**ESPECIALIZAÇÃO DATA SCIENCE E BIG DATA**

**ANÁLISE DE TEXTOS COM R E PYTHON**

**Professor Leandro ESCOBAR**

**PARTE 0 – Operações com texto**

Os conteúdos na forma de sentenças ou de tokens devem ser processados, uma vez que :

* O processamento requer preparação dos textos, de forma a obedecer os requisitos do algoritmo a ser utilizado;
* Após processamento, surge a necessidade de tratar os resultados, seja na busca de outros padrões ou na otimização dos padrões obtidos. Por exemplo, criar novas sentenças ou localizar elementos dentro dos resultados.

# Operações com strings

Crie um novo programa Py com o nome operações\_strings.py

As operações estão demonstradas no código, acompanho a prática com o professor

*#OPERAÇÕES COM STRINGS*lista\_nomes = [**'Ana'**,**'João'**,**'Maria'**,**'Pedro'**]  
print(type(lista\_nomes)) *#exibe a classe (tipo) do objeto. Nesse caso string  
  
  
#Join*nomes = **';'**.join(lista\_nomes)  
print(type(nomes), **':'**, nomes) *# lista de nomes foi convertida em uma string com os elementos separados por ';'  
  
#Split*sentenca = **'O cachorro está no jardim'**lista\_sentenca = sentenca.split(**' '**)  
print(type(lista\_sentenca), lista\_sentenca)

#Imprimir o número de elementos na lista

print(len(lista\_sentenca))

*#*sentenca = **'teste-algum-dia-sol-chuva'**lista\_sentenca = sentenca.split(**'-'**)  
print(lista\_sentenca)  
  
*#crie uma lista com a string nomes  
  
#Adicionando e multiplicando strings*texto\_add = **'Itaipu'** + **'PTI'** + **'UP'**print(texto\_add)  
  
texto\_multi = **'Itaipu'** \* 10  
print(texto\_multi)  
  
texto\_multi = **'Itaipu'** +**';'**\*10  
print(texto\_multi)  
  
texto\_multi = (**'Itaipu'** +**';'**)\*10  
print(texto\_multi)  
  
*#Índices nas strings*texto = **'Python NLTK'**print(texto[1])  
print(texto[-3])  
*#O que aconteceu?  
# Índice posito = da esquerda para a direita  
# Índice negativo = da direita para a esquerda*

*#Subtituir parte da string*str1 = **'Curitiba é uma grande cidade, com muitas opções de turismo e compras'**str2 = str1.replace(**'Curitiba'**,**'Foz do Iguaçu'**)  
print(str2)

*#####################################  
# APROFUNDANDO OPERAÇÕES COM TEXTOS  
  
  
#como tratar intervalos  
#em listas*lista\_nomes = [**'Ana'**,**'João'**,**'Maria'**,**'Pedro'**]  
print(lista\_nomes[0])  
print(lista\_nomes[0:3]) *#[0:3] é um intervalo fechado à esquerda e aberto à direita*print(lista\_nomes[1:3])  
  
*#em strings*str = **'Aula NLTK com Python'***#Recuperar (ou coletar ou carregar) parte de uma string = substring*print(str) *#string completa*print(str[:5])*# até o 5º caractere*print(str[2:5])*#do 2º até o 5º caractere*print(str[-3])*# o que aconteceu???  
  
#Verificar se um termo existe dentro da string***if 'nltk' in** str:  
 print(**'NLTK encontrado'**)  
**else**:  
 print(**'NLTK não encontrado'**)  
  
*#teste com o termo nltk (minúsculas) e diga o que aconteceu  
#Cuidado: as operações são 'Case sensitive', ou seja, minúsculas são diferentes de maiúsculas  
  
#Acessar cada caractere da string*print(**'Acessar cada caractere da string'**)  
**for** c **in** str:  
 print(c)  
*# O comando for retira cada elemento dentro de str e o coloca, individualmente, no objeto c*

*#Salvando um arquivo com o texto*f = open(**'saidaDOC.txt'**,**'w'**,encoding=**'utf8'**)*#w para escrita*texto = **'Eu nunca penso no futuro. Ele não tarda a chegar. (Albert Einstein)'**f.write(texto)  
f.close()

# Arquivos DOCX e PDF

### Manipulando documentos DOCx

Para a manipulação de documentos do tipo DOCX, será utilizada a biblioteca python-docx.

