# Introdução ao Processamento de Linguagem Natural

Pipeline geral para trabalhar com textos

# O que é “NLP”?

Processamento de Linguagem Natural, ou “NLP” (sigla em inglês), é uma área interdisciplinar da linguística com a computação que visa estudar como o computador recebe, interpreta/processa e reproduz a língua natural humana.

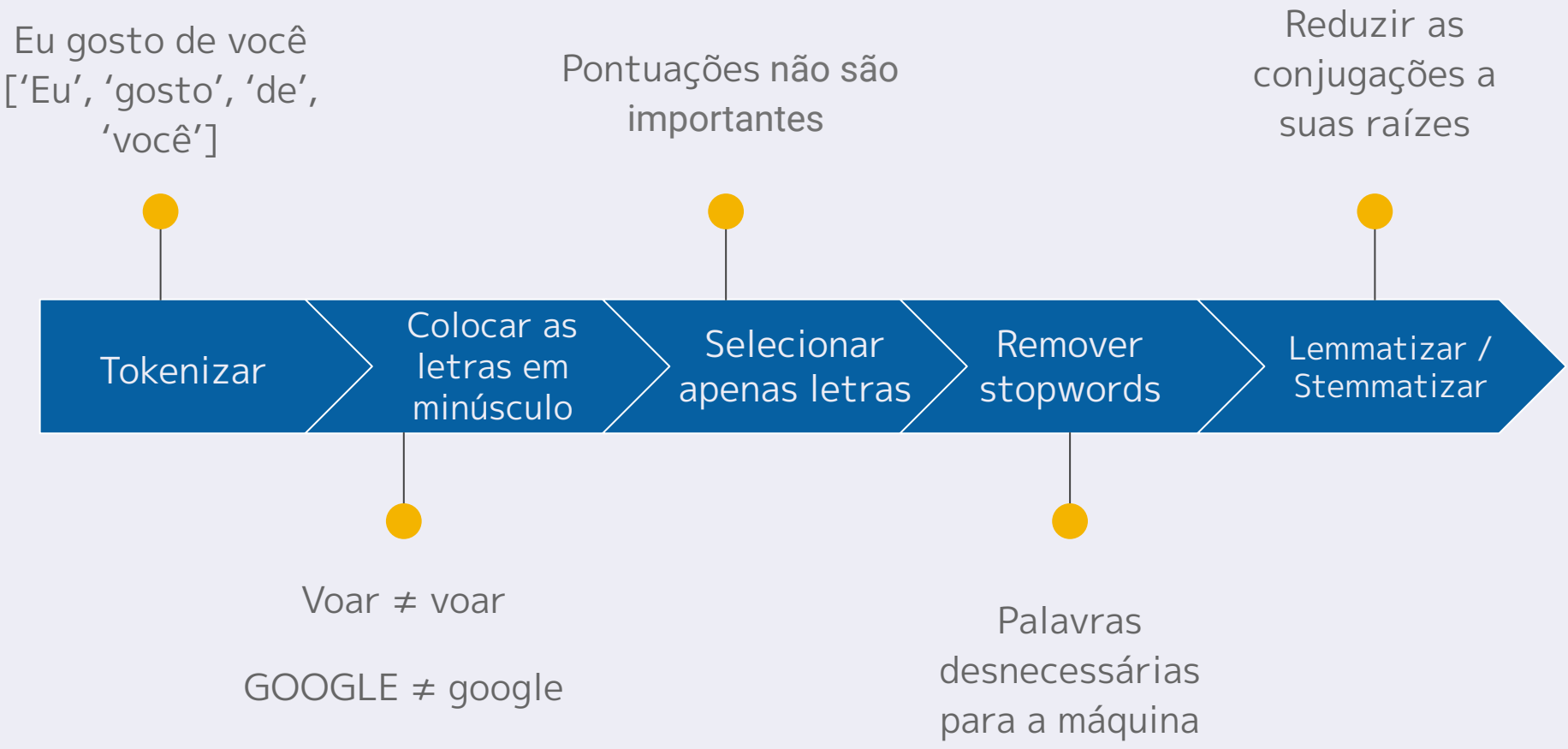
# Onde usar NLP?

