ANALIZA SENTIMENTA

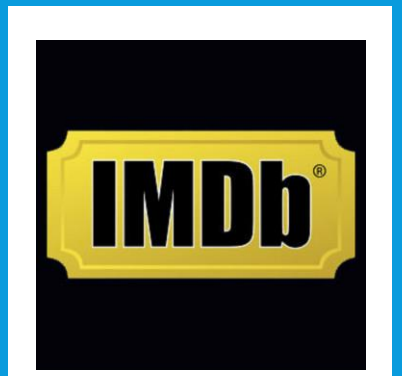
Završni projekat kursa Mašinsko učenje 2023. godine na Matematičkom fakultetu, Univerziteta u Beogradu

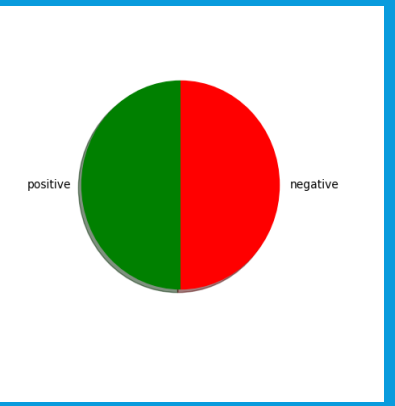
Pavle Savić

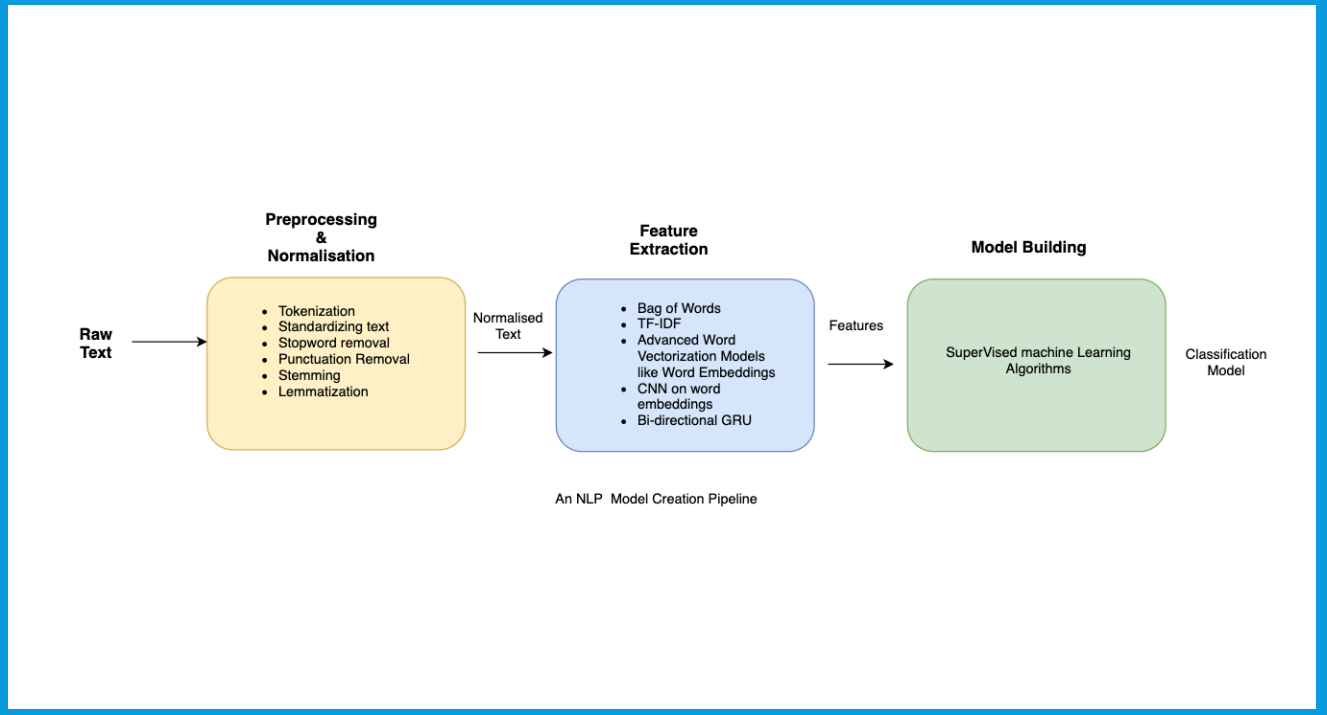
