UM CHATBOT PARA RESPONDER FAQs[1]

Sérgio Ricardo Josino da Silva Júnior, Yuri de Almeida Malheiros Barbosa

Departamento de Ciências Exatas - Universidade Federal da Paraíba (UFPB)
Rio Tinto - Paraíba, Brasil

{sergio.ricardo, yuri}@dcx.ufpb.br

Abstract. *Frequently Asked Questions (FAQs) are used to answering the most common questions made by users, with that the support service receive less calls. However, the number of questions presented in a FAQ could be large, what makes difficult the search for an answer. This paper presents the creation of a chatbot which is capable of FAQ answering using Latent Semantic Indexing (LSI), its use and evaluation.*

Keywords: *Chatbot, Vectors, LSI, FAQs*

Resumo. *Frequently Asked Questions* (FAQs) no português Perguntas Mais Frequentes são usadas para responder às perguntas mais comuns feitas por usuários, de forma que o suporte de um produto ou serviço seja menos contatado. Contudo, a grande quantidade de perguntas e respostas exibidas ao usuário podem confundí-lo ou tornar difícil o que seria a simples busca por uma resposta. Diante disso, este artigo apresenta a criação de um *chatbot* capaz de responder FAQs, utilizando *Latent Semantic Indexing* (LSI) assim como seu uso e avaliação.

Palavras-chave: Chatbot, Vetores, LSI, FAQs

[1] Trabalho de Conclusão de Curso do discente Sérgio Ricardo Josino da Silva Júnior, sob a orientação do docente Yuri de Almeida Malheiros Barbosa submetido ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal da Paraíba, Campus IV como parte dos requisitos necessários para obtenção do grau de Bacharel em Sistemas de Informação.

1. Introdução

Frequently Asked Questions (FAQs), em português, Perguntas Mais Frequentes é a forma de responder perguntas que foram feitas repetidamente por usuários de um determinado produto ou serviço. A adoção de uma FAQ para um fornecedor de serviço poderá fazer com que o seu suporte seja cada vez menos contatado, assim diminuindo os custos de suporte aos clientes. Atualmente, grandes companhias tais como bancos e lojas, utilizam FAQs e as dividem em categorias facilitando a busca do cliente por uma solução, tendo como principais objetivos a rapidez na resposta dada e a clareza do conteúdo apresentado na mesma.

Devido a vasta quantidade de perguntas que são apresentadas dentro de uma FAQ, adicionando o fato de que nem todas as pessoas conseguem encontrar com facilidade a resposta desejada é interessante pensar em uma forma de automatizar o processo de busca.

É observável o constante crescimento do uso de *chatbots,* são diversos os propósitos para o desenvolvimento de um *chatbot*, seja para tutoria, assistentes virtuais em atividades do dia a dia, auxiliar no aprendizado de um novo idioma, orientação de um usuário na web enquanto navega e até esclarecer dúvidas quanto a serviços oferecidos por uma organização. Em todas as necessidades citadas, é esperada a interação com o *chatbot* através de linguagem natural [Junior e Netto, 2014][Souza e Moraes, 2015].

Um *chatbot* caberia como solução viável para buscas de respostas em FAQs. As pessoas querem se comunicar com os computadores da mesma forma que se comunicam umas com as outras [Shawar e Atwell, 2007]. Com base nesta informação, observamos a importância da inserção dos *chatbots* no ramo das FAQs, considerando que pode ocorrer uma mudança, fazendo com que algo cansativo torne-se bastante intuitivo, ao invés de buscar em uma vasta lista de perguntas, o usuário apenas digitará sua dúvida e o *chatbot* responderá, tornando assim, um diálogo natural entre homem e máquina.

