Disciplina:

Processamento de linguagem natural

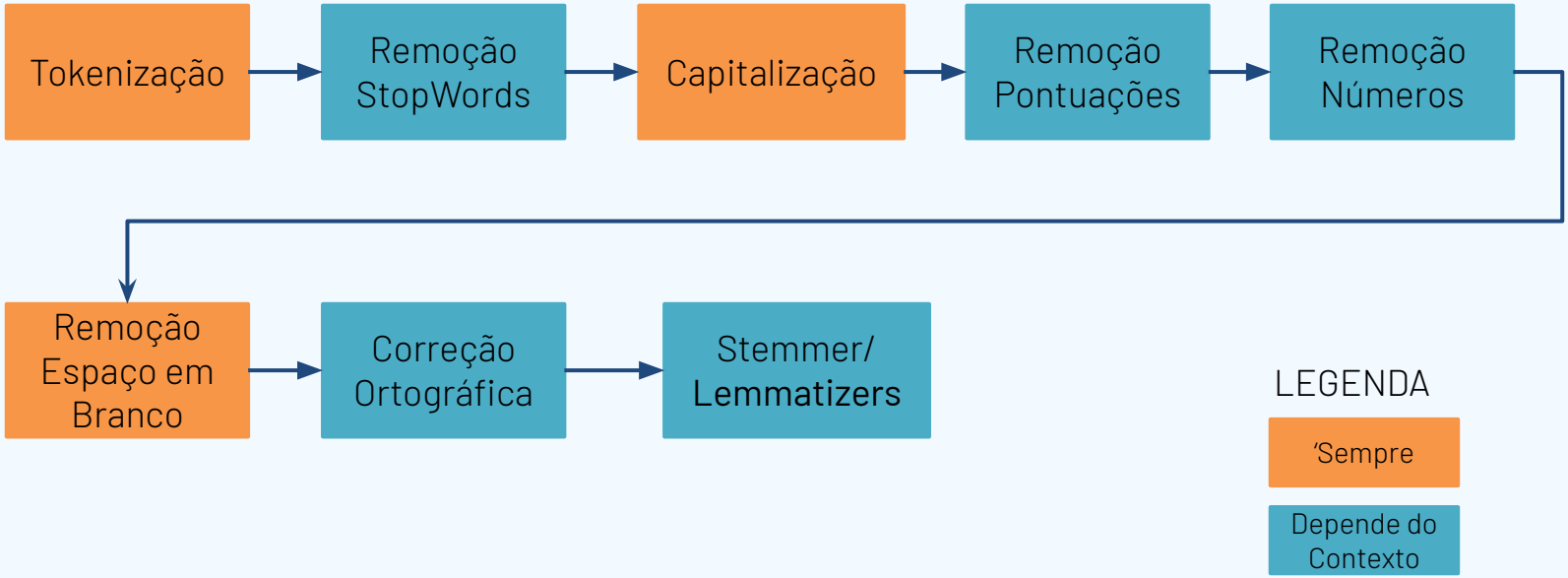
Professor: Gabriel Assunção

Pré-processamento

O que é pré-processamento?

- Normalização dos dados para análises.
- Melhorar a qualidade do dado.
- Reduzir ruídos nas análises

Pipeline de processamento



Tokenização

Tokenização

- **Objetivo:** Extrair unidades mínimas do texto.
- Tokens podem ser palavras, números ou frases.
- Processar o dado para que possamos extrair a informação/significado dele.

Ferramentas para tokenização:

Split, regex, NLTK, spacy, keras, gensim.

Tokenização

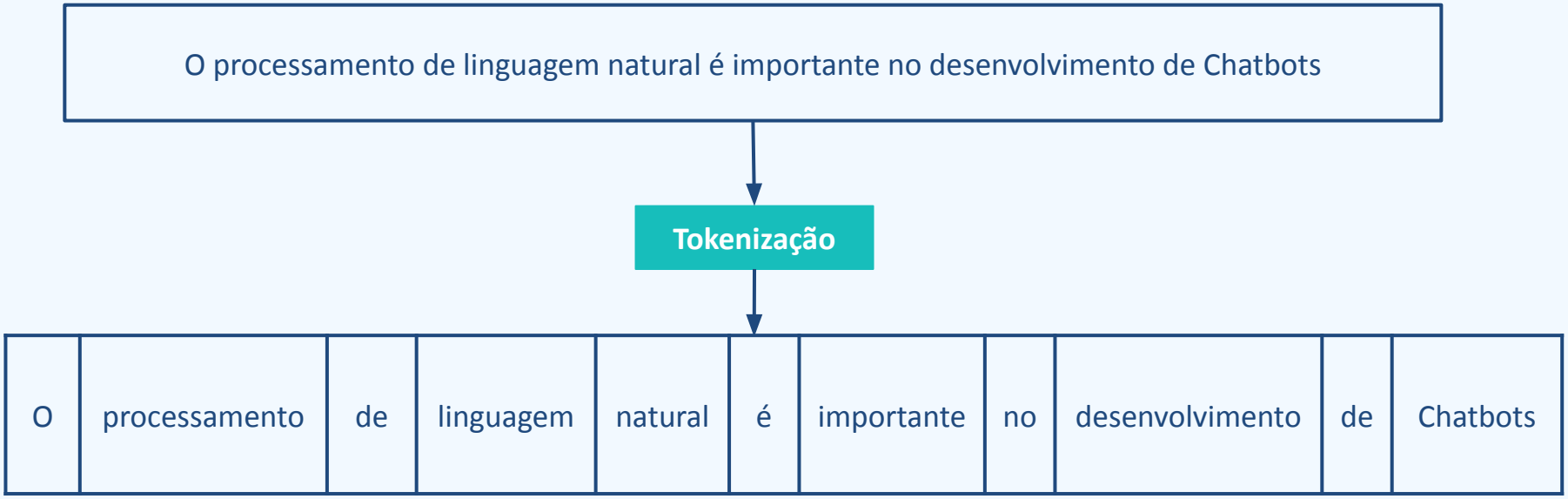
O processamento de linguagem natural é importante no desenvolvimento de Chatbots. O NLP pode ser usado em diversas áreas.

