UNIVERSIDADE MUNICIPAL DE SÃO CAETANO DO SUL - USCS
PRÓ-REITORIA DE GRADUAÇÃO
ESCOLA POLITÉCNICA

André Picaro Michelli
Fernando Navarro de Souza
Mauricio Carvalho Pinheiro
Marcelo Rodrigues Pascoal

Expl.AI.n ChatBot: Sistema de Recuperação de Informação em Arquivos Textuais.

São Caetano do Sul, 2021.

UNIVERSIDADE MUNICIPAL DE SÃO CAETANO DO SUL USCS PRÓ-REITORIA DE GRADUAÇÃO
ESCOLA POLITÉCNICA - USCS

André Picaro Michelli
Fernando Navarro de Souza
Mauricio Carvalho Pinheiro
Marcelo Rodrigues Pascoal

Expl.AI.n ChatBot: Sistema de Recuperação de Informação em Arquivos Textuais.

Trabalho de Conclusão de Curso apresentado à Escola Politécnica da Universidade Municipal de São Caetano do Sul como requisito obrigatório para obtenção do título de bacharel em Ciência da Computação

ORIENTADOR : Prof° Raphael Lopes de Souza.

São Caetano do Sul, 2021.

Resumo

O seguinte projeto para obtenção do título de bacharel em ciência da computação, curso vinculado a Escola Politécnica da Universidade Municipal de São Caetano do Sul tem como objetivo criar um software que processa arquivos textuais, para facilitar o acesso à determinada informação através de um algoritmo de busca, onde por meio de um *chatbot* o usuário insere uma busca textual, e o mesmo faz a tratativa dos dados de entrada com um algoritmo de processamento de linguagem natural em um arquivo *pdf ou docx*. Ou seja, a aplicação realiza o processamento dos documentos previamente salvos e dá uma resposta com base em uma busca do usuário.

Futuramente, pretendemos criar outras funcionalidades na aplicação, e assim poderemos também inserir buscas por voz.

Com os avanços da tecnologia, o usuário expõe um problema que busca solução e a tecnologia em questão deverá buscar a resposta esperada, sendo necessário refinarmos ainda mais os processos e tecnologias empregadas no fluxo, para que as grandes empresas de diversos ramos possam agilizar suas produções e maximizar seus lucros.

O crescimento dos dados atualmente é exponencial, sendo assim, existe a necessidade e procura por novas tecnologias que possam suprir essa demanda, e assim, o mercado se torna bastante competitivo e em busca de ideias inovadoras.

Palavras-Chave: IBM Watson, PLN (Processamento de Linguagem Natural), Algoritmo, ChatBot, Arquivos Textuais.

Abstract

The following project to obtain a bachelor's degree in computer science, a course linked to the Polytechnic School of the Municipal University of São Caetano do Sul, aims to create a software that processes text files, to facilitate understanding through a search algorithm. Where, through a chatbot, the user enters a textual search, and it handles the input data, and thus the application processes previously saved documents, such as a PDF or DOCX file. In the future, we intend to create other functionalities in the application, so that we will also be able to enter data by voice.

In the future, we intend to create other functionalities in the application, so that we willalso be able to enter data by voice.

With the advances in technology, the user exposes a problem that seeks a solution and the technology in question must seek the expected answer, making it necessary for us to further refine the processes and technologies used in the flow, so that large companies in different fields can streamline their production and maximize your profits.

The data growth is currently exponential, therefore, there is a need and demand for new technologies that can meet this demand, and thus, the market becomes very competitive and in search of innovative ideas.

Keywords: IBM Watson, algorithm, natural language processing,textual files, chatbot.

Lista de Siglas

- PLN: Processamento da Linguagem Natural;
- SQL: Structured Query Language;
- PHP: HyperText PreProcessor;
- PYTHON: Python is an interpreted, object-oriented, high-level programming language with dynamic semantics.
- Rank-BM25 Search algorithms
- JS: JavaScript;
- PDF: Portable Document Format;
- ML: Machine Learning;
- HTML: HyperText Markup Language;
- CSS: Cascading Style Sheets;
- WWW: World Wide Web;
- SQL: Structured Query Language;
- SGBD: Sistema Gerenciador de Banco de Dados.

