Universidade Federal da Paraíba Centro de Informática
Disciplina: Processamento de Linguagem de Natural
Professores: Yuri Malheiros
Grupo :Paulo Marcelo Ribeiro Soeiro da Silva
Thiago Vanucci Corrêa de Araujo

Reconhecimento de Entidades Nomeadas a partir de textos para extração de informações de notícias ou artigos

Índice :

1. Apresentação do Problema

A extração de informações de notícias ou artigos é uma tarefa desafiadora no processamento de linguagem natural (NLP) e na análise de texto. Com o crescimento exponencial da quantidade de dados disponíveis na internet, torna-se cada vez mais difícil analisar manualmente grandes volumes de conteúdo.

A análise manual de notícias e artigos demanda tempo e esforço consideráveis, e está sujeita a erros e inconsistências. Imagine ter que identificar informações específicas, como nomes de pessoas, locais, organizações ou eventos relevantes, em milhares de documentos. Essa tarefa se torna inviável sem o auxílio de técnicas automatizadas.

A problemática da extração de informações reside na necessidade de desenvolver métodos e algoritmos capazes de identificar e extrair automaticamente as informações relevantes contidas em um texto. Isso inclui a detecção e classificação de entidades nomeadas, como pessoas, locais e organizações, bem como a compreensão das relações e estruturas presentes nas informações extraídas.

A relevância desse problema é evidente em diversos setores, como jornalismo, finanças, medicina e segurança. A capacidade de extrair informações precisas e relevantes de notícias e artigos de forma automática e eficiente pode impulsionar a tomada de decisões, fornecer insights valiosos e automatizar tarefas repetitivas.

Nesse contexto, o desenvolvimento de técnicas de extração de informações se torna fundamental para lidar com a vasta quantidade de dados textuais disponíveis, permitindo uma análise mais eficiente e precisa do conteúdo.

2.Objetivos

O objetivo deste projeto é desenvolver um modelo de extração de informações capaz de identificar e classificar entidades nomeadas em textos de notícias ou artigos. As entidades nomeadas podem incluir nomes de pessoas, organizações, locais, datas, entre outros.

Nossa solução tem como principais objetivos:

- Automatizar o processo de extração de informações: A extração manual de entidades nomeadas em grandes volumes de texto é um trabalho demorado e propenso a erros. Nosso modelo visa automatizar esse processo, tornando-o mais eficiente e confiável.

- Aumentar a precisão e a velocidade da extração: Com o uso de técnicas de processamento de linguagem natural e aprendizado de máquina, buscamos criar um modelo que seja capaz de identificar e classificar as entidades nomeadas com alta

precisão e em tempo hábil. Isso permitirá uma análise mais eficiente e rápida dos textos.

- Facilitar a organização e recuperação de informações: Ao extrair as entidades nomeadas, nosso modelo fornecerá informações estruturadas e categorizadas, o que facilitará a organização e a recuperação de informações relevantes em um conjunto de documentos.

Em resumo, nosso objetivo é desenvolver um modelo de extração de informações de notícias ou artigos que seja preciso, eficiente e versátil. Isso possibilitará uma análise mais ágil e precisa do conteúdo textual, proporcionando benefícios significativos para pesquisadores, jornalistas, profissionais de análise de dados e outros envolvidos na busca por informações relevantes.

3.Dados Utilizados e Pré-processamento dos Dados

Dados Utilizados:

Para treinar e avaliar nosso modelo de extração de informações, utilizamos um conjunto de dados passados por arquivos txt do CoNLL003. Esses dados consistem em frases e suas respectivas tags de entidades nomeadas. As frases representam trechos de notícias ou artigos, e as tags indicam a classe das entidades presentes, como pessoas, locais, organizações, entre outros.

Pré-processamento dos Dados:

Antes de alimentar os dados para o modelo de rede neural, foi necessário realizar o pré-processamento para preparar os dados de acordo com os requisitos do modelo. O pré-processamento envolveu as seguintes etapas:

- **Tokenização:**
  Utilizamos a biblioteca NLTK (Natural Language Toolkit) para tokenizar os textos em palavras individuais. Isso envolve a separação dos textos em unidades significativas, removendo espaços em branco e pontuações.

- **Criação de Vocabulário:**
  Construímos um vocabulário de palavras únicas encontradas nos textos. Isso nos permitiu mapear cada palavra para um índice único, facilitando o processamento subsequente.

- **Codificação das Tags:**
  As entidades nomeadas são rotuladas com tags específicas. Criamos um vocabulário de tags e mapeamos cada tag para um índice único. Em seguida, codificamos as tags em formato one-hot, transformando-as em vetores binários para facilitar o treinamento do modelo.

- **Preenchimento de Sequências:**
  Para garantir que todas as sequências de palavras tenham o mesmo comprimento, preenchemos as sequências com zeros no final, usando a função pad_sequences do Keras. Isso é necessário para criar uma matriz de entrada com dimensões consistentes.

O pré-processamento dos dados nos permite representar os textos de maneira adequada para alimentar o modelo de rede neural.

4.Rede Neural - Arquitetura e Treinamento

A extração de informações de notícias ou artigos requer o uso de modelos de aprendizado de máquina, especificamente redes neurais, para realizar a tarefa de identificar e classificar entidades nomeadas. Neste projeto, utilizamos uma arquitetura de rede neural recorrente com camadas LSTM (Long Short-Term Memory) para realizar essa tarefa.

A arquitetura da rede neural utilizada consiste em três camadas principais:

1. Camada de Embedding: A primeira camada da rede é uma camada de embedding, que mapeia as palavras do vocabulário para vetores densos de representação. Essa camada permite que as palavras sejam representadas em um espaço vetorial contínuo, capturando relações semânticas entre elas. Utilizamos um tamanho de saída de 32 para os vetores de embedding.

