1. Prva laboratorijska vježba iz kolegija Duboko učenje

Cilj i opis laboratorijske vježbe

Cilj laboratorijske vježbe je upoznavanje s osnovnim bibliotekama za obradu numeričkih tipova podataka kroz nešto kompleksniji primjer nego što je to bio na predavanjima. Primjer raščlanjivanja kakav je dan u nastavku se često koristi u praksi i na njega će se vezati vježba kada će se raditi generiranje teksta.

Upute za rješavanje laboratorijske vježbe

Potrebno je učitati podatke nasumično odabranih 10 knjiga iz Gutenberg seta podataka knjiga koji je moguće pronaći na: https://shibamoulilahiri.github.io/gutenberg_dataset.html.

Nakon navedenog, tekst svih knjiga potrebno je spojiti i nakon toga raščlaniti na slova slično kao što je prikazano u trećem predavanju i četvrtom primjeru. Nakon što se knjige raščlane na slova, potrebno je kreirati ulaze za treniranje neuronske mreže. Niz je potrebno raščlaniti na način kao da neuronsku mrežu želimo trenirati na nizu podataka od 50 znakova. Svaki novi ulaz je pomaknut za 5 znakova dalje.

Dakle, ako postoji sljedeći niz (koristimo samo brojeve od 0-9 zbog preglednosti):

```
12345678946547687416876499842284842165987626595568541267841154878125656823 2
```

Prvi niz će biti:

```
12345678946547687416876499842284842165987626595568
```

Drugi će biti:

```
67894654768741687649984228484216598762659556854126
```

I tako dalje dok se ne dođe do kraja seta podataka. Ukoliko se na kraju dogodi preljev, slobodno se može ukloniti zadnji niz.

Nakon što se nizovi pohrane u Python listu, iste je potrebno preoblikovati u NumPy ndarray. To je potrebno napraviti na nešto drugačiji način. Svako slovo, odnosno ulaz u svaku neuronsku mrežu je vektor. Potrebno je pretvoriti broj slova u specifičan vektor koji se zove one hot encoded vektor. Taj vektor nije ništa pametnije nego binarni vektor koji na određenoj poziciji označava neko slovo. One hot encoded vektor je jedan od načina za označavanje kategoričkih podataka. U ovom slučaju svako slovo u tekstu je jedna kategorija, odnosno kategorički podatak. Ukoliko postoji ukupno 100 različitih znakova, taj vektor mora biti dugačak 100, a na poziciji gdje je zapravo to slovo mora biti vrijednost True, dok su ostale vrijednosti False.

Na primjer:

Imamo ukupno 10 različitih slova:

```
abcdefghij
```

Slova su reprezentirana indeksima:

```
0123456789
```

Vektori će izgledati kao u tablici:

a	b	c	d	e	f	g	h	i	j
False	False	False	False	False	False	False	False	False	True
False	False	False	False	False	False	False	False	True	False
False	False	False	False	False	False	False	True	False	False
False	False	False	False	False	False	True	False	False	False
False	False	False	False	False	True	False	False	False	False
False	False	False	False	True	False	False	False	False	False
False	False	False	True	False	False	False	False	False	False
False	False	True	False	False	False	False	False	False	False
False	True	False	False	False	False	False	False	False	False
True	False	False	False	False	False	False	False	False	False

Ako želimo tim istim vektorima napisati riječ abeceda, to će izgledati kao u tablici:

a	b	e	c	e	d	a
False	False	False	False	False	False	False
False	False	False	False	False	False	False
False	False	False	False	False	False	False
False	False	False	False	False	False	False
False	False	False	False	False	False	False
False	False	True	False	True	False	False
False	False	False	False	False	True	False
False	False	False	True	False	False	False
False	True	False	False	False	False	False
True	False	False	False	False	False	True

Dakle, iste takve tablice (odnosno ndarray u NumPy biblioteci) je potrebno dobiti za tekstove odabranih knjiga. U potpunosti je svejedno koje će knjige biti odabrane.

Ndarray mora imati sljedeću veličinu: (n, 50, m)

- n – označava broj nizova od po 50 koji ćete dobiti iz podataka
- m – označava broj različitih slova, odnosno duljinu vektora

Nakon navedenog, potrebno je navedenu višedimenzionalni vektor pohraniti na tvrdi disk. Prilikom pokretanja programa ukoliko postoji pohranjena datoteka, podatke je potrebno pročitati iz nje, a ukoliko nije napraviti postupak raščlanjivanja koji je opisan prethodnim koracima. Primijetiti razliku u vremenu potrebnom za raščlanjivanje i učitavanje iz datoteke.