nedeljainformatike v6.0

Obrada prirodnih jezika (NLP)

Mina Šekularac

Matematička gimnazija

22. 04. 2021.

Pitanja

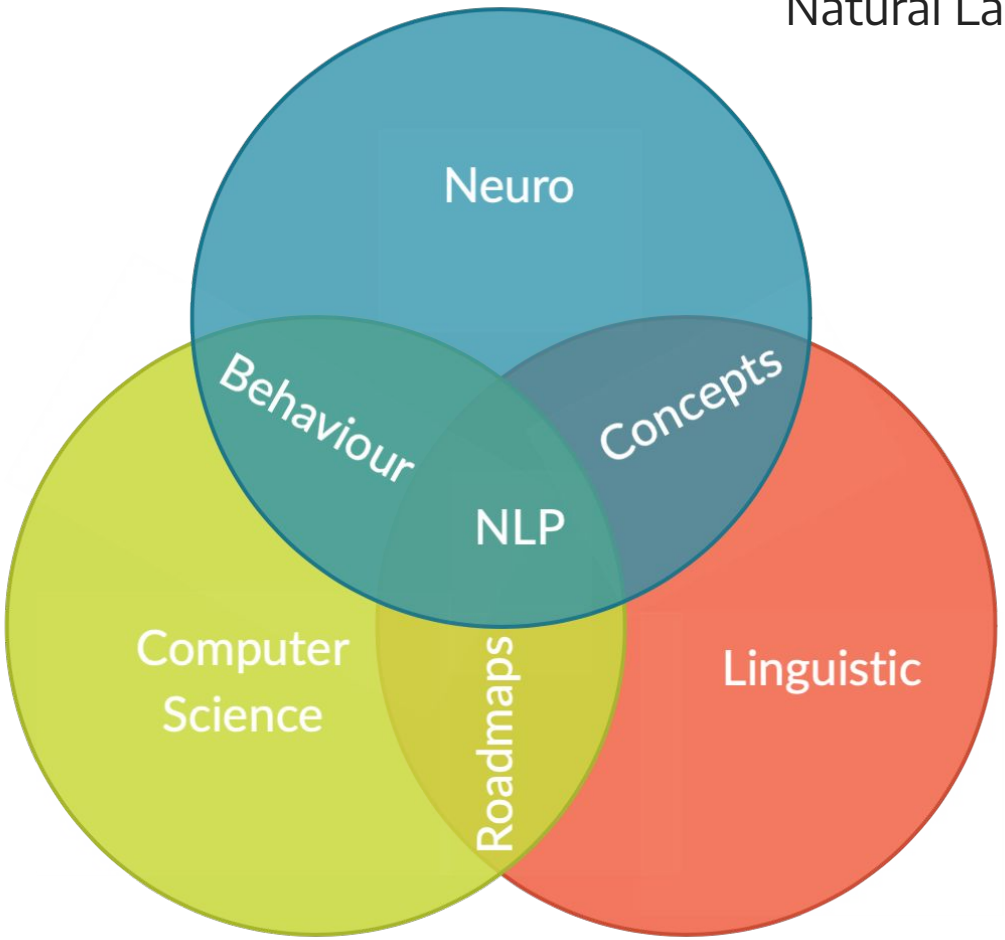
1. Šta su prirodni jezici?
2. Čemu služi NLP?
3. Graf od rečenice, kako?
4. Koje je rastojanje između četvrtka i petka?
5. Zašto baš NLP?

Agenda

1. Pojmovi
2. Problemi
3. Lingvistička analiza teksta
4. Algoritmi
+ Bonus

Šta je NLP?