O objetivo deste trabalho é criar um *chatbot* capaz de responder FAQs, assim um usuário entra com uma pergunta, o *chatbot* analisa e consulta seu banco de perguntas e respostas para tentar responder. O algoritmo por trás do *chatbot* utilizará o princípio LSI (*Latent Semantic Indexing*) no português Indexação Semântica Latente, técnica capaz de medir a similaridade entre documentos e consultas baseando-se nos vetores matemáticos que

os representam. Se a consulta for considerada semelhante a um documento, então a aproximação será expressada através de um coeficiente de similaridade [Silva, 2011].

O algoritmo será treinado através de um *dataset* criado com as perguntas mais frequentes dos usuários, presentes no site de FAQ da Microsoft[2] e irá buscar a similaridade entre as palavras digitadas pelo usuário com as que foram recebidas durante o treinamento utilizando-se da LSI, retornando assim, a possível resposta para o usuário. A avaliação de qualidade do *chatbot* será medida através de perguntas feitas por usuários, onde serão medidas quantas perguntas são respondidas adequadamente.

2. Fundamentação Teórica

2.1. Chatbot

Chatbots são descritos como sistemas capazes de conversar com usuários de maneira natural. Oferecem auxílio ao usuário em uma interação homem-máquina. Possuem capacidade de examinar e até mesmo influenciar o comportamento do seu usuário, perguntando e respondendo às suas perguntas [Abdul-Kader e Woods, 2015].

Existe uma plataforma por trás de qualquer *chatbot* que é responsável por mapear as palavras do usuário para uma resposta apropriada. Quando codificado à mão, a construção do conhecimento de um *chatbot* pode ser tornar cansativa e difícil de se adaptar a novos domínios [Huang, et al, 2007].

Segundo Dale (2006), se chamadas de assistentes virtuais, interfaces conversacionais ou *chatbots*, o conceito básico seria o mesmo: conseguir resultados conversando com uma máquina através de um diálogo inteligível através de linguagem natural. Siri (Apple), Cortana (Microsoft), Alexa (Amazon) e Google Assistant (Google) são as mais conhecidas do mercado e utilizam na maior parte do tempo por comandos de voz. Porém, ainda existem muitos *chatbots* que utilizam-se de texto como seu método de conversa principal.

Levando em consideração a definição de *chatbot*, percebemos que os mesmos já encontram-se presentes desde a década de 1960, onde Joseph Weizenbaum desenvolveu a Eliza, este *chatbot* tinha como objetivo conversar como uma psicoterapeuta, foi desenvolvida utilizando *pattern-matching*. Devido muitas pessoas não terem se dado conta de que Eliza

[2] https://www.microsoft.com/en-us/software-download/faq

não era humana deu inspiração para que um dia fosse possível construir um *chatbot* que passasse no Teste de Turing [Dale, 2016].

.

2.1.1. AIML

AIML (*Artificial Intelligence Markup Language*) é uma linguagem derivada do XML que se tornou base para o desenvolvimento de *chatbots*, além de ser baseada na tecnologia desenvolvida para a A.L.I.C.E. [Alice, 2015].

O propósito da AIML é simplificar o trabalho da modelagem conversacional, em relação à um processo de estímulo-resposta. Consiste em uma linguagem de marcação baseada no XML e depende dos seus identificadores (*tags*) responsáveis pelos seus trechos de código para enviar instruções ao *chatbot*. A classe é definida em AIML como um objeto AIML e este objeto tem como responsabilidade modelar padrões conversacionais. Isto significa que cada objeto AIML possui uma *tag* associada à um comando [Abdul-Kader e Woods, 2015].

Dentro da AIML os objetos mais importantes são: *category*, *pattern* e *template*. A tarefa da *tag category* é definir o conhecimento de uma conversação, ou seja, a categoria que será tratada pelo *chatbot*. A *tag pattern* identifica a entrada do usuário e o objetivo da *tag template* é responder à entrada de usuário específica [Abdul-Kader e Woods, 2015 apud Marietto et al., 2013].

