

Fundamentos do Processamento de Linguagem Natural

A decorative graphic spanning the width of the slide. It features a blue line graph with circular markers, some of which are highlighted with a white border. The graph is set against a background of green and yellow areas that resemble a stylized landscape or data visualization. The entire graphic is positioned below the main title and above the course information.

SIN260 - SISTEMAS INTELIGENTES

Isabela Neves Drummond

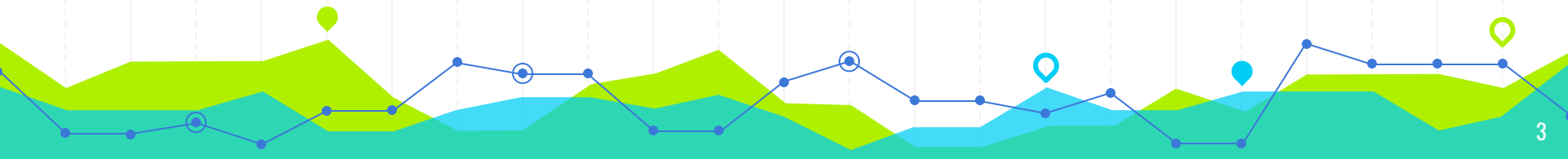
Flávio Belizário da Silva Mota

Introdução

- O que é linguagem natural?
 - É a linguagem que usamos para nos expressar
 - Um conjunto de protocolos acordados mutuamente envolvendo palavras e sons que usamos para nos comunicarmos

Introdução

- O PLN tem como objetivo permitir que máquinas possam compreender e produzir expressões em língua humana
- Surgiu do estudo de campos como a inteligência artificial, linguística, linguagens formais e compiladores
- A maior parte dos trabalhos começaram durante os anos 1980

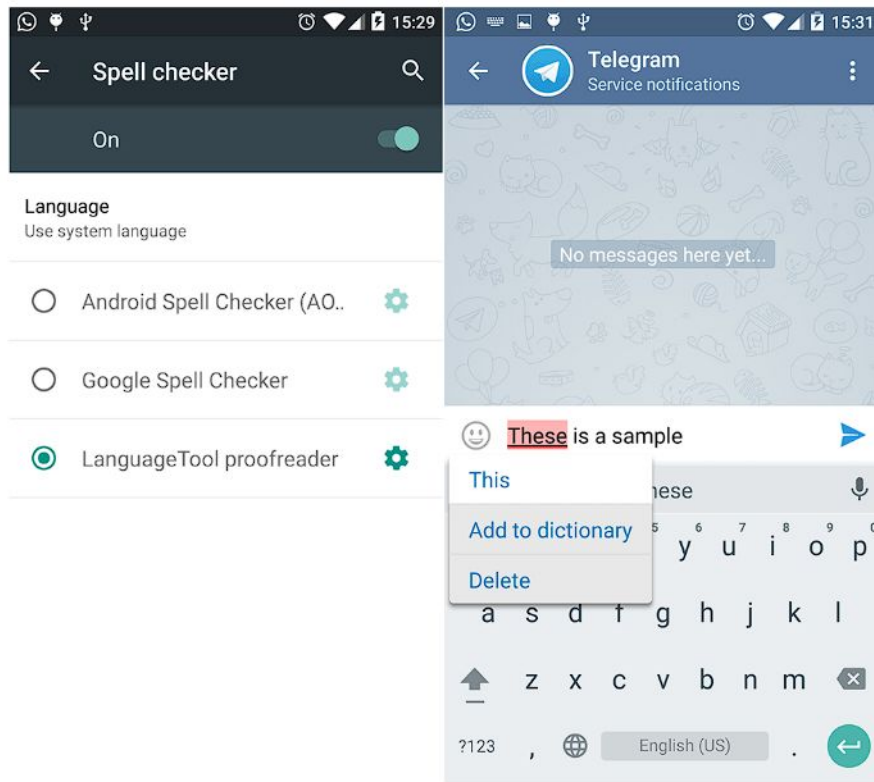


Introdução

- O PLN pode ser categorizado de forma abrangente em dois tipos:
 - Compreensão de Linguagem Natural (NLU): habilidade de uma máquina compreender a linguagem natural
 - Geração de Linguagem Natural (NLG): habilidade de uma máquina se manifestar utilizando uma língua que os humanos estão aptos a entender

