Projeto Spark

Gabriella Cukier
Manuel Castanares

Descrição do projeto

Neste projeto vocês irão analisar um *dataset* grande usando o Spark (https://spark.apache.org/). O *dataset* contém vários *reviews* de produtos da Amazon (https://s3.amazonaws.com/amazon-reviews-pds/readme.html).

O *dataset* foi explorado em aula, e encontra-se em s3://megadados-alunos/dados/all_reviews_clean_tsv/. Trata-se de um dataset no formato texto, onde as colunas estão separadas pelo caractere '\t' (*tab-separated value*).

Para saber o significado de cada coluna, consulte https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt .

Tarefa 1

Quantos reviews existem?

- Existem **150962278** reviews

Quantos clientes existem?

- Existem **33497620** clientes

Quantos produtos existem?

- Existem **21390118** produtos

Quantos reviews existem para cada "star_rating" (de 1 a 5 estrelas)?

- * **12099639**
- ** **7304430**
- *** **12133927**
- **** **26223470**
- ***** **93200812**

Tarefa 2

Analise os *bots*: certos clientes tem um número extremamente alto de reviews – provavelmente são *bots*. Construa uma caracterização dos bots: para qual tipo de produto eles fazem *review*, se eles são sempre positivos ou podem ser mais negativos, quantos são (como definir?), etc.

Como definir os Bots?

Para a definição e filtragem dos reviews que foram feitos por bots, usamos a função histogram do pyspark. Essa função recebe como argumento um número, correspondente à quantidade de buckets (intervalos para separação) e retorna uma tupla dos intervalos dos buckets e da quantidade de valores em cada um. No caso, foi utilizada uma separação em 100 intervalos.

Percebemos pelos resultados que a grande maioria dos clientes (33495417) está no primeiro intervalo, de valores menores que 527. Este valor pareceu condizente com uma quantidade alta o suficiente de reviews para ser pertencente a um bot. Desta forma, escolhemos este limiar para a determinação de bots.

Quantos bots existem?

Com esta definição acima, foi estabelecido que existem 2203 bots ao total.

Para qual tipo de produto eles fazem *review?*

Os produtos com mais reviews de bots são livros físicos, seguidos de ebooks.

Vídeos e músicas também tiveram uma quantidade de reviews bastante significativa.

Os produtos com menos reviews foram Softwares e Gift Cards

A quantidade de reviews por cada categoria é listada abaixo, juntamente com um gráfico de distribuição das mesmas.

