PROCESSAMENTO DE LINGUAGEM NATURAL

Sequências de caracteres, *tokens* e palavras

TÓPICOS

1. Sequências de caracteres
2. *Tokens*
3. Palavras

LINGUAGENS

O QUE É A LINGUAGEM?

“Sistema de **símbolos de um vocabulário** que, quando colocados numa determinada **ordem** e expressos num determinado **contexto**, emitem um **significado**.”

SEQUÊNCIAS DE CARACTERES

EXPRESSÕES REGULARES

- **São um mecanismo muito simples, porém muito poderoso, para manipulação de sequências de caracteres**
- **São úteis para**
  - Encontrar padrões em texto e auxiliar no fluxo de um chatbot, por exemplo
  - Encontrar e remover sequências de caracteres indesejadas, como *emojis*
  - Encontrar e substituir sequências de caracteres para diversas funcionalidades, como tokenização

SEQUÊNCIAS DE CARACTERES

EXPRESSÕES REGULARES

Colab - ER

Padrão	Função	Resultado
José	casamento exato da string	ocorrência de “José”
[Ee] agora, José?	disjunção	ocorrências de “E agora, José?” e “e agora, José?”
[a-z]	intervalo	caractere minúsculo
[a-z]+	repetição	um ou mais caracteres minúsculos
[^a-z]	negação	o que não é caractere minúsculo

SEQUÊNCIAS DE CARACTERES

EXPRESSÕES REGULARES

Colab - ER

Padrão	Função	Resultado
\^	escape	trata um caractere especial como “normal”
[?|!]	múltiplo padrão de busca	ocorrências de ?” ou “!”
^	início	padrão será buscado no início do texto
$	fim	padrão será buscado no final do texto

SEQUÊNCIAS DE CARACTERES

EXPRESSÕES REGULARES

Colab - ER

Padrão	Função	Resultado
\b	o que não é alfanumérico	separador de palavras
?	ocorrência opcional	0 ou 1 vez do padrão
*	opcional ou várias	0 ou mais vezes do padrão
+	uma ou várias	1 ou mais vezes do padrão
.	curinga	qualquer caractere
{<NUM>}	NUM repetições	NUM ocorrências de um determinado padrão

SEQUÊNCIAS DE CARACTERES E *TOKENS*

TOKENIZAÇÃO

O menino foi para a escola de ônibus.

O	menino	foi	para	a	escola	de	ônibus	.

Usando expressão regular e sua “memória”, que “salva” o padrão encontrado para uso posterior

SEQUÊNCIAS DE CARACTERES E *TOKENS*

Colab - ER

TOKENIZAÇÃO

Padrão	Função	Resultado
\w	alfanumérico	qualquer caractere alfanumérico
\W	não alfanumérico	qualquer caractere não alfanumérico
\d	dígito	qualquer dígito
\D	não dígito	qualquer caractere que não seja dígito
\s	espaço	qualquer caractere de espaço
\S	não espaço	qualquer caractere que não seja espaço

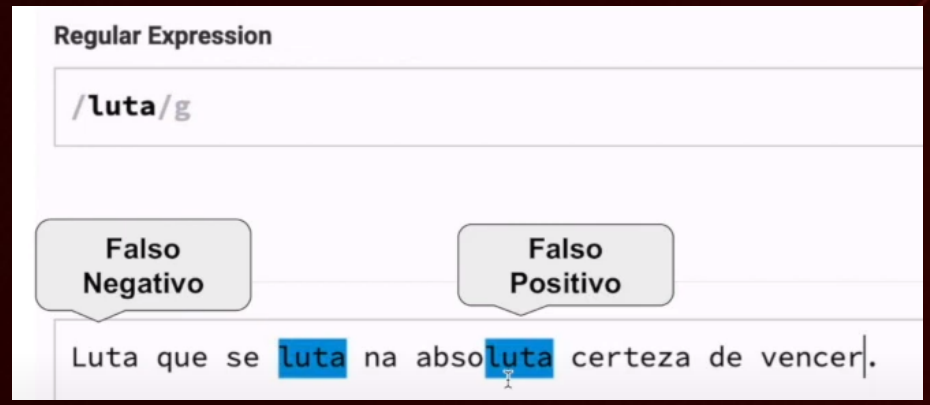
SEQUÊNCIAS DE CARACTERES E *TOKENS*

EXPRESSÕES REGULARES – CUIDADOS

Definir um conjunto de teste (*corpus*, padrão ouro, *gold standard*) que contenha tanto casos de falsos positivos como de falsos negativos