1075/2022

SKUP PODATAKA

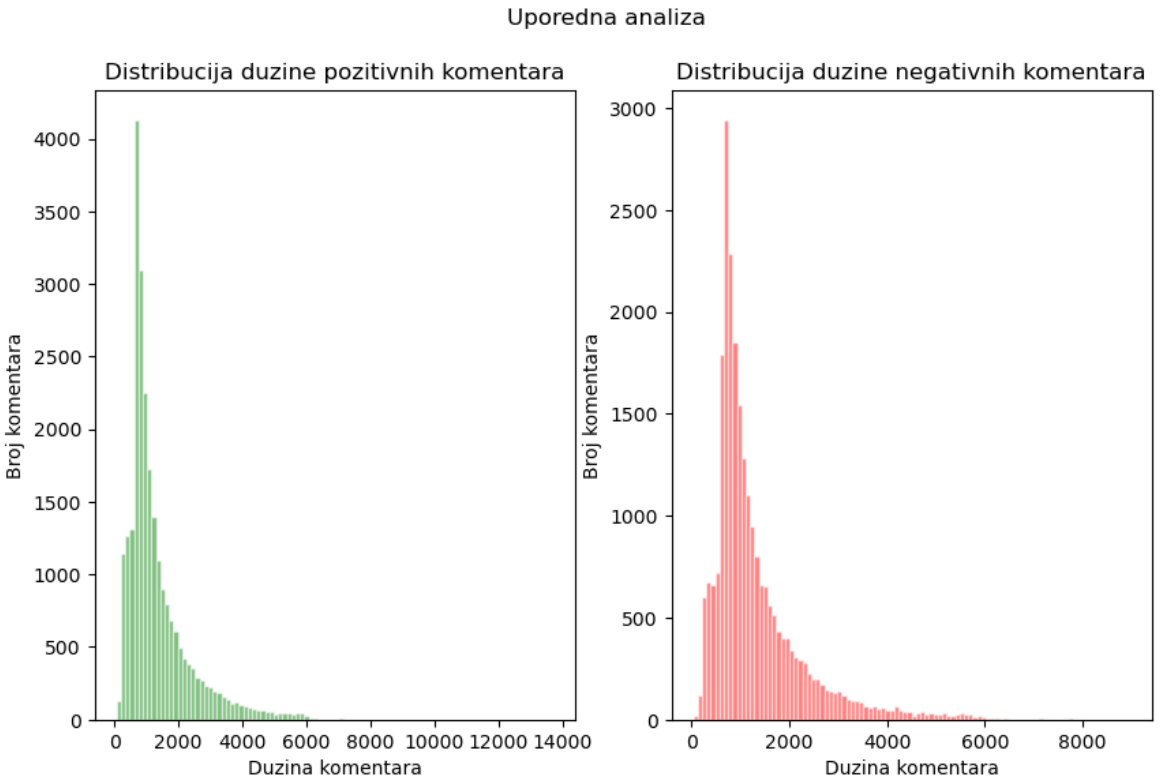
- 50 000 komentara na filmove sa IMDB sajta
- zadaci: osnovna analitika nad sirovim podacima, analiza prirodnog jezika i binarna klasifikacija komentara na pozitivne i negativne

