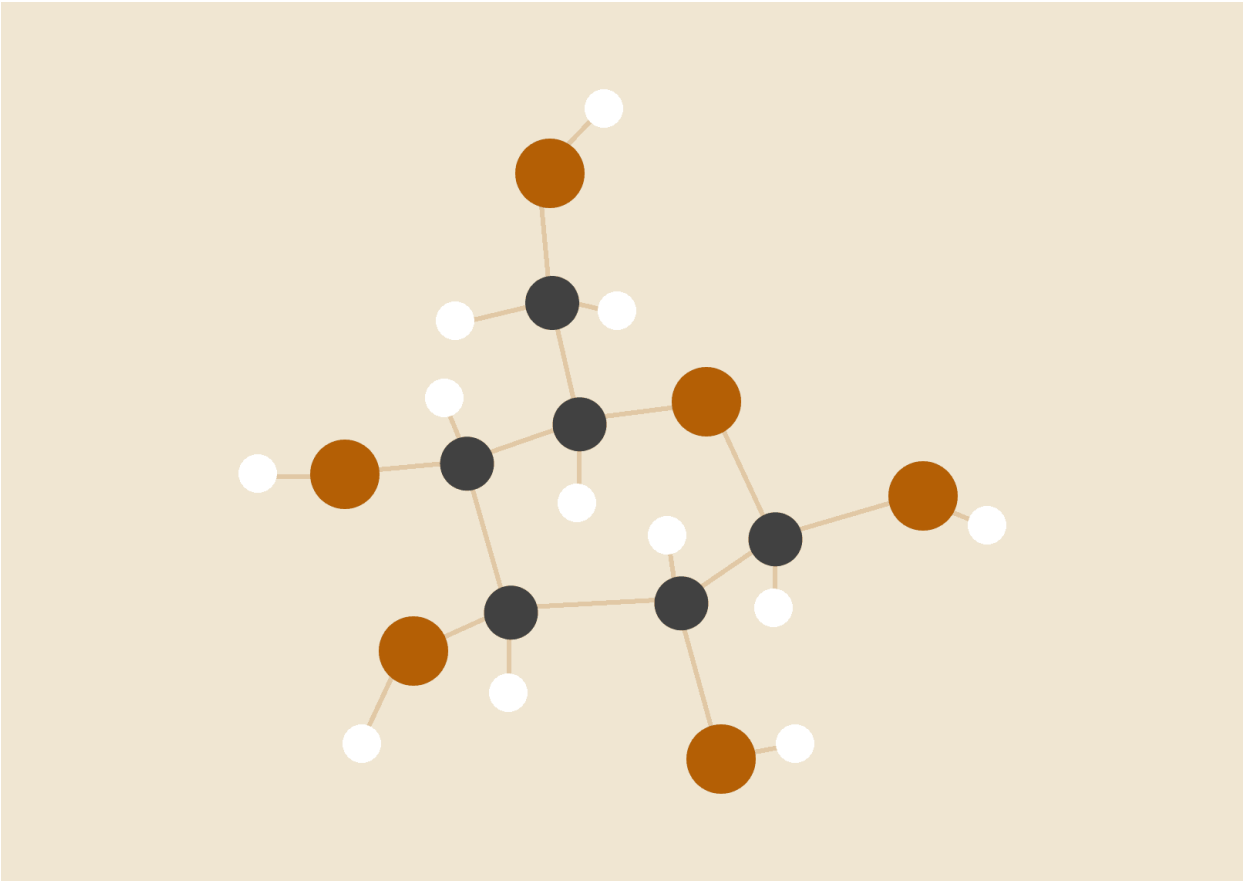
Govorne Tehnologije

Generisanje sadržaja wikipedije korišćenjem GPT-2 modela



Veljko Todorović E1 51/2023

Jovan Konjević E1 56/2023

29.04.2024

UVOD

Namera ovog projekta je da iskoristi arhitekturu modela GPT-2 za treniranje na sadržaju srpske Wikipedije, s fokusom na korišćenje ćiriličnog pisma. Cilj je istraživanje primene savremenih metoda generisanja teksta na srpskom jeziku, posebno na ćirilici.

IMPLEMENTACIJA

Preuzimanje podataka

Kao izvor sadržaja na srpskom jeziku i ćiriličnom pismu koristimo wikipediju. Inicijalni članak koji se koristi kao početna tačka za preuzimanje je Nikola Tesla. Sve veze na originalnom članku se obrađuju kako bi se preuzeli svi povezani članci na srpskoj Wikipediji. Na ovaj način ne dobijamo samo članak o Nikoli Tesli nego i sve referencijalne članke, kao da smo odradili pretragu u širinu za jedan korak. Ovako se dobija veliki skup povezanih artikala. Dobija se skup podataka od 4.1 milion tokena, što je za današnje modele izuzetno malo, ali za naš primer dovoljno.

Tokenizer

Tokenizer je ključni deo procesa obrade teksta u mašinskom učenju i obradi prirodnog jezika. On je odgovoran za razbijanje ulaznog teksta na diskretne jedinice, koje se obično u literature zovu tokeni, koje kasnije predstavljamo kao vektore i proslećujemo **GPT-2** modelu.

Za tokenizaciju teksta koristili smo **BPE (Byte Pair Encoding)** tokenizer, koji je efikasno obrađivao tekst na ćirilici. Ova tehnika se zasniva na iterativnom procesu spajanja najčešćih parova karaktera, čime se formiraju novi tokeni. Ovaj novi token se stavlja u listu svih pronadjenih tokena. Ovaj proces se ponavlja određeni broj iteracija ili dok se ne dostigne zadata veličina vokabular, u našem slučaju veličina vokabulara je **512 tokena**.