Falso positivo: termo que deveria ser retornado e não foi

Falso negativo: termo que foi retornado e não deveria ter sido



TOKENS

- **SEQUÊNCIAS DE CARACTERES GANHAM SENTIDO**
- **PRÉ-PROCESSAMENTO (NORMALIZAÇÃO)**
  - **Nível: morfologia**
  - **Objetivo: Sequências de caracteres são padronizadas para representarem algo que faça sentido na linguagem**
  - **Tokenização**
  - **Lematização**
  - **Radicalização**

TOKENS

TOKENIZAÇÃO

O menino foi para a escola de ônibus.

O	menino	foi	para	a	escola	de	ônibus	.

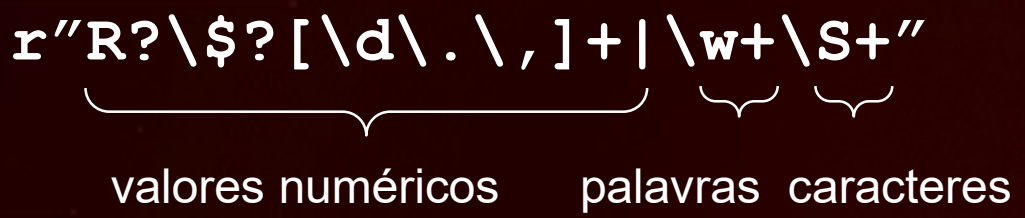
Visa distinguir as diferentes unidades linguísticas de um texto

Unidades linguísticas = *tokens*

TOKENS

TOKENIZAÇÃO

- Tradicionalmente, os tokenizadores são definidos com base nas regras linguísticas de uma língua
- Regras codificadas por meio de expressões regulares
- Por exemplo, para o português:



TOKENS

TOKENIZAÇÃO

- Com NLTK
- Quantidade de *tokens* e *types*
  - *Tokens* – conta todas as ocorrências
  - Types – conta apenas uma ocorrência = tamanho do vocabulário
- Subpalavras
  - Segmentar palavras raras em partes mais frequentes
  - Técnicas: Byte-Pair Encoding (BPE), Byte-level BPE, WordPiece, Unigram

Colab - NLTK

PALAVRAS

- **LEXEMA**
  - Unidade (abstrata) de significado
  - Corresponde a um conjunto de formas relacionadas
  - Ex.: menino, menina, meninão, menininha, meninos

- **LEMA**
  - Forma canônica, dicionarizada, escolhida por convenção para representar um lexema
  - Ex.: menino

- **RAIZ**
  - Morfema básico, sem afixos derivacionais ou flexionais
  - Ex.: menin

PALAVRAS

Colab - Spacy

LEMATIZAÇÃO

- Converte palavras em lemas

O	menino	foi	para	a	escola	de	ônibus	.

```
O  menino ~~ser~~ ~~parir~~ o escola  de ônibus .
          ir  para
```

PALAVRAS

Colab - NLTK

RADICALIZAÇÃO (STEMIZAÇÃO OU *STEMMING*)

- Converte as palavras para suas raízes

O	menino	foi	para	a	escola	de	ônibus	.

```
o  menin foi par  a  escol de ônibu .
```

PALAVRAS

STOPWORDS

- Em diversas aplicações de PLN é interessante desconsiderar algumas palavras que pouco acrescentam ao conteúdo do texto, como preposições, determinantes, conjunções etc.

- Essas palavras são conhecidas como *stopwords.*

PALAVRAS

REMOÇÃO DE *STOPWORDS*

Colab - Spacy

Colab - NLTK

O	menino	foi	para	a	escola	de	ônibus	.

```
menino foi          escola   ônibus
```

O QUE VIMOS?

- Sequências de caracteres
- Tokens
- Palavras

PRÓXIMA VIDEOAULA

- Prática de PLN com Python

REFERÊNCIAS

- Curso de Processamento de Linguagem Natural
  - Profa. Helena Caseli (UFSCar)

- Curso de Processamento de Linguagem Natural
  - Prof. Thiago Pardo (ICMC-USP)

- Curso de Linguística Computacional
  - Prof. Thiago Castro Ferreira (UFMG)