Segundo Abdul-Kader e Woods (2015), estas são as tags mais frequentes que servem como base para construir um *chatbot* AIML capaz de responder inteligentemente às conversões de fala. As tags acima citadas possuem a seguinte ordem:

A imagem 1 demonstra a estrutura básica da AIML.

```
<category>
    <pattern>User Input</pattern>
    <template>
        Corresponding Response to Input
    </template>
</category>
```

Imagem 1. Estruturação do AIML

2.2. Latent Semantic Indexing (LSI)

Latent Semantic Indexing (LSI) em português, Indexação Semântica Latente é uma técnica criada baseada em representação de textos em vetores [LSI, 2004]. Tem como principal objetivo, captar a similaridade dos textos comparando seus vetores. Foi desenvolvido como uma tarefa de (RI) Recuperação da Informação, isto é, selecionando os documentos mais importantes dentro de uma vasta coleção de documentos, onde será selecionada a mais similar à uma pergunta feita.

O LSI abrange a aproximação baseada em vetor, utilizando a Singular Value Decomposition (SVD) [SVD, 1994] para reconfigurar os dados. O SVD é uma técnica de matriz algébrica que reorienta e classifica as dimensões dentro de um espaço vetorial. Com isso, as dimensões em um espaço vetorial computadas pelo SVD são ordenadas de mais para menos significantes.

Baseando-se em um número reduzido de representações, palavras que estão presentes em contextos similares, tendem a possuir vetores similares, por isso, podem conseguir alto índice de similaridade.

Muitas das aplicações foram desenvolvidas utilizando dicionários de palavras muito pequenos, em uma das aplicações bem-sucedidas, o LSI foi treinado com um dicionário que continha algumas centenas de kilobytes, sendo cerca de 2000 palavras, 30.000 tokens de palavras e 325 documentos, enquanto em outros lugares conseguiram mostrar bons resultados com dicionários de palavras muito grandes, uma pesquisa feita pela Universidade do Colorado reportou que fizeram um treinamento de LSI que continha cerca de 750.000 palavras, 550 milhões de tokens de palavras e 3.6 milhões de documentos [Wiemer-Hastings, 2004].

Baseando-se nisto, caso a quantidade de palavras seja pequena, será necessário aumentar também o número de dimensões, caso seja uma baixa quantidade de palavras, poucas dimensões podem dar conta com facilidade.

3. Metodologia

Nesta seção será descrito o passo a passo para o desenvolvimento do *chatbot.*

3.1 Desenvolvimento do Chatbot

O *chatbot* foi desenvolvido utilizando a linguagem Python, com auxílio da biblioteca gensim[3]. Ele recebe como entrada uma pergunta de um usuário e compara com um banco de perguntas cadastradas para fornecer uma resposta. Exemplos das perguntas podem ser vistos na Tabela 1.

Tabela 1 - Perguntas presentes no *dataset* extraído da FAQ da Microsoft

Perguntas
How do I find my Windows Product Key?
My Windows 7 product key won't verify. What's the problem?
I purchased my copy of Windows through a university. Can I download it here?
How do I tell if my computer can run a 64-bit version of Windows?

O processo de funcionamento do *chatbot* ocorre de acordo com a figura 2.

