Normalização

Transcrição

O valor total de `393916` palavras retornado anteriormente em nosso corpus não representa a quantidade de palavras únicas que podem ser corrigidas, afinal muitas delas aparecerão *_repetidas_* nos textos.

Portanto, precisaremos calcular quantos vocábulos **únicos** existem **sem repetição**.

Por exemplo, se pegarmos a frase “Se você chegou neste artigo, muito provavelmente você já deu o seu primeiro passo”, veremos que existem 14 palavras ao total, sendo que "você" se repete duas vezes.

Logo, o NLP precisará saber apenas quantos **tipos de palavras** únicas temos, e neste caso seriam 13 contando o artigo, pois o vocábulo "você" já teria sido contabilizado na primeira ocorrência, e não precisaria ser novamente.

Será essa quantidade de tipos de palavras que nosso corretor ortográfico aprenderá a corrigir.

O primeiro passo será verificarmos se temos as mesmas palavras escritas de formas diferentes. Para entendermos melhor essa questão, imprimiremos as cinco primeiras palavras da `lista_palavras`.

```
print(lista_palavras[:5])
```

COPIAR CÓDIGO

O retorno deste comando será `['imagem', 'Temos', 'a', 'seguinte', 'classe']`.

Neste caso, a palavra "Temos" inicia com letra maiúscula, mas também poderemos ter a mesma palavra "temos" iniciando com letra minúscula no meio do texto, então as contabilizaremos como dois tipos diferentes.

Portanto, deveremos **normalizar** o corpus e **transformar** todo o texto em **letras minúsculas**.

Definiremos a função **`normalizacao()`** com `def`, em seguida passaremos a `lista_palavras` e retornaremos a **`lista_normalizada`**, a qual será composta apenas das palavras cujos caracteres foram todos transformados em letras minúsculas. Isso será realizado com `palavra.lower()` sendo o parâmetro de `.append()` da `lista_normalizada`.

Porém, ainda não fizemos uma **iteração**, e precisaremos acessar a `lista_palavras`. Fazendo `for` para a `palavra` dentro da lista, teremos realizado esta operação.

Antes do `for`, criaremos a variável `lista_normalizada` sendo igual a uma lista vazia.

Ao final da célula, faremos a normalização de fato chamando a `lista_normalizada` sendo igual a `normalizacao()` com a `lista_palavras`.

Por fim, verificaremos o funcionamento do código imprimindo as mesmas cinco primeiras palavras da `lista_normalizada`.


```
def normalizacao(lista_palavras):
    lista_normalizada = []
    for palavra in lista_palavras:
        lista_normalizada.append(palavra.lower())
    return lista_normalizada

lista_normalizada = normalizacao(lista_palavras)
print(lista_normalizada[:5])
```

COPIAR CÓDIGO

Como retorno, receberemos a mesma lista de palavras anterior, mas a palavra `'temos'` aparecerá com letra minúscula ao invés de maiúscula como estava antes.

Com nosso texto normalizado, poderemos contar os tipos de vocábulos.

A seguir, aprenderemos uma forma de **eliminar palavras repetidas** para fazermos a contagem definitiva que precisamos.