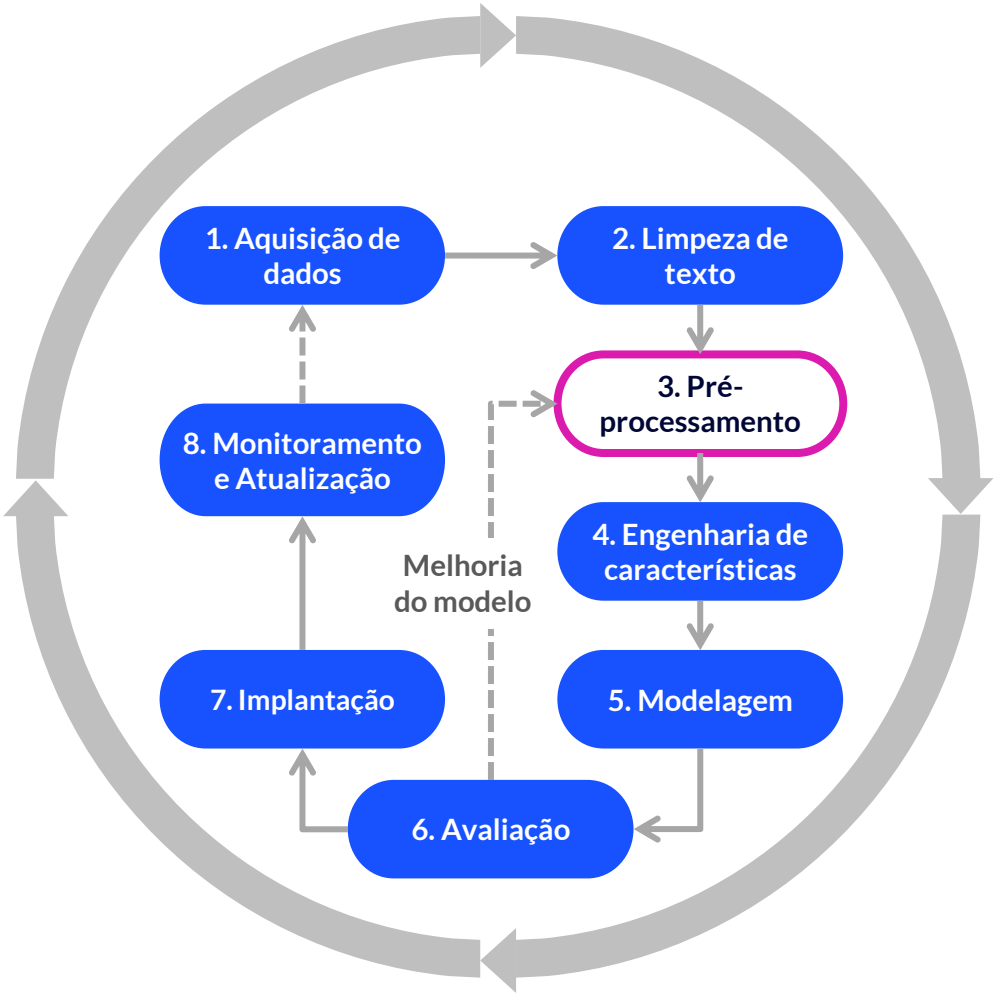
PUC Minas Virtual

PROCESSAMENTO DE LINGUAGEM NATURAL

PRÉ-PROCESSAMENTO

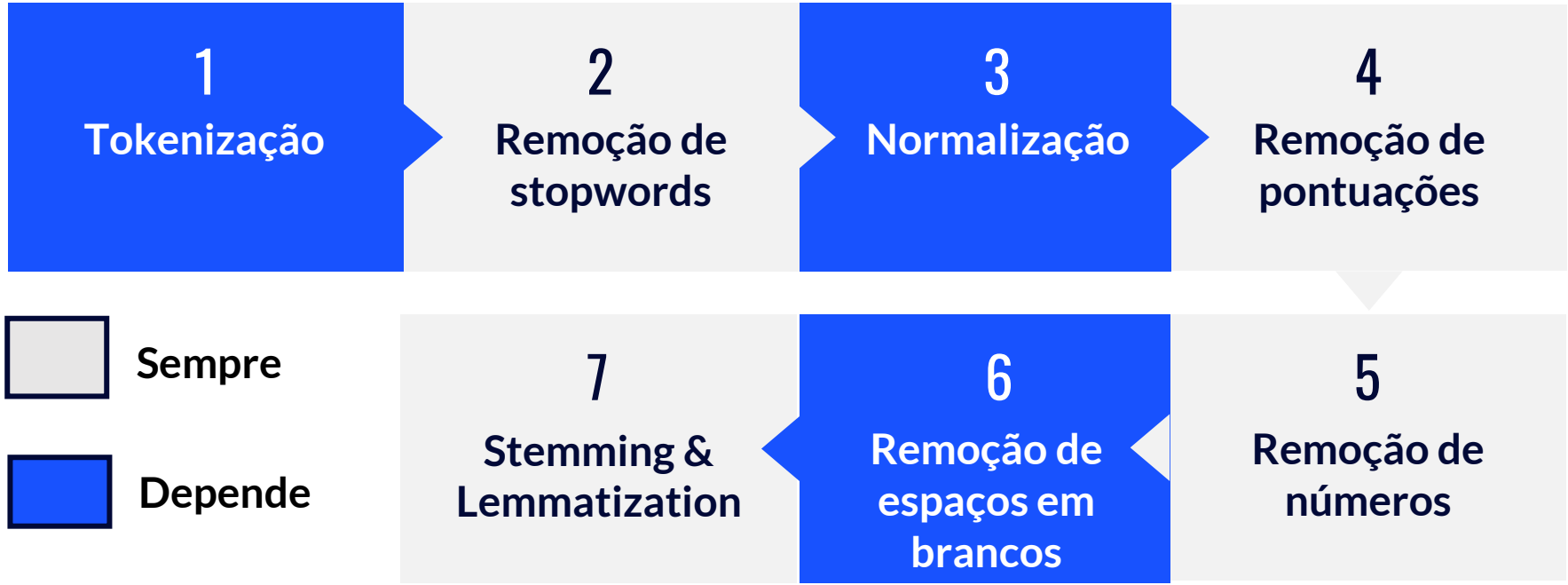
Pré-processamento



Pré-processamento de Textos

Técnicas aplicadas para organizar dados textuais

Pipeline de pré-processamento



Tokenização

- Extrair unidades mínimas de texto
- Tokens podem ser palavras, números ou sinais de pontuação
- Identificar as palavras que constituem uma sequência de caracteres.

Tokenização

Exemplo

O sol brincava feliz.
As crianças brincavam no parque.

O sol brilhava forte .

As Crianças Brincavam no Parque .

Tokenização

Exemplo

O sol brincava feliz.
As crianças brincavam no parque.

O	sol	brincava	feliz	.

As	Crianças	Brincavam	no	Parque	.

Remoção de stopwords

- Palavras comuns em um idioma
- Cada idioma tem seus stopwords
- Ocorrem com muita frequência
- Não contém informações importantes

Remoção de stopwords

- Palavras comuns em um idioma
- Cada idioma tem seus stopwords
- Ocorrem com muita frequência
- Não contém informações importantes

Remoção de stopwords

- Palavras comuns em um idioma
- Cada idioma tem seus stopwords
- Ocorrem com muita frequência
- Não contém informações importantes

Remoção de stopwords

- Palavras comuns em um idioma
- Cada idioma tem seus stopwords
- Ocorrem com muita frequência
- Não contém informações importantes

Exemplo

Remoção de stopwords

O sol brincava feliz .

As Crianças Brincavam no Parque .

sol brincava forte .

Crianças Brincavam

Parque .

Normalização

- Texto varia entre letras maiúsculas e minúsculas
- Prática comum: reduzir tudo para minúsculo
- Vocabulário é reduzido, mas alguns significados podem ser perdidos

Normalização

- Texto varia entre letras maiúsculas e minúsculas
- Prática comum: reduzir tudo para minúsculo