O arquivo PrimeiroDOC.docx será utilizado (na pasta da disciplina, em Dados), baixe-o em seu computador.

Instalar:

pip install python-docx

Ou, no Pycharm:

File/Settings/Project Interpreter / Adicionar o python-docx

Inicie um novo programa PY com o nome leitor\_docx.py

*#Programa para leitura de documentos (DOCx)***import** docx  
  
*#O arquivo PrimeiroDOC.docx será utilizado  
  
#C:\Users\Convexa\Documents\Aulas\NaturalLanguageProcessing\_Python\Dados*arquivo = **'C:/Users/Convexa/Documents/Aulas/NaturalLanguageProcessing\_Python/Dados/PrimeiroDOC.docx'**doc = docx.Document(arquivo)  
texto\_completo = [] *#Lista para receber o texto***for** paragrafo **in** doc.paragraphs:  
 texto\_completo.append(paragrafo.text)  
  
print(texto\_completo)*#imprime a lista*

*#como essa lista pode ser convertida em texto/string?*

### Manipulando arquivos PDF

Para a manipulação de documentos do tipo DOCX, será utilizada a biblioteca pypdf2.

O arquivo PrimeiroPDF.pdf será utilizado (na pasta da disciplina, em Dados), baixe-o em seu computador.

Instalar:

pip install pypdf2

Ou, no Pycharm:

File/Settings/Project Interpreter / Adicionar o pypdf2

Inicie um novo programa PY com o nome leitor\_pdf.py

**from** PyPDF2 **import** PdfFileReader  
arquivo = **'C:/Users/Convexa/Documents/Aulas/NaturalLanguageProcessing\_Python/Dados/PrimeiroPDF.pdf'**pdf\_arquivo = open(arquivo, **'rb'**)  
read\_pdf = PdfFileReader(pdf\_arquivo)  
print(**'Número de páginas'**)  
print(read\_pdf.getNumPages())  
  
texto\_completo = []  
**for** i **in** range(0,read\_pdf.getNumPages()-1): *#percorre o documento até a última página* texto\_completo.append(read\_pdf.getPage(i).extractText())  
  
print(texto\_completo)  
  
texto\_corrido = **'\n'**.join(texto\_completo)  
print(texto\_corrido)

## EXERCITANDO 1

**Execute o que se pede. Logo após, cole a solução do exercício neste documento.**

**Lembre-se de salvar seu programa, para estudos posteriores**

**Tempo estimado 15 minutos**

* Crie uma string com o conteúdo ‘Ainda que falasse as línguas dos homens e falasse a língua dos anjos, sem amor eu nada seria.’
* Imprima cada caractere da string
* Segmente a string em uma lista
* Quantas palavras há na lista?
* Imprima cada palavra da string
* Substitua o termo ‘dos homens’ por ‘do mundo’
* Imprima o fragmento que vai do 21º até o 30º caracteres
* Imprima os últimos 15 caracteres
* Salve a sentença em um arquivo do tipo txt

## EXERCITANDO 2

**Execute o que se pede. Logo após, cole a solução do exercício neste documento.**

**Lembre-se de salvar seu programa, para estudos posteriores**

**Tempo estimado 15 minutos**

Abra o arquivo Romance.docx e faça o que se pede

* Crie uma lista com os parágrafos do documento
* Quantos parágrafos o documento possui?
* Imprima o conteúdo do 1º parágrafo do texto
* Imprima os parágrafos 3 a 6, inclusive
* O termo ‘Machado’ está no documento?
* Crie um texto corrido a partir dos parágrafos lidos
* Substitua o termo ‘Batista’ por ‘João Batista’