Programa Atlântico Academy Future

Ciência de Dados

Grupo 01 - Data DiscoveryOne

Trabalho Prático

Entrega 1/3

Professor: Madson Luiz Dantas Dias

Alunos:

- Arildo Alves
- Germano Fenner
- Maurício Moura
- Nator Júnior
- Rodolfo Ricardo

Fortaleza, CE, 05 de novembro de 2021

Análise de Compras da Olist

A Olist é uma startup brasileira que atua no segmento de vendas pela internet (e-commerce). Através da sua plataforma de serviços, ela é um mediador e possibilita que outras empresas se inscrevem e vendam seus produtos. Em e-commerce, isso é conhecido como *marketplace,* ou seja, trata-se de uma loja virtual (a Olist) em que o cliente acessa a um site e compra produtos de diversos varejistas, pagando tudo junto, em um único lugar. Sendo assim, a Olist oferta diversas vitrines, como e fosse um *shopping*, vendendo os mais diversos itens.

Objetivos e resultados chave

O *dataset* da Olist, conjunto de dados, concentra diversas informações relacionadas as compras e vendas realizadas por meio da sua plataforma de serviços.

Através dos *reviews*, análise crítica de dados, do *dataset* da Olist em relação as compras e vendas realizadas, este trabalho tem os seguintes objetivos:

1. Obter os *reviews* das compras e vendas realizadas;
2. Identificar variáveis e suas influências nos resultados;
3. Classificar os *reviews* por análise de sentidos por meio da técnica de processamento de linguagem natural (PLN);
4. Por meio da análise dos resultados gerados, buscar compreender como é a experiência da compra/uso dos serviços da Olist.

O resultado pretendido por esse trabalho busca conhecer/compreender a experiência dos usuários após utilizarem os serviços da Olist considerando:

- Satisfação de uso dos serviços disponibilizados na plataforma (boa, ruim);
- Quantitativo e percentual de grupos de respostas;
- Orientação para tomada de com base nos resultados gerados.

Conteúdo

O repositório está organizado seguindo a ideia básica do *git-flow*, sendo assim, algumas *branchs* estão disponíveis, sendo elas: a *main*, onde estão as versões estáveis do código, a *develop*, que é a linha do tempo principal de desenvolvimento, ou seja, as novas *features* deverão ser incluídas nela. Por fim, todas as *features*/atividades são realizadas em uma *branch* separada e assim que for finalizada é mesclada na *develop*.

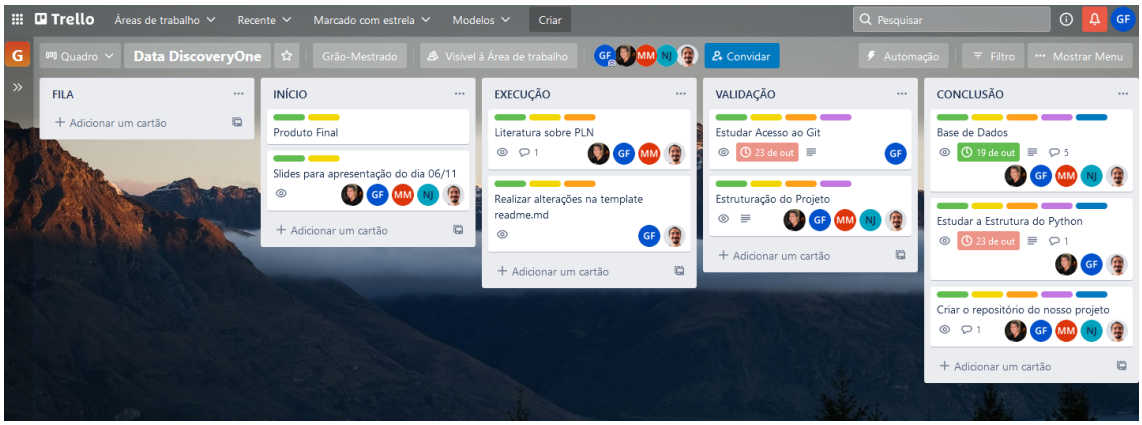
Neste sentido, no diretório notebook na *branch main*, está o arquivo referente a primeira entrega, com a exploração de dados inicial, que contem uma avaliação primaria, sobre a quantidade de dados disponíveis e as condições iniciais, está incluído também uma versão inicial da classificação, usando *Gaussian Naive Bayes* e transformação dos dados usando *bag of words.*

Utilização

Os comandos necessários para a execução do *jupyter* são descritos abaixo, os mesmos estão inicialmente disponibilizados via o gerenciador de pacotes PIP:

- pip install wordcloud;
- pip install pandas;
- pip install tqdm;
- pip install sklearn;
- pip install matplotlib;
- pip install plotly.