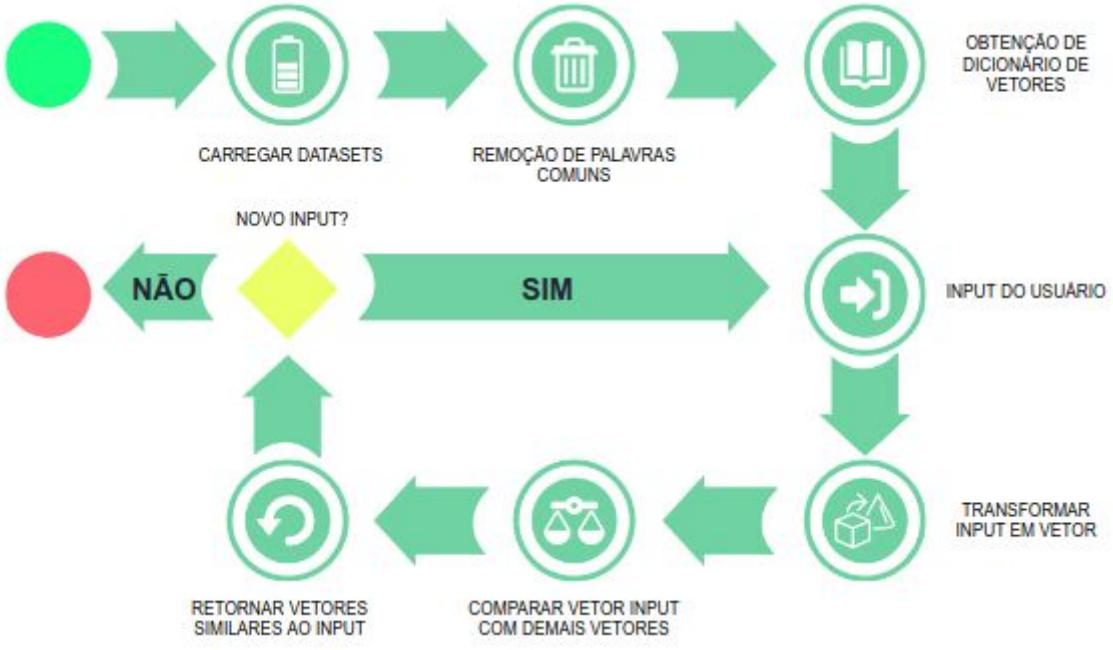

Figura 2. Processo de Funcionamento do Chatbot

[3] https://radimrehurek.com/gensim/

Após a leitura das perguntas acima citadas, o seguinte passo foi remover *stopwords*, que são palavras muito comuns presentes em diálogos em determinados idiomas [Stopwords, 1992]. Ao finalizar os dois primeiros passos, foi possível obter um dicionário de perguntas, que logo em seguida foram transformadas em vetores através do LSI.

Com o *bag-of-words*, podemos representar cada documento como um vetor de palavras que ocorrem no documento ou em outras representações, como frases ou sentenças [Matsubara, Edson et al., 2003]. Após aplicado o *bag-of-words* foi possível contabilizar 77 palavras dentro do documento que estamos trabalhando, cada palavra assume um id de 0 a 76, como por exemplo: `{'do': 0, 'had': 18, 'difference': 70, …}`, onde a palavra *do* possui (id 0), respectivamente *had* e *difference* possuem (id 18) e (id 70).

Feito isso, obtivemos o seguinte modelo de representação em vetor: `[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1)]` onde em cada parênteses o primeiro número representa o id de uma palavra e o outro representa o número de ocorrência desta mesma palavra, o vetor representado acima equivale à pergunta "*How do I find my Windows Product Key?*", onde `{'do': 0, 'find': 1, 'how': 2, 'key?': 3, 'my': 4, 'product': 5, 'windows': 6}`, a falta das palavras *How* e *I* dá-se devido a ambas serem *stopwords*, esta técnica repete-se para todas as sentenças presentes no *dataset*.

No próximo passo, foi realizada a conversão do tipo de vetor, foi utilizada a *tfidf* (*term frequency-inverse document frequency*), esta técnica tem como objetivo contar a frequência que um determinado termo aparece nos diálogos, fazendo com que os diálogos mais frequentes possuam peso de representação menor [Matsubara, Edson et al., 2003].

Após aplicar a *tfidf*, obtivemos o seguinte vetor: `[(0, 0.28516984562326425), (1, 0.5703396912465285), (2, 0.39041755082143187), (3, 0.5703396912465285), (4, 0.1525734185012273), (5, 0.28516984562326425), (6, 0.10524770519816758)]`, este vetor também equivale a pergunta "*How do I find my Windows Product Key?*", porém, desta vez estão distribuídos de acordo com a peso de representação que cada um obteve através do cálculo feito pela *tfidf*, onde o (id 6) que equivale a palavra *Windows* possui a menor peso de representação, o que é justificável, por ser um produto da Microsoft, a palavra repete-se inúmeras vezes na FAQ fazendo com que a sua peso de representação baixe.

