

Specifikacija projekta iz predmeta Softcomputing

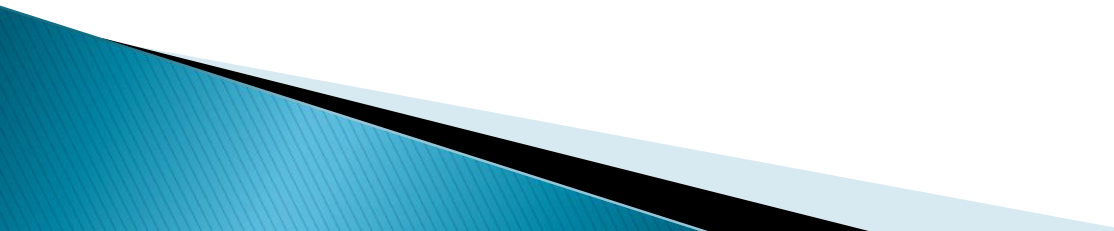
TEMA: Klasifikacija tekstova po
sentimentu

Luka Pavlica RA8/2012

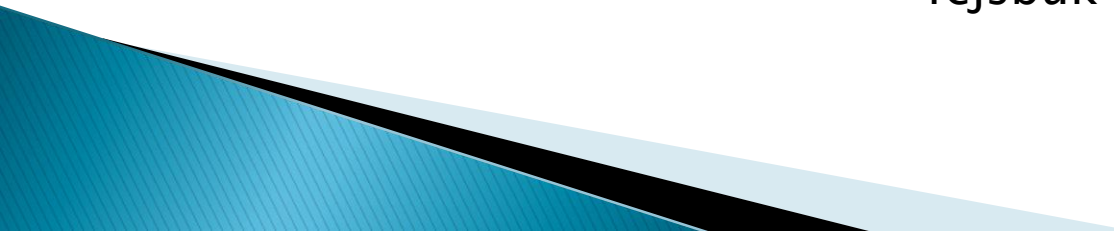
UVOD

Danas neuronske mreže predstavljaju veoma atraktivnu oblast istraživanja i postoje brojne oblasti u kojima se koriste.

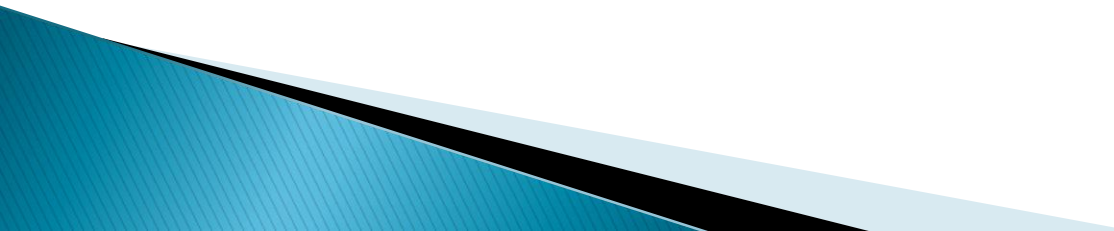
Primenjuju se za:


- prepoznavanje oblika
 - prepoznavanje rukopisa
 - prepoznavanje govora
 - upravljanju robota
 - analizi tekstova
 - i u raznim drugim oblastima
- 

Projektni zadatak

- Ovaj projekat je jedna vrsta analize teksta. Na osnovu analize datog teksta, klasifikuje se po sentimentu
 - Test po sentimentu se deli na pozitivan, negativan i neutralan tekst
 - Slična rešenja u ovoj oblasti su:
 1. klasifikacija tekstova po tematici (politika, sport, zabava),
 2. o kom sportu je reč,
 3. koji film, ili glumac se opisuje u tekstu,
 4. klasifikacija da li je novinski članak, ili fejsbuk status, itd...
- 

Koraci implementacije

- ❑ Za početak treba slikati par kratkih tekstova i te slike prebaciti u onaj repozitorijum gde će se nalaziti i kod
 - ❑ Tekst će se sastojati od proizvoljnog broja redova, a pismo će biti srpska latinica, ali će postojati i engleska slova(q, x, w, y)
 - ❑ Postojeće implementacija rotiranja teksta, tj svakog slova za određeni ugao radi lakšeg prepoznavanja slova od strane neuronske mreže
 - ❑ Pored teksta se mogu naći i fotografije koje treba da budu izbačene, tj. neuronska mreža neće obraćati pažnju na njih
 - ❑ U slučaju da se neko slovo sastoji od dve konture, postojaće funkcionalnost koja spaja određene konture
- 

- ❑ Tekst neće biti idealno fotografisan, tj. postojaće razni šumovi koji će se morati odstraniti da bi se izdvojile konture od značaja
 - ❑ Alfabet će se sastojati od slova, kao i od svih mogućih znakova interpunkcije
 - ❑ Moraće se utvrditi tačan razmak između slova, reči i novog reda
 - ❑ Na kraju će biti ispisano da li je tekst pozitivan, negativan, ili neutralan
 - ❑ Biće prikazano par slika tekstova jednog fonta. Otprilike za svaku vrstu po 2 teksta
 - ❑ Osim srpskog jezika bice omogućen i engleski (zato će neuronska mreža i učiti slova q, y, w...)
- 

IDEJA

Napomena: Tekstovi neće biti komplikovani i teški za razumevanje.

Postojeće 2 brojača čije su inicijalno vrednosti nula. Na osnovu k-means algoritma utvrđivaće se razmak između reči. Kada se gleda reč, na osnovu prvih par slova će se vršiti analiza. Takođe će se vršiti analiza za par poslednjih slova

Na primer ako imamo neku reč i ako su prva 3 slova te reči “pob” velika verovatnoća da je reč vrednosti “pobeda”, ili neka slična reč. Dalje uzimamo par poslednjih slova i testiramo da li se završava reč na “da”. Ukoliko se završava povećavamo prvi brojač

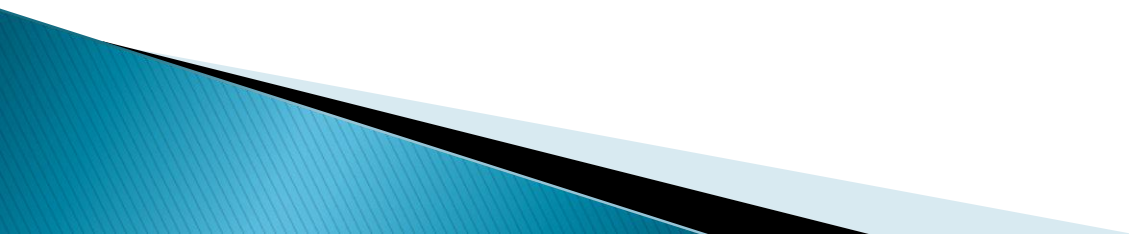
Sa druge strane ako su prva 2 slova neke reči “ne” verovatnoća je da je ta rečenica u negativnom sentimentu i time se povećava drugi brojač

U zavisnosti kolika je dužina rečenice i koji je jezik (engleski, ili srpski), uzimaće se u obzir određeni broj prvih, ili poslednjih slova



Ako je rečenica od samo 3 karaktera verovatno će se analizirati cela reč i vršiti komparacija sa nekim rečima, npr. "win"

Na kraju u zavisnosti od vrednosti brojača će biti utvrđeno o kakvom tekstu se radi. Ukoliko su vrednosti brojača dosta približnih vrednosti radiće se o neutralnom tekstu.



KRAJ