

Analiza teksta

Jelena Matković,
Novi Sad, 2024

Šta je analiza teksta?

- Proces ekstrakcije značajnih informacija iz tekstualnih sadržaja
- Obuhvata različite tehnike iz oblasti:
 - lingvistike
 - mašinskog učenja
 - statistike
 - obrade prirodnog jezika
 - dubokog učenja

Šta je analiza teksta?

- Ciljevi:
 - otkrivanje šablona
 - prevođenje nestrukturiranih podataka u strukturirane podatke
 - upotreba podataka za dalju mašinsku obradu i statistiku
 - izvođenje zaključaka
- Upotreba u oblasti prava?

Text analysis ↔ Text mining ↔ Text analytics?

- Analiza teksta obuhvata i manuelnu analizu tekstualnih sadržaja
- Rudarenje teksta odnosi se na primenu tehnika za obradu prirodnog jezika sa ciljem:
 - Otkrivanja obrazaca
 - Ekstrakcije podataka
- Analitika teksta obuhvata i kvantitativnu analizu i statističke prikaze
- Pojmovi danas gotovo izjednačeni

Šta je Natural Language Processing (NLP)?

- Interdisciplinarna oblast računarstva i lingvistike koja, između ostalog, omogućava računarima da razumeju, interpretiraju i generišu sadržaje prirodnog jezika
- Obuhvata tehnike za obradu prirodnog jezika u pisanoj ili audio formi

Šta je Natural Language Processing (NLP)?

- Rule-based metode:
 - Definisani setovi pravila
 - Bez učenja podataka
 - Eksplicitna uputstva
- Primer:
 - Reč završava na “ti” —> infinitiv glagola
 - Niz karaktera odvojen od drugih nizova karaktera razmakom ili znakovima interpukcija —> reč
 - Tokenizacija

Šta je Natural Language Processing (NLP)?

- Statističke metode:
 - Bazirane na matematičkim modelima i verovatnoći
 - Analiza većih količina podataka za identifikovanje ključnih podataka
- Primer:
 - Bag of Words (BoW)
 - TF-IDF
 - Latent Dirichlet Allocation (LDA)

Šta je Natural Language Processing (NLP)?

- Metode mašinskog učenja:
 - Primena tehnika i algoritama za učenje iz podataka i obavljanje određenih zadataka
 - Koristi velike količine podataka (označenih ili neoznačenih)
- Primer:
 - Support Vector Machine (SVM)
 - Regresija

Šta je Natural Language Processing (NLP)?

- Metode dubokog učenja (Deep Learning):
 - Upotreba neuronskih mreža za prepoznavanje složenih obrazaca
 - Koristi velike količine podataka
- Primer:
 - RNN, LSTM, CNN
 - Transformeri

Text Analysis ↔ NLP

- Ne mora svaka analiza da koristi tehnike obrade prirodnog jezika
 - Regex
 - Manuelna analiza
- Ne koristi se svaka tehnika obrade prirodnog jezika za analizu teksta
 - Speech to text
 - Speech Recognition

Koje se osnovne tehnike koriste u analizi teksta?

- Tokenizacija
- Named Entity Recognition ([NER](#))
- Lematizacija i korenovanje
- Klasterizacija
- Pattern Recognition
- Part-of-Speech ([POS](#))

Koje se osnovne tehnike koriste u analizi teksta?

- Topic Modeling
- Klasifikacija teksta
- Sumarizacija teksta
- Embedding

Šta su konkretni ciljevi analize teksta?

- Ekstrakcija informacija
- Klasifikacija teksta
- Prepoznavanje obrazaca
- Razumevanje sentimenta
- Generisanje sažetaka
- Optimizacija pretrage
- Generisanje šablona

Šta su konkretni ciljevi analize teksta u pravu?

