Nome: Fernando Buligon Antunes **Data:**21/03/2025

Introdução ao NLP

NLP, Natural Language Processing, se refere ao ramo da inteligência artificial que faz com que as máquinas sejam capazes de ler e atribuir um significado a aquelas palavras lidas, então ela combina visão computacional com o campo da linguística, permitindo que o modelo decifre a estrutura da linguagem.

Todo dia milhares de humanos se comunicam uns com os outros através de redes sociais públicas, e essas conversas geram dados, dados esses que são preciosos para conseguir compreender o comportamento humano. Então quem trabalha com machine learning ou análise de dados pode acabar usando esses dados para fazer com que NLP exista, fazendo com que as máquinas consigam imitar o comportamento humano.

NLP hoje está presente no dia a dia de todo mundo, por exemplo, assistentes virtuais, auto corretor de texto, detecção de spam, e etc.

O processo possui algumas etapas:

- Segmentation: Separar todo o texto em partes, geralmente essa separação é marcada por pontos ou vírgulas;
- Tokenization: Atribuir um valor a cada palavra das partes separadas;
- Streaming (derivação): Ensinar a máquina que palavras diferentes podem possuir o mesmo significado a depender dos prefixos e sufixos;
- Lemmatization: Ensinar a origem das palavras, por exemplo, “am”, “are” e “is” vem do “be”;
- Speech Tagging: Ensinar o conceito de verbos, substantivos, advérbios, preposições … para a máquina, adicionando tags nas palavras;
- Named Entity Tagging: Adicionar nomes populares que podem ocorrer no texto, como por exemplo, nomes de filmes, pessoas, lugares e etc.

Tendo todas essas etapas concluídas, depois é usado um algoritmo de machine learning como o naive bayes para ensinar a máquina o comportamento humano.

Seção 1 - Introdução

O autor faz as boas vindas ao curso e também explica o conceito de processamento de linguagem natural, ele define como a “implementação de funcionalidades de aprendizado de máquina que dependem de interpretação ou produção de uma linguagem natural”, exemplo, ferramentas de tradução de texto, classificação de textos, sintetização de fala, previsão de digitação, chatbots, reconhecimento de autoria de documentos, análise sintática, busca de similaridade, produção de resumos. Também é falado que NLP é uma das áreas mais

importantes da ciência de dados, e também está em constante evolução, estando presente em milhares de produtos no mercado de trabalho.

Ele também faz a separação do NLP da ciência de dados como um “mundo a parte”, tendo suas próprias pesquisas, bibliotecas e técnicas. Define NLP com uma área extremamente difícil, por isso seus maiores avanços são com o auxílio da inteligência artificial por sua capacidade de aprender. Alguns motivos de ser tão difícil são: linguagens são ambíguas (exemplo: “ele pegou o dinheiro do banco” não há como saber se ele pegou o dinheiro de um banco de sentar ou uma instituição financeira), mesmas palavras podem significar coisas diferentes (exemplo: laranja pode ser a cor ou a fruta), ironias, contextos e também tem o fato de que a linguagem natural está em constante mudança.

Depois é feita uma introdução e apresentação do Google Colab para os que nunca trabalharam com ele antes.

Seção 2 - Fundamentos de Processamento de Linguagem Natural

Ele começa explicando alguns conceitos que são específicos do NLP que são necessários ter conhecimentos antes de ir para a prática.

- Corpus, consiste em um conjunto de documentos, um texto não estruturado, ou seja, não vem apresentado através de uma estrutura de linhas e colunas;
- Anotações ou annotations, é uma técnica de localizar e classificar elementos específicos de um texto, exemplo, anotar sentimentos para treinar um modelo, exemplos de ferramentas: Doccano e brat;
- Tokenization, processo de separar a sentença em suas partes;
- Tagging, adiciona uma tag a cada token, informando do que se trata, exemplo, verbo, substantivo e etc;
- Lemma, traz a palavra na sua forma raiz, por exemplo, “é” é “ser”.
- Stemming, corta as palavras buscando ter uma única representação, exemplo, “amigo, amigos, amiga, amigas viraria amig”;
- Dependency Parsing, encontra relação entre as palavras;
- NGRAM, é quando tem um processo de NLP que vai tratar de palavras consecutivas.