('Books', 945457),

('Digital_Ebook_Purchase', 420044),

('Video DVD', 275148),

('Music', 258477),

('Toys', 42652),

('Beauty', 41400),

('Health & Personal Care', 40694),

('Home', 40086), ('Grocery', 39597),

('Kitchen', 38719),

('PC', 34630),

('Wireless', 31585),

('Video', 29282),

('Digital_Music_Purchase', 26950),

('Digital_Video_Download', 25225),

('Office Products', 22458),

('Electronics', 18975),

('Sports', 18440),

('Video Games', 17897),

('Apparel', 17373),

('Home Improvement', 16530),

('Pet Products', 15877),

('Lawn and Garden', 15718),

('Mobile_Apps', 13940),

('Camera', 10894),

('Outdoors', 9829),

('Tools', 9112),

('Automotive', 8500),

('Shoes', 8304),

('Jewelry', 7825),

('Baby', 7819),

('Software', 5777),

('Musical Instruments', 4004),

('Home Entertainment', 3494),

('Watches', 2941),

('Furniture', 2787),

('Luggage', 1781),

('Digital_Video_Games', 901),

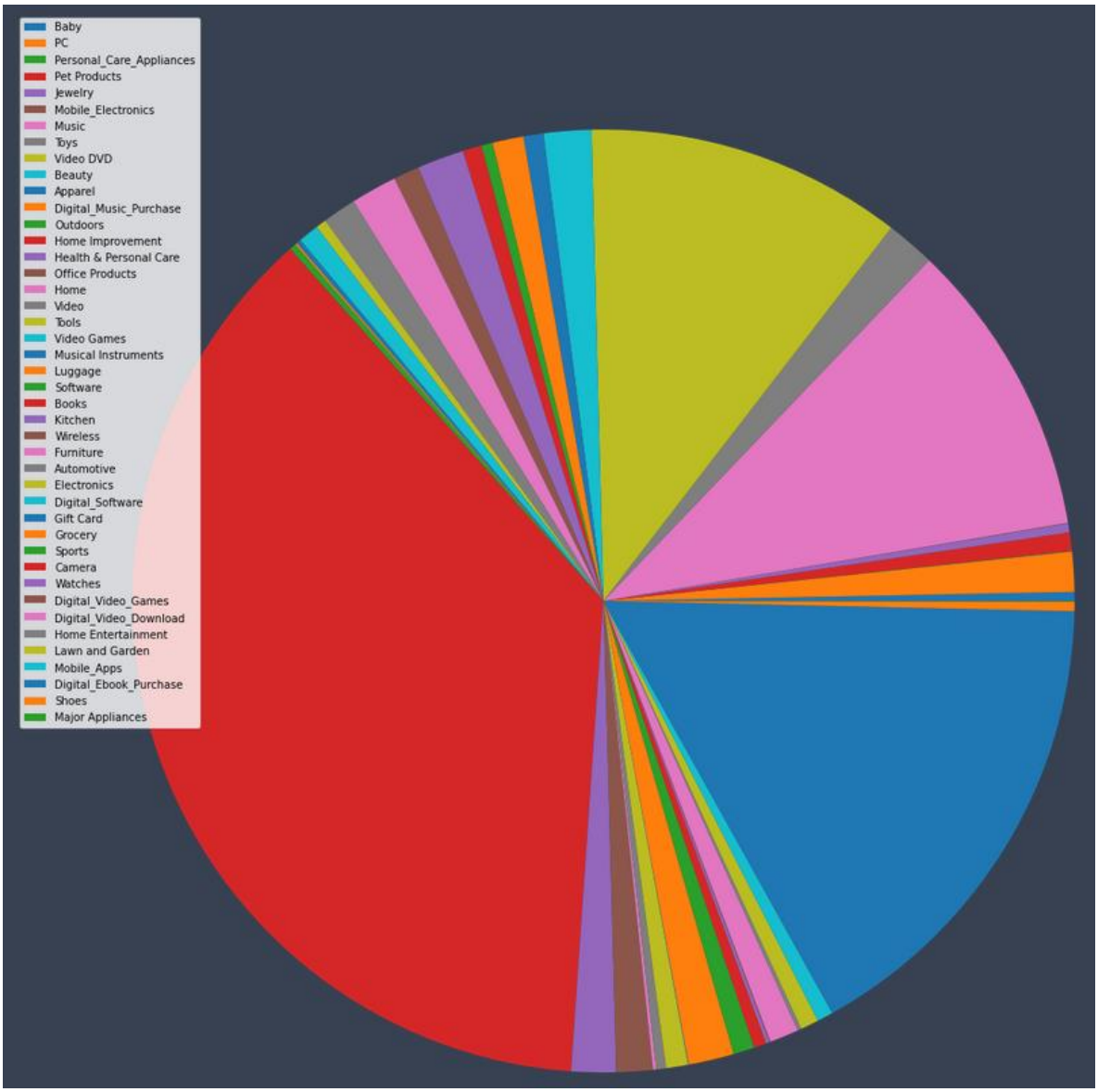
('Personal_Care_Appliances', 603),

('Mobile_Electronics', 452),