- Anonimizacija/pseudonimizacija
- Predstava dokumenata u mašinski čitljivom formatu
- Optimizacija pretrage dokumenata
- Prepoznavanje pravne oblasti
- Pronalaženje i povezivanje precedenata
- Generisanje šablona pravnih dokumenata
- Klasifikacija na tipove pravnih dokumenata

Koje faze postoje u analizi teksta?

- Prikupljanje podataka
 - Dobavljanje dokumenata u papirnom obliku
 - Web scraping
 - Korišćenje API-ja
- Preprocesiranje teksta
 - Tokenizacija
 - Uklanjanje stop reči
 - Lematizacija i korenovanje
- Odabir tehnika za obradu prirodnog jezika koji odgovara konkretnom zadatku

Na koji način možemo prikupiti podatke?

- Manuelno prikupljanje podataka
- Prikupljanje podataka sa interneta (Web scraping)
- Prikupljanje podataka korišćenjem API-ja
- Prikupljanje podataka direktno iz baza podataka
- Otvoreni podaci

Koje tehnike se koriste za preprocesiranje teksta?

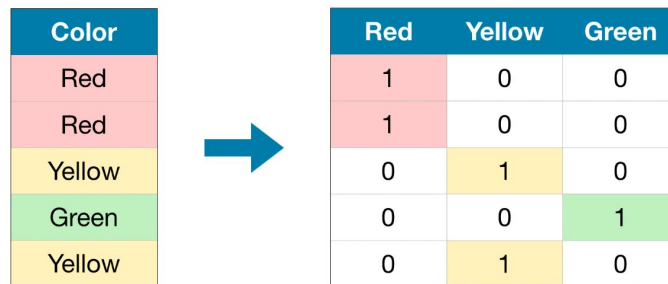
- Tokenizacija
- Brisanje stop reči
- Lematizacija i korenovanje
- Normalizacija
- Izbacivanje specijalnih karaktera i brojeva
- Ekstrakcija ključnih reči
- Prepoznavanje entiteta

Šta je reč?

- Osnovna jedinica jezika koja nosi značenje
- Ključni element koji se analizira, prepoznaje i interpretira
- Tipovi reči?
- Značenje reči?
- Uloga reči?

Kako predstavljamo reč za potrebe analize teksta?

- Reč mora biti razumljiva “mašini”
- One-hot encoding



The diagram illustrates the process of converting a list of words into a numerical format suitable for machine learning. On the left, a table lists five color words: Red, Red, Yellow, Green, and Yellow. A blue arrow points to the right, where a corresponding one-hot encoding matrix is shown. This matrix has three columns labeled Red, Yellow, and Green. Each row in the matrix corresponds to a word in the first table, with a '1' in the column corresponding to the word and '0' in the other columns.

Color
Red
Red
Yellow
Green
Yellow

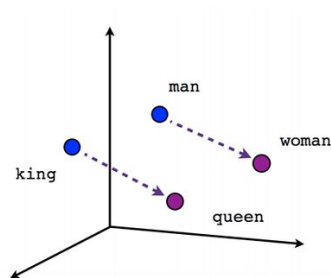
Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

Kako predstavljamo reč za potrebe analize teksta?

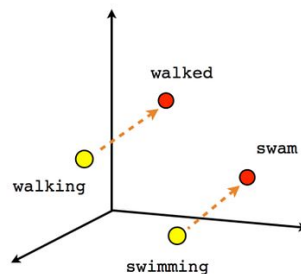
- Word Embeddings
 - Word2Vec
 - GloVe

Kako predstavljamo reč za potrebe analize teksta?

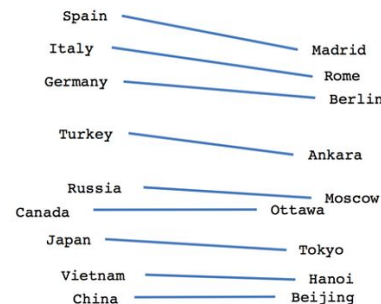
- Word Embeddings



Male-Female



Verb tense



Country-Capital

Šta je word embedding?