Aplicações

● Corretores Automáticos



Aplicações

● Auto-Complete



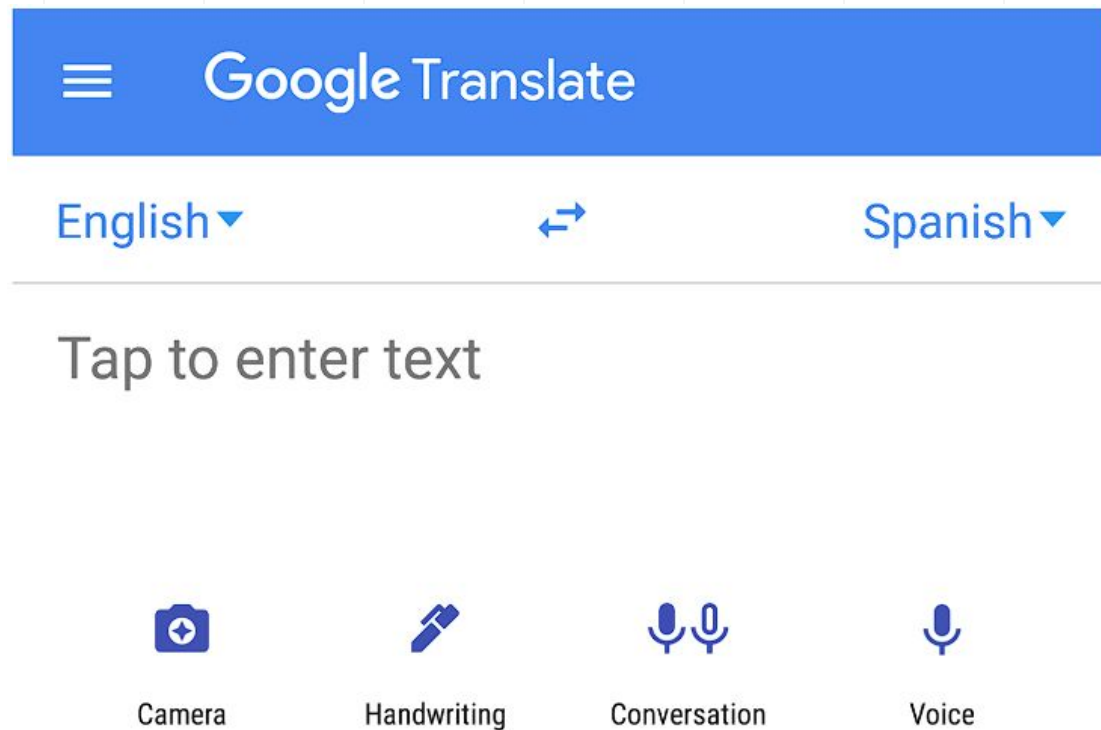
san f

- san francisco weather
- san francisco
- san francisco giants
- san fernando valley
- san francisco state university
- san francisco hotels
- san francisco 49ers
- san fernando
- san fernando mission
- san francisco zip code