Lista de Figuras

Sumário

Introdução

Com o avanço da quantidade de dados na sociedade, o ser humano sente a necessidade de, cada vez mais, adquirir informações de forma mais rápida. Segundo Ángel F. Villarejo-Ramos (2020, p. 1) e Juan-Pedro Cabrera-Sánchez (2020, p. 1):
1) “Atualmente, a sociedade gera dados sobre nossas atividades a uma taxa exponencial de crescimento”. Ou seja, a quantidade total de dados duplica a cada dois anos. As organizações também seguem esse mesmo padrão. Através de dados textuais em arquivos pdf ou docx extensos, e estão em constante alteração. Com issosurge a necessidade dos atores estarem consultando esses materiais constantemente,a fim de manterem-se atualizados em relação aos protocolos e ou informações existentes nesses arquivos.

O processo de comunicação com o chatbot consiste em quatro componentes: emissor, receptor, mensagem e canal, onde o emissor cria uma busca para ser transmitida por uma interface de chat para um receptor. O IBM Watson Assistant, sendo este responsável por encaminhar a mensagem para a API do chatbot, iniciando o processamento da busca e devolvendo o status de "aprendendo", caso a api ainda não teve contato com a mesma busca anteriormente. Após alguns instantes, o usuário poderá fazer a mesma busca, pois a api já processou a requisição, isso permite à API emitir uma resposta relacionada à busca, ao Watson Assistant, que por sua vez devolve ao emissor a resposta "aprendida". Sendo assim, utilizaremos o processo de *PLN (Processamento de Linguagem Natural).*

Motivação

A motivação deste trabalho é a necessidade de obtenção de informações de forma rapida em arquivos, desenvolver um software que estruture esses dados de arquivos textuais e auxilie as mesmas a agilizar o processo de elucidação de novas informações que precisam ser acessadas de forma rápida e prática. A demanda que visamos atingir nas organizações ou na sociedade são os indivíduos que optam por sintetizar informações contidas em tais arquivos, como pdf ou arquivos docx, para acessa-las de forma direta e rápida.

Objetivos Gerais

- Criação de um software que processa arquivos textuais.
- O software auxiliará o usuário a fazer tais buscas em arquivos.

Objetivos Específicos

- Utilizar técnicas de processamento de linguagem natural para criar a api que vai receber as buscas.
- Fazer integração da api com o IBM Watson Assistant.
- O usuário deverá ser capaz de inserir uma query a ser processada intuitivamente seguindo as instruções do próprio software dentro de do chatbot,o IBM Watson Assistant.
- O software deverá capturar os dados inseridos e fazer o processamento realizando o processo de PLN para fornecer uma resposta.
- O software será treinado com as respostas previamente aprendidas e fornecerá a resposta que melhor se adequar à busca feita pelo usuário feita anteriormente.
- Utilizar o Heroku para fazer o deploy. Heroku é uma plataforma em nuvem como um serviço que suporta várias linguagens de programação.

Técnica de Processamento de Linguagem Natural

O software de forma geral pode ser denominado com um sistema de recuperação deinformação que usa um algoritmo probabilístico.

O algoritmo adota palavras-chave (termo de indexação) para indexar e recuperar sentenças. Um termo de indexação éuma palavra que aparece no texto de um documento em uma coleção. O sistema de recuperação apresenta os resultados à uma consulta do usuário, com base no score, e cabe ao sistema identificar qual a sentença é relevante ou não-relevante à solicitação.

Na construção do algoritmo de processamento utilizamos como técnica principal o uso do algoritmo Okapi BM25 ou simplesmente BM.

Okapi BM25 é um algoritmo de classificação usado pelos mecanismos de pesquisa para estimar a relevância de sentenças para uma determinada consulta de pesquisa.