- Predstavljanje reči u numeričkom obliku
- Očuvana semantika
- Manja dimenzionalnost
- Učenje iz konteksta
- Slične reči su blizu jedna druge

Šta je word2vec i kako se dobija?

- Numerička predstava reči
- Mapira reči u vektorski prostor
- Dva osnovna modela za učenje reči:
 - Skip-gram model
 - Continuous Bag of Words

Šta je word2vec i kako se dobija?

- Skip-Gram model
 - predviđamo okolne reči na osnovu date reči
- Continuous Bag of Words
 - predviđamo centralnu reč na osnovu okolnih reči
- Trenira se model za predikciju centralnih ili okolnih reči

Koja je razlika između nadgledanog, polunadgledanog i nenadgledanog učenja?

- Nagledano učenje koristi označene podatke (labele)
- Svaki ulazni podatak ima poznat očekivan izlaz
- Polunadgledano učenje koristi mali skup označenih podataka i veliki skup neoznačenih podataka
- Mali označeni skup se koristi za inicijalno treniranje modela

Koja je razlika između nadgledanog, polunadgledanog i nenadgledanog učenja?

- Nenadgledano učenje koristi neoznačene podatke
- Cilj je identifikacija obrazaca, klastera
- Prednosti i mane svakog pristupa?
- Korišćenje svakog pristupa u pravu?

Mašinsko učenje ↔ duboko učenje?

- Oblasti veštačke inteligencije koje obuhvataju algoritme i tehnike za obučavanje modela u cilju rešavanja različitih problema
- Duboko učenje se smatra podskupom ili posebnom granom koja obuhvata modele bazirane na neuronskim mrežama
- Sličnosti i razlike?

Mašinsko učenje ↔ duboko učenje?

Mašinsko učenje	Duboko učenje
Manji skupovi podataka	Veliki skupovi podataka
Jednostavniji modeli	Složeni modeli
Zahteva manje resursa	Zahteva više resursa

Modeli/algoritmi mašinskog učenja

- Naive Bayes
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Decision Trees
- Random Forests
- Logistic Regression

Modeli/algoritmi dubokog učenja

- Convolutional Neural Networks (CNN)
 - Prostorni raspored
 - Konvolucija
 - Pooling
- Recurrent Neural Networks (RNN)
 - Sekvencijalni podaci
 - Rekurzivne veze
 - Vremenski koraci
 - Vanishing gradient

Modeli/algoritmi dubokog učenja

- *Long Short-Term Memory (LSTM)*
 - Uvođenje ćelija
 - Dugoročno pamćenje
 - *Input, output i forget gate*
- ELMo
 - Bidirekcionni LSTM
 - Generisanje vektora reči u zavisnosti od konteksta

Modeli/algoritmi dubokog učenja

- *Transformers*
 - Arhitektura koja omogućava paralelnu obradu podataka
 - Koristi mehanizam pažnje (*attention*)
 - Encoder
 - Decoder
 - BERT
 - GPT

LLM

- Šta je Large Language Model (LLM)?
- Koje vrste LLM-a postoje?
- Koja je najčešća arhitektura LLM-a?
- Šta je prompt engineering?
- Za rešavanje kojih zadataka može da se koristi LLM?
- Kako izgleda preprocesiranje teksta kod LLM-a?
- Šta je ChatGPT?
- Kako je treniran ChatGPT?
- Sa kojim problemima se susreću Large Language modeli?

Upotreba ChatGPT-a za analizu pravnih dokumenata

- Za koje probleme u pravu se može koristiti?
 - Klasifikacija
 - Ekstrakcija metapodataka
 - Ekstrakcija referenci
 - Anotiranje dokumenata
 - Anonimizacija
 - Generisanje šablona

Upotreba ChatGPT-a za analizu pravnih dokumenata

- Problemi?