Organização do Projeto



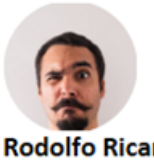
Desenvolvedores

Arildo
Celular: (61) 9 8189-5600
E-mail: arildoalves@gmail.com

Nator Júnior
Celular: (88) 9 9422-0929
E-mail: natorjuniorccc@gmail.com

Germano Fenner
Celular: (85) 9 9959-5900
E-mail: germanofenner@gmail.com

Maurício Moura
Celular: (85) 9 9936-0320
E-mail: mouramauricio99@gmail.com



Rodolfo Ricardo
Celular: (85) 9 9612-6313
E-mail: rodolforicardotech@gmail.com

Organização de diretórios

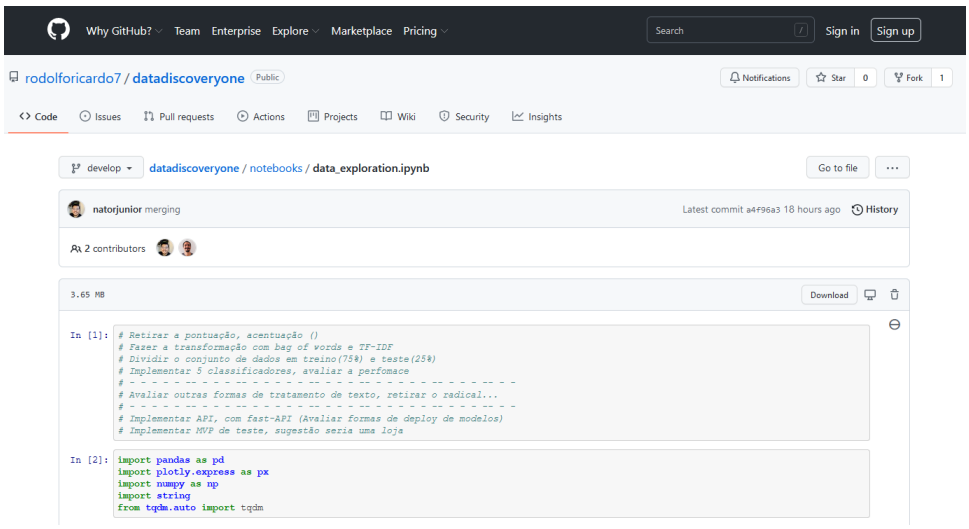
README.md

Project Organization

```
├── LICENSE
├── Makefile           <- Mal
├── README.md          <- Th
├── data
│   ├── external       <- Da
│   ├── interim        <- In
│   ├── processed      <- Th
│   └── raw            <- Th
│
├── docs               <- A
│
├── models             <- Tr
│
├── notebooks          <- Ju
│                         th
│                         `1
│
├── references         <- Da
│
├── reports            <- Ge
```

README.md

```
│   └── figures        <- Ge
│
├── requirements.txt   <- Th
│                         ge
│
├── setup.py           <- mal
├── src                <- So
│   ├── __init__.py    <- Mal
│   │
│   ├── data           <- Sc
│   │   └── make_dataset.py
│   │
│   ├── features       <- Sc
│   │   └── build_features.p
│   │
│   ├── models         <- Sc
│   │   │                 pr
│   │   ├── predict_model.py
│   │   └── train_model.py
│   │
│   └── visualization  <- Sc
│       └── visualize.py
│
└── tox.ini            <- to
```



In [2]:
```
import pandas as pd
import plotly.express as px
import numpy as np
import string
from tqdm.auto import tqdm
```

PLN Transformation

- bag of words
- TF-IDF

In [3]:
```
df = pd.read_csv('../data/raw/olist_order_reviews_dataset.csv')
```

In [4]:
```
df.head()
```

Out[4]:

	review_id	order_id	review_score	review_comment_title	review_comment_message	review_c
0	7bc2406110b926393aa56f80a40eba40	73fc7af87114b39712e6da79b0a377eb	4	NaN	NaN	2018-01-1
1	80e641a11e56f04c1ad469d5645fdfde	a548910a1c6147796b98fdf73dbeba33	5	NaN	NaN	2018-03-1
2	228ce5500dc1d8e020d8d1322874b6f0	f9e4b658b201a9f2ecdecbb34bed034b	5	NaN	NaN	2018-02-1
3	e64fb393e7b32834bb789ff8bb30750e	658677c97b385a9be170737859d3511b	5	NaN	Recebi bem antes do prazo estipulado.	2017-04-2
4	f7c4243c7fe1938f181bec41a392bdeb	8e6bfb81e283fa7e4f11123a3fb894f1	5	NaN	Parabéns lojas lannister adorei comprar pela I...	2018-03-0

In [5]:
```
df['review_score'].value_counts()
```

Out[5]:
```
5    57328
4    19142
1    11424
3     8179
2     3151
Name: review_score, dtype: int64
```