Implementação de um Protótipo de Chatbot Usando Recursos de Processamento de Línguas Naturais

Bruno Queiroz Santos, William dos Santos Abreu

[1] Departamento de Ciência da Computação
Universidade Federal de Lavras (UFLA) – Lavras, MG – Brasil

{bruno.santos2, william.abreu}@estudante.ufla.br

Abstract. *This article aims to describe a chatbot prototype implementation, that is a computational artifact that interacts with the user through a primitive dialogue, consisting of the human requesting questions and the computer answering them. For the chatbot model building, natural language processing tools were used to process the human text in a standardized form that can be handled by the machine. In this process, the work takes advantage of the numerous resources available in programming toolkits for the development of the prototype. The results obtained at the end of the implementation show that the model is quite simple and slow, but it can answer some simple information present in the corpus.*

Resumo. *Este artigo tem como objetivo descrever a implementação de um protótipo de chatbot, artifício computacional que interage com o usuário através de um diálogo primitivo, consistindo do humano fazendo perguntas ao computador e este o respondendo. Para a construção do modelo do chatbot, foram usadas ferramentas de processamento de língua natural para processar o texto humano em uma forma normalizada que possa ser tratada pela máquina. Nesse processo, o trabalho aproveita dos inúmeros recursos disponíveis em toolkits de programação para o desenvolvimento do protótipo. Os resultados obtidos ao final da implementação mostram que o modelo é bastante simples e lento, mas consegue responder algumas informações simples presentes no córpus.*

1. Introdução

Um chatbot é um software de inteligência artificial em um dispositivo, aplicativo ou website, que tenta avaliar as necessidades do usuário e ajudá-lo a executar uma tarefa específica a partir da interação de diálogo entre humano e computador. Hoje é comum encontrar em websites de inúmeras empresas chatbots para ajudar o usuário a encontrar a informação da qual ele precisa, além das assistentes pessoais digitais disponíveis, mas que interagem através de comandos de voz.

O chatbot tenta simular um ser humano na conversação. O objetivo é responder as perguntas de tal forma que as pessoas tenham a impressão de estar conversando com outra pessoa e não com um programa de computador. Após o envio de perguntas em linguagem natural, o programa consulta uma base de conhecimento e, em seguida, fornece uma resposta que tenta imitar o comportamento humano.

O primeiro programa para processamento de linguagem natural da história foi o software Eliza, que foi e foi criado por Joseph Weizenbaum no laboratório de Inteligência

Artificial do MIT, entre os anos de 1964 e 1966. A principal implementação do programa mostra a simulação da conversa entre um paciente e seu psicólogo, na qual o usuário é o paciente e o software o psicólogo.

Atualmente, o mercado dos chatbots está muito em alta. A finalidade e as possibilidades de um chatbot são ilimitados, pois pode ser usado para diversas formas de interação com usuários e clientes, como para obtenção de informações de um serviço, do status do pedido de uma encomenda, para solicitar suporte e etc., tudo de maneira automatizada.

2. Revisão Bibliográfica

O termo Chatbot surgiu da junção das palavras chatter (a pessoa que conversa) e da palavra bot (abreviatura de robot), ou seja, um robô (em forma de software) que conversa com as pessoas. A palavra foi inventada por Michael Mauldin em 1994, para descrever estes robôs de conversação na Twelfth National Conference on Artificial Intelligence.

Existem basicamente duas variantes de chatbots: baseada em regras e em auto-aprendizagem. Em uma abordagem baseada em regras, um bot responde a perguntas com base em algumas regras nas quais ele é treinado. As regras definidas podem ser desde muito simples até muito complexas. Os bots podem lidar com consultas simples, mas não conseguem responder as perguntas mais complexas. Já na auto-aprendizagem, são utilizadas abordagens baseadas no Aprendizado de Máquina, que são definitivamente mais eficientes do que os bots baseados em regras devido a maior capacidade de responder a requisições mais elaboradas.

O campo de estudo que se concentra nas interações entre a linguagem humana e os computadores é chamado Processamento de Linguagem Natural (PLN). Ele está na intersecção entre ciência da computação, inteligência artificial e linguística computacional. O PLN é uma forma de os computadores analisarem, compreenderem e derivarem o significado da linguagem humana. Através do PLN é possível a construção do chatbot, por ser uma tarefa de processamento do texto de entrada para obtenção de uma resposta para a saída.