- Vocabulário é reduzido, mas alguns significados podem ser perdidos

Normalização

- Texto varia entre letras maiúsculas e minúsculas
- Prática comum: reduzir tudo para minúsculo
- Vocabulário é reduzido, mas alguns significados podem ser perdidos

Exemplo

Normalização

O sol brincava feliz .

As Crianças Brincavam no Parque .

sol brincava forte .

crianças brincavam

parque .

Remoções úteis

- Remoção de números
- Remoção de pontuação
- Remoção de caracteres especiais
- Remoção de excesso de brancos

Exemplo

Remoção
úteis

O sol brincava feliz .

As Crianças Brincavam no Parque .

sol brincava feliz

crianças brincavam

parque

- Redução da palavra do léxico até a raiz
- Eliminação de sufixos
- Stemming é um corte bruto de afixos

Stemming

7

Stemming

- Redução da palavra do léxico até a raiz
- Eliminação de sufixos
- Stemming é um corte bruto de afixos

7

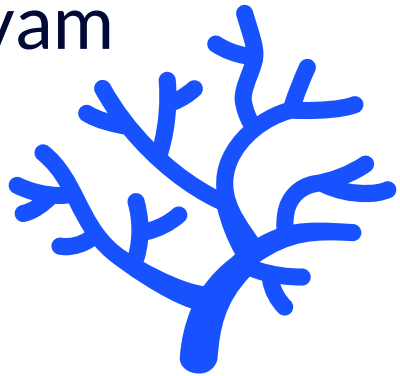
Stemming

- Redução da palavra do léxico até a raiz
- Eliminação de sufixos
- Stemming é um corte bruto de afixos

Stemming

Exemplo

brincavam



brincava

brinc

8

Lemmatization

- Reduz as palavras para sua forma básica linguisticamente correta
- Transforma a palavra raiz com dicionário e análise morfológica

8

Lemmatization

- Reduz as palavras para sua forma básica linguisticamente correta
- Transforma a palavra raiz com dicionário e análise morfológica

Exemplo

Lematization

brincavam

brincava

brincar

Comparação

Stemming

01 Corte bruto de afixos

02 Análise morfológica com dicionário

03 Busca rápida e tratamento de grandes volumes de texto

VS

Lemmatization

Análise morfológica com dicionário 01

Forma de dicionário gramaticalmente correta 02

Análise precisa em tarefas de NLP 03

NLTK

https://pypi.org/project/nltk/

PUC Minas
Virtual