A estrutura da função de processamento consiste em receber um identificador do arquivo que será analisado a quantidades de páginas.

Assim, criamos um loop para interagir em cada página, criando assim as sentençascom base em cada parágrafo.

Sentenças são criadas em cada final de linha, designado por um ponto final. Com as sentenças criadas, passamos para a classe BM25Okapi do Python, que por sua vez fazos devidos cálculos através do método get_scores.

Com os scores podemos resgatar as sentenças mais relevantes através do método get_top_n. Após esse processamento fazemos um de-para entre os os 4 scores mais altos e as 4 sentenças mais relevantes.

Tecnologias Utilizadas

Durante o desenvolvimento da aplicação, foram utilizadas algumas tecnologias de front-end, back-end e para tratamento dos dados, dentre elas temos:

- HTML e CSS para a estilização.
- Next JS: Next.js é uma estrutura da web de desenvolvimento front-end React de código aberto criada por Vercel que permite funcionalidades como renderização do lado do servidor e geração de sites estáticos para aplicativos da web baseados em React para a estrutura da aplicação.
- Python para fazer o processamento de linguagem natural.
- IBM Watson Assistant e IBM functions para o chatbot para fazer conexão entre a API que faz o processamento da linguagem natural PLN hospedada na plataforma em nuvem Heroku.
- MySQL para armazenamento das respostas processadas e obtidas pela API, e dos devidos indentificadores dos documentos de entrada inseridos na base de dados pelo usuário.

JavaScript

Consiste em uma linguagem de alto nível, com tipagem dinâmica fraca e multiparadigma. Com o HTML e CSS, o JavaScript é uma das três maiores tecnologias para WWW. Por aceitar páginas web interativas, a linguagem se tornou essencial no desenvolvimento de aplicativos Web, sendo assim, grandes provedores possuem um mecanismo dedicado á execução dessa linguagem.

Por ser uma linguagem multiparadigma, a linguagem suporta estilos de programação funcional e imperativa, possuindo recursos como closures e high-order functions, geralmente indisponíveis em em linguagens mais comuns como Java e C++.

Next.JS

Framework usado para react, sendo essa uma biblioteca para construção de interfaces, e com isso, o next.js adiciona várias funcionalidades baseadas em react.

Algumas funcionalidades do Next.js:

- Renderização estática;
- Possui suporte ao Typescript;
- Serviço de tratamento de rotas.

MySQL

MySQL é o sistema gerenciador de banco de dados que utiliza a linguagem SQL como interface. Atualmente é o SGBD mais utilizado e conhecido do mundo com mais de 10 milhões de instalações. Foi criado pela Oracle Corporation em 1995, sendo utilizado por mais de 26 anos e funcional para diversas empresas realizando a organização dos dados produzidos pelo sistema ERP conectado pelo servidor, este podendo ser customizado para cada empresa que comprar uma licença comercial.

Por ser um banco de dados relacional, sua manipulação é mais simples, porém requer cuidado com sua utilização. Para que o banco seja integro e nenhum dado seja violado, é importante levar em consideração as quatro formas normais para alcançarmos a normalização dos dados, retirando quaisquer redundâncias e mantendo a ligação das chaves primárias de forma correta.

A ferramenta utilizada foi o MySQL WorkBench, que juntamente com a criação das tabelas e relacionamentos, cria o diagrama entidade-relacionamento, facilitando a visualização do fluxo de dados dentro do SGBD.

Python

É uma linguagem de alto nivel fortemente tipada Open-Source utilizada em data science, machine learning, desenvolvimento para web e aplicativos, automação de scripts e muito mais.

Por ter sua vertente voltada para analise de dados, a linaguagem foi escolhida para que a inteligência artificial do IBM Watson possa reconhecer os dados de entrada e fazendo uso dos algoritmos de machine learning, a máquina possa aprender as respostas encontradas e fornecer dados adicionais como quantidade de buscas por mês, gráficos, histórico de buscas e temas mais procurados, sendo essas funcionalidades inseridas em

novas atualizações do produto.