O principal problema com dados de texto é o fato dos dados serem cadeias de caracteres (strings). No entanto, os algoritmos de aprendizado de máquina precisam de algum tipo de recurso numérico para executar a tarefa. Então, antes de qualquer projeto de PNL, é necessário fazer o pré-processamento para normalizar o formato para entrada em um algoritmo. O pré-processamento de texto básico inclui:

- Converter todo o texto em caixa baixa (**lowercase**), para que o algoritmo não diferencie palavras iguais devido a mudança, dando caráter de não case-sensitive.
- **Tokenizar** o texto, que é o processo de converter as cadeias de texto normais em uma lista de tokens, ou seja, palavras que realmente são desejáveis. O tokenizador de sentença pode ser usado para encontrar a lista de frases e o tokenizador de palavras pode ser usado para encontrar a lista de palavras em sequências de caracteres.
- Remover **ruído** do texto, ou seja, tudo o que não está seja um número ou uma letra padrão, como pontuações.
- Remover as **stopwords**, que são palavras extremamente comuns e que não carregam um significado útil para o processamento.

- Fazer **stemming**, processo de reduzir as palavras para a sua raiz, convertê-la em seu radical.
- **Lematizar**, que é uma ligeira variante de stemming. A principal diferença entre eles é que, muitas vezes, o stemming pode criar palavras inexistentes, enquanto os lemas são palavras reais, pois o lema é uma palavras que você pode procurar em um dicionário, enquanto o radical pode ser apenar uma parte da palavra, que não significa nada sozinha.

Após o pré-processamento, é necessário transformar o texto em um vetor significativo de números, conforme limitação citada anteriormente. A bag-of-words é uma representação de texto que descreve a ocorrência de palavras dentro de um documento, envolvendo:

- O vocabulário das palavras conhecidas.
- A medida de ocorrência da palavra conhecida.

A tradução do termo remete a saco de palavras. O nome é dado pelo fato de que qualquer informação sobre a ordem ou estrutura das palavras no documento é descartada, deixando o modelo se preocupar apenas com o fato de as palavras conhecidas ocorrerem no documento e não onde elas ocorrem no documento. A intuição por trás do bag-of-words é que os documentos são semelhantes se tiverem conteúdo semelhante. Além disso, podemos aprender algo sobre o significado do documento apenas a partir do seu conteúdo.

Um problema com a abordagem do bag-of-words é que palavras altamente frequentes começam a dominar no documento (por exemplo, pontuação maior), mas podem não conter tanto conteúdo informativo. Além disso, ele dará mais peso a documentos mais longos do que documentos mais curtos. Uma abordagem é redimensionar a frequência das palavras pela frequência com que aparecem em todos os documentos para que as pontuações de palavras frequentes como “o”, que também são frequentes em todos os documentos, sejam penalizadas. Esta abordagem de pontuação é chamada Term Frequency Inverse Document Frequency (TF-IDF), em que:

- Term Frequency é o valor de frequência de uma palavra no documento.
- Inverse Document Frequency é o valor de quão rara é uma palavra através dos documentos.

O valor de TF-IDF é um valor frequentemente usado na recuperação de informações e na mineração de texto. Esse peso é uma medida estatística usada para avaliar a importância de uma palavra para um documento em um córpus.

Modelos de PLN funcionam a partir de um conjunto de dados, o córpus, assim como outros recorrentes modelos de inteligência artificial. O córpus linguístico é o conjunto de textos escritos em uma determinada língua e que serve como base de análise. É a partir dele que um chatbot, no caso, pode buscar respostas para cada pergunta que é feita a ele.

O TF-IDF é uma transformação aplicada a textos para obter dois vetores reais no espaço vetorial. Pode ser então obtida a semelhança de cosseno de qualquer par de vetores, pegando seu produto escalar e dividindo-o pelo produto de suas normas. Isso produz o cosseno do ângulo entre os vetores. Semelhança de cosseno é uma medida de similaridade entre dois vetores não nulos. Usando essa fórmula, pode-se descobrir a semelhança entre quaisquer dois documentos d1 e d2.

3. Método

A construção do protótipo é feita a partir do toolkit em Python NLTK. O NLTK (Natural Language Toolkit) é uma plataforma líder para criar programas em Python para trabalhar com dados em linguagem natural. Ele fornece métodos fáceis de usar para diversos recursos corpora e lexicais, junto com um conjunto de bibliotecas de processamento de texto para classificação, tokenização, stemming, tagging, análise e raciocínio semântico, wrappers para bibliotecas de PLN de nível industrial.

Também é necessário o uso do toolkit Scikit-learn, para análise de dados em nível de inteligência artificial, também implementado em Python. Nele está disponível os modelos de conversão TF-IDF e o método de cálculo da semelhança de cosseno.

O córpus utilizado na construção do chatbot é na verdade uns corpora (conjunto de córpus, plural) compilados pelos próprios autores através de webcrawling, sobre os assuntos de mercado financeiro, saúde, tecnologia e telefonia. Deste modo, o chatbot agrega capacidade de responder a esses determinados assuntos, pois são as informações que ele conhece.

