FACULDADE IBRATEC

Normalização de Texto

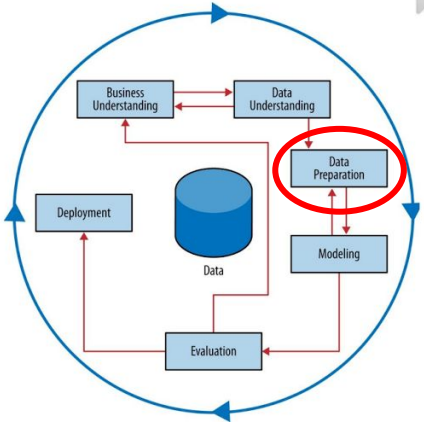
Prof. Marcelo Pita

Pós em Inteligência Artificial

Processamento de Linguagem Natural

Normalização de texto

Normalização é o conjunto de rotinas de tratamento e transformação aplicadas sobre o texto a fim de deixá-lo adequado para a modelagem analítica.



Normalização de texto

Rotinas que vamos explorar:

- Transformação para minúsculas
- Remoção de palavras por tamanho
- Remoção de *stop words*.
- Remoção de sinais de pontuação
- Estemização e lematização
- Remoção de sinais de acentuação
- Remoção de palavras por frequência

Transformação para minúsculas

Consiste em uniformizar o padrão de escrita das palavras, eliminando as diferenças tipográficas no que diz respeito à escrita com caixa alta ou caixa baixa.

Exemplo: **“Dai a César o que é de César”**

Texto transformado: “dai a césar o que é de césar”

Remoção de palavras por tamanho

Visa diminuir a quantidade de palavras erradas inseridas no texto com base no tamanho (erros de digitação, ruídos de OCR, etc).

Exemplo: Eliminar palavras com menos de 3 caracteres da frase **“A pressa é inimiga da perfeição”**

Texto transformado: “pressa inimiga perfeição”

Remoção de *stop words*

Stop words são as palavras mais comumente usadas em uma linguagem, removidas no pré-processamento por não acrescentarem muito valor semântico ao texto.

Exemplos de stop words comuns na língua portuguesa:

- Artigos (a, o, um, uma)
- Preposições (para, sobre, com, em)
- Alguns verbos (ser, estar, ter)
- Pronomes (meu, seu, quem, quando, que, algum).

Remoção de sinais de pontuação

Consiste em permitir no texto apenas caracteres alfanuméricos (i.e. números e letras do alfabeto). Excluir todo tipo de pontuação e outros caracteres especiais (?!:,;~{}[]&#@=+…).

Exemplo: **“Até tu, Brutus!”**

Texto transformado: “Até tu Brutus”

Estemização

A **estemização** (do inglês, *stemming*) consiste em reduzir palavras flexionadas ou derivadas à sua raiz.

Exemplos:

- As palavras “gato”, “gata”, “gatos”, “gatas”, “gatuno”, se reduzem a **“gat”**.
- As palavras “menino”, “meninas”, “meninice”, “meninada”, se reduzem a **“menin”**.

Lematização

Consiste em *deflexionar* as palavras, reduzindo-as aos seus **lemas**.

Exemplos:

- As palavras “gato”, “gata”, “gatos”, “gatas”, “gatuno”, se reduzem a **“gato”**.
- As palavras “menino”, “meninas”, “meninice”, “meninada”, se reduzem a **“menino”**.
- As palavras “tiver”, “tenho”, “tinha”, “tem” se reduzem a **“ter”**.

Remoção de sinais de acentuação

Visa uniformizar o tratamento de palavras que eventualmente foram escritas sem acentuação.

Exemplo: **“A pressa é a inimiga da perfeição”**

Texto transformado: “A pressa e a inimiga da perfeicao”

Remoção de palavras por frequência

Decisão de remover palavras pouco ou muito frequentes no texto.

Decisão de remoção deve levar em consideração o objetivo da análise.

Exemplo: palavras muito frequentes em um processo de classificação de texto podem ser eliminadas por não serem discriminativas.

O que é “muito”?