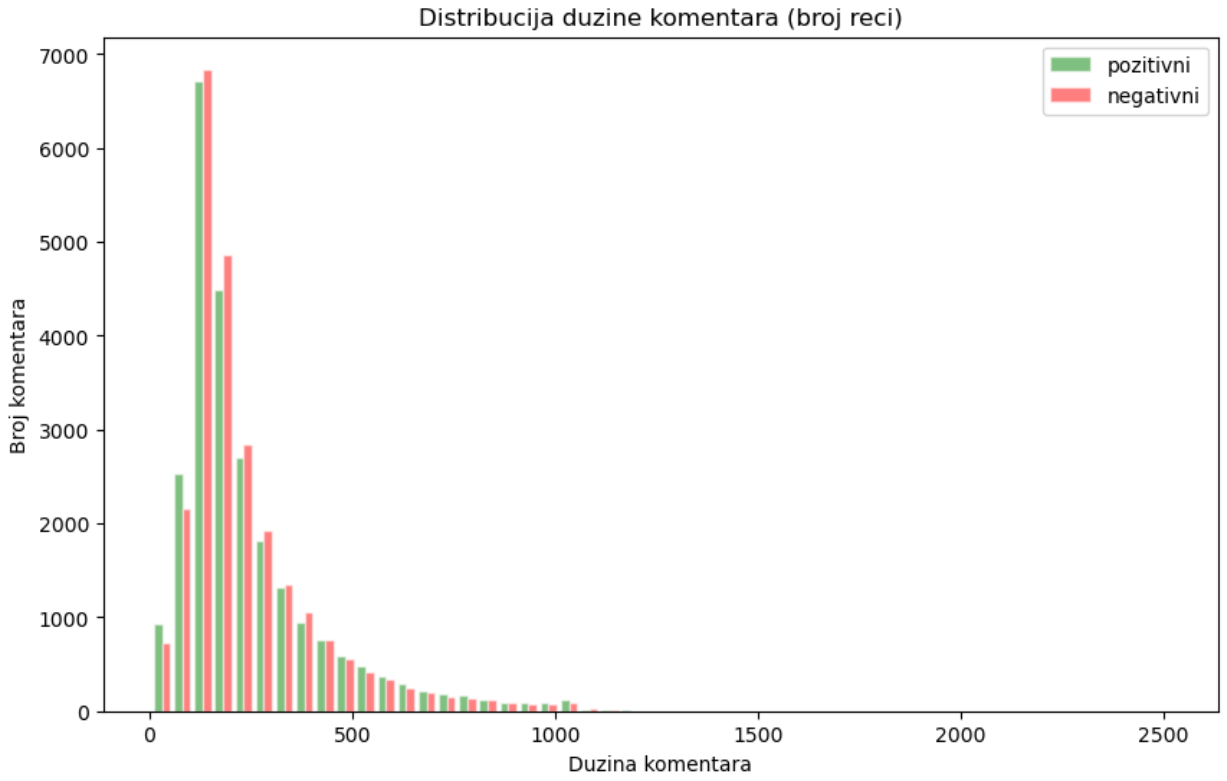






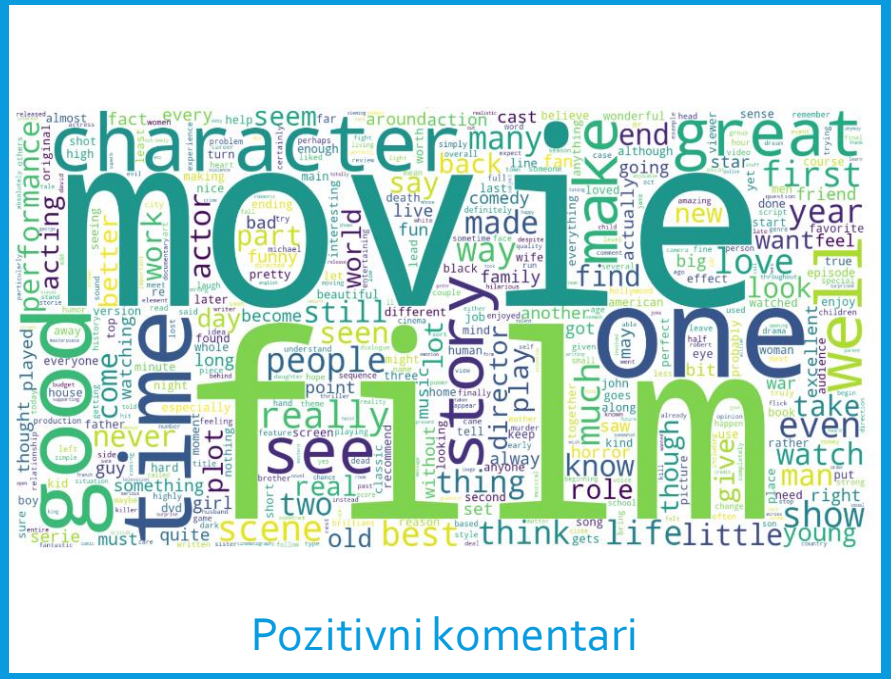
INICIJALNA ANALIZA SKUPA



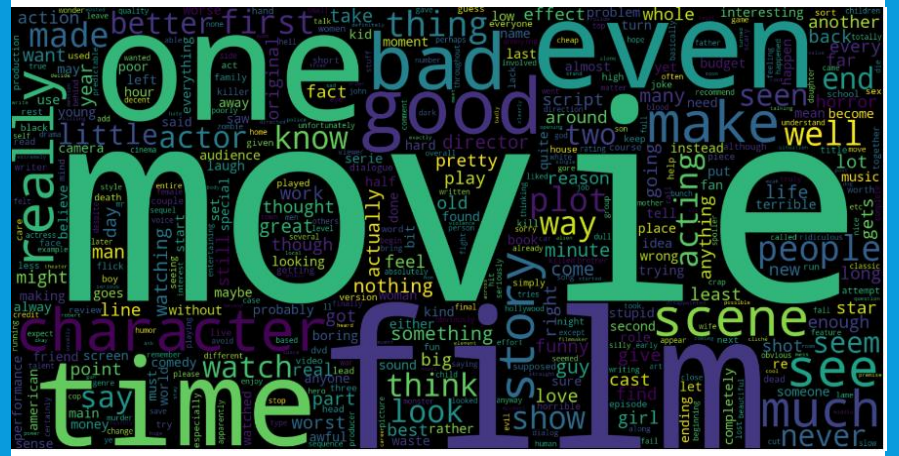


Statistika	Pozitivni	Negativni
Prosek	1324.80	1294.06
Medijana	968.00	973.00
Standardna devijacija	1031.47	945.87
Maks vrednost	13704.00	8969.00
Min vrednost	65.00	32.00

Statistika	Pozitivni	Negativni
Prosek	241.84	239.48
Medijana	179.00	182.00
Standardna devijacija	184.33	172.02
Maks vrednost	2515.00	1620.00
Min vrednost	10.00	6.00

