**Nanodegree Engenheiro de Machine Learning**

**Projeto final**

Yuri Martins Campolongo

03 de abril de 2018

1. **Definição**

**Visão geral do projeto**

Uma das grandes utilizações de machine Learning nos últimos tempos, e que se mostra cada vez mais promissora é a interpretação de linguagem natural. Essa técnica se tornou muito útil e procurada nos últimos tempos devido ao grande aumento do uso de chatbots, que são robôs de atendimento que interpretam textos digitados pelo cliente, de maneira natural, para realizar a busca de possíveis soluções para aquele problema.

De acordo com grandes influenciadores do mercado de tecnologia como Mark Zuckerberg (Facebook) o autoatendimento trará mais eficácia das interações entre usuários e empresas, diminuindo o tempo de resposta e auxiliando clientes na resolução de problemas, essa definição foi dada por ele em 2016 durante um evento para desenvolvedores.

A Criação de um algoritmo para a análise de linguagem natural surgiu pois atualmente trabalho em uma companhia que cria sistemas de atendimento para empresas, e após análise da área comercial, verifiquei que as dúvidas dos possíveis novos clientes que entram em contato via chat, podem ser classificadas e posteriormente ensinadas para um algoritmo de aprendizagem, que pode também utilizar essas classificações para exibir a melhor resposta, sem a necessidade de uma pessoa para responder dúvidas frequentes e recorrentes.

**Descrição do problema**

O objetivo é criar um algoritmo de processamento de linguagem natural capaz de classificar corretamente entradas de dúvidas de clientes em um chat, esse algoritmo poderá se tornar útil para a empresa na qual eu trabalho pois trabalhamos diariamente com atendimentos e com grandes volumes, com o algoritmo funcionando e com uma taxa alta de acuracidade será possível fornecer esse aprendizado para outros sistemas que poderão responder automaticamente para os clientes, sem a necessidade de uma atendente humana respondendo questões repetitivas que consomem tempo, liberando essa atendente para atendimentos mais personalizados e específicos.

Para atingir o objetivo citado acima, foi obtido um dataset com dúvidas que possíveis futuros clientes da empresa tem, e que entram em contato utilizando o site da empresa. Cada dúvida dessa foi classificada manualmente como um tipo de dúvida relacionada a um dos produtos que a empresa fornece, com esse dataset, os seguintes passos podem ser executados para a criação de um algoritmo de linguagem natural que atenda a necessidade:

* A coluna com a dúvida do cliente passará por um pré-processamento de texto, com o objetivo de remover palavras que não são importantes para o contexto, ou seja, palavras sem significância semântica, essas palavras estarão disponíveis em um arquivo .csv que o algoritmo utilizará para removê-las dos textos em questão.
* As palavras restantes passarão por um processo chamado de TF-IDF que é uma sigla em inglês para term-frequency-inverse document frequency, ou seja, frequência do termo–inverso da frequência nos documentos, o intuito dessa técnica é indicar a importância de cada palavra contida em um documento, analisando todos os documentos disponíveis, dessa forma, as palavras com menos importância geral, serão descartadas.
* Será aplicada a técnica ‘Bag of Words’ para transformar os textos restantes em entradas numéricas para o algoritmo. Essa técnica funciona iniciando um vetor em que cada posição representa uma palavra do dataset geral, e para cada texto de entrada, será contada quantas vezes uma determinada palavra ocorre naquele texto e esse número será adicionado na posição que aquela palavra é representada no vetor, exemplo:

Textos:

|  |
| --- |
| O curso de machine learning da Udacity é muito bom! |
| Processamento de linguagem natural é muito interessante. |

Textos após tratamento:

|  |
| --- |
| CURSO MACHINE LEARNING UDACITY MUITO BOM |
| PROCESSAMENTO LINGUAGEM NATUAL MUITO INTERESSANTE |

O vetor do bag of words conterá 10 posições, pois temos ao total 10 palavras diferentes no exemplo, já que a palavra MUITO se repete 2x, assim sendo, segue a tabela de como ficará os vetores para os dois exemplos acima:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CURSO | MACHINE | LEARNING | UDACITY | MUITO | BOM | PROCE... | LINGUAGEM | NATURAL | INTERESSANTE |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 |

Com o exemplo acima é possível ver que com essa técnica é possível utilizar um algoritmo de classificação que espera entradas numéricas.

* Treinar um classificador bayesiano com os vetores de entrada.