Obrada prirodnih jezika
Natural Language Processing

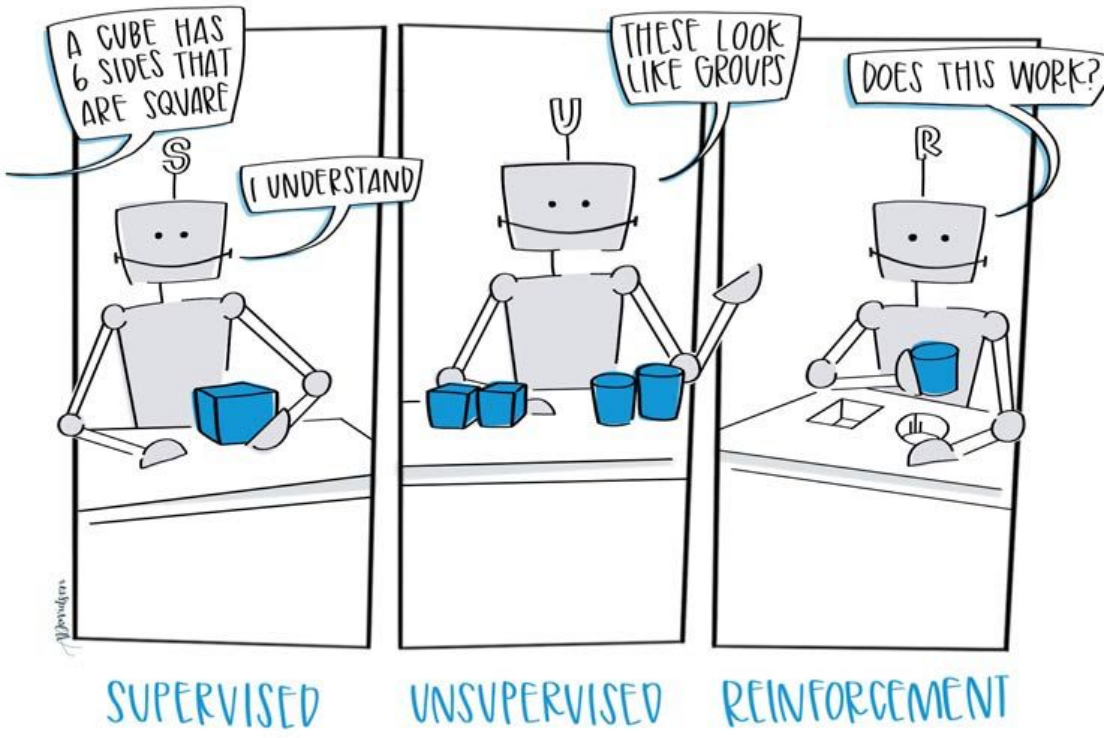


Šta je NLP?

- Prirodni jezici
  - srpski, engleski, italijanski, hindi, ...
- Popularni primeri:
  - Masinsko prevođenje
    - Google Translate
  - Licni asistenti
    - Siri/ Google Assistant/ Cortana/ Alexa
  - Ispravka pogrešno otkucanih reči (*spell checker*)
  - Detekcija neželjene pošte (*spam*-a)

Mašinsko učenje

- Podoblast **veštačke inteligencije** (*AI*)
- Računar sam uči iz datih podataka
- Vrste učenja:
  - Nadgledano učenje (*supervised learning*)
  - Nenadgledano učenje (*unsupervised learning*)
  - Učenje uslovljavanjem (*reinforcement learning*)



NLP pojmovi

- **Korpus** - trening set
  - SrpKor - 4925 tekstova
  - Članci iz dnevnih novina, magazina, delovi knjiga, tekstovi sa internet portala
  - Paralelni korpusi - Eng/srp (Bibliša)
- **Vreća reči**
- **Lemma** reči
- **Token**
- **Etiketirati** (*tagging*)
- **Parsirati**
- **Vektori značenja reči** (*word embeddings*)

Agenda

1. Pojmovi
2. Problemi
3. Lingvistička analiza teksta
4. Algoritmi

\+ Bonus

Čime se bavi NLP?

- Podaci:
  - Sirovi text u prirodnom jeziku (*raw text*)
  - Knjige, časopisi, novine, društvene mreže, e-mailovi
- Problem kompleksnosti jezika:
  - Dvosmislene reči (čas)
  - Idiomatski izrazi (sekund)
  - Razumevanje teksta pomocu šireg znanje o svetu (*world knowledge*)
- Razumevanje texta je AI-potpun problem
  - Zahteva da računar dostigne ljudski nivo inteligencije

14 ŠOU Sreda, 11. novembar 2009.

