

Separando palavras de tokens

ATIVIDADES 4 de 10

Transcrição

DISCORD ALURA

FÓRUM DO CURSO

VOLTAR PARA DASHBOARD

MODO NOTURNO

16.1k xp

3

Se fizermos o len() da variável palavras_separadas, o resultado será 5.

Ou seja, a frase "Olá, tudo bem?" possui cinco tokens, sendo três palavras e duas pontuações, mas precisamos descobrir apenas a quantidade de vocábulos no corpus.

Para fazermos a separação dos tokens, criaremos uma função chamada separa_palavras() com def . Esta pegará justamente a lista_tokens como parâmetro, e retornará a lista_palavras com return .

def separa_apalavras(lista_tokens):
return lista_palavras

COPIAR CÓDIGO

Com isso, deixaremos de lado as pontuações, números ou qualquer outra coisa que não seja uma palavra de fato.

Percorreremos a lista de tokens e acrescentaremos apenas os itens que são realmente vocábulos.



Para descobrirmos isso, usaremos um método chamado .isalpha() em uma nova célula antes da definição da função anterior, o qual percorrerá cada caractere da string chamada 'palavra' e selecionará apenas as letras do alfabeto, retornando-as como verdadeiro. Caso encontre um que não seja alfabético como em palavral ou apenas './' com um número ou pontuação por exemplo, nos retornará falso.

Então, em nossa nova função, perguntaremos para cada um dos nossos tokens se todos os seus caracteres são alfabéticos ou não. Para isso usaremos um for para token na lista_tokens, perguntando se isalpha() com if.

ALURA

Se o retorno for verdadeiro, ou seja, se for uma palavra de fato, o adicionaremos à lista palavras com .append().

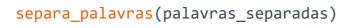
Como ainda não criamos a lista de palavras para adicioná-lo, diremos que lista_palavras é igual a uma lista vazia. Em seguida, a retornaremos.

PARA

Testaremos a função passando a frase de exemplo "Olá, tudo bem?" dentro de palavras_separadas como parâmetro de separa palavras().

MODO **NOTURNO**

```
def separa palavras(lista tokens):
lista palavras = []
for token in lista tokens:
    if token.isalpha():
        lista_palavras.append(token)
return lista palavras
```



COPIAR CÓDIGO

16%

Com isso, teremos separado a lista de tokens somente em palavras.

ATIVIDADES 4 de 10 A seguir, faremos o mesmo processo com o nosso corpus textual.

DISCORD ALURA

FÓRUM DO CURSO

VOLTAR PARA DASHBOARD

MODO NOTURNO



16.1k xp

