UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Classificação de Anúncios da OLX

Álvaro Cândido, Ana Clara Samarino, Israel de Melo, Matheus Schimieguel

[1]Departamento de Ciência da Computação - DCC - Universidade Federal de Minas Gerais

1. Introdução

Para anunciar na OLX você precisa selecionar uma entre várias categorias disponíveis no site, mas esse processo manual pode levar com que o anunciante defina uma categoria errada para o produto. A figura 1 ilustra como é feito a seleção da categoria.

Figura 1. Menu de seleção de categoria

Diante disso, surge a necessidade de criar um meio para classificar as categorias de forma automática. Para isso, foi feito um web scraping que pega os dados de anúncios da OLX, dentre os dados, o título e o preço dos anúncios, foram utilizados para treinar um modelo de classificação utilizando o naive Bayes e Bernoulli.

2. Web Scraping

Foi feito um web scraping no site da OLX para pegar os dados dos anúncios, divulgados pelos usuários da plataforma. O funcionamento básico do scraping consiste em, inicialmente, buscar os links presentes nas divs (html) da OLX. Inicia-se a busca a partir da div class = jx24x3-2, posteriormente pegam-se os links de redirecionamento presente nessa div. Um dos links capturados, por exemplo, é referente ao ícone “imóveis”. Já dentro de imóveis, captura todos os links dentro dessa mesma área retangular. Em seguida, entra-se dentro deste link, captura-se todos os outros links dentro dessa div e entra-se neles até não encontrar mais links. Esse processo é ilustrado na Figura 2.

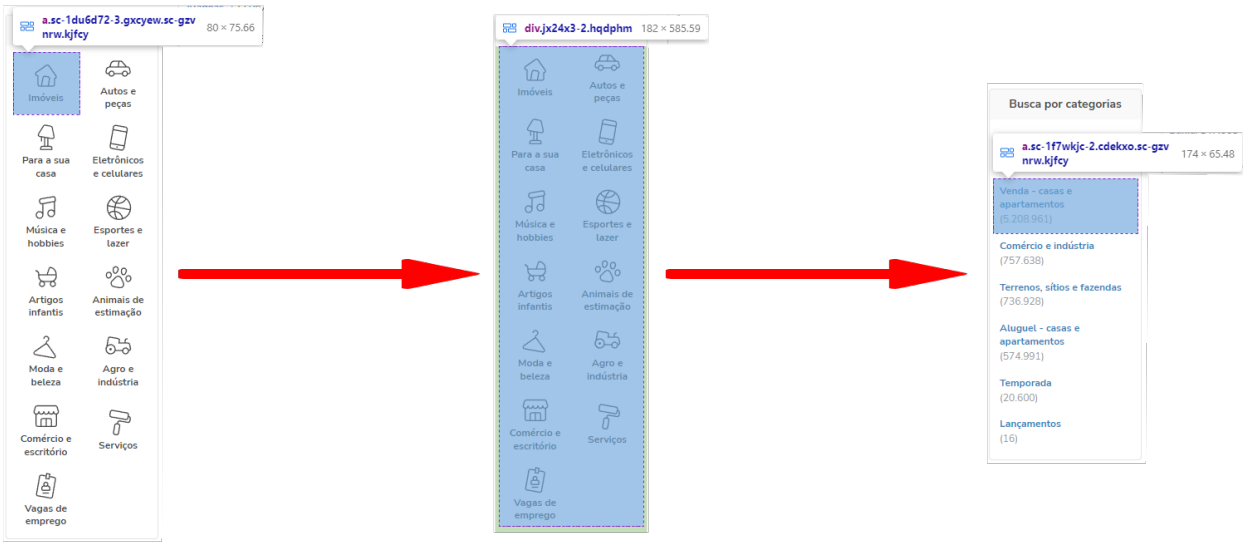

Figura 2. Processo captura de links

Com isso, é possível pegar todos os links referentes às categorias de forma recursiva. A partir da obtenção de todos esses links, verificam-se os links das categorias finais (última subcategoria possível). Por exemplo, no caso de imóveis, passamos por todas as subcategorias:

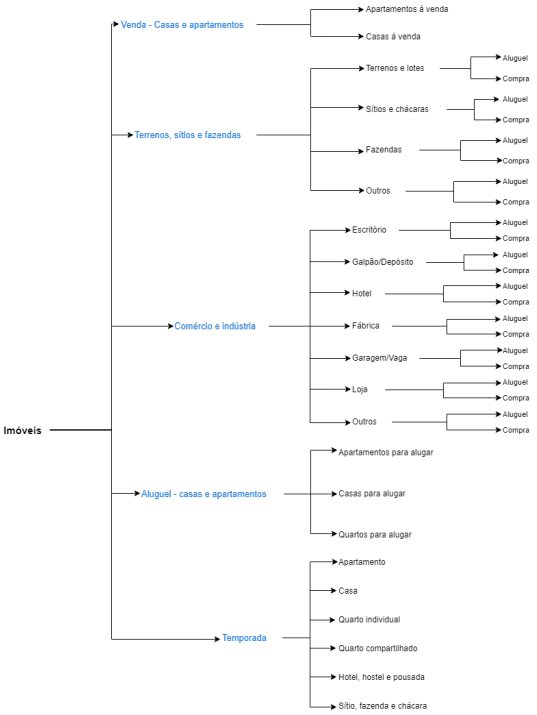

Figura 3. Subcategorias imóveis

No caso da primeira linha da imagem IMÓVEIS → VENDA CASAS E APARTAMENTOS → APARTAMENTOS A VENDA/CASAS À VENDA, obtemos os seguintes links:

https://mg.olx.com.br/imoveis
https://mg.olx.com.br/imoveis/venda
https://mg.olx.com.br/imoveis/venda/apartamentos
https://mg.olx.com.br/imoveis/venda/casas

Os dois últimos links são links de interesse, pois todos os anúncios estão classificados em alguma subcategoria final e ambos os links referem-se a subcategorias finais (nesses links não há referência a outros links).

Faz-se um processo análogo para capturar links mais específicos (links finais), relacionadas às localidades presentes no filtro de região apresentado abaixo

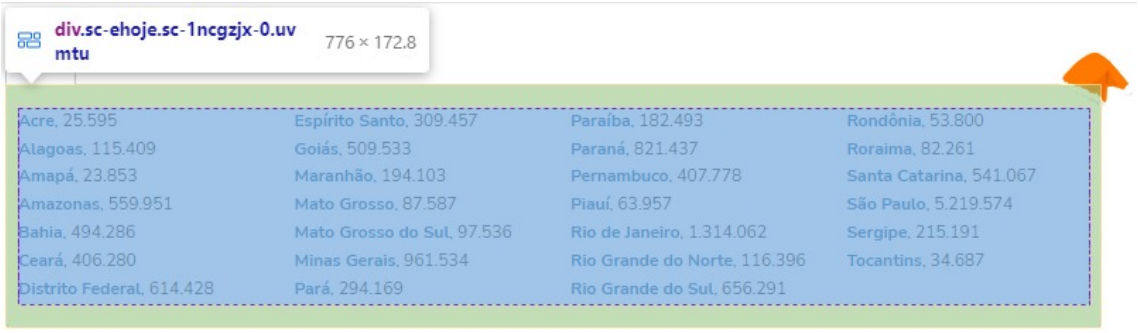

Figura 4. Links dos estados

Após capturar todos esses links, obtemos todos os links das listas de anúncios presentes no site da olx para, a partir deles, iniciar o scraping dessas listas.