VEST

Dajan Kruger glumi Srpkinju

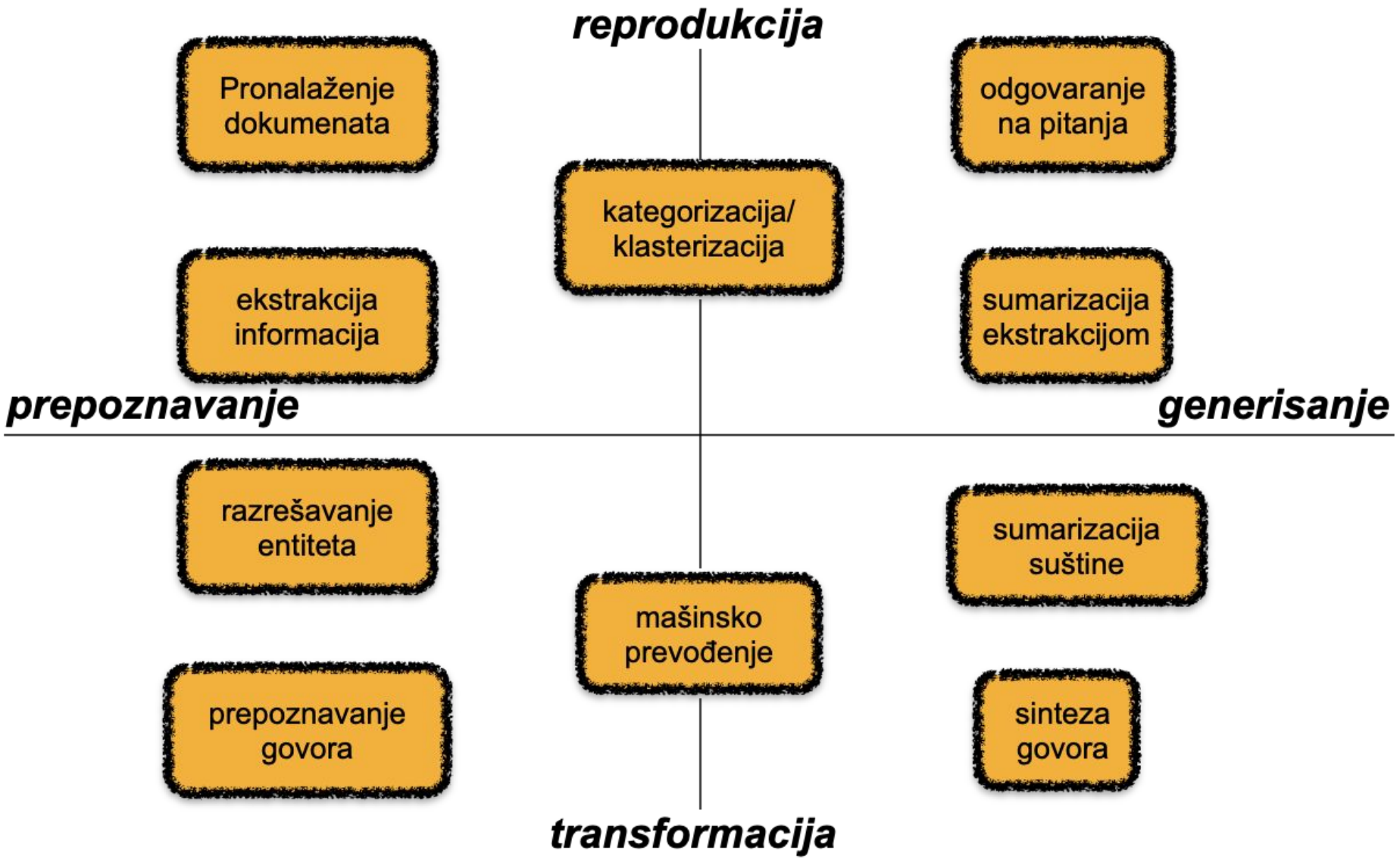
AKCENAT Posle uloge u filmu Kventina Tarantina „Inglorious Basterds", u kom je glumila na nemačkom i engleskom jeziku, glumica Dajan Kruger se sprema za snimanje trilera „Unknown White Male" u kom će tumačiti imigrantkinju iz Srbije. Partner

Greh njene majke preti da obori Ranjenog orla

Prvu epizodu nove televizijske serije „Greh njene majke"

Kompleksnija priča od prethodne

Zadaci kojima se bavi NLP?



