AULA 11

Processamento de Linguagem Natural - Parte II

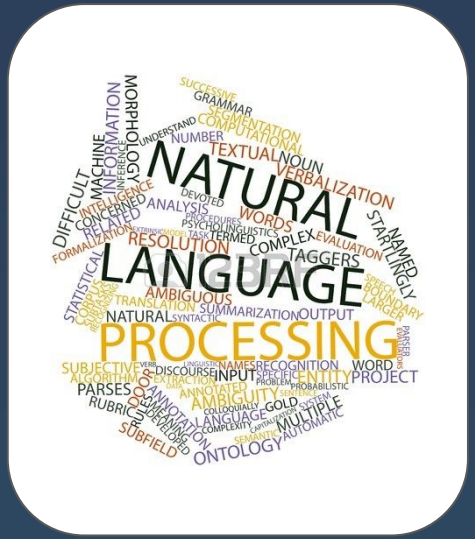
ROTEIRO DA AULA

1. Parts of Speech - POS
2. Named Entity Recognition - NER
3. Extração de Features: Vetorização

Tarefas Importantes no PLN

- ***Part of Speech Tagging***: Etiquetagem de elementos textuais, com o fim de evidenciar a estrutura gramatical de um determinado trecho de texto.

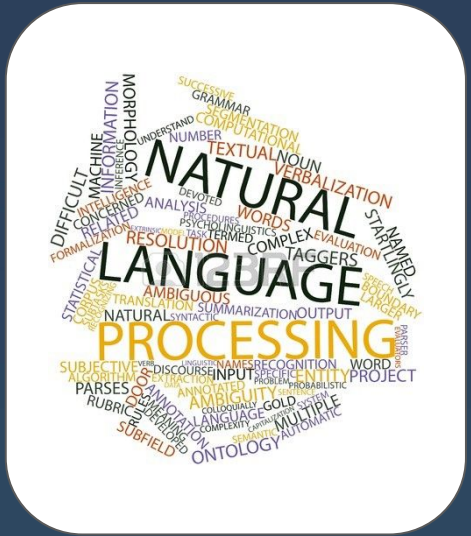
- **Reconhecimento de entidade nomeada (*Named Entity Recognition* — NER)**: Permite identificar e classificar nomes em textos.



Tarefas Importantes no PLN

- **Extração de informação**: Método para converter dados não estruturados (texto) em em dados estruturados (atributos).

- **Classificação de textos**: Classifica textos em classes (ex: Análise de sentimentos, Classificação de Spam, Identificação de Autenticidade).



Part of Speech - POS

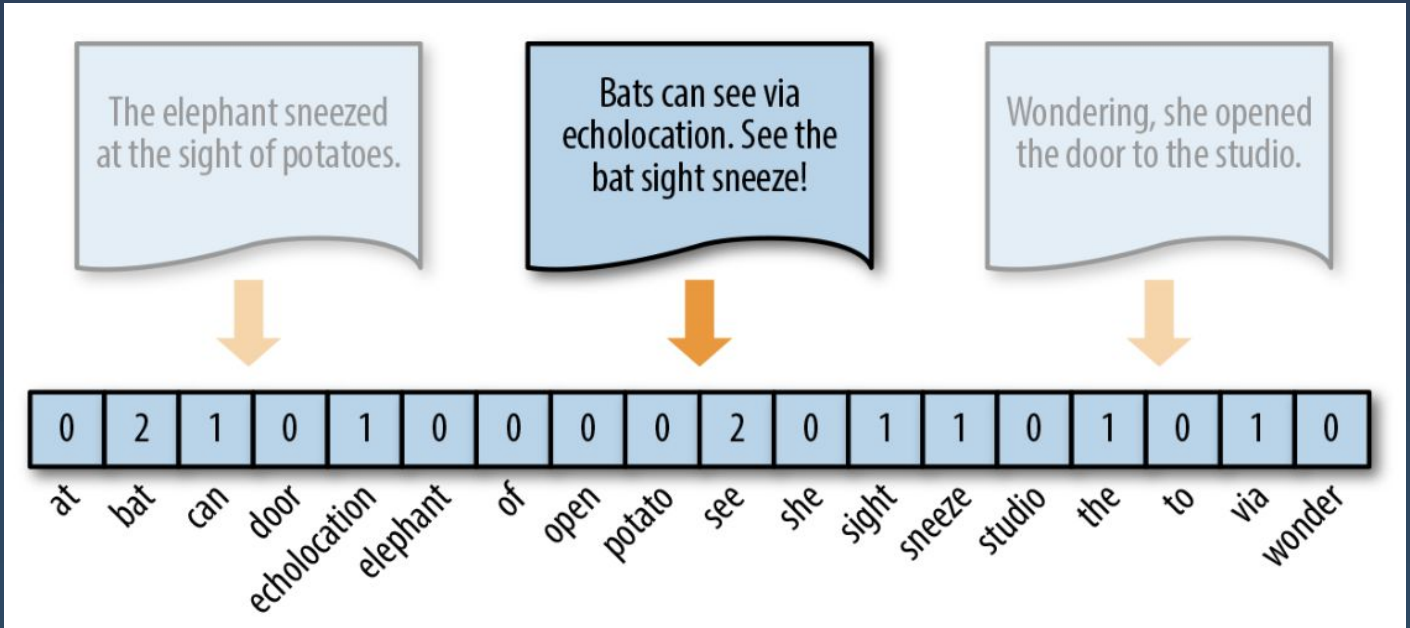
- Técnica que lê um texto, em algum idioma, e assinala, para cada palavra, a classe gramatical a qual ela pertence.
- Pode ser classificada em:
  - **Coarse-grained POS tags:** são tags mais genéricas, representando as grandes classes gramaticais (substantivos, verbos, adjetivos).
  - **Fine-grained POS tags:** são tags mais específicas, contendo detalhes morfológicos do token (substantivo no plural, verbo no futuro do pretérito, adjetivo superlativo).

Extração de Features

- Transformação do texto em um conjunto de dados estruturados (vetores de números) que pode ser utilizado em técnicas analíticas e aprendizado de máquina. Também conhecido como **Vetorização**. Tipicamente a saída da modelagem recebe o nome de Word Embedding.

- Bag of Words (Counter Vectorization)
- TF-IDF

Bag of Words


The elephant sneezed at the sight of potatoes.
Bats can see via echolocation. See the bat sight sneeze!
Wondering, she opened the door to the studio.
0 2 1 0 1 0 0 0 0 2 0 1 1 0 1 0 1 0
at bat can door echolocation elephant of open potato see she sight sneeze studio the to via wonder

TF-IDF

- ***Term Frequency - Inverse Document Frequency***
- **Term Frequency**: Frequência da palavra no documento atual
  - TF = (# vezes que o termo t aparece no documento) / (# de termos no documento)
- **Inverse document frequency (IDF)**: o quão raro é o termo no documento
  - IDF = log(N/n), N é o # de documentos e n é o # de documentos que o termo t aparece.
- **TF-IDF:** Importância de um termo para um documento para uma coleção.

TF-IDF

- **Exemplo:**

1. **Documento com 100 termos, no qual o termo “cachorro” aparece 5 vezes**
   TF = 5 / 100 = 0.05

2. **Temos 100 documentos (N) e o termo “cachorro” aparece em 20 desses documentos (n)**
   IDF = log(100/20) = 0.69
   TF-IDF = 0.05 * 0.69 = 0.034