A seguir, o vetor que foi gerado pelo *tfidf* foi analisado pelo algoritmo do LSI, com o número de dimensões 10, devido a grande proximidade com a quantidade de perguntas presentes no *dataset* que eram 9. Estas dimensões servem para procurar em determinada

pergunta a correlação entre as palavras presentes no mesmo, neste exemplo podemos observar o tópico 0 equivalente à dimensão 0: `[(0, '-0.285*"how" + -0.253*"do" + -0.227*"key?" + -0.227*"find" + -0.212*"windows?" + -0.212*"64-bit" + -0.190*"run" + -0.189*"if" + -0.189*"computer" + -0.189*"tell"')]` é notável que as palavras acima estão de acordo com LSI relacionadas de alguma forma.

O próximo passo é obter a similaridade entre uma consulta, no nosso caso a entrada do usuário com os demais diálogo presentes em nosso *dataset*, este passo utiliza-se dos vetores gerados pelas dimensões do LSI anteriormente citadas juntamente com uma técnica conhecida como *Cosine Similarity* [Cosine Similarity, 2010].

A técnica irá retornar similaridades em um alcance de -1 à 1, sabendo que quanto maior a similaridade, melhor e mais parecido é o contexto. Utilizando novamente da frase "*How do I find my Windows Product Key?*" e simulando a entrada como "*What do I do to find my product key?*", é obtido o resultado: `[(0, 0.9810546)]`, onde o 0 equivale a frase "*How do I find my Windows Product Key?*" que foi utilizada durante todo o decorrer do texto e 0.9810546 é o grau de similaridade entre a sentença presente no *dataset* e a digitada pelo usuário.

4. Resultados

Um *chatbot* foi criado através da metodologia aqui proposta, foi possível treinar o algoritmo com um *dataset* de perguntas e respostas de uma FAQ real e por fim responder às perguntas que foram feitas pelos usuários.

Para a fase final de testes, convidamos quatro pessoas para utilizarem o *chatbot*, explicando-lhes que os resultados ali obtidos iriam fazer parte de um trabalho acadêmico, a seguir, era lançada uma problemática: "*Digamos que você é um usuário que instalou o Windows, porém, precisa da chave do produto, como você a encontraria e faria a ativação do Windows?*", feito isso, as pessoas enviaram cada uma a sua forma de interação com o *chatbot* e obtiveram a resposta.

Neste processo avaliativo, foram elegidas as três perguntas mais similares à pergunta feita pelo usuário, o *chatbot* respondeu adequadamente a maioria das perguntas, devido serem perguntas fáceis de serem feitas, visto que todo o conteúdo está no idioma inglês e não é grande o número de pessoas que consegue manter uma boa comunicação através deste.

Quando as respostas falharam, foi checado qual o índice de similaridade da pergunta correta e na maioria das vezes, estava presente na segunda posição.

As perguntas que não foram respondidas de forma adequada, foram reformuladas as perguntas, onde foram obtidas as resposta esperada, ou seja, em alguns casos é necessário que as perguntas sejam idênticas às perguntas existentes no *dataset*.

5. Trabalhos Relacionados

Durante o período de estudo para desenvolvimento deste trabalho, foi possível encontrar diversos trabalhos relacionados à *chatbot* com temáticas diferentes, dando forte destaque ao ensino-aprendizado.

Parab et al (2017) utilizando Processamento de Linguagem Natural propuseram a criação de um *chatbot* capaz de auxiliar na escolha de uma carreira pelo usuário, respondendo perguntas como a melhor área de escolha, qual curso pode ser considerado como tendência, etc. Um dos métodos utilizados por eles foi a correspondência de padrões (*pattern-matching*), onde o bot irá procurar o padrão da entrada com um padrão já existente na base de conhecimento.

