



16%

ATIVIDADES
4 de 10

DISCORD
ALURA

FÓRUM DO
CURSO

VOLTAR
PARA
DASHBOARD

MODOS
NOTURNO



16.1k xp

a

04

Separando palavras de tokens

Transcrição

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.



16%

ATIVIDADES
4 de 10

DISCORD
ALURA

FÓRUM DO
CURSO

VOLTAR
PARA
DASHBOARD

MODOS
NOTURNO



16.1k xp

a

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 `palavra1` 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`.

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.

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

```
def separa_palavras(lista_tokens):  
    lista_palavras = []  
    for token in lista_tokens:  
        if token.isalpha():  
            lista_palavras.append(token)  
    return lista_palavras
```



16%

`separa_palavras`(palavras_separadas)

COPIAR CÓDIGO

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

A seguir, faremos o mesmo processo com o nosso corpus textual.

ATIVIDADES
4 de 10

DISCORD
ALURA

FÓRUM DO
CURSO

VOLTAR
PARA
DASHBOARD

MODOS
NOTURNO



16.1k xp

a