05

Tokenizador

Nesta aula, exploramos e desenvolvemos as funções auxiliadoras para utilizar na criação do classificador. Entre as funções desenvolvidas, temos:

- a responsável pelo tratamento e tokenização dos textos;
- a que combina os vetores de cada palavra e cria uma representação vetorial para a frase;
- a função que cria a representação vetorial de toda base de dados.

Sobre a função de tokenização e tratamento responda:

Qual das funções a seguir recebe como entrada:

```
"Rio de Janeiro 1231231 ***** @#$ é uma cidade maravilhosa!"
```

e retorna a seguinte lista:

```
['rio', 'janeiro', 'cidade', 'maravilhosa']
```

P.S: Considere nlp os modelos treinados em Português:

```
nlp = spacy.load("pt_core_news_sm", disable=["paser", "ner", "tagger", "textcat"])
```

Selecione uma alternativa

A

```
def tokenizador(texto):

    doc = nlp(texto)
    tokens_validos = []
    for token in doc:
        e_valido = not token.is_stop and token.is_alpha
        if e_valido:
            tokens_validos.append(token.text.lower())

    return  tokens_validos
```

B

```
def tokenizador(texto):

    doc = nlp(texto)
    tokens_validos = []
    for token in doc:
        e_valido = not token.is_stop and token.is_alpha
        if e_valido:
            tokens_validos.append(token)

    return  tokens_validos
```

C

```
def tokenizador(texto):

    doc = nlp(texto)
    tokens_validos = []
    for token in doc:
        e_valido = not token.is_stop and token.is_alpha
        if e_valido:
            tokens_validos.append(token.text.lower())

    return  “ “.join(tokens_validos)
```