Google Search I'm Feeling Lucky

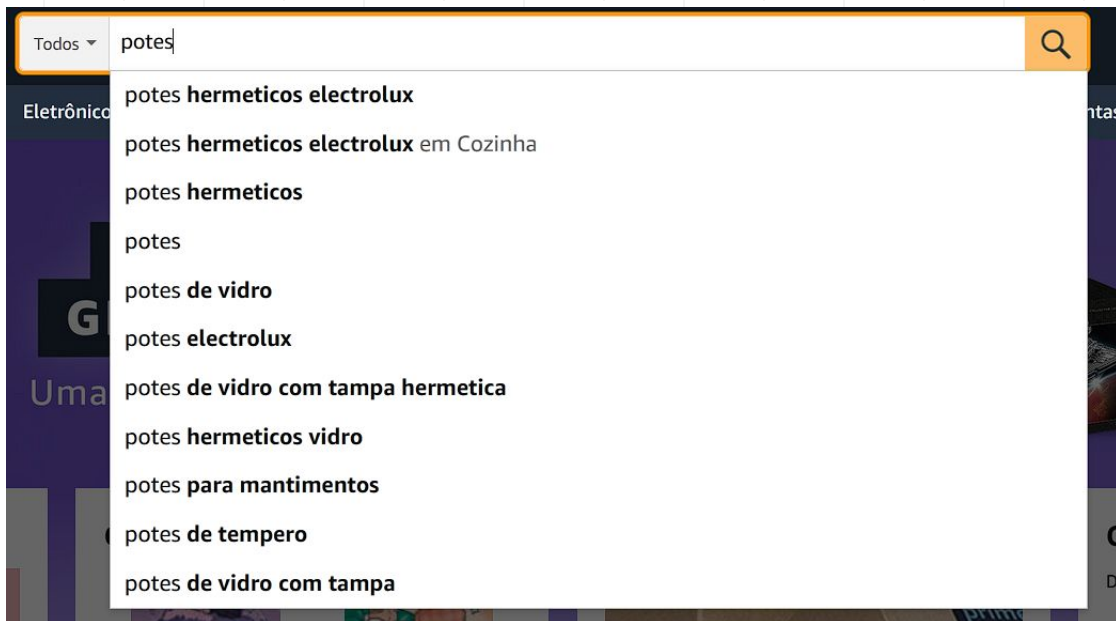
Aplicações

● Tradução



Aplicações

● Busca Inteligente

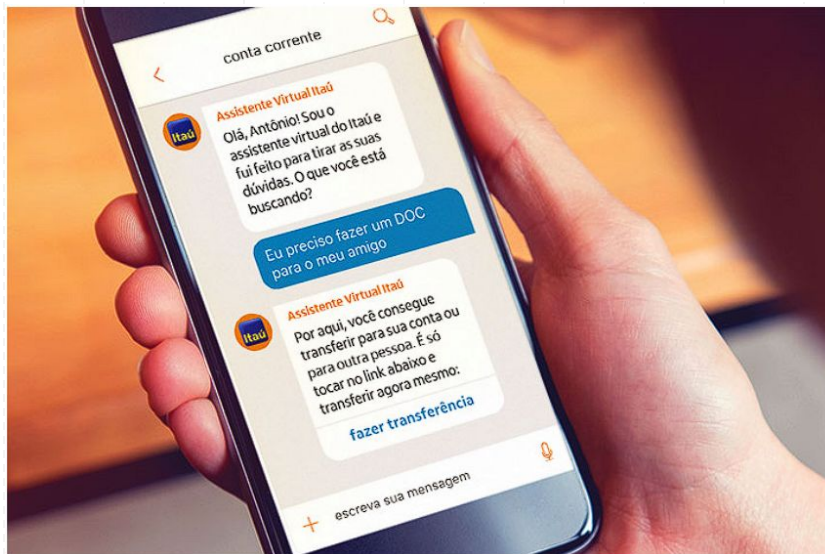


Aplicações

- Bots de Mensagens
- Assistentes Pessoais



"Alexa, quando é a minha próxima reunião?"



Introdução

- O que é um texto?
- Um texto é uma sequência de palavras, ou frases, ou sentenças, ou parágrafos
- Podemos pensar no texto como sendo apenas uma sequência de caracteres mas há uma limitação nessa definição que não interessa para o contexto de PLN

Introdução

- O que é uma palavra?
- É natural pensar em um texto como uma sequência de palavras
Uma palavra é uma sequência significativa de caracteres
- Como encontrar os limites das palavras? No Português podemos dividir uma palavra por espaços ou pontuação

Prática - Google Colab

The logo for Google Colab, featuring the word "colab" in a bold, orange, sans-serif font.

Introdução ao Processamento de Linguagem Natural

- A NLTK (**N**atural **L**anguage **T**ool**K**it) é uma biblioteca do python que oferece uma infinidade de recursos para o tratamento de dados textuais. Ela conta com uma excelente documentação, que inclui passo a passo a utilização dos recursos. Um diferencial dessa biblioteca é a disponibilidade de recursos linguísticos em Português.

Tarefas do PLN

- *Tokenização*
- Limpeza do texto
- Etiquetagem
- Contagens de palavras
- *Stemming*



Classificação de textos

- Extração de características

- Bag of Words

- Tem como objetivo construir uma matriz onde cada coluna representa uma palavra do vocabulário de um documento e cada linha um texto desse documento

Texto 1: “Eu gosto do detetive Sherlock Holmes.”

Texto 2: “Sherlock Holmes não é um buscador da verdade.”

	Eu	gosto	do	detetive	Sherlock	Holmes	não	é	um	ele	buscador	da	verdade
Texto 1	1	1	1	1	1	1	0	0	0	0	0	0	0
Texto 2	0	0	0	1	1	1	1	2	2	1	1	1	1

Classificação de textos

● Extração de características

● TF-IDF

- O problema do BoW é que a frequência de um termo no texto não representa totalmente quanto de informação ele agrega
- Termos que aparecem pouco (mais raros) podem agregar muito mais informação
- O TF-IDF é uma estratégia que permite quantificar a informação que é agregada por um termo

Classificação de textos

- Extração de características
 - TF-IDF = Term Frequency - Inverse Document Frequency
 - $tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t$
 - TF => Frequência do Termo:
 - A Frequência de um termo t em um documento $d = df_t$
 - IDF => Frequência Inversa em Documentos
 - $idf_t = \log_{10} (N/df_t)$
 - onde N é o número total de documentos

Classificação de textos

df_t	N/df_t	idf_t	relevância
1000	1	0	nenhuma
100	10	1	pouca
10	100	2	média
1	1000	3	muita

N = 1000 documentos

Fases de um projeto de PLN



Referências

1. FERREIRA, M.; LOPES, M. **Linguística Computacional**. 1. ed. Contexto, 2020.
2. LUGER, G. **Inteligência artificial**. 6. ed. Pearson, 2013.
3. RASCHKA, S.; MIRJALILI, V. **Python Machine Learning**. 2. ed. Packt, 2017.
4. GHOSH, S.; GUNNING, D. **Natural Language Processing Fundamentals**.

