



AULA 10

Processamento de Linguagem
Natural – Parte 1

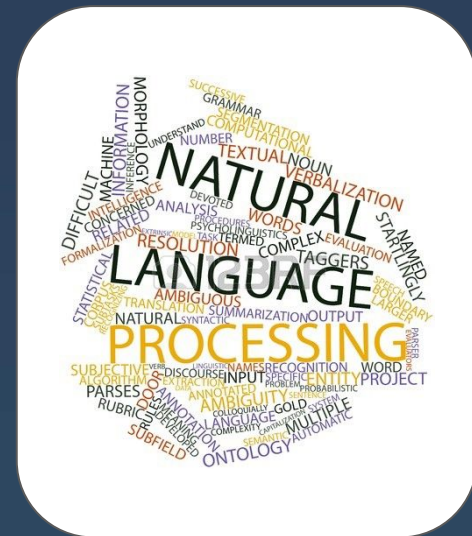
ROTEIRO DA AULA

1. Quem sou eu?
2. Apresentação dos Alunos
3. Dados não estruturados
4. Processamento Digital de Sinais
5. Processamento Digital de Imagens

Introdução

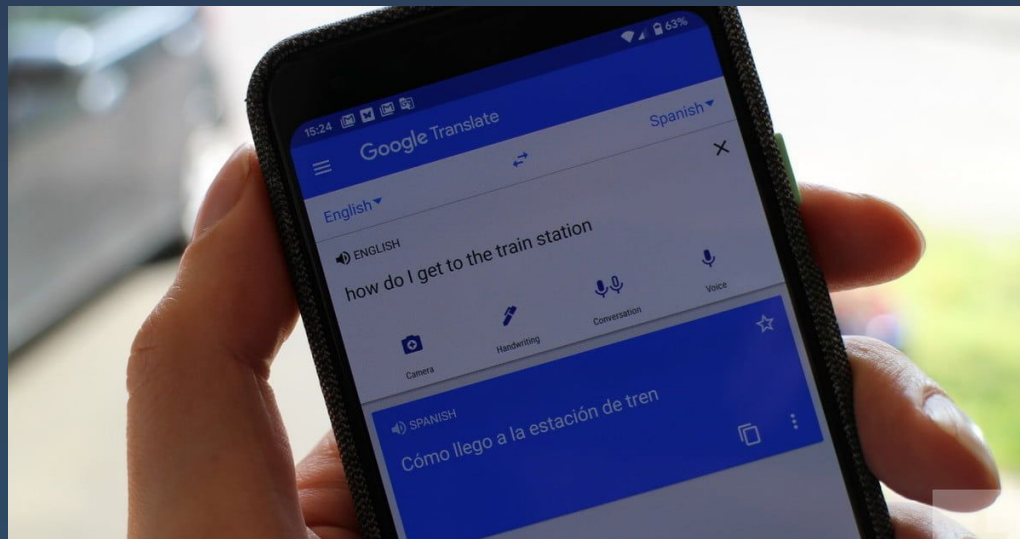
DESAFIOS NO PLN

- Linguagem informal
- Problemas de segmentação (diferentes sentidos em diferentes contextos)
- Diferentes idiomas
- Neologismos
- Nomes de entidades



Aplicações do PLN

1. Tradução Automática

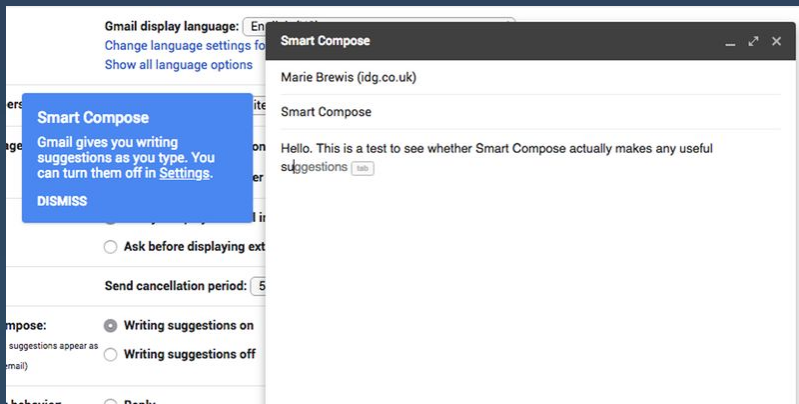


Aplicações do PLN

2. Extração contextual

Hi Dan, we've now scheduled the curriculum meeting.
It will be in Gates 159 tomorrow from 10:00-11:30.
-Chris

