Card22_Processamento_de_LinguagemNatural

Davi Bezerra Barros

Introdução

O curso **Natural Language Processing(NLP) Zero to Hero** ensina

Natural Language Processing(NLP) Zero to Hero

O curso, oferecido pelo canal oficial do TensorFlow, ensina a trabalhar com classificadores de texto usando a biblioteca TensorFlow/Keras, abordando os conceitos teóricos necessários para a compreensão da tecnologia e desenvolvendo aplicações práticas ao longo das aulas.Os conceitos abordados estão descrito a seguir.

Tokenização: Processo de codificar as palavras contidas em uma string de texto as convertendo para índices inteiros, chamados de tokens. As frases são convertidas em vetores com sequências de tokens, permitindo o processamento dos dados por modelos de rede neural.

Padding: Técnica utilizada para garantir que todas as entradas(listas de token) tenham o mesmo tamanho, independente do seu tamanho original. Isso é necessário por que modelos de rede neural requerem entradas de tamanho fixo. O tensorflow utiliza a função *pad_sequences* para preencher a menor sequência com zeros, a deixando com o mesmo tamanho da maior sequência de entrada.

Embedding: Técnica utilizada para mapear cada token a um vetor espacial, permitindo o processamento semântico das palavras ao relacionar seu significado com posições espaciais. É um processo fundamental para o processamento de linguagem natural, por que transforma dados categóricos, como palavras, em uma forma que pode ser processada por uma rede neural. O tensorflow utiliza a função *padding_sequences* para realizar esta operação.

Rede Neural Recorrente(RNN): Um tipo de rede neural desenvolvida para processar dados sequenciais, como linguagem natural ou outros tipos de dado em que a ordem dos elementos é importante.

- **Memória:** As RNN tem a capacidade de "lembrar" informações das entradas anteriores. Os neurônios das RNN tem conexões recorrentes nas camadas ocultas, o que significa que a saída de um neurônio é passada novamente para a entrada juntamente com o próximo elemento da sequência, criando algo semelhante a uma memória.

- **Long Short-Term Memory(LSTM):** Uma variante das RNN que retém informações importantes em sequências mais longas, permitindo uma maior janela de contexto. Diferente das RNNs tradicionais, que têm dificuldade em lembrar dependências de longo prazo, as LSTMs utilizam um estado de célula para transportar informações ao longo da sequência. Isso é gerenciado por gates que definem o que manter ou esquecer, garantindo que dados importantes sejam preservados.

Natural Language Processing with spaCy & Python

O spcy é um biblioteca de processamento de linguagem natural projetada para lidar com grandes volumes de dados, utilizada em grandes aplicações de indústria. Conta com diversas ferramentas, e as principais estão descritas abaixo:

Docs: A principal estrutura de dados utilizada par armazenar os textos processados. É composto por uma sequência de tokens, onde cada token armazena uma informação relevante.

- **Tokens:** Unidade básica de texto, como palavras e pontuações, e armazena informações sobre a sua posição no texto, classe gramatical, etc. seus atributos são:
  - `text`: O texto original do token.
  - `lemma_`:A forma básica do token.
  - `pos_`: A parte do discurso.
  - `dep_`: A dependência sintática.
  - `is_stop`: Se o token é uma palavra de parada.

Displacy: É uma ferramenta de visualização do spacy, utilizada para identificar as entidades nomeadas diretamente no texto, e também as árvores de dependência sintática.

Word Vectors: São representações numéricas das palavras em um espaço dimensional e capturam suas semelhanças semânticas.

Pipelines: É a sequência de funções/componentes que processa o texto, como tokenização, reconhecimento de entidades, etc.

EntityRulers: Permite adicionar funções e regras manualmente ao pipeline para rotular entidades específicas em um texto. É usado para melhorar a precisão do modelo ao realizar ajustes finos em suas funções.

Matcher: Localiza os padrões de tokens dentro de um texto de acordo com o padrão estabelecido em regras textuais.

Multi-Word Token: São tokens compostos por várias palavras e tratados como uma única entidade.

Atividade prática: Análise financeira

A atividade prática implementada no curso simula um cliente da área financeira, que precisa de uma aplicação para analisar artigos e notícias relevantes para seus investimentos. A aplicação encontra automaticamente todas as empresas, ações, vendas e índices presentes em um texto.