A primeira tarefa é carregar as sentenças dos corpora um uma estrutura de dados de lista, deixando os textos tokenizados em sentenças. A partir de então, é feito o pré-processamento das sentenças para normalizar os tokens gerados, deixando os tokens em uma forma padrão de caixa baixa e removendo sinais gráficos de pontuação.

A interação inicial do bot com o humano é simples, bastando apenas fazer o reconhecimento de palavras de saudação para gerar a resposta de cumprimento. A geração de respostas para perguntas de fato feitas pelo usuário é baseada em uma regra simples de inferência no córpus. A entrada digitada pelo humano primeiramente é pré-processada (normalizada) deixando em caixa baixa, removendo pontuações e convertendo os tokens para seus respectivos lemas. A partir disso, a sentença representando a pergunta do usuário o chatbot é convertida em um vetor TF-IDF e, através desse vetor, é feita uma busca sequencial nas sentenças do córpus aplicando semelhança de cosseno de cada sentença com a pergunta. O casamento de melhor resultado é retornado como resposta do bot na interação entre humano e compuatdor. Se nenhuma sentença for semelhante suficiente, será respondida uma mensagem informando a incapacidade de responder aquela pergunta.

4. Resultados

Após a implementação do protótipo do chatbot, são executados testes quanto ao seu comportamento, funcionalidade e eficiência, observando como o modelo responde a determinadas perguntas e quanto tempo ele gasta para processar essas respostas.

Um teste rápido foi executado e foram obtidos os seguintes resultados, destacando a pergunta de entrada, a resposta do chatbot e o tempo gasto para processar a requisição:

1. **Pergunta:** politica
   Resposta: Se a sua resposta for apelo ao comunismo, eu não discuto politica.
   Tempo: 15,0 segundos

2. **Pergunta:** o que você vai fazer diante a situação da saúde brasileira?
   Resposta: Como fazer?

Tempo: 12,7 segundos

3. **Pergunta:** Qual o melhor sistema operacional, linux ou windows?
   Resposta: ”Usuário Windows desde meu primeiro computador, nunca tive grande interesse em usar outro sistema operacional.
   Tempo: 12,6 segundos

4. **Pergunta:** linux ou ruindows?
   Resposta: Muito provavelmente o Linux.
   Tempo: 12,6 segundos

5. **Pergunta:** fala papai, preveja meu futuro
   Resposta: Chinês não fala.
   Tempo: 12,7 segundos

6. **Pergunta:** bitcoin e promissor?
   Resposta: “É promissor, mas ainda muito inicial.
   Tempo: 12,6 segundos

7. **Pergunta:** como vai?
   Resposta: Não vai.
   Tempo: 12,7 segundos

Além do exemplo de execução mostrado anteriormente, a partir das observações no protótipo, pontuam-se os seguintes aspectos:

- O chatbot responde a mesma resposta para a mesma pergunta, característica da regra de inferência sequencial aplicada ao algoritmo.
- O chatbot demora por volta de 12 segundos para gerar uma resposta, isso porque não foi aplicada alguma técnica de otimização de busca para contornar isso.
- O chatbot pode ter a mesma resposta pra duas perguntas distintas, isso acontece porque as vezes as palavras que estão na pergunta não ocorrem muitas vezes no córpus.

O comportamento do chatbot apresenta-se bastante fraco, mas é possível extrair informações do córpus desde que se tenha a destreza de fazer a pergunta certa, por causa da regra de inferência. Mesmo assim, tentar interagir com o bot é uma experiência cômica, já que ele pode retornar algumas respostas parecendo uma pessoa tola falando.

5. Considerações Finais

Embora seja um bot muito simples, com quase nenhuma habilidade cognitiva, é um bom protótipo que utiliza os principais recursos aprendidos dos toolkits utilizados na disciplina de Introdução ao Processamento de Línguas Naturais.

6. Referências

1. Disponível em: ⟨https://medium.com/analytics-vidhya/building-a-simple-chatbot-in-python-using-nltk-7c8c8215ac6e⟩.

Acesso em: 21 jun. 2019.

2. Disponível em: ⟨https://blog.cedrotech.com/exemplos-de-chatbots-o-que-as-empresas-estao-fazendo/⟩. Acesso em: 22 jun. 2019.

3. Disponível em: ⟨https://en.wikipedia.org/wiki/Natural_language_processing⟩. Acesso em: 22 jun. 2019.

4. Disponível em: ⟨http://www.nltk.org/book/⟩. Acesso em: 22 jun. 2019.

5. Disponível em: ⟨https://pt.wikipedia.org/wiki/Chatterbot⟩. Acesso em: 22 jun. 2019.

6. Disponível em: ⟨https://pt.wikipedia.org/wiki/ELIZA⟩. Acesso em: 22 jun. 2019.

7. Disponível em: ⟨https://chatbotsmagazine.com/chatbot-report-2018-global-trends-and-analysis-4d8bbe4d924b⟩. Acesso em: 22 jun. 2019.