Veličina vokabulara određuje koliko različitih tokena može biti prisutno u našem modelu. Veći vokabular omogućava modelu da nauči širi spektar reči i pojmova, ali istovremeno zahteva više resursa za treniranje i obradu. Važno je pažljivo odabrati veličinu vokabulara i parametre tokenizacije kako bismo postigli balans između efikasnosti, performansi i resursa potrebnih za treniranje i obradu modela. Takođe, možemo eksperimentisati sa različitim tehnikama tokenizacije kako bismo optimizovali kvalitet generisanog teksta. Kao što je podela sekvence teksta na pod sekvence koje zadovoljavaju neka pravila. Mi delimo tekst na reči sa i bez razmakom na početku, specijalne karaketere kao što su otvorene zagrade, crtice,.., pojedinačne

cifre i sve ostale sekvence koje se ne svrstavaju u ove tri kategorije. Ovime imamo kontrolu nad spajanjem karaktera u token, tako što se ne mogu spojiti sekvence kao što cifre i karakteri u zaseban token.

GPT-2 model

Model koji je izabran za ovaj projekat je GPT-2 zbog svoje jednostavnosti i dostupnosti na internetu. Predstavlja i jednu od ključnih tačaka u upotrebi transformer arhitekture u generisanju pisanog sadržaja.

GPT-2 (Generative Pre-trained Transformer 2) je model dubokog učenja koji je razvila kompanija OpenAI. On predstavlja unapređenje prethodnog modela GPT-1, a njegova revolucionarnost leži u nekoliko ključnih karakteristika:

1. **Transformer arhitektura**: GPT-2 se zasniva na Transformer arhitekturi, koja je pokazala izuzetnu efikasnost u obradi prirodnog jezika. Ova arhitektura koristi mehanizam pažnje kako bi model mogao efikasno da obradi duge sekvence podataka.
2. **Pre-treniranje na velikom skupu podataka**: Pre GPT-2, modeli su obično trenirani na relativno malim skupovima podataka. Međutim, GPT-2 je treniran na izuzetno velikom skupu podataka, koji sadrži stotine gigabajta teksta sakupljenog sa interneta. Ovo omogućava modelu da nauči bogatije i opštije reprezentacije jezika.
3. **Unsupervised pre-treniranje**: GPT-2 se pre-trenira na neoznačenim podacima, što znači da se koristi neoznačeni tekst sa interneta za učenje modela. Ovo omogućava modelu da samostalno nauči reprezentacije jezika bez potrebe za ručnim označavanjem podataka.
4. **Generisanje teksta**: GPT-2 je model generativnog tipa, što znači da može generisati novi tekst na osnovu ulaznog teksta. Ovo je postignuto tako što je model treniran tako da predviđa sledeću reč u sekvenci teksta na osnovu prethodnih reči.
5. **Različite primene**: GPT-2 može biti korišćen za različite zadatke u obradi prirodnog jezika, kao što su generisanje teksta, prevođenje jezika, odgovaranje na pitanja, analiza sentimenta, i druge.

PARAMETRI MODELA I TRENINGA

Veličina vokabulara je postavljena na 512 tokena, što omogućava modelu da obuhvati raznovrsnost srpskog jezika. Sam model ima arhitekturu sa 3 sloja, svaki sa 4 glave i vektorskom reprezentacijom tokena od 256 dimenzija. Kako bi se izbeglo prenaučenost, primenjen je dropout sa verovatnoćom od 0.1. Takođe, u modelu je isključen bias u slojevima Linears i LayerNorms, što doprinosi bržem i efikasnijem učenju.

Što se tiče konfiguracija za treniranje, skup podataka je podeljen na trening (80%), test (10%) i validaciju (10%). Veličinu jednog batch-a je 64. Broj iteracija za treniranje je 20000, uz postepeno povećanje i smanjenje stope učenja (learning rate) sa maksimalnom vrednošću od 6e-3 i minimalnom vrednošću od 6e-4.

REZULTATI

Kao metriku za evaluaciju modela koristimo **Perplexity**. Perplexity (perpleksnost) je mera koja se često koristi u evaluaciji jezičkih modela, uključujući i modele za generisanje teksta kao što je GPT-2. Ova mera pruža uvid u to koliko dobro model predviđa sledeću reč u sekvenci teksta.

Formalno, perplexity je recipročna vrednost verovatnoće modela za generisanje testnog skupa podataka. Niža vrednost perplexity-ja ukazuje na to da je model bolje generalizovao i predstavio distribuciju verovatnoća reči u skladu sa stvarnim podacima.

Perplexity se izračunava na sledeći način:

$$PPL = exp(\frac{1}{N}\sum_{i=1}^{N} logP(w_i))$$

Niža vrednost perplexity-ja ukazuje na to da model bolje predviđa sledeće reči u sekvenci, dok viša vrednost može ukazivati na to da model nije dobro generalizovao distribuciju reči u skladu sa stvarnim podacima.

U ovom istraživanju, perplexity se koristi kao ključna metrika za evaluaciju performansi GPT-2 modela na srpskoj Wikipediji. Model treniran nad našim podacima i testiran nad test skupom podataka daje **perplexity od 13.198796.**

REFERENCE

1. https://github.com/karpathy/nanoGPT
2. https://github.com/karpathy/minbpe
3. https://huggingface.co/docs/transformers/en/perplexity