('Major Appliances', 427),

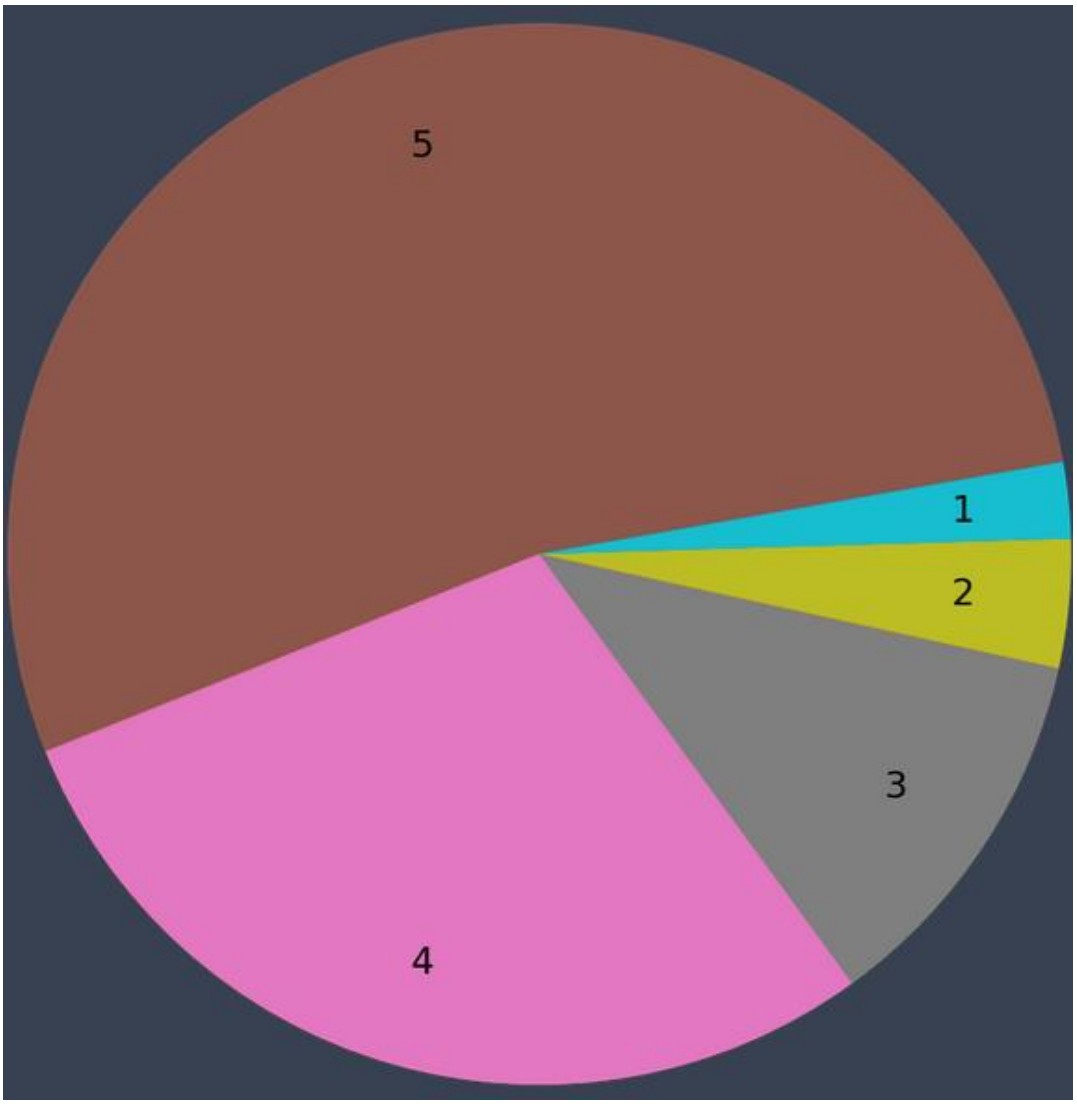
('Gift Card', 305),

('Digital_Software', 255)]



Os reviews são em sua maioria positivos ou negativos?

[('5', 1349931), ('4', 733239), ('3', 291908), ('2', 99453), ('1', 58633)]



Como mostra o gráfico, a grande maioria dos reviews é positiva (4 ou 5 estrelas). Isso mostra que os Bots possuem como prioridade influenciar o cliente a comprar determinado produto. Desta forma, provavelmente os poucos reviews negativos se referem aos produtos concorrentes daqueles em que se quer promover.

Os reviews dos bots recebem votes? Se sim, muitos ou poucos?

A porcentagem de reviews que receberam ao menos um vote (considerando apenas reviews de bots) foi de aproximadamente 70.33%. Destes, cerca de 19.3% tiveram muitos votes (mais de 10).

Uma possível explicação é que os próprios bots podem dar votes em reviews com características semelhantes, visando a aumentar a visibilidade do review, assim influenciando mais usuários a visualizarem o produto.