- Construção de chatbots
- Análise de sentimentos
- Tradução automática
- Assistentes digitais
- Corretores / resumos automáticos de texto
- Muito mais!

# Pré-processamento

## Limpando e preparando seu texto

- ➔ Pré-processar um texto é preparar ele para ser recebido pela máquina
- ➔ Esse passo é importante quando usamos algoritmos mais simples de Machine Learning, e deixa de ser tão necessário com algumas técnicas de Deep Learning (modelos robustos)
- ➔ Nem toda tarefa requer o mesmo pré-processamento


Eu gosto de você
['Eu', 'gosto', 'de', 'você']
Pontuações não são importantes
Reduzir as conjugações a suas raízes
Tokenizar
Colocar as letras em minúsculo
Selecionar apenas letras
Remover stopwords
Lemmatizar / Stemmatizar
Voar ≠ voar
GOOGLE ≠ google
Palavras desnecessárias para a máquina

# Vamos ver cada um desses passos no código

# Feature Extraction

Para usarmos um modelo precisamos de features: **uma forma estruturada de armazenar informações.** Porém, textos são um tipo de dado não estruturado, assim, é difícil para o computador entendê-los e analisá-los.

Por isso, realizamos a chamada feature extraction, ou seja, transformamos o texto em uma informação numérica de modo que seja possível utilizá-lo para alimentar um modelo. Uma das maneiras mais populares e simples de fazer isso é com Bag of Words (BoW).

# Bag of Words

## O que é?

BoW é uma forma de representar o texto de acordo com a ocorrência das palavras nele. O “saco de palavras” recebe esse nome porque não leva em conta a ordem ou a estrutura das palavras no texto, apenas se ela aparece ou a frequência com que aparece nele.

Por exemplo, se a palavra AULA aparece muito num texto, ela se torna mais central e importante para a máquina. Portanto, BoW pode ser um ótimo método para determinar as palavras significativas de um texto com base no número de vezes que ela é usada.

# Passos para aplicar o BoW

- Selecionar os dados
- Gerar o vocabulário
- Formar vetores a partir do documento

# Vamos ver cada um desses passos no código

# TFIDF

## O que é?

Outro método de feature extraction, no qual, diferente do BoW, não consideramos somente a frequência de uma palavra em seu documento, mas também a repetição dela ao longo de todos os documentos que estão sendo analisados.

Exemplo:

Analisando três documentos: uma revista de futebol, uma de vôlei e uma de basquete.

- “esporte”: aparece nas 3, pouco relevante para a análise;
- “cesta”: se repete muito na revista sobre basquete, mas não nas outras, tende a se tornar mais importante para o TF-IDF.

# TFIDF

## O que é?

Cada palavra no documento recebe uma pontuação TF-IDF, feita multiplicando duas métricas diferentes:

- TF = Term Frequency (a frequência do termo), que mede a frequência com que um termo ocorre num documento;
- IDF = Inverse Document Frequency (inverso da frequência nos documentos), que mede o quão importante um termo é no contexto de todos os documentos.

$$TFIDF = TF \times IDF$$

OU

$$TFIDF = \frac{\text{nº de vezes que uma palavra aparece em documento}}{\text{nº de palavras do documento}} \times \log \frac{\text{Total de documentos}}{\text{Nº de documentos com o respectivo termo}}$$

# Vamos ver cada um desses passos no código

# Novos modelos de Processamento de Linguagem Natural

Novos modelos de Deep Learning não requerem mais exatamente os mesmos passos. Por exemplo, stopwords não fazem mais diferença no processamento de um texto. Entretanto, essas técnicas que vimos hoje ainda são usados para modelos simples de Machine Learning.

# Obrigado a todos!

## Alguma dúvida?