Kompleksnost jezika

- Morfološka
  - Više različitih oblika jedne iste reči
  - Padezi, menjanje glagola po vremenu i licu
- Sintaktička
  - Različite sintaktičke strukture za jedan skup reči
- Semantička
  - Različita značenja za jednu istu reč/iskaz
    - Idiomatski izrazi
    - Frazeologizmi

Višesmislenost reči

- Gore gore gore gore.
  - Brda iznad lošije gore.
    - Imenica, glagol, pridev i prilog
    - Kategorijska dvosmislenost
  - Lošija brda gore iznad.
    - Sintaktičke dvosmislenosti

- Polisemija - glava
- Homonimija - sto

Agenda

1. Pojmovi
2. Problemi
3. Lingvistička analiza teksta
4. Algoritmi

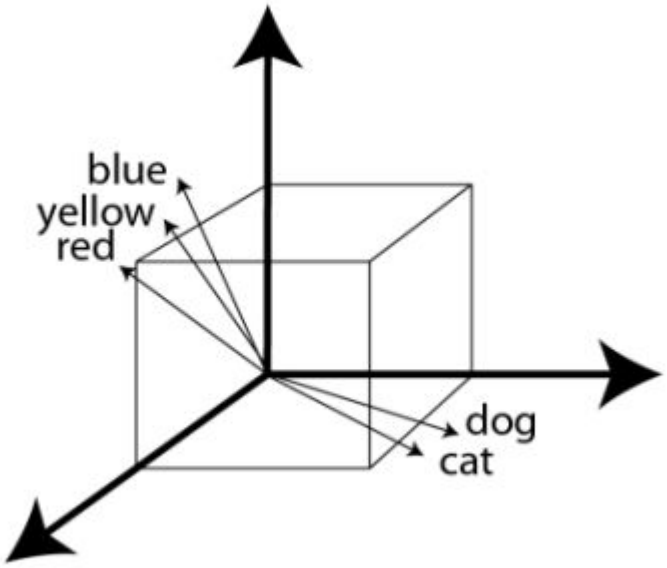
\+ Bonus

Lingvistička analiza

1. **Tokenizacija**
   - Podela sirovog teksta u pasuse
   - Podela pasusa u rečenice
   - Podela rečenice u **tokene** (pojedinačne reći i simbole)
   - Regulatni izrazi
2. **Etiketiranje** (*tagging*)
   - Vrsta reči
3. **Parsiranje**
   - Gramatička analiza

Token

- Raspon (*span*)
- Lemma ili osnovni oblik reči (*lemma*)
  - Padeži
  - Glagoli po vremenima
- Vektor reči (*word embedding*)
- Tag vrste reči (*Part-of-speech tag*)



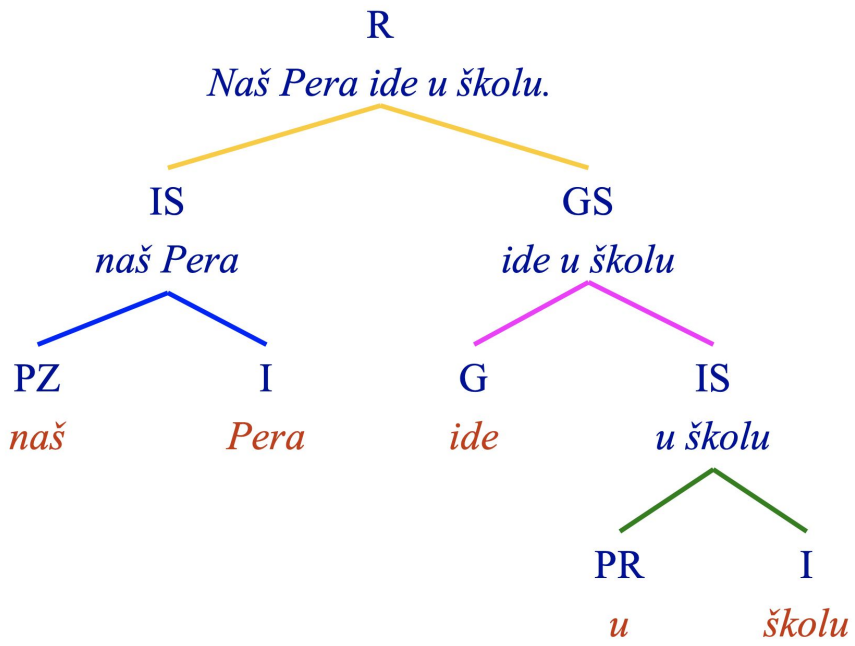
word2vec

Naš	pridevska zamenica
Pera	imenica
ide	glagol
u	predlog
školu	imenica

PoS tag

Obrada rečenice

- Frazne strukture
  - imeničke i glagolske **sintagme** (*noun-verb phrase*)
- Parsiranje
  - Stablo parsiranja (*parsing tree*)