No scraping das listas de anúncios foram capturados dados do título do anúncio, url do anúncio, url da lista onde se encontra o anúncio, a categoria completa a que o anúncio pertence, a categoria final a que o anúncio pertence, os detalhes do anúncio (quando existirem), o preço, a url da imagem principal, a quantidade de imagens, a localização completa do anunciante, o complemento da localização (quando existir), a classificação entre profissional ou não profissional, a data de publicação do anúncio e a data de coleta dos dados pelo scraping. Alguns desses dados é possível visualizar apenas no html, e alguns na interface do site:

Figura 5. Dados obtidos pelo scraping

3. Características da Base

3.1. Volume de Anúncios Obtidos no Scraping

O gráfico abaixo apresenta a quantidade de anúncios por categoria no scraping realizado. Somando todos os valores temos um total de 12.021.707 de anúncios com urls distintas.

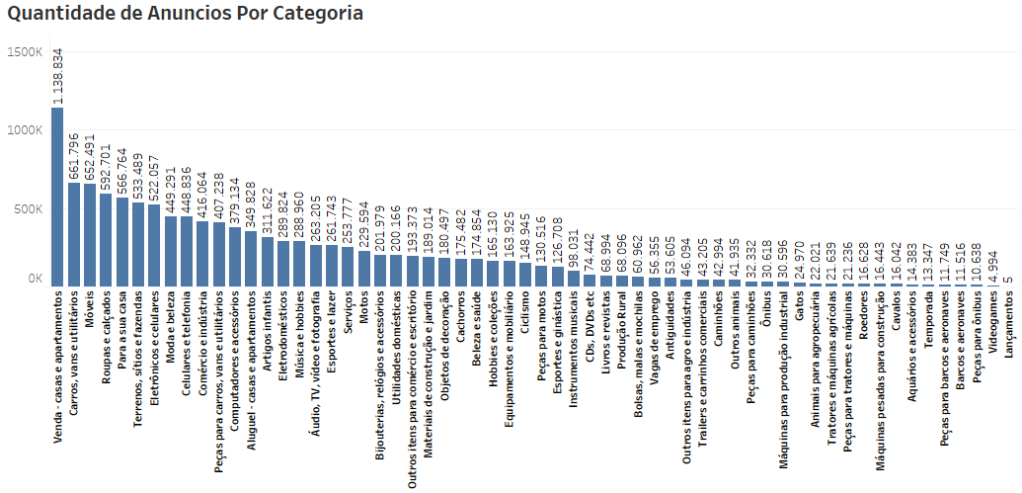

Figura 6. Quantidade de anúncios por categoria (tableau)

3.2. Média de Preço Obtidos no Scraping

O gráfico abaixo apresenta o preço médio dos anúncios por categoria no scraping realizado, sendo "Terrenos, sítios e fazendas" a categoria com o maior preço médio (R$2.450.368) e "Serviços" a categoria com o menor preço médio (R$670).

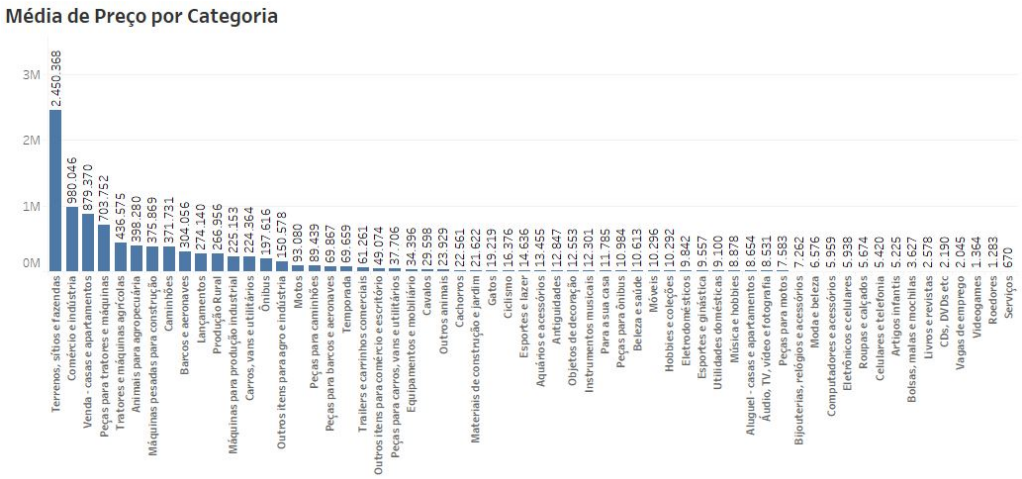

Figura 7. Preço médio por categoria (tableau)

O preço é uma variável importante para o modelo, pois pode atuar como um fator de distinção entre as categorias de aluguel de casas/apartamentos e venda de casas/apartamentos que podem possuir um título de anúncio semelhante, mas o preço é muito diferente conforme o gráfico apresentado.

