



UNIVERSIDADE
FEDERAL DO CEARÁ

Atividade 01 - NLP

Arielly Gonçalves

1) Apresentação sobre a usabilidade de outra biblioteca usando língua portuguesa.

Stanza (Stanford NLP)

```

1  import stanza
2  import pandas as pd
3
4  # Download do modelo em português
5  #stanza.download('pt')
6
7  nlp = stanza.Pipeline('pt')
8
9  texto = "Hoje eu fui ao mercado e comprei frutas frescas."
10 doc = nlp(texto)
11
12 # tokens, lemas e classes gramaticais
13 dados = []
14 for sent in doc.sentences:
15     for word in sent.words:
16         dados.append({
17             "Texto": word.text,
18             "Lema": word.lemma,
19             "Classe gramatical (UPOS)": word.upos
20         })
21
22 # tabela
23 tabela = pd.DataFrame(dados)
24 print(tabela)
25

```

Texto	Lema	Classe gramatical (UPOS)
Hoje	hoje	ADV
eu	eu	PRON
fui	ser	AUX
a	a	ADP
o	o	DET
mercado	mercado	NOUN
e	e	CCONJ
comprei	comprar	VERB
frutas	fruta	NOUN
frescas	fresco	ADJ
.	.	PUNCT

Apple PROP
está AUX
querendo VERB
comprar VERB
uma DET
startup NOUN
do ADP
Reino PROP
Unido PROP
por ADP
100 NUM
milhões NUM
de ADP
dólares NOUN

POS	DESCRIPTION	EXAMPLES
ADJ	adjective	big, old, green, incomprehensible, first
ADP	adposition	in, to, during
ADV	adverb	very, tomorrow, down, where, there
AUX	auxiliary	is, has (done), will (do), should (do)
CONJ	conjunction	and, or, but
CCONJ	coordinating conjunction	and, or, but
DET	determiner	a, an, the
INTJ	interjection	psst, ouch, bravo, hello
NOUN	noun	girl, cat, tree, air, beauty
NUM	numeral	1, 2017, one, seventy-seven, IV, MMXIV
PART	particle	's, not,
PRON	pronoun	I, you, he, she, myself, themselves, somebody
PROPN	proper noun	Mary, John, London, NATO, HBO
PUNCT	punctuation	., (,), ?
SCONJ	subordinating conjunction	if, while, that
SYM	symbol	%, \$, ©, +, -, ×, ÷, =, :), 🤔
VERB	verb	run, runs, running, eat, ate, eating
X	other	sfpksdpsxmsa
SPACE	space	

2) Tabela
com os
significados
das siglas
pos tagging
calculadas
pela spaCy.

3) Implementação para receber um texto com erros gramaticais em português.

```
from enlvo.normaliser import Normaliser
from nltk.stem import RSLPStemmer
import pandas as pd
import spacy
import nltk

nltk.download("stopwords")
nltk.download('rslp')

nlp = spacy.load("pt_core_news_sm")

texto = "hj eu fui para a igreja, foi muito legau"
print("Texto original:\n", texto)

normalizar = Normaliser(tokenizer="readable", capitalize_acs=True,
                        capitalize_inis=True, capitalize_pns=True,
                        sanitize=True)

correcao = normalizar.normalise(texto)
print("\nTexto corrigido:\n", correcao, "\n")

doc = nlp(correcao)
for token in doc:
    print(token.text)

stopwords = nltk.corpus.stopwords.words("portuguese")
```

```
sem_stopwords = " ".join([palavra for palavra in correcao.split() if palavra.lower() not in stopwords])

print("\nTexto sem stopwords:\n", sem_stopwords, "\n")

stemmer = RSLPStemmer()

tabela = []
for token in doc:
    token_text = token.text
    token_stem = stemmer.stem(token_text)
    token_lemma = token.lemma_
    tabela.append([token_text, token_stem, token_lemma])

# DataFrame
df = pd.DataFrame(tabela, columns=["Token", "Stemming", "Lematização"])
print(df)
```

a) Realize a correção do texto

Texto original:

hj eu fui para a igreja, foi muito legau

Texto corrigido:

Hoje eu fui para a igreja foi muito legal

b) Tokenize o texto

Hoje
eu
fui
para
a
igreja
foi
muito
legal

c) Remova as stop words

Texto sem stopwords:
Hoje igreja legal

d) Tabela mostrando 3
colunas para todos os
tokens: token , stemming
e lematização.

Token	Stemming	Lematização
Hoje	hoj	hoje
eu	eu	eu
fui	fui	ser
para	par	para
a	a	o
igreja	igrej	igreja
foi	foi	ser
muito	muít	muito
legal	legal	legal



UNIVERSIDADE
FEDERAL DO CEARÁ

Obrigado!

Arielly Gonçalves