Agenda

1. Pojmovi
2. Problemi
3. Lingvistička analiza teksta
4. Algoritmi
+ Bonus

Rastojanja između stringova

	NLP	Bioinformatika
Problem	Rastojanje između stringova	Najveća sličnost između sekvenci
Sekvence	Stringova	DNK
Alfabet	Slova i simboli	Nukleotide
Algoritam	Levenshtein algoritam	Needleman-Wunsch algoritam

- Poređenje reči:
  - Informatike
  - Informacijama
- Rastojanje između slova se računa pomoću funkcije rastojanja pri čemu:
  - Ista slova - 0
  - Razmak - 1
  - Različita slova - [0, 1]

i	n	f	o	r	m	a	t	i	k	e	–	–
|	|	|	|	|	|	|	*	|	*	*		
i	n	f	o	r	m	a	c	i	j	a	m	a

DP algoritam

- Dinamičko programiranje
- Pseudo kod:
  - Poređenje stringa x i y (len(x) = N i len(y) = M)
  - Inicijalizacija:

$$D(0,0) = 0$$

$$D(i,0) = D(i-1,0) + f(brisanja);\ 1 < i \leq N$$

$$D(0,j) = D(0,j-1) + f(dodavanja);\ 1 < j \leq M$$

  - Rekurentna relacija (indukcioni korak):

$$D(i,j) = min \begin{cases} D(i-1,j) + f(brisanja) \\ D(i,j-1) + f(umetanja) \\ D(i-1,j-1) + f(x_i, y_j) \end{cases}$$

  - Završetak:

$D(N,M)$ je distanca

DP algoritam

	–	č	e	t	v	r	t	a	k
–	0	1	2	3	4	5	6	7	8
p	1								
e	2								
t	3								
a	4								
k	5								

f(x,x) = 0
f(x,y) = 1, x≠y
f(umetanja) = 1
f(brisanja) = 1

DP algoritam

	–	č	e	t	v	r	t	a	k
–	0	1	2	3	4	5	6	7	8
p	1	1	2	3	4	5	6	7	8
e	2	2							
t	3	3							
a	4	4							
k	5	5							

f(x,x) = 0
f(x,y) = 1, x≠y
f(umetanja) = 1
f(brisanja) = 1

DP algoritam

	–	č	e	t	v	r	t	a	k
–	0	1	2	3	4	5	6	7	8
p	1	1	2	3	4	5	6	7	8
e	2	2	1	2	3	4	5	6	7
t	3	3	2						
a	4	4	3						
k	5	5	4						

f(x,x) = 0
f(x,y) = 1, x≠y
f(umetanja) = 1
f(brisanja) = 1

DP algoritam

	–	č	e	t	v	r	t	a	k
–	0	1	2	3	4	5	6	7	8
p	1	1	2	3	4	5	6	7	8
e	2	2	1	2	3	4	5	6	7
t	3	3	2	1	2	3	4	5	6
a	4	4	3						
k	5	5	4						

f(x,x) = 0
f(x,y) = 1, x≠y
f(umetanja) = 1
f(brisanja) = 1

DP algoritam

	–	č	e	t	v	r	t	a	k
–	0	1	2	3	4	5	6	7	8
p	1	1	2	3	4	5	6	7	8
e	2	2	1	2	3	4	5	6	7
t	3	3	2	1	2	3	4	5	6
a	4	4	3	2	2	3	4		
k	5	5	4	3	3	3	4		

f(x,x) = 0
f(x,y) = 1, x≠y
f(umetanja) = 1
f(brisanja) = 1

DP algoritam

	–	č	e	t	v	r	t	a	k
–	0	1	2	3	4	5	6	7	8
p	1	1	2	3	4	5	6	7	8
e	2	2	1	2	3	4	5	6	7
t	3	3	2	1	2	3	4	5	6
a	4	4	3	2	2	3	4	4	5
k	5	5	4	3	3	3	4	5	