IBM Watson Assitant

O IBM Watson Assistant é um serviço em nuvem de marca branca que permite aos desenvolvedores de software de nível corporativo incorporar um assistente virtual de inteligência artificial ao software que estão desenvolvendo e marcar o assistente como seu. O serviço, que dá aos consumidores acesso ao Watson AI , é fornecido por meio da IBM Cloud.

- Os usuários interagem com o assistente por meio de um ou mais destes pontos de **integração**:
    - Um assistente virtual que você publica diretamente em uma plataforma de sistema de mensagens de mídia social existente, como o Slack ou o Facebook Messenger.
    - Um bate-papo da web que você integra ao seu website da empresa que pode responder às perguntas do cliente diretamente e pode transferir solicitações complexas para um representante de suporte ao cliente.
    - Um aplicativo customizado que você desenvolve, como um aplicativo móvel ou um robô com uma interface de voz.
- O **assistente** recebe a entrada do usuário e a roteia para a qualificação de diálogo.
- A **qualificação de diálogo** interpreta a entrada do usuário mais a fundo e, em seguida, direciona o fluxo da conversa. O diálogo reúne quaisquer informações necessárias para responder ou executar uma transação no nome do usuário.
- Todas as perguntas que não podem ser respondidas pela qualificação de diálogo são enviadas à **qualificação de procura**, que localiza respostas relevantes procurando as bases de conhecimento da empresa configuradas para esse propósito.

Ilustração de Funcionamento do IBM Watson

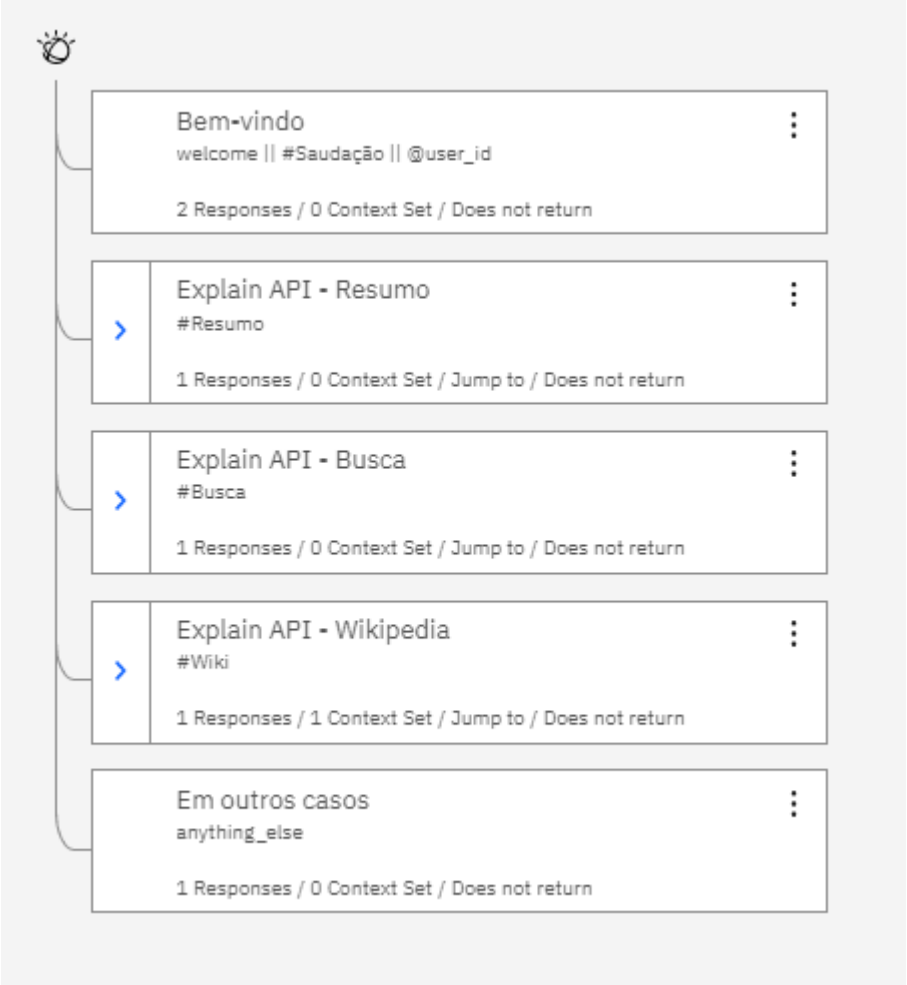

Figura 1: Diagrama de funcionamento do IBM Watson.