Computadores no geral são capazes de processar apenas números, então possuem certa dificuldade em tratar dados não estruturados, dados não estruturados seriam textos, que diferentemente de uma tabela não vem estruturados em linhas em colunas, por isso, é preciso criar maneiras para representar esse texto de forma com que o computador seja capaz de processar, transformando ele em representação numérica e estruturada.

Por esta razão, é falado de “world embedding”, que se trata da representação computacional de texto, a forma mais simples de world embedding é o one hot encoding, em que cada palavra possui uma representação única, no caso se

houvesse palavras repetidas, ambas teriam a mesma representação. Um dos problemas do one hot encoding é que se houver um texto muito grande a matriz vai se tornar gigantesca, e outro problema é que as palavras são representadas independente de contexto. Outra forma de world embedding é o TF-IDF, é uma forma de representar uma palavra de acordo com a frequência que aparece no documento, então essa forma de representação mostra o quanto a palavra é importante para o documento colocando um peso nela. Outra forma que é bem popular é a Word2Vec, que produz um vetor que mostra a ocorrência da palavra no Corpus e a relação entre as palavras. Também existem algumas outras formas como o FastText, Glove e Bert, que o autor cita mas não explica pois não vão ser usadas ao longo do desenvolvimento do curso.

Pipeline ou Workflow de como funciona o processo de um projeto de NLP, a primeira etapa é a ingestão dos dados, ou seja, coleta dos dados que vão ser usados, depois vem o pré-processamento, em que vão ser aplicadas técnicas para deixar os dados mais limpos, como remover partes sem importância, fazer anotações e deixar no formato requisitado, após isso vem a representação textual, onde vai ser aplicado alguma técnica de world embedding, por penúltimo, a modelagem em que o modelo vai ser treinado e testado, e por último, o deploy. Importante lembrar de que de acordo com o objetivo do projeto esse pipeline pode variar, mas ele pode ser visto de uma maneira geral

Seção 3 - NLP com Spacy

Spacy é uma biblioteca moderna em python que possui várias técnicas de pré-processamento de texto faladas anteriormente. Por ser uma arquitetura moderna, ele tem uma pipeline, então quando o método NLP é chamado, por padrão já executa várias etapas de pré-processamento, como o tagging e o parser. Um fator interessante é que esse pipeline padrão pode ser alterado de acordo com a vontade de quem estiver usando, podendo adicionar ou remover etapas.

Com essa biblioteca carregamos um modelo pré treinado em português, e com esse modelo fizemos algumas coisas como criar um documento, checar o tamanho do vocabulário, o tipo do documento.

O documento é separado por tokens, que são possíveis acessar por index, foram feitos alguns testes como, checar a quantidade total, checar se são, stop words, alfanumérico, maiusculas, pontuação, numero, sentença inicial, seu formato, sua classe gramatical, dependência, maneira “raiz”, morfologia, seu texto, sua tag, entidades nomeadas.

