**Corretor Ortográfico em Python – Aplicando técnicas de NLP**

Link do google drive:

<https://drive.google.com/drive/folders/1wnumjuGByvZKLdVXVk39JLhO5RNyNUdt>.

Link das libs:

- <https://www.nltk.org/>

1. **Aula 1 – Explorando Um Projeto De NLP:**
   1. NLP: do inglês significa processamento de linguagem natural. Responsável por fazer o intermédio da comunicação entre humanos e máquinas, tendo diversas aplicações como o google assistente, siri, alexa dentre outras possibilidades como análise de sentimento, tradutores, buscadores e corretores ortográficos.
   2. Importando Um Corpus Textual:
      1. Precisamos ter uma base de dados para que o nosso algoritmo conheça as palavras para que possa corrigir.
      2. Quando trabalhamos NLP a nossa base de dados se chama corpus, sendo um conjunto de documentos e, para ler um arquivo de texto em python precisamos fazer um código simples, onde dizemos que com o nosso arquivo aberto em formato de leitura, como f, atribua a leitura de f em artigos:

with open('artigos.txt', 'r') as f:

  artigos = f.read()

* + 1. Assim, quando printarmos artigos, todo o nosso corpus será exibido.
  1. Tokenização:
     1. A primeira coisa para saber se nosso corpus é ideal, é saber a quantidade de palavras que ele possui, além de saber se elas se repetem muito ou não.
     2. Tokens são pequenos pedaços do nosso corpus composto por palavras e pontuações, uma vez que quando utilizamos o método split() para quebrar todo o nosso texto em uma lista com palavras, ele não separa somente as palavras, mas a concatenação delas com a pontuação também:

texto\_exemplo = 'Olá, tudo bem?'

token = texto\_exemplo.split()

print(palavras\_separadas)



* 1. O que aprendemos:
     1. O que é o processamento de linguagem natural (NLP);
     2. O que está por trás de um corretor ortográfico;
     3. Como ler um arquivo de texto utilizando python;
     4. Como contar a quantidade de caracteres de um texto usando len().

1. **Aula 2 – Utilizando NLTK Para Tokenizar Um Texto:**
   1. Refinando a Tokenização: natural linguage toolkit
      1. Importamos e fizemos a instalação do punkt para podermos utilizar o método nltk.tokenize.word\_tokenize(texto\_exemplo) separando as palavras da pontuação:

import nltk

nltk.download('punkt')

palavras\_separadas = nltk.tokenize.word\_tokenize(texto\_exemplo)

palavras\_separadas



* 1. Separando Palavras De Tokens:
     1. Podemos criar uma função para separar fazendo um for que percorra toda a nossa lista de tokens e, se a palavra atual Isalpha(), ou seja, for alfabética, não contendo números, pontuação e afins, então adiciona a uma nova lista, retornando ela no final:

def separa\_palavras(lista\_tokens: list):

  lista\_palavras = []

  for token in lista\_tokens:

    if token.isalpha():

      lista\_palavras.append(token)

  return lista\_palavras

separa\_palavras(palavras\_separadas)



* 1. Contando Palavras Do Corpus:
     1. Fizemos exatamente a mesma coisa que acima, mas com o nosso artigo, e fizemos a contagem da quantidade de palavras. Podemos ver que temos um número BEM significativo de palavras no nosso corpus:

lista\_tokens = nltk.tokenize.word\_tokenize(artigos)

lista\_palavras = separa\_palavras(lista\_tokens)

print(f'O número de palavras é: {len(lista\_palavras)}')



* 1. Normalização:
     1. Apesar de termos essa quantidade gigante de palavras no nosso corpus, sabemos que na realidade não temos tudo isso de palavras diferentes, uma vez que em um único texto repetimos a mesma palavra várias vezes. Considerando isso precisamos fazer a normalização do nosso corpus, mantendo somente palavras diferentes.
     2. A normalização consiste em manter somente 1 tipo de cada palavra, ou seja, escrita de uma única forma e apenas uma vez, por exemplo: ‘Temos’ e ‘temos’. São considerados 2 tipos pois começam com letras maiúscula e minúscula. Para fazer essa normalização criamos uma função que acessa cada palavra da nossa lista e adiciona ela em outra lista, mas com todas as letras minúsculas, ou seja, outro for:

def normalizacao(lista\_palavras: list):

  lista\_normalizada = []

  for palavra in lista\_palavras:

    lista\_normalizada.append(palavra.lower())

  return lista\_normalizada

* 1. Tipos De Palavras:
     1. Para remover as repetições de uma lista utilizamos o set(). Após fazer isso com nossa lista e verificarmos a quantidade de tipos que temos, podemos ver que o número diminui consideravelmente, indo de 393k para 17k. Contudo, após uma pesquisa, vimos que a quantidade de palavras necessárias para uma pessoa ser considerada fluente é de 8k, ou seja, 17k é mais do que o suficiente:

lista\_unica = set(lista\_normalizada)

len(lista\_unica)



* 1. O que aprendemos:
     1. O que são tokens;
     2. O que é e como utilizar a biblioteca NLTK;
     3. Como tokenizar um texto usando NLTK;
     4. Como separa apenas caracteres alfabético de um token;
     5. O que são tipos de palavras.