Tokenização de Sentença

O processamento de linguagem natural é importante no desenvolvimento de Chatbots.

O NLP pode ser usado em diversas áreas.

Tokenização



StopWords

StopWords

- **Objetivo:** remover palavras que muitas vezes não adicionam informação ao texto.
- Stopword são palavras que são: comuns em um idioma, palavras que conectam sentenças. Exemplos: as, e, de, oi.
- Cada idioma tem seu conjunto de stopwords.

Ferramentas para remoção de stopwords:

NLTK.

Capitalização

Capitalização

- **Objetivo:** padronizar as palavras para evitar case-sensitive.
- Prática usual é passar todas as palavras para minúsculo.
- Podemos perder informação como nome próprio (inicial em maiúsculo); siglas (USA e usa);

Ferramentas para capitalização:

Funções nativas do Python.

Remoção de caracteres

Remoção de caracteres

- **Objetivo:** remover caracteres que não irão acrescentar informação ao texto ou substituir o caracter para padronização.
- Remoção de pontuação
- Remoção/substituição de números
- Remoção de espaços em brancos
- Remoção de caracteres especiais
- Remoção/substituição de emoji

Ferramentas para remoção de caracteres:

Regex, emoji

Stemming

Stemming

- **Objetivo:** reduzir a palavra até o seu radical eliminando sufixos.
- Exemplos:
    - automate(s), automatic, automation→automat
    - Estou estudando muito -> estou estud muit

Ferramentas para stemming:

NLTK

Lemmatization

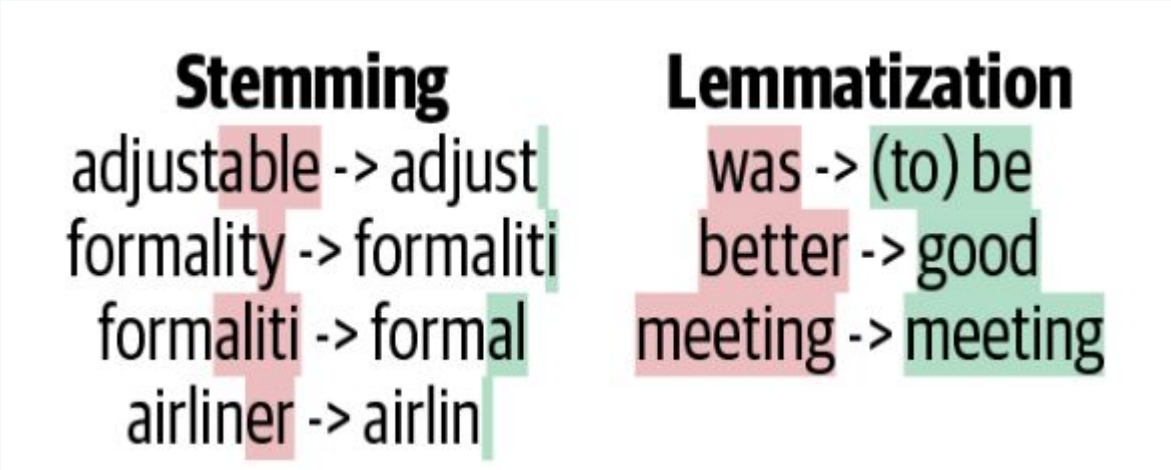
Lemmatization

- **Objetivo:** reduzir a palavra até a sua forma básica.
- Usa de dicionários e análise morfológica.
- Exemplos:
    - am, are, is⟶ be
    - car, cars, car's, cars' ⟶ car
    - Estou estudando muito -> Estar estudar muito

Ferramentas para Lemmatization:

NLTK

Lemmatization x Stemming



Lemmatization x Stemming

- Stemmers usam uma abordagem algorítmica para remover os prefixos e sufixos. O resultado pode não ser uma palavra real
- Stemmers são mais rápidos do que lemmatizers.
- Se você só quer garantir que o seu sistema é tolerante a variações de palavras, use Stemmers.
- Se você precisa de palavras existentes num dicionário, use um Lemmatizer.