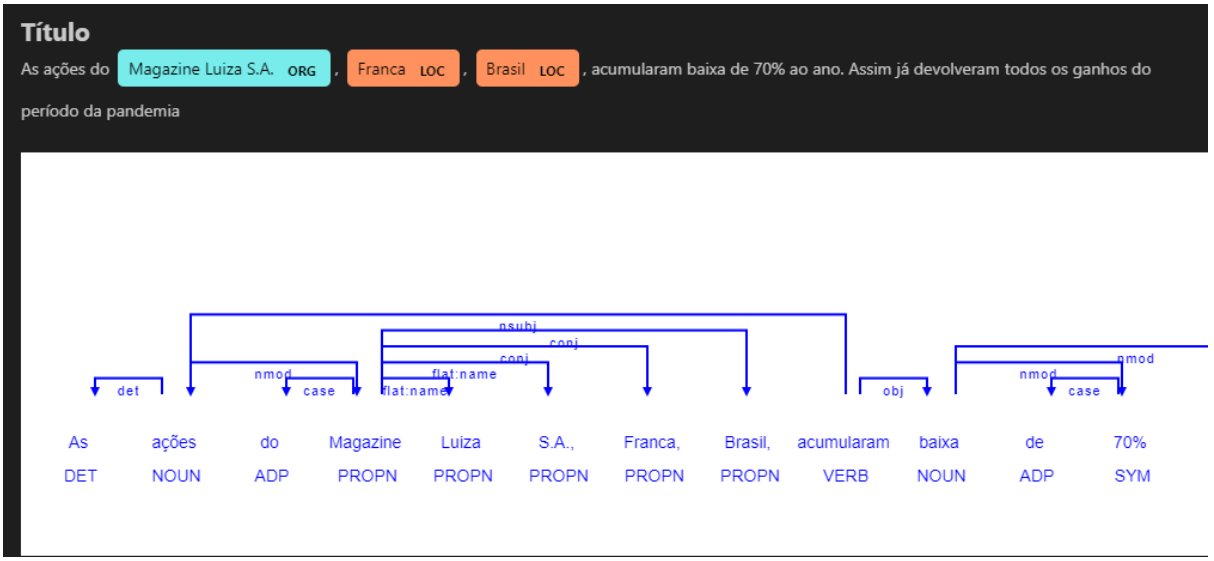
Usando as stop words, temos como acessar todas as stop words padrões do modelo e adicionar ou remover alguma que seja de nossa vontade.

Na parte do vocabulário, nós vimos como acessar a hash de uma string, como acessar uma string pela sua hash e como ver o vetor que representa as relações de uma palavra com as outras.

Também foi visto como ver a similaridade, importante lembrar que o spacy calcula a similaridade baseada em contexto, retornando um valor entre 0 e 1, quanto mais próximo de 1, mais similar é. Foram feitos testes com frases inteiras e com palavras únicas.

O matcher é capaz de encontrar padrões dentro de textos, também é possível que nós mesmos adicionamos esse padrão.

Depois foi visto o discplacy, que é o módulo de visualização do spacy, com ele nós conseguimos ter a representação visual do documento que está sendo trabalhado, na imagem abaixo é possível ver as entidades e as dependências.



Seção 4 - Nlp com NLTK

Natural Language Toolkit (NLTK), é uma vasta biblioteca de processamento de linguagem natural em Python, sendo a mais antiga e uma das mais usadas, possuindo várias ferramentas para NLP.

A principal diferença para do NLTK para o Spacy é que é preciso fazer download de cada pacote que for usar, mas um ponto interessante sobre isso é que é possível gerenciar os downloads com uma interface interativa, podendo ver todos.

Depois de ter baixado cada pacote, foram feitas as mesmas coisas que no Spacy, tokenização, stop words, pontuação, frequência de palavras, palavras mais comuns, diferentes tipos de stemmer, tags, pós tag de palavra, pós tag a nível de sentença, lematização e reconhecimento de entidades nomeadas.

Modelos de Machine Learning para NLP

Há uma breve introdução sobre o que é NLP, e depois define o que são modelos de machine learning como sistemas do mundo real que são treinados com uma base de dados para prever relações simples e complexas entre as entradas e

saídas. Também cita que o uso de modelos pré treinados facilita pois reduz os esforços de ter que treinar um. Depois é feito uma linha do tempo mostrando alguns modelos, 1992 (**SVM**), 2000 (**RNNs**), 2014 (**CNNs**), 2017 (**Transformer model**), 2018 (**Pre-trained Language Model**), depois é feita uma curta descrição de cada ponto.

SVM: Support Vector Machine, na maioria das vezes é usado para tarefas de NLP como classificação de texto e análise sentimental, foi ele que introduziu o hiperplano que separava as classes ao máximo.

RNNs: Recurrent Neural Networks, inovou com tarefas de NLP avançadas como modelação, tradução e análise de sentimento.

CNNs: Convolutional Neural Networks, eram reconhecidos por processamento de imagens, mas depois foram usados para classificação de texto e análise de sentimentos, esses modelos tinham uma capacidade muito alta de encontrar características no texto.

Transformer model: Alguns exemplos são BERT e GPT, trouxeram avanços significativos para o NLP, são capazes de aprender de forma eficiente as relações entre as palavras.

Pre-trained Language Model: São treinados com uma grande base de dados e por isso podem ser usados para tarefas específicas.