Prática: Fundamentos de NLP (I)

Eduardo Prasniewski

1 Video

Apresenta visão geral, casos de uso e introdução sobre o que se trata LLM's, assim como de forma didática e superficial, explica os processos de: segmentation, tokenization, remoção de stop words, stemming, lemmatization, speech tagging e name entity tagging.

2 Minicurso Udemy

2.1 Seção 1

Introdução sobre o Processamento de Linguagem Natural suas vantagens e problemas na criação de modelos NLP, tais como ambiguidade, iroia etc. Assim como os motivos de não poder usar a gramática formal.

Aborda as principais aplicações da NLP: tradução de sentenças, chatbots, análise de sentimentos, sintentização de fala, previsão de digitação, classificação de textos e documentos, reconhecimento de autoria de documentos, análise sintática, reconhecimento de entidades nomeadas etc.

2.2 Seção 2

Definição de:

- Corpus: conjunto de documentos não estruturado.
- Annotations: Localizaree classificar elementos no texto.
- Tokenization: serparar a sentença em suas partes: palavras, pontos, símbolos etc.
- Parts-of-Speech Tagging (POS): adiciona tags a cada token (substantivo, adjetivo etc.)
- Lemmatizing: contrai a palavra na sua flexão.
- Stemming: corta a palavra para ter a representação base. Remove os sufixos e prefixos.
- Depending Parsing: processo de encontrar relação entreee as palavras.
- NGRAM: Processo que trata palavras consecutivas (normalmente 2 ou 3).
- Model: Um banco de dados linguísticos.

A forma mais simples de Word Embedding é One Hot Encoding, porém apresenta muitos problemas, principalmente relacionada ao uso de memória. Outra forma é o TF-IDF, atribui um peso para as palavras, ou seja, o quanto ela é importante. Já a forma mais popular é *Word2Vec*, no qual através de um processo de treinamento, produz um vetor demonstrando matematicamente a relação entre as palavras (um vetor para cada palavra).

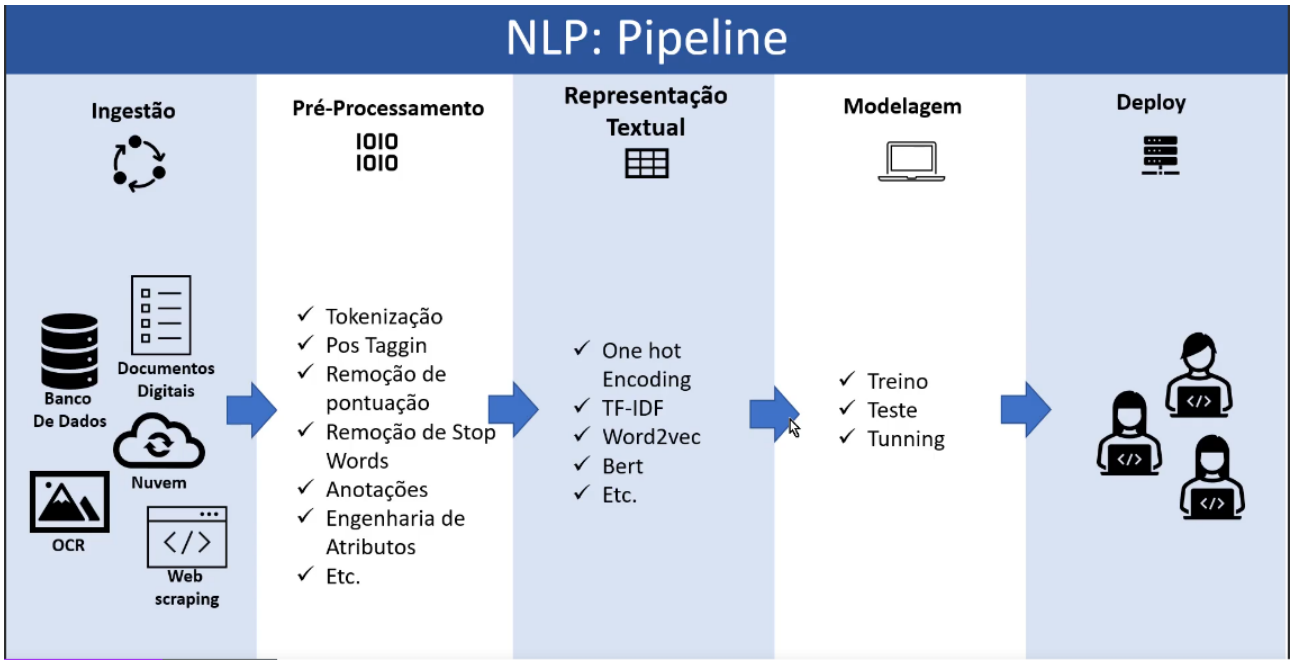

Figura 1: Pipeline da NLP

2.3 Seção 3

Primeiramente instala e importa a biblioteca Spacy versão 3 e a introduz, analisando seu pipeline base e realizando testes. Aborda os conceitos práticos de:

- Tokenização
- Pos-taggin
- Morfologia
- NER (Entidades nomeadas)
- Stop words
- Vocab (lexemas)
- Similaridade com word embeddings
- Matching
- Displacy
- Pipelines

2.4 Seção 4

Nesta seção aborda a NLP usando o NLTK (Natural Language Toolkit). Assim como na seção anterior, de forma prática os mesmos conceitos. Dentre as diferenças apresentadas com Spacy, as que se destacam são: a produção de métricas, vários tipos de Stemming (Porter, Snowball e Lancaster)

3 Artigo

O artigo destaca a evolução dos modelos de Processamento de Linguagem Natural (NLP). Primeiramente com Suppoer Vector Machines (SVMs) na década de 1990 para tarefas como classificação de texto. Posteriornte as redes neurais recorrentes (RNNs) e convolucionais (CNNs) ganharam popularidade na década de 2000 para tradução automática e análise de sentimentos. Transformadores, como BERT e GPT, introduzidos em 2017, revolucionaram o NLP com mecanismos de self-attention. Modelos pré-treinados, como GPT-3 e T5, desde 2018, permitem ajustes para tarefas específicas, economizando tempo e custos computacionais.

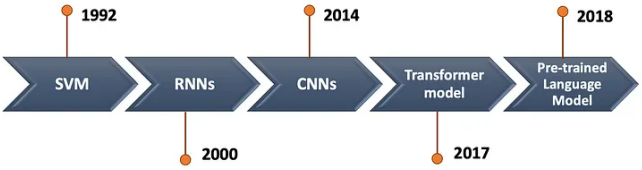

Figura 2: Linha do tempo NLP