f(x,x) = 0
f(x,y) = 1, x≠y
f(umetanja) = 1
f(brisanja) = 1

DP algoritam

	–	č	e	t	v	r	t	a	k
–	0	1	2	3	4	5	6	7	8
p	1	1	2	3	4	5	6	7	8
e	2	2	1	2	3	4	5	6	7
t	3	3	2	1	2	3	4	5	6
a	4	4	3	2	2	3	4	4	5
k	5	5	4	3	3	3	4	5	4

f(x,x) = 0
f(x,y) = 1, x≠y
f(umetanja) = 1
f(brisanja) = 1

DP algoritam

	–	č	e	t	v	r	t	a	k
–	0	1	2	3	4	5	6	7	8
p	1	1	2	3	4	5	6	7	8
e	2	2	1	2	3	4	5	6	7
t	3	3	2	1	2	3	4	5	6
a	4	4	3	2	2	3	4	4	5
k	5	5	4	3	3	3	4	5	4

č	e	t	v	r	t	a	k
*	|	|	–	–	–	|	|
p	e	t				a	k

DP algoritam

	-	č	e	t	v	r	t	a	k
-	0	1	2	3	4	5	6	7	8
p	1	1	2	3	4	5	6	7	8
e	2	2	1	2	3	4	5	6	7
t	3	3	2	1	2	3	4	5	6
a	4	4	3	2	2	3	4	4	5
k	5	5	4	3	3	3	4	5	4

	č	e	t	v	r	t	a	k
*	|	–	–	–	|	|	|
p	e				t	a	k	

Rastojanje između stringova

Primeri:

- Autocomplete
- Spell check

Zadatak:

- Napravi svoj spell checker

Agenda

1. Pojmovi
2. Problemi
3. Lingvistička analiza teksta
4. Algoritmi

\+ Bonus

Napredak NLP-a

Primer mašinskog prevođenja (Google Translate) odlomka kroz godine:

- Original:
  but perhaps it was because they were habituated to the finer performances of the London stage, which she knew, on Isabella's authority, rendered everything else of the kind "quite horrid."

- 2013:
  ali možda je to zato što su bili habituated da finije performanse Londonu fazi, koja je znala, na vlasti je Izabela, donosi sve ostalo tog tipa "prilično strašan."

- 2015:
  ali možda je to bilo zato što su navikli na finije performansi na sceni Londonu, koja je poznavala, na Isabella autoritet, donio sve ostalo te vrste "prilično užasni."

Napredak NLP-a

Primer mašinskog prevođenja (Google Translate) odlomka kroz godine:

- Original:
  but perhaps it was because they were habituated to the finer performances of the London stage, which she knew, on Isabella's authority, rendered everything else of the kind "quite horrid."

- 2017:
  ali možda je to bilo zato što su boravile u tačnija performansi na londonskoj sceni, koja je znala, na Isabella autoritet, doneo sve drugo te vrste "prilično užasni.„

- 2019:
  ali možda je to bilo zbog toga što su bili naviknuti na bolje predstave londonske scene, na kojoj su se, kako je znala iz Izabelinog uveravanja, davali svakakvi "užasni" komadi.

Napredak NLP-a

Primer mašinskog prevođenja (Google Translate) odlomka kroz godine:

- Original:
  but perhaps it was because they were habituated to the finer performances of the London stage, which she knew, on Isabella's authority, rendered everything else of the kind "quite horrid."

- 2021:
  ali možda je to bilo zato što su bili naviknuti na finije predstave na londonskoj sceni, za koje je ona znala da su, po Isabelinom autoritetu, sve ostalo te vrste učinile „prilično užasnim“.

tl;dr - Zašto NLP?

- Simplifikacija teksta
- Mašinski prevod
- Detekcija plagijata
- Odgovori na pitanja
- Ekstrakcija ključnih reči
- Četbotovi
- Sumarizacija suštine
- Pametni asistenti
- Prediktivni tekst
- Analiza sentimenta

Literatura

- “Speech and Language Processing”
  - Daniel Jurafsky i James H. Martin
  - https://web.stanford.edu/~jurafsky/slp3/

Hvala na pažnji!

Pitanja?