

NLP

NATURAL

LANGUAGE

PROCESSING

INTRODUÇÃO

Tópicos especiais em gestão de TI.

O QUE É PROCESSAMENTO DE LINGUAGEM NATURAL?

- Podemos defini-lo como um algoritmo que passa a compreender funções da linguagem humana.
- Ex.: Tradutor de textos.
- Tem uma relação grande com a ciência de dados, porém é um mundo à parte.
- Extensa área de pesquisa, com grandes oportunidades e grande inovações.

PRINCIPAIS PROBLEMAS DE NLP

- Existem alguns problemas de NLP que são extremamente difíceis, sendo eles:
 - Ambiguidade - José encontrou sua namorada com seu primo.
 - Ironia – Estudou 10 minutos, certamente vai passar no ENEM.
 - Sensibilidade ao contexto – Ele assiste TV a cabo graças a um gato.
 - Ordem - Joana é uma nova pessoa. | Joana é uma pessoa nova.

CONCEITOS

- Letras
- Palavras
- Sentenças
- Documentos - Texto
- Corpus – Conjunto de documentos;

CONCEITOS

- Anotação – Localizar e classificar elementos no texto, muito utilizado para análise de sentimentos.
- Tokenização – Separação das sentenças em partes menores, sendo elas: palavras, pontos, símbolos

Destacados representantes del Parlamento y la prensa rusos criticaron hoy el "belicism ha definido como posible blanco de su lucha antiterrorista.

El presidente de la Duma (cámara baja), Guennadi Seleznirov, calificó de "claramente ap del Kremlin para Chechenia, Serguéi Yastrzhembski.

El asesor presidencial dijo que Rusia puede lanzar un ataque preventivo contra los camp

| A chuva estragou os planos. | | | | | |
|-----------------------------|-------|----------|----|--------|---|
| TOKENIZATION | | | | | |
| A | chuva | estragou | os | planos | . |

CONCEITOS

- Parts-of-Speech tagging (POS) – Adiciona a classe gramatical(verbo, substantivo, adjetivo, etc) a cada token.
- As principais bibliotecas de NLP já fazem esse trabalho automaticamente.
- Vamos comer **porco**. (Substantivo)
- Ele não toma banho, é um **porco**! (Adjetivo)
 - Algumas bibliotecas conseguem realizar corretamente essa classificação.

CONCEITOS

- Lemmatizing (Lemma) - A lematização em linguística é o processo de agrupar as formas flexionadas de uma palavra para que possam ser analisadas como um único item, identificado pelo lema da palavra, ou forma de dicionário.

| Original | Lemma |
|----------|---------|
| Amigos | Amigo |
| Amigas | Amigo |
| Amizade | Amizade |

CONCEITOS

- Stemming – Corta a palavra, trazendo apenas o radical da mesma.
- Lemmatizing é mais sofisticado.

| Original | <u>Stemming</u> |
|----------|-----------------|
| Amigos | <u>Amig</u> |
| Amigas | |
| Amigo | |

CONCEITOS

- NGRAM – N palavras consecutivas.
- Muito comum utilizar bigram ou trigram, sequencias maiores não é recomendado, pois podem aumentar muito o volume dos dados.

APLICAÇÕES DE NLP

- Tradutor de idiomas
- Análise de sentimento
- Sintetização de fala (Alexa, etc.)
- Previsão de digitação
- Classificação de textos
- Chatbots
- Identificar autoria de textos
- Similaridade entre documentos
- Reconhecimento de entidades nomeadas
- Produção de resumos
- Perguntas e respostas