Diagrama de Caso de Uso

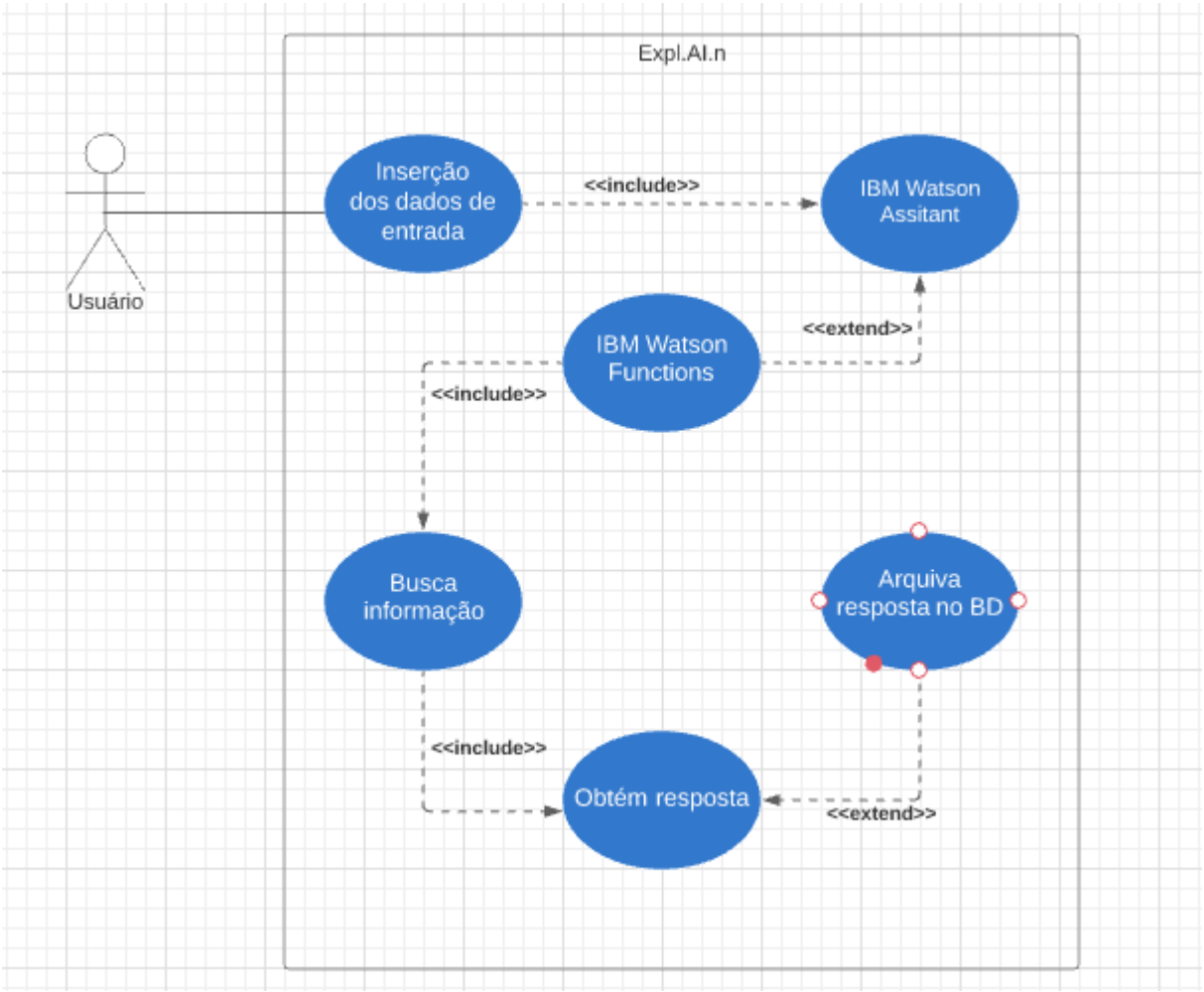

Figura 2: Diagrama de caso de uso referente ao Expl.AI.n

Considerações Finais do Grupo

Durante o desenvolvimento do trabalho, todo grupo pode perceber que a quantidade de dados que são obtidos é exponencial. Segundo Fosse, 2019, os dados crescem exponencialmente à medida que se criam mais sensores e a interação entre pessoas e coisas aumenta. Conversas geradas nas mídias sociais, dados de localização, sinais fisiológicos de uma pessoa, logs de acesso a websites –todos são dados dignos de valor para alimentar soluções inteligentes que visam aprimorar a experiência do cliente e criar novos serviços.

O usuário necessita a cada dia obter a resposta a dado um problema de forma mais rápida e ágil, sendo necessário utilizarmos várias tecnologias com a finalidade de refinarmos cada vez mais a performance das buscas e assertividade dos resultados obtidos, pois assim, grandes corporações, empresas e entidades podem se beneficiar do produto para que suas operações possam ser agilizadas e maximizem seus lucros.

Desde os primórdios, o ser humano aperfeiçoa o processo de comunicação, usando como meios de comunicação a linguagem falada, linguagem de sinais, braile ou linguagem não falada, por meio da arte. Sendo assim, a linguagem é baseada em comandos ou instruções que chegam ao destinatário daquele comando, gerando a comprrensão do assunto da mensagem e feedback. Para as linguagens de programação, podemos utilizar o mesmo conceito, o desenvolvedor gera comandos para serem enviados ao compilador para tradução e compreensão da máquina, sempre em linguagem de máquina ou de baixo nível.