4. Machine Learning

O modelo de machine learning utilizado foi o Naive Bayes.

4.1. Naive Bayes

O Naive Bayes é um método de algoritmo de aprendizado supervisionado baseado na aplicação do teorema de Bayes com a suposição "ingênua"de independência condicional entre cada par de recursos, dado o valor da variável de classe. O teorema de Bayes estabelece a seguinte relação, dada a variável de classe y e o vetor de características dependente x:

$$P(y \mid x_1, ..., x_n) = \frac{P(y)P(x_1,...,x_n|y)}{P(x_1,...,x_n)}$$

Usando a suposição de independência condicional ingênua que

$$P(x_i \mid y, x_1, ..., x_i - 1, x_i + 1, ..., x_n) = P(x_i \mid y)$$

Para todo i, essa relação é simplificada para

$$P(y \mid x_1, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1,...,x_n)}$$

Desde que $P(x_1, ..., x_n)$ é uma constante dada a entrada, podemos usar a seguinte regra de classificação:

$$\hat{y} = \text{arg}\max_y P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

4.2. Como foi feito o treinamento

No modelo foi utilizado como base para o aprendizado, o título e preço do produto anunciado pelo usuário.

Foi feito um dicionário de palavras com todas as palavras que constam na base de dados gerada pelo scraping, tendo como chave as palavras, e como valor, a quantidade de vezes que cada palavra aparece na base. Antes de popular o dicionário, foram feitos os seguintes tratamentos:

- Retiraram-se acentos, números e caracteres especiais das palavras, e
- Trataram-se espaços entre palavras (casos com mais de um espaço entre duas palavras passa a ter apenas um)

Após populado o dicionário, retirou-se palavras com 3 letras ou menos, além de todas as palavras que apareceram menos de 30 vezes em toda a base de dados. Utilizando uma biblioteca para tratamento de linguagem natural, retiraram-se também preposições e artigos, visto que palavras chaves são mais úteis para treinar o modelo.

Foi construída uma lista de todas as palavras dos títulos, outra de todas as categorias finais e uma de preços para se utilizar no modelo, uma vez que podemos utilizar os índices das listas para popular dados e construir a matriz exigida para o treinamento do modelo naive bayes.

Para o preço entrar como uma variável no modelo, ele foi classificado em índices conforme a escala logarítmica na base 10, ou seja, um produto de 1000 reais terá um índice 3. A escolha da base logarítmica está relacionada à grande variação de preços na base de dados e também ao fato de não gerar uma matriz muito grande para o modelo, já que cada índice referente ao preço será uma coluna da matriz necessária para treinar o modelo. O valor máximo do log foi limitado a 10 para descartar valores exorbitantes.

O modelo foi treinado em partes, (por categoria), devido ao grande volume de dados. A quantidade de colunas da matriz usada para treinar o modelo corresponde à soma da quantidade total de palavras (presente na lista com todas as palavras da base de dados) com a quantidade de índices de preços, gerando um total de 29.055 colunas.

Para cada categoria em treinamento, foram obtidos os índices das palavras utilizadas nos títulos, segundo a lista de todas as palavras (abordado anteriormente), com o índice da linha desse título (repetido a quantidade índices da coluna preenchidos). Ao fazer isso com todas as linhas da categoria em questão, foi utilizado uma matriz esparsa para construir a matriz de entrada para o naive bayes.

5. Teste do Modelo

A partir de um novo scraping obteve-se uma nova base de dados para testar o modelo. Nesse teste foram utilizadas 2 métricas: probabilidade de pertencimento à classe e rank médio.

