

Generiranje imena naselja pomoću LSTM mreže

Čogelja, Granić, Lubina, Jurković, Juvančić, Logarušić

10. siječnja 2025.

prijedlog

(I.) Ishod projekta

Ishod projekta je LSTM rekurzivna neuronska mreža na razini znakova koja generira realistična imena hrvatskih naselja.

Mreža radi sa vektorima koji predstavljaju slova hrvatske abecede proširene specijalnim znakovima $\Sigma = \{\text{hrv. abeceda}\} \cup \{\langle start \rangle, " \setminus 0" \}$.
Ulaz mreže je one-hot vektor $\mathbf{x}^{(t)}$ dimezije $|\Sigma| = 30 + 2$.

$$\mathbf{x}_i^{(t)} = \begin{cases} 1, & \text{ako } i = j \\ 0, & \text{inače} \end{cases} \quad (1)$$

Izlaz dobiven na kraju pojedinog vremenskog koraka t je vektor vjerojatnosti pojave pojedinog znaka abecende.

$$\hat{\mathbf{y}}^{(t)} = \begin{bmatrix} p(c_0) \\ p(c_1|c_0) \\ \vdots \\ p(c_{|\Sigma|-1} | \bigcap_{i=0}^{|\Sigma|-2} c_i) \end{bmatrix} \quad \text{Gdje } c \in \Sigma \quad (2)$$

Vjerojatnosti su dobivene softmax funkcijom parametriziranom hiperparametrom temperature τ .

Na temelju tih vjerojatnosti se uzorkuje konačni izlazni vektor $\mathbf{y}^{(t)}$, odnosno t-ti znak u imenu naselja.

$$\mathbf{y}^{(t)} \sim \hat{\mathbf{y}}^{(t)} = \sigma_{\tau}(f(\mathbf{x}^{(t)}; \boldsymbol{\theta})) \quad (3)$$

$f(\mathbf{x}; \boldsymbol{\theta})$ predstavlja ukupno djelovanje ćelija modela nad njenim ulazom parametrizirano hiperparametrima modela $\boldsymbol{\theta} = [|\mathbf{a}| \quad \mu \quad \tau]$ (opisani u poglavlju ??)
 Temperaturno uzorkovanje je izabrano, jer omogućava eksperimentiranje i generiranje zanimljivih toponima.

Izlaz mreže je niz znakova $\{\mathbf{y}^{(t)}\}_{t=0}^{T-1}$, odnosno ime naselja.

Željena točnost modela η je $\lim_{\tau \rightarrow 0} \eta = 0.5$

(II.) Tema i kratki opis

Fokus projekta je treniranje i razvijanje neuronske mreže za generiranje realističnih imena hrvatskih naselja. Korištenjem LSTM mreže, koja je prilagođena za analizu sekvencijskih podataka, cilj je razviti model sposoban za učenje jezičnih obrazaca i struktura iz postojećih imena naselja. Svrha mreže je generiranje novih imena temeljenih na tim naučenim obrascima, pri čemu se zadržavaju jezične i strukturne zakonitosti specifične za taj kontekst. U planu je karakterizirati mrežu sa sljedećim hiperparametrima:

1. Dimenzija skrivenog stanja: $|\mathbf{a}|$
2. Stopa učenja: μ
3. Temperatura: τ
4. Broj LSTM ćelija

LSTM ćelija i mreža će biti implementirane u radnom okviru pyTorch.
 Dizajn mreže i podešavanje hiperparametara se odvija paralelno sa implementacijom mreže u radnom okviru Keras.
 Točan izgled ćelije i dizajn mreže će biti određeni naknadno.

BPTT će biti korišten kao algoritam učenja.
 Funkcija pogreške će biti određena naknadno.

(III.) Zadatci na projektu i raspodjela posla

Ostvarenje projekta podrazumijeva slijedeće zadatke:

Zadatak	Podzadaci	Članovi
Dokumentacija	<ul style="list-style-type: none"> • usporedba rezultata • arhitektura • ažurirat po potrebi 	Lubina, Čogelja
Prezentacija		–
Tokenizacija	<ul style="list-style-type: none"> • Proširiti Σ start tokenima pojedinih jezika 	Granić
Implementacija	<ul style="list-style-type: none"> • Implementacija modela u Kerasu • Treniranje • Optimizacija hiperparametara ■ Naci kako se u kerasu može tokenizirat ulazna abeceda ■ Naci kako se može mijenjati dimenzija skrivenog stanja LSTM ćelije ■ Output mora biti vjerojatnost slova, a ne one-hot. 	Jurković, Juvančić, Logarušić
Custom ćelija	<p>Moguće je da ćemo morat raditi custom ćeliju, ako:</p> <ol style="list-style-type: none"> 1. se ne mogu drugačije postići zahtjevi iz prethodnog retka 2. nam kaže asistent da moramo još "zakomplikirati" projekt 	–

Tablica 1: Podjela posla

(IV.) Vremenski plan rada

U priloženoj tablici prikazan je okviran plan rada koji obuhvaća ključne datume i zadatke koji su planirani u sklopu projekta. Rokovi su estimirani uzimajući u obzir praznike i ispitne rokove (MI i ZI) kako bi se mogao predvidjeti realan tok aktivnosti. Rokovi su isto tako fleksibilni zbog akademskih obaveza članova grupe što osigurava aktivno sudjelovanje svih članova. Mogući iterativni postupci promjene dizajna mreže i/ili ćelije te ispravljanje grešaka nisu bili mogući za opisati, ali su uzeti u obzir kao i, ispravljanje raznih grešaka.

Rok	30.10.	15.12.	6.1.	15.1.
Zadaci	1. Početak rada	1. Obrada ulaznog skupa podataka 2. Dizajn modela u kerasu 3. Implementacija modela u TensorFlow-u 4. Treniranje modela	1. Validacija modela 2. Podešavanje hiperparametara u kerasu	1. Pisanje dokumentacije 2. Priprema prezentacije

Tablica 2: Planirani tok rada na projektu