Com o uso da PLN, o sofware ganhou interatividade com o usuário e seu nível de usabilidade, conforme as Heurísticas de Nielsen, aumenta significativamente. Portanto, o uso das linguagens descritas durante esse trabalhos e tornou mais eficiente e assertiva.

Como dificuldades que tivemos, o tempo foi um fator determinante, pois tinhamos muitas ideias para este trabalho, mas fizemos aquilo que era de suma importância para sua conclusão, porém podemos adicionar mais funcionalidades e aprimorar ainda mais essa tecnologia.

A escolha das tecnologias também foi muito importante para o desenvolvimento do trabalho, pois com a elicitação dos requisitos da aplicação, levantar as tecnologias necessárias sempre foi um desafio grande, para que a requisição do usuário seja atendida.

Referências Bibliográficas

“FATORES QUE AFETAM A ADOÇÃO DE ANÁLISES DE BIG DATA EM EMPRESAS Rev. adm. empres. vol.59 no.6 São Paulo Nov./Dec. 2019 Epub Jan 10, 2020”Juan-Pedro Cabrera-Sánchez, Ángel F Villarejo-Ramos, 2020 Disponível em: https://www.scielo.br/scielo.php?pid=S0034-75902019000600415 HYPERLINK "https://www.scielo.br/scielo.php?pid=S0034-75902019000600415&script=sci_arttext&tlng=pt"tlng=pt. Acesso em 25/11/2020.

“O desenvolvimento da computação cognitiva”, 2015. Yasmim Man Tchin Chang Lee, Andrea Martins Cristóvão, Rogerio MatheusGrillo,Carlos Rafael Melo de Lira, 2015
Acesso em 10/11/2020

Rank-BM25: A two line search engine. Disponível em: https://pypi.org/project/rank-bm25/. Acesso em: 23/11/2021.

“Sobre o Watson Assistant”, IBM, Acesso em 25/11/2021, https://cloud.ibm.com/docs/assistant?topic=assistant-inde.