Desafio de NLP

No link abaixo há um arquivo csv com documentos médicos. Alguns deles concluem que o paciente não apresenta nenhuma anormalidade, outros descrevem alguma patologia ou alteração médica. Usamos esse tipo de documento para classificar o paciente em quais doenças eles tem e até mesmo a gravidade da doença.

1 - Faça uma análise descritiva em cima dos documentos. O que você conclui apenas olhando para eles? Qual a forma de escrita dos documentos? Há tipos diferentes? Olhando os textos consegue identificar alguma alteração comum entre os pacientes?

A análise dos documentos médicos revela uma grande diversidade de formatos, incluindo textos em HTML, RTF e texto simples. A escrita médica apresenta características consistentes como estilo telegráfico, uso extensivo de abreviações e terminologia técnica específica, além da frequente organização por tópicos bem definidos. Os documentos são estruturados em vários tipos, como notas de evolução, prescrições médicas, anotações de enfermagem, resumos de alta e notas de admissão, cada um com propósito e estrutura próprios. Quanto ao conteúdo clínico, observa-se a recorrência de quadros infecciosos, evidenciada pelo uso frequente de antibióticos como Ceftriaxona; monitoramento constante de sinais vitais; manejo da dor como sintoma prevalente; utilização de dispositivos médicos como acessos venosos e cateteres de oxigênio; alterações metabólicas que requerem monitoramento; e presença significativa de comorbidades crônicas como hipertensão e diabetes. Esta análise descritiva justifica a abordagem adotada no projeto, que desenvolveu métodos específicos para normalização e extração de informações relevantes, adaptados às peculiaridades da documentação médica brasileira.

2 - Como você extrairia informações de medicamentos, exames e doenças do paciente? Analise os dados que tem em mãos e descreva o pipeline de desenvolvimento para um modelo como esses, levando em consideração todo o processo de análise, anotação, treino até a validação do modelo.

A extração de informações médicas de documentos clínicos não estruturados requer uma abordagem sistemática e robusta. Com base nos dados analisados, o pipeline ideal combina técnicas de processamento de linguagem natural com aprendizado de máquina, implementando um sistema de reconhecimento de entidades nomeadas (NER) especializado para o domínio médico. O desenvolvimento inicia-se com uma etapa crucial de pré-processamento, onde os diferentes formatos de documentos (HTML, RTF, texto simples) são normalizados através de funções específicas para remoção de tags, padronização de caracteres e tokenização adaptada ao contexto médico. Esta etapa é fundamental devido à heterogeneidade dos formatos encontrados nos registros clínicos. Após a normalização, procede-se à anotação dos dados, inicialmente utilizando uma abordagem baseada em dicionários para identificar entidades como medicamentos, procedimentos, dispositivos e condições. Esta anotação semi-automática gera um conjunto de dados rotulados no formato BIO (Beginning-Inside-Outside), permitindo a identificação precisa do início e continuação de cada entidade. Com os dados anotados, inicia-se o desenvolvimento do modelo propriamente dito, utilizando uma arquitetura de transformers pré-treinada em português (como o BERT) e adaptando-a para a tarefa específica de NER médico. O treinamento é realizado com divisão estratégica de dados (80% para treino, 20% para validação), aplicando técnicas de tokenização especializadas para alinhar corretamente as etiquetas aos tokens gerados pelo modelo. A etapa de validação inclui avaliação em exemplos representativos e métricas específicas para cada categoria de entidade. Para complementar o modelo NER, implementa-se também uma abordagem baseada em prompts estruturados para modelos de linguagem grandes, permitindo extrações mais detalhadas e contextualizadas. A combinação das duas abordagens – NER para identificação básica de entidades e prompts para extrações estruturadas – proporciona um sistema robusto capaz de extrair informações precisas sobre medicamentos, exames e doenças, mesmo diante da complexidade e variabilidade da documentação médica.

3 - Se fizesse as extrações via prompt, que tipo de prompt você usaria para extrair informações desses documentos médicos? Mostre exemplos.

Para realizar extrações eficientes via prompt a partir de documentos médicos, é necessário desenvolver prompts estruturados que orientem claramente o modelo de linguagem na identificação e formatação das informações desejadas. O prompt ideal deve especificar com precisão o tipo de informação médica a ser extraída (medicamentos, exames ou diagnósticos), o formato esperado da resposta (geralmente estruturado como JSON) e

incluir exemplos ilustrativos que demonstrem o nível de detalhe desejado. Um prompt eficaz para extração de medicamentos começaria estabelecendo o contexto para o modelo, identificando-o como especialista em análise de documentos médicos e solicitando a extração de todos os medicamentos mencionados em um formato específico: "Você é um assistente especializado em extrair informações médicas de textos clínicos em português. Analise o seguinte documento médico e extraia todos os medicamentos mencionados com suas respectivas informações no formato JSON: {'medicamentos': [{'nome': 'nome do medicamento', 'dose': 'dosagem', 'via': 'via de administração', 'frequencia': 'frequência de administração'}]}. Inclua apenas medicamentos claramente mencionados no texto. Se alguma informação estiver ausente, use null." Para extração de resultados de exames, um prompt semelhante adaptaria a estrutura solicitada: "Extraia todos os exames e seus resultados mencionados no texto no formato JSON: {'exames': [{'nome': 'nome do exame', 'resultado': 'resultado', 'unidade': 'unidade de medida', 'referencia': 'valor de referência'}]}." Para uma extração completa, o prompt combinaria múltiplas categorias: "Extraia as seguintes informações do texto: medicamentos (nome, dose, via, frequência), dispositivos utilizados, sinais vitais (tipo, valor, unidade), diagnósticos e procedimentos realizados. Estruture a resposta em formato JSON com categorias separadas." Esta abordagem baseada em prompts demonstra particular eficácia quando complementada com orientações específicas para lidar com abreviações médicas comuns (como "EV" para endovenoso), formatos de dosagem e expressões temporais frequentes em contextos clínicos. A estruturação precisa do prompt e a solicitação de formatos consistentes de resposta garantem extrações mais confiáveis e padronizadas, facilitando a integração posterior com sistemas de processamento de dados clínicos.

Não se preocupe muito com os resultados do algoritmo, estamos mais preocupados em ver como você desenvolve a solução. Você pode desenvolver a solução no Colab do Google e nos enviar o link direto ou se preferir commitar o notebook em um repositório do github.