Os votes dos bots são classificados como úteis?

Cerca de 91,97% dos votes de reviews de bots foram classificados como não sendo úteis. Isso pode ser interpretado como uma comprovação de que vários dos votes de reviews são feitos por bots, na medida que outros clientes provavelmente não acham os reviews de bots pertinentes.

Hipótese e considerações finais

A grande maioria dos bots deixou reviews positivos, estimulando os consumidores a comprar livros, ebooks, músicas e filmes. Como estes produtos possuem um caráter mais subjetivo referente à qualidade, é esperado que para eles uma quantidade alta de reviews positivos estimule a compra de forma significativa. Os votes também podem ser usados para dar mais visibilidade à mercadoria, mesmo não sendo em sua grande maioria classificados como úteis. Desta forma, os bots podem de fato influenciar positivamente a venda deste tipo de produto.

Tarefa 3

Vamos definir que avaliações 5-estrelas são positivas, 4-estrelas são neutras, e 3-estrelas ou menos são negativas. Construa um classificador *naive-Bayes* que determina se um review é positivo, neutro ou negativo. Um exemplo de classificador *naive-Bayes* com Spark pode ser lido em https://ai.plainenglish.io/build-naive-bayes-spam-classifier-on-pyspark-58aa3352e244 (nota: usa API DataFrames do Spark).

Metodologia

Para a construção do classificador Naive Bayes foram seguidos os seguintes passos:

1. Preparação dos dados
    a. Criação de um DataFrame com as informações do dataset de reviews e criação de uma nova coluna indicativa se o review é bom (B), neutro (N) ou ruim (R)

```
DataFrame[_c0: string, _c1: int, _c2: string, _c3: string, _c4: int, _c5: string, _c6: string, _c7: string, _c8: int, _c9: int, _c10: string,
_c11: string, _c12: string, _c13: string, _c14: timestamp, label_string: string]
```

2. Construção dos estágios de Pipeline

```
%pyspark
stages
```

```
[RegexTokenizer_225834c06b09, CountVectorizer_90a31effd386, StringIndexer_8bb18f1eec49, VectorAssembler_54e0ba53a94d]
```

    a. RegexTokenizer: limpeza de strings (pontuação, maiúsculas, etc.)
    b. CountVectoriser: conversão de texto em vetores
    c. Conversão para label numérica
    d. VectorAssembler: Concatena os resultados em um vetor
3. Fit do Pipeline para transformação de dados
4. Separação do dataset em treino e teste
5. Implementação do Naive Bayes

```
+-----+----------+--------------------+
|label|prediction|         probability|
+-----+----------+--------------------+
|  0.0|       0.0|[0.86108740449468...|
|  1.0|       1.0|[2.35419578545685...|
|  2.0|       0.0|[0.77663674346206...|
|  0.0|       0.0|[0.83367000786591...|
|  0.0|       0.0|[0.99996697224669...|
|  0.0|       0.0|[0.98251932318851...|
|  2.0|       0.0|[0.94946237552155...|
|  0.0|       2.0|[1.85935385668654...|
|  1.0|       1.0|[0.31916969294290...|
|  0.0|       0.0|[0.96665214947837...|
|  0.0|       0.0|[0.64850112306616...|
|  0.0|       0.0|[0.98839930141158...|
|  0.0|       0.0|[0.88775837309154...|
|  0.0|       0.0|[0.82234781517899...|
```

6. Smoothing e cross-validation para melhorar os resultados

Resultados

O classificador teve uma acurácia inicial de aproximadamente 74.17%

Após a implementação de smoothing e cross-validation, a acurácia continuou a mesma, 74.17849245888665%, com o grid com os valores 1 e 2

Foi feita também uma análise do modelo, por meio de uma matriz de confusão, em que a diagonal principal mostra is acertos por cada categoria. Os outros valores indicam as atribuições que não corresponderam ao esperado. As colunas e linhas estão na ordem (0,1,2) ou seja, bom, neutro e ruim.

```
DenseMatrix([[23597703.,  2477489.,  1873470.],
             [ 1674461.,  6777690.,  1007904.],
             [ 4238409.,  1893017.,  1735840.]])
```