Create new Calendar entry



Aplicações do PLN

3. Análise de sentimentos



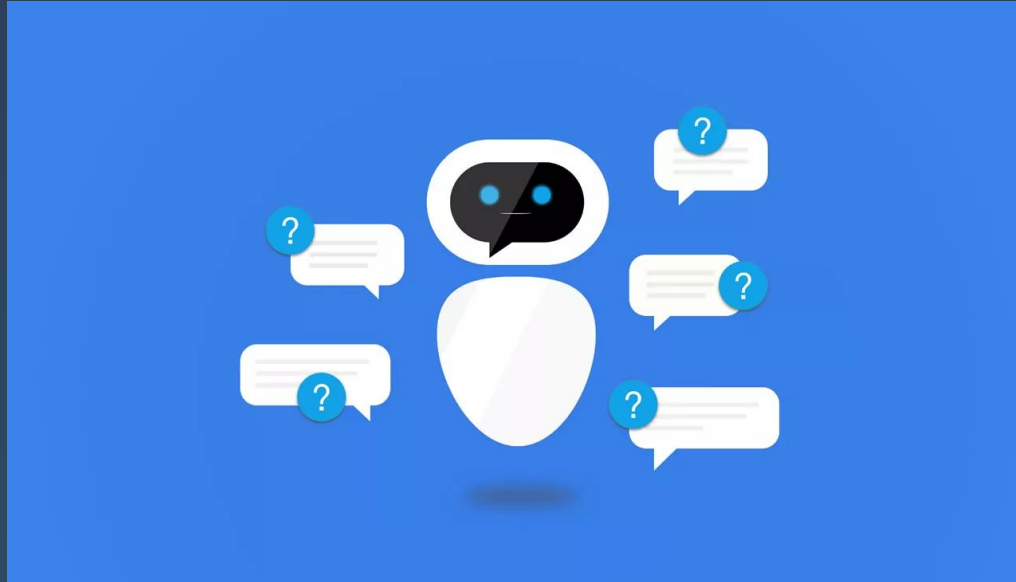
Aplicações do PLN

4. Detecção de Spam




Aplicações do PLN

5. Chatbots



Estado da Arte

Você pode acompanhar o desenvolvimento de pesquisas na área...

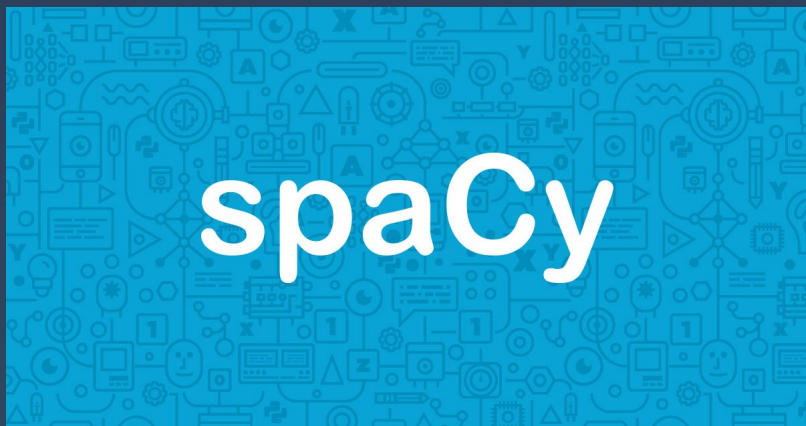
View on GitHub 

NLP-progress

Repository to track the progress in Natural Language Processing (NLP), including the datasets and the current state-of-the-art for the most common NLP tasks.

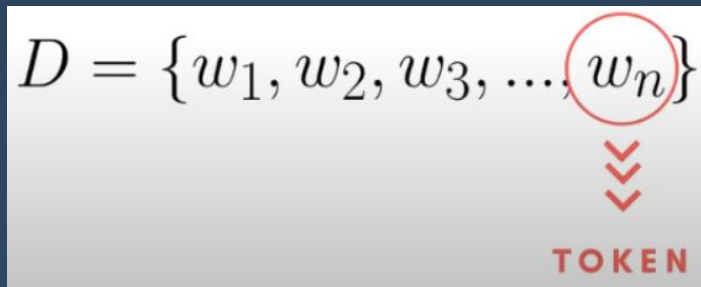
<https://nlpprogress.com/>

Na prática!



Documento

- Conjunto de palavras e caracteres especiais que compõem os objetos de estudo de uma atividade de NLP
- Exemplos: uma frase, uma resposta a um questionário, um texto de blog, uma página web, dentre outros.



The diagram shows a mathematical set definition $D = \{w_1, w_2, w_3, \dots, w_n\}$ on a light gray background. The term w_n is circled in red. Below the circle, three red chevrons point downwards to the word "TOKEN" in red capital letters.

$$D = \{w_1, w_2, w_3, \dots, w_n\}$$

TOKEN

Tokens

- Um documento (pensando no Spacy) é uma sequência de objetos do tipo token e possui diversas informações sobre o texto que ele contém.
- Dividindo a frase em tokens, o documento é uma estrutura iterável e portanto, deve ser acessada como tal.
- Portanto, o token é uma parte da estrutura e pode ser uma frase, palavra, uma pontuação, um espaço em branco, etc.
- Se o nosso documento é uma frase, os tokens serão constituídos de palavras e pontuações.

Part of Speech

- Técnica que lê um texto, em algum idioma, e assinala, para cada palavra, a classe gramatical a qual ela pertence.
- Um dos POS mais conhecidos foi desenvolvido pela Universidade de Stanford.



Stemmização

- Técnica que reduz palavras flexionadas/conjugadas, de um determinado idioma, para a sua raiz.
- A raiz (ou radical) de uma palavra é a menor parte da mesma, que contém o seu significado léxico, sem os seus afixos ou flexionais.

EXEMPLO

encontrar encontrarão encontraram encontrei encontraríamos

Radical: encontr

go going goes gone went

Radical: go gone went

Lemmatização

- Técnica que reduz palavras ao seu lema, sua forma que aparece no dicionário.
- Difere do processo de stemmização por considerar o contexto no qual a palavra está inserida, bem como sua classe gramatical.

EXEMPLO

encontrar encontrarão encontraram encontrei encontraríamos

Radical: encontr

go going goes gone went

Radical: go gone went