2. Camada Bidirecional LSTM: A camada seguinte é uma camada bidirecional LSTM. As LSTMs são um tipo de célula de memória recorrente que ajuda a capturar dependências de longo prazo em sequências de dados. Ao tornar a camada bidirecional, a rede pode aprender padrões em ambas as direções do texto. Isso permite capturar tanto o contexto anterior quanto o posterior de cada palavra. Utilizamos unidades LSTM com uma dimensão de 32 e definimos o parâmetro "return_sequences" como True para obter saídas em sequência.

3. Camada TimeDistributed Dense: A última camada é uma camada densa "TimeDistributed", que atribui um rótulo a cada palavra na sequência de entrada. Essa camada aplica uma operação de classificação softmax em cada posição da sequência para prever a tag correspondente a cada palavra. O número de neurônios na camada densa é igual ao tamanho do vocabulário de tags.

Durante o treinamento, usamos o otimizador "adam" e a função de perda "categorical_crossentropy". Essa função de perda é adequada para problemas de classificação multiclasse, como o nosso, em que cada palavra é atribuída a uma tag específica. Além disso, monitoramos a métrica de acurácia durante o treinamento para avaliar o desempenho do modelo.

O modelo foi treinado com os dados de treinamento, utilizando um tamanho de lote de 32 e um total de 10 épocas. Validamos o modelo usando os dados de teste para acompanhar o desempenho em dados não vistos durante o treinamento.

5.Resultados

Após o treinamento do modelo de extração de informações de notícias ou artigos, avaliamos seus resultados com base em métricas de desempenho. As principais métricas utilizadas são a acurácia e a perda durante o treinamento e a validação.

Durante as 10 épocas de treinamento, observamos uma melhoria gradual do desempenho do modelo. A acurácia aumentou a cada época, indicando que o modelo estava aprendendo a identificar corretamente as entidades nomeadas nos dados de treinamento. Além disso, a perda diminuiu progressivamente, o que indica que o modelo estava ajustando seus parâmetros para minimizar a discrepância entre as previsões e as tags reais.

Ao final do treinamento, aplicamos o modelo aos dados de teste para avaliar seu desempenho em dados não vistos anteriormente. Observamos uma acurácia satisfatória nos dados de teste, o que indica que o modelo generalizou bem e é capaz de extrair informações de notícias ou artigos de forma precisa.

Além das métricas de desempenho, também podemos avaliar visualmente os resultados do modelo. Ao aplicá-lo a um texto de exemplo, como "New box, John Doe Johnson and Chris is from New York and works at Google", podemos extrair as entidades nomeadas presentes no texto. Essas entidades podem incluir nomes de pessoas, locais, empresas e outras informações relevantes contidas no texto.

Em resumo, os resultados obtidos com o modelo treinado para extração de informações de notícias ou artigos foram promissores. O modelo foi capaz de aprender com os dados de treinamento, alcançando uma boa acurácia nos dados de teste. Isso demonstra sua capacidade de identificar e classificar corretamente as entidades nomeadas em textos, o que pode ser útil para diversas aplicações, como análise de conteúdo, indexação de informações e geração de insights.

6.Conclusão

Ao longo deste projeto, abordamos a problemática da extração de informações de notícias ou artigos. Essa tarefa é de grande relevância, uma vez que lidamos com volumes cada vez maiores de dados textuais e a análise manual se torna inviável.

Apresentamos uma solução baseada em uma rede neural para realizar a extração de entidades nomeadas nesses textos. Utilizamos um modelo de processamento de linguagem natural, treinado com dados rotulados, que nos permitiu identificar entidades como locais, organizações e nomes de pessoas.

Realizamos o pré-processamento dos dados, convertendo as palavras e as tags em sequências de índices, além de criar vocabulários para mapear esses índices. Em seguida, treinamos a rede neural com os dados de treinamento, ajustando os pesos dos neurônios para que o modelo pudesse aprender a reconhecer as entidades nas frases.

Após o treinamento, testamos o modelo com dados de teste e avaliamos sua acurácia. Os resultados obtidos mostraram uma boa performance na extração de informações, com uma alta taxa de acerto na identificação das entidades.

Em resumo, a extração de informações de notícias ou artigos por meio de técnicas de processamento de linguagem natural e redes neurais representa uma abordagem promissora para lidar com o crescente volume de dados textuais. A capacidade de identificar e extrair automaticamente entidades nomeadas em textos traz benefícios significativos para a análise de informações e a tomada de decisões em diversas áreas, como jornalismo, pesquisa e inteligência de negócios.

Com base nos resultados, concluímos que o modelo proposto é eficaz na extração de informações de notícias ou artigos, contribuindo para automatizar essa tarefa e agilizar o processo de análise de grandes volumes de dados textuais.

7.Referencias

https://paperswithcode.com/dataset/conll-2003

https://paperswithcode.com/sota/named-entity-recognition-ner-on-conll-2003

https://paperswithcode.com/paper/automated-concatenation-of-embeddings-for-1

https://medium.com/data-hackers/reconhecimento-de-entidades-nomeadas-entidades-subentidades-relacionamentos-e-ambiguidade-bc4f302d0f9b

https://learn.microsoft.com/pt-br/azure/cognitive-services/language-service/named-entity-recognition/overview

https://www.techtarget.com/whatis/definition/named-entity-recognition-NER#:~:text=NER%20involves%20detecting%20and%20categorizing,times%2C%20monetary%20values%20and%20percentages

https://www.turing.com/kb/a-comprehensive-guide-to-named-entity-recognition

https://repositorio.ufc.br/bitstream/riufc/49711/1/2019_tcc_nsaraujo.pdf