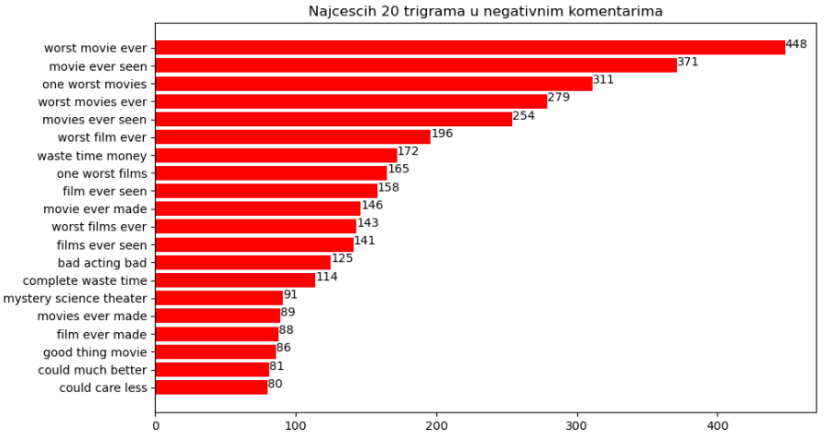


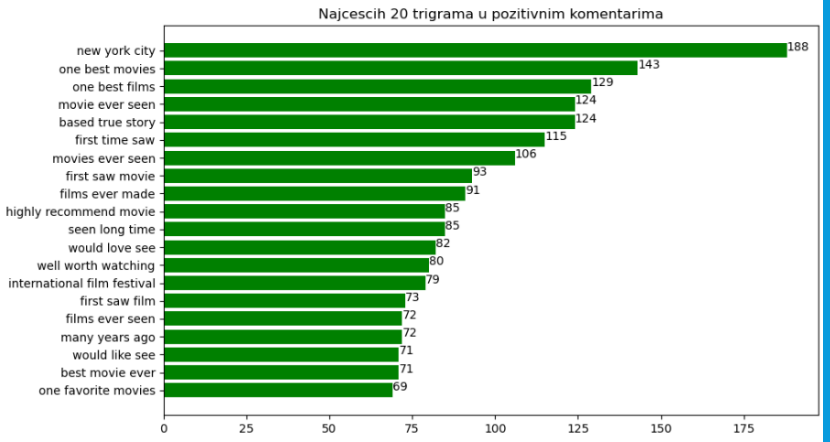
Pozitivni komentari

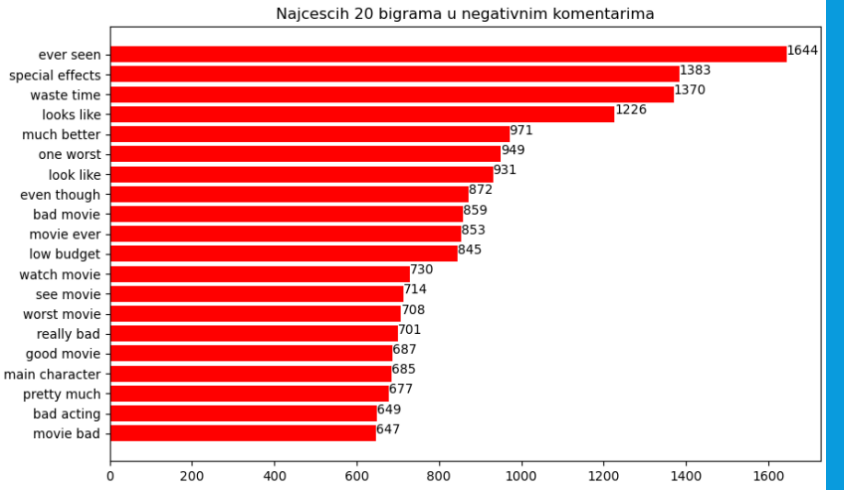


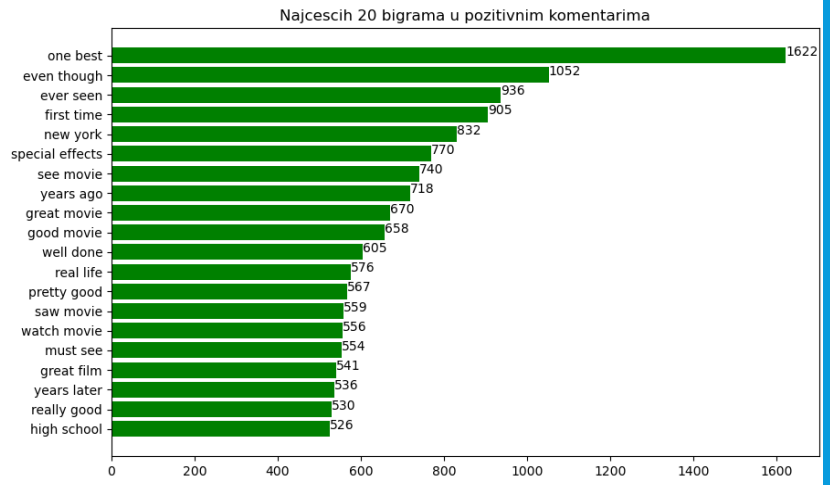
Negativni komentari

WORD CLOUD – NAJZASTUPLJENIJE REČI









ANALIZA BIGRAMA I TRIGRAMA

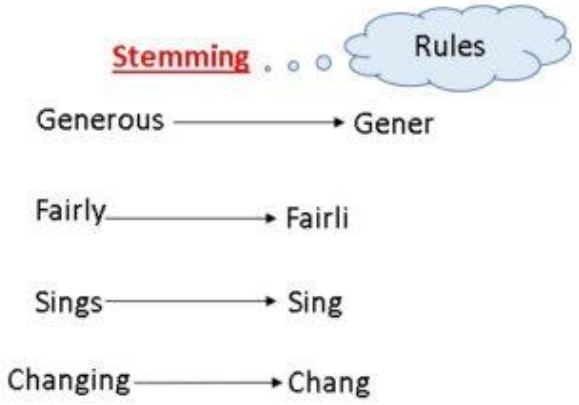
VEKTORIZACIJA


```python
def review_preprocessor(text):
    text = text.lower()
    text = short_form_transform(text)
    text = strip_html(text)
    text = strip_url(text)
    text = full_stop_abbrev_elim(text)
    return text
```

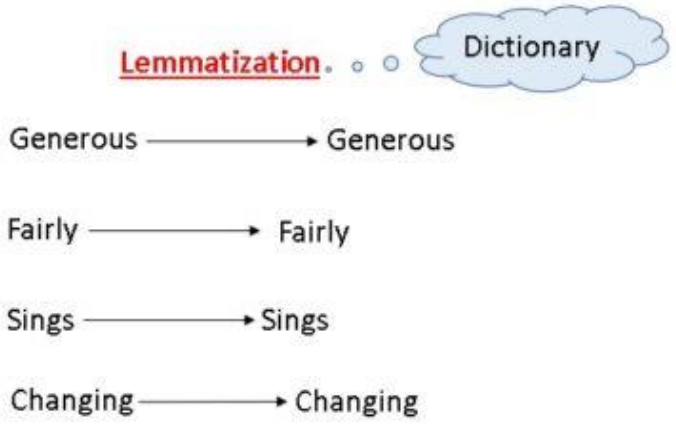
```python
def simple_tokenization(review):
    tokens = nltk.tokenize.word_tokenize(review)
    tokens_without_punctuation = [token for token in tokens if token not in string.punctuation]
    return tokens_without_punctuation
```

```python
def review_tokenizer(stemming, text):
    tokens = simple_tokenization(text)
    tokens = remove_stop_words(tokens)

    stems = []

    for token in tokens:

        token_pattern = re.compile(r'\b[^\W\d_]+\b')
        if not token_pattern.match(token) or len(token) <= 2:
            continue

        stem = stemming.stem(token)
        stems.append(stem)
    return stems
```

