4ª EDIÇÃO Science & Business Connection by colmeia
COLMEIA
PIT

Análise , Projeto e Criação de Datasets para Assistente Virtual do Programa Queimadas

Fillipe Pereira Bueno de Almeida[1]; Abner Rodrigo da Silva Costa[2]; Fabrício Galende Marques de Carvalho[3]

[1] INPE. (fillipepereira8@gmail.com).
[2] INPE. (abner.costa@inpe.br).
[3] INPE. (fabricio.galende@inpe.br).

Resumo: Com a crescente necessidade por sistemas que incorporam o Processamento de Linguagem Natural (PLN) em função do suporte e atendimento ao cliente, é de suma importância o desenvolvimento de uma boa base de dados para que algoritmos como assistentes virtuais apresentem bom desempenho. Este projeto tem como objetivo a criação e análise de uma base de dados para sistemas de PLN voltados ao desenvolvimento de um assistente virtual para auxílio ao usuário de sistemas desenvolvidos no programa Queimadas do Instituto Nacional de Pesquisas Espaciais (INPE). Para tanto, foi projetado e implementado um protótipo de pipeline de criação e análise de dataset aplicando-se diferentes técnicas de PLN e análise de dados. A base gerada foi utilizada para o treinamento de dois modelos de representação, sendo eles: Word2Vec e Bag Of Words (BOW). A metodologia adotada baseou-se também nas frases retiradas da "seção de perguntas frequentes" (i.e., FAQ), do Programa Queimadas, que foram categorizadas em classes como: Informações sobre satélite, especificações de dados, tipos de focos, dados para contato e outros exemplos considerados como "fora de contexto". Após a categorização foi realizado o balanceamento destas classes para que o modelo de representação não se tornasse enviesado. Montado o conjunto de dados, parte-se para o treinamento do modelo de representação Word2Vec, utilizando-se a biblioteca Gensim, que é capaz de calcular embeddings que consideram o contexto em que cada palavra é utilizada (i.e., palavras antes e depois da palavra que está sendo convertida para embedding). Cabe ressaltar que normalmente palavras que são utilizadas nos mesmos contextos terão uma de similaridade maior, ou seja, representações semelhantes no espaço latente. No segundo caso, usou-se Bag Of Words com a transformação TF-IDF (*Term Frequency-Inverse Document Frequency*) que cria vetores de representação por contagem ponderada por pesos de ocorrência considerando as frases de treinamento. Para o BOW, cabe também ressaltar a ocorrência de vetores de alta dimensionalidade. A abordagem de avaliação dos vetores de representação consistiu na visualização aplicando dendrogramas hierárquicos e PCA (*Principal Component Analysis*). Ao se proceder com esse tipo de avaliação, é possível realizar ajustes para que a representação vetorial das classes se agrupem de uma melhor forma. Os resultados obtidos mostram que, para os exemplos do dataset criado, o Word2Vec obteve vetores representativos melhores, em termos de proximidade no espaço de representação, dos que os obtidos a partir do BOW. Utilizando-se a representação Word2Vec e aplicá-los a um modelo de classificação, foi possível desenvolver um assistente virtual que conseguiu classificar corretamente as

perguntas do usuário e responder de modo adequado com elevada precisão. Portanto, a estratégia de projeto de dataset para treinamento de assistentes virtuais discutidas nesse trabalho mostra-se adequada e eficaz e potencialmente pode ser estendida a outras aplicações. Em trabalhos futuros pretende-se melhorar e automatizar a geração de elevados volumes de dados de treinamento e, também, avaliar a aplicação em modelos baseados em LLM (i.e., aplicar, por exemplo, ao fine tuning)

Palavras-chave: Processamento de Linguagem Natural; Assistente Virtual; Projeto de Dataset; Word2Vec; Bag of Words.