A probabilidade de pertencimento à classe é fornecida pelo algoritmo após o treinamento. Foi utilizada a média dessa probabilidade como métrica para averiguar se o treinamento foi eficaz e se existe curva de separação bem definida. O rank médio faz referência a posição da categoria dado às categorias que podem ser escolhidas ao se cadastrar um novo produto na OLX.

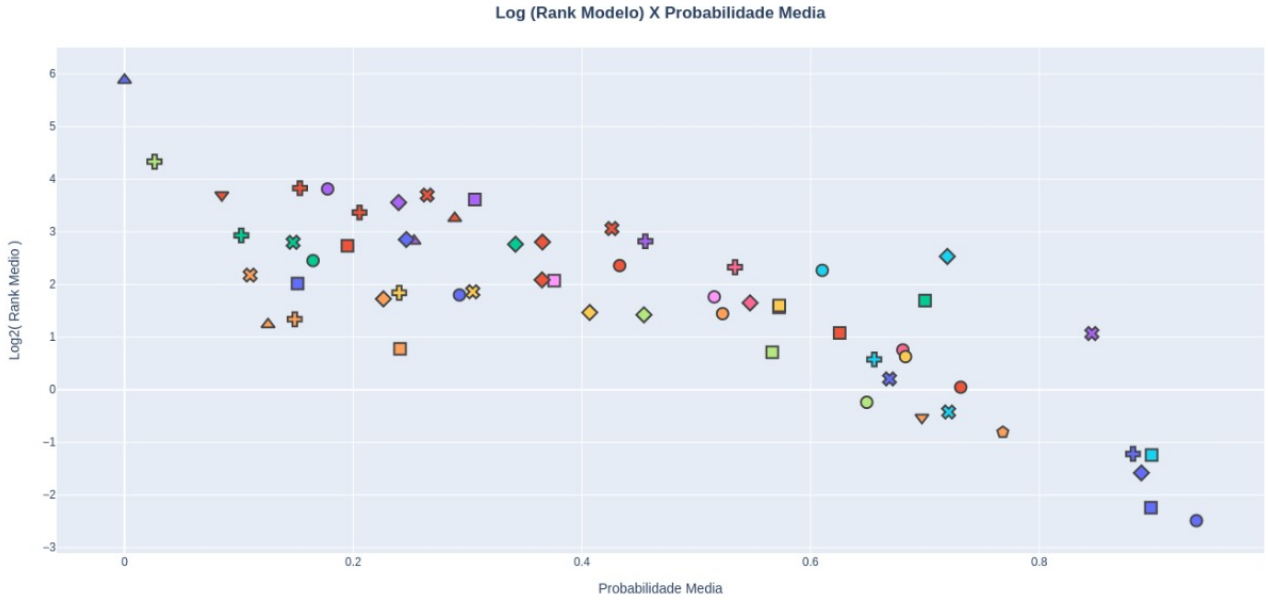

Figura 8. Plot dos resultados dos testes

Acesse a versão interativa clicando aqui’

Figura 9. Legenda

Como pode ser visto, muitas categorias apresentaram curvas de separação mal definidas. Uma explicação plausível para esse comportamento é o fato que muitos dos anúncios de categorias específicas usam palavras muito genéricas, dificultando a caracterização de determinado anúncio pelo modelo, uma vez que o modelo utilizado não usa a semântica do título. Acreditamos que uma rede neural profunda seria muito mais eficaz para esse tipo de problema dado que temos uma amostra de treinamento relevante (cerca de 13M de anúncios). Por fim, acreditamos que o modelo utilizado naive bayes é um modelo muito leve e de simples compressão, o mesmo poderia ser melhorado em nosso problema se considerássemos outras características, porém nesse modelo foi utilizado apenas o preço e as palavras do título, não foi utilizado para modelo a ordem das palavras.

Referências

[1] Naive Bayes - Scikit Learn
[2] Repositório do trabalho