Junior e Netto (2014) propuseram dar mais características humanas ao seu *chatbot* educacional ao criarem aspectos e parâmetros que possam ser considerados durante a sua conversa com o seu usuário. Para isso, utilizaram o AIML juntamente com o EmotionML, batizaram-na de MindML.

6. Conclusão

Este trabalho apresenta a criação de um *chatbot* desenvolvido com o intuito de responder FAQs. Foram demonstrados os processos e passos necessários para a sua criação, utilizando a linguagem de programação Python, assim como a técnica principal o LSI (*Latent Semantic Indexing*). Mostramos como foram feitas as avaliações do *chatbot* que conseguiu dar uma resposta ao comparar as similaridades entre perguntas feitas pelos usuários e perguntas presentes no banco de perguntas.

No entanto, durante o processo de teste foi possível observar que melhorias podem realizadas dentro da aplicação. Uma delas é estender o entendimento do *chatbot* para que o mesmo consiga captar a mesma pergunta feitas de formas diferentes, utilizando palavras diferentes.

REFERÊNCIAS

Abdul-Kader, S., Woods, J. (2015) "Survey on Chatbot Design Techniques in Speech Conversation Systems", https://thesai.org/Downloads/Volume6No7/Paper_12-Survey_on_Chatbot_Design_Techniques_in_Speech_Conversation_Systems.pdf

AbuShawar, B., Atwell, E. (2015) "ALICE chatbot: trials and outputs.", http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-55462015000400625

Dale, Robert, (2016) “The return of the chatbots”, https://www.cambridge.org/core/services/aop-cambridge-core/content/view/0ACB73CB66134BFCA8C1D55D20BE6392/S1351324916000243a.pdf/return_of_the_chatbots.pdf

De Lathauwer, L., et al. "Singular value decomposition." *Proc. EUSIPCO-94, Edinburgh, Scotland, UK*. Vol. 1. 1994

Dehak, Najim, et al. "Cosine similarity scoring without score normalization techniques." *Odyssey*. 2010.

Huang, J., Zhou, M., Yang, D. (2007) "Extracting Chatbot Knowledge from Online Discussion Forums", http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-066.pdf

Jonco, C. e Silveira, S., (2015) “Hey Siri: Inteligência artificial e a humanização dos assistentes pessoais”, http://www.repositorio.jesuita.org.br/bitstream/handle/UNISINOS/6407/Camila%20Medeiros%20Jonco_.pdf?sequence=1

Junior, R., Netto, J. (2014) "Um Chatterbot Educacional Baseado em EmotionML", http://www.br-ie.org/pub/index.php/sbie/article/view/3054

Matsubara, E., Martins C., Monard, M. (2003) "PreTexT: uma ferramenta para pré-processamento de texto utilizando a abordagem bag-of-words", http://conteudo.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_209.pdf

Parab, A., Palkar, S., Maurya, S., Balpande, S. (2017), "An Intelligent Career Counselling Bot", https://www.irjet.net/archives/V4/i3/IRJET-V4I3604.pdf

Shawar, B., Atwell, E. (2007) "Chatbots: Are they Really Useful?", https://www.researchgate.net/publication/220046725_Chatbots_Are_they_Really_Useful

Silva, A. (2011) "A Influências dos Parâmetros de Análise por Semântica Latente Aplicada a Localização de Defeitos de Software", https://repositorio.ufu.br/bitstream/123456789/12505/1/InfluenciaParametrosAnalise.pdf

Souza, L., Moraes, S. (2015) "Construção automática de uma base AIML para Chatbot: um estudo baseado na extração de informações a partir de FAQS", http://www.lbd.dcc.ufmg.br/bdbcomp/servlet/Evento?id=832

Wiemer-Hastings, P. (2004) "Latent Semantic Analysis", http://reed.cs.depaul.edu/peterh/class/csc578/Papers/lsa-encyc.pdf

Wilbur, W. John, and Karl Sirotkin. "The automatic identification of stop words." *Journal of information science*18.1 (1992): 45-55.