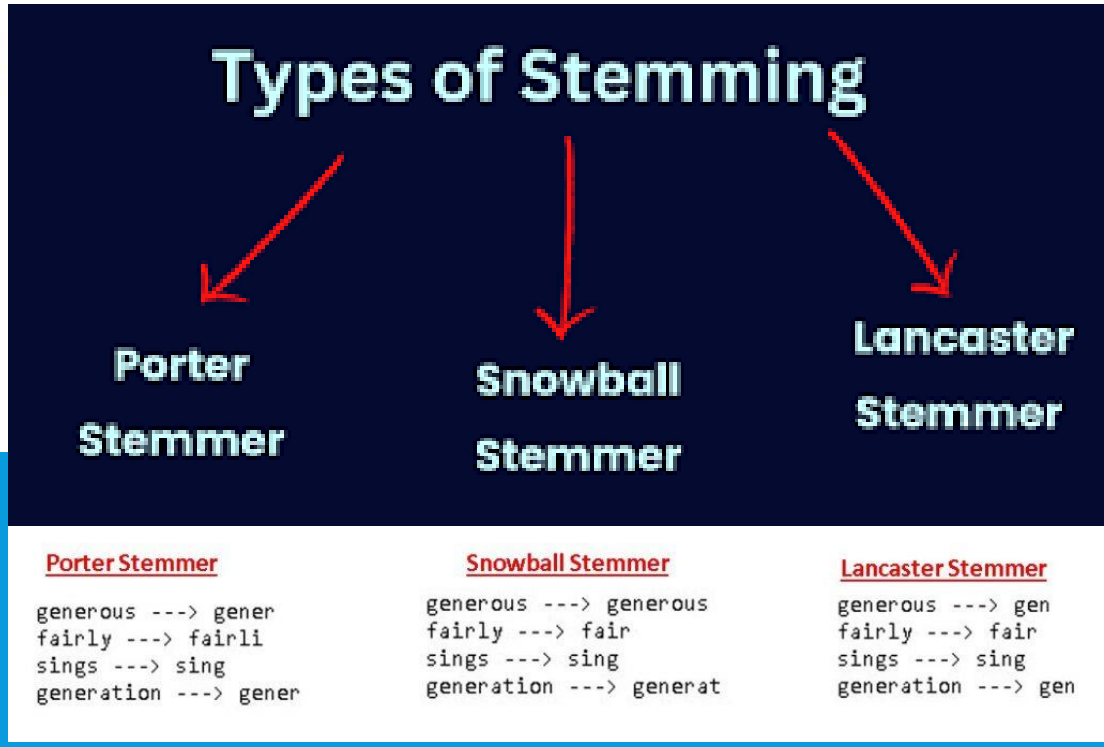
STEMOVANJE ILI LEMATIZACIJA



- Dosledna primena niza pravila kako bi se dobio `stem` - veštački koren reči
- Manja preciznost
- Značajno brže
- Stemovi mogu biti reči bez značenja (ograničena primena)



- Reči se pridružuje njen gramatički koren (lema)
- Veća preciznost
- Značajno sporije
- Leme zadržavaju značenje polazne reči



Povećanje agresivnosti i brzine

PODELA SKUPA

LITERATURA

- https://www.analyticsvidhya.com/blog/2021/11/an-introduction-to-stemming-in-natural-language-processing/
- https://www.datacamp.com/tutorial/stemming-lemmatization-python
- https://medium.com/@Mirza_Yusuf/using-a-bert-model-for-sentiment-analysis-6c6fcc106843
- https://www.geeksforgeeks.org/sentiment-classification-using-bert/