CATEGORIZANDO PRODUTOS

Usando Machine Learning da AWS

PROBLEMA

- Produtos não associados à categorias
  - Aproximadamente 217.000 produtos
- Só encontrados através da busca do próprio site
- Não são encontrados pelo google
- Venda menor destes produtos já que não tem como navegar até eles

PROPOSTA

- Identificar o departamento e as sub-categorias a partir da base de dados atual (aproximadamente 466 Mil produtos classificados com exceção de livros)
- Apresentar uma forma de classificação automática a partir de informações básicas do produto

MACHINE LEARNING

- É o processo onde uma máquina aprende, ou seja, usa alguns algoritmos de análise de grandes massas de dados a fim de detectar regras e padrões
- Após o aprendizado é possível automatizar tomadas de decisão baseadas nos dados de treinamento.
- A AWS fornece um serviço que encapsula a complexidade deste processo, tornando-o muito mais simples. O cliente só precisa fornecer a base de dados para treinamento.

MACHINE LEARNING

Algoritmos de Aprendizado:

- O AWS ML utiliza uma técnica de otimização chamada SGD e os seguintes algoritmos de aprendizado:
  - Classificação Binária:
    - logistic regression algorithm
  - Classificação Multi-classe:
    - multinomial logistic regression algorithm

DESENVOLVIMENTO DA SOLUÇÃO

1. Mineração dos dados:
   - ➤ Decidir os campos mais relevantes
     - ➤ nome do produto, cor, descrição, marca, …
   - ➤ Gerar uma base de dados para aprendizado
     - ➤ produtos categorizados, excluindo livros

DESENVOLVIMENTO DA SOLUÇÃO

2. Geração de modelo multi-classe
   - ➤ Tentativa de descobrir 3 níveis de categoria:
     - ➤ Departamento, Linha e Classe
   - ➤ Baixa porcentagem - 67%
   - ➤ Limitação do serviço

DESENVOLVIMENTO DA SOLUÇÃO

3. Geração de modelo Binário
   - ➤ Para cada departamento existiria um modelo
   - ➤ Automotivo e Informática > 98%
   - ➤ Alguns modelos tiveram baixa precisão pois tinham poucos dados na base de aprendizado

DESENVOLVIMENTO DA SOLUÇÃO

4. Geração de modelo multi-classe (1 nível de categorização)
   - ➤ Precisão de 88%
   - ➤ Melhor tratamento dos dados para base de treinamento
     - ➤ remoção de tags html
     - ➤ remoção de stopWords (artigos, preposições)
     - ➤ adicionadas receitas:
       - ➤ n-gram de 2 na descrição
- ➤ Melhoria da precisão para 90%!

DESENVOLVIMENTO DA SOLUÇÃO

5. Geração de novo multi-classe(2º e 3º nível de categorização)

   ➤ Para cada um dos 15 departamentos com mais produtos em estoque foi gerado um modelo multi-classe

   85% - Beleza e Saúde

   85% - Instrumentos musicais

   87% - Bebês

   91% - Móveis de cozinha

   94% - Cama, Mesa e Banho

DESENVOLVIMENTO DA SOLUÇÃO

6. Modelo binário para decisão de Livros

- Foi gerado um modelo para analisar se um produto é um livro de acordo com os livros categorizados atualmente
- Precisão: 99%

DESENVOLVIMENTO DE API

Foi criada uma API para prever as categorias de um produto utilizando as seguintes informações:

nome, preço, cor, fornecedor, marca e descrição

Exemplo de resultado:


```
{
  "departId": 14,
  "departName": "BRINQUEDOS",
  "departScore": "0.9960",
  "lineDepartId": 1122,
  "lineDepartName": "PELUCIA",
  "lineDepartScore": "0.9757",
  "classDepartId": 2231,
  "classDepartName": "BICHINHOS",
  "classDepartScore": "0.9757"
}
```

CUSTOS

- Serviço de predição:
  - $ 0.015/h
  - $ 0.10 cada mil predições
- Na pesquisa foi gasto um total de pouco mais de cem dólares

RESUMO

- Existem aprox. 92 mil produtos não categorizados e que não pertencem aos nossos parceiros que vendem livros

- Estes produtos representam 13% de todos os produtos exclusivos do marketPlace

RESUMO

- Somando com livros existem 217 mil produtos (32% dos exclusivos no marketPlace)

- A grande maioria de produtos não categorizados são livros.

RESUMO

A API e os modelos criados conseguem categorizar o departamento destes 13% com taxa de acerto médio de 90% e as sub-categorias com média de 78%