**Projeto TOK&STOK**

**Manual de Utilização**

**Código para Análise de Reclamações**

****

****

**Consultores:**

Esthevão Marttioly

Thiago Oliveira

**Coordenador:**

Eduardo Balint

**Sumário**

[**Introdução**](#_cw690qvc6ioc) **3**

**Utilização do Software 3**

**Importação da base de dados 4**

**Análise de Palavras 5**

**Análises por setores 6**

**Conclusão 7**

# **Introdução**

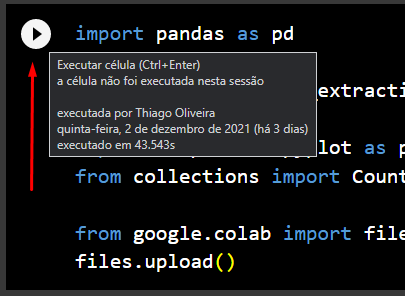
O objetivo deste documento é explicar como os funcionários da TOK&STOK poderão, em ocasiões futuras, utilizar o código em Python realizado pelos consultores da EJFGV para analisar as palavras mais repetidas nas reclamações dos clientes da TOK&STOK.

# 

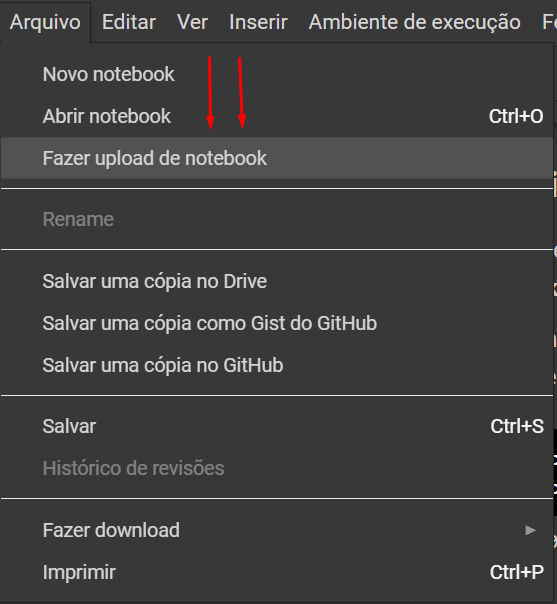
# **Utilização do *Software***

O Google Colab permite a utilização do Python no navegador de modo fácil e compartilhado.

Esse *software* pode ser encontrado por meio do *link:* <https://colab.research.google.com/notebooks/welcome.ipynb?hl=pt-BR&authuser=0&pli=1> ou da pesquisa pelo termo “*Google Colaboratory*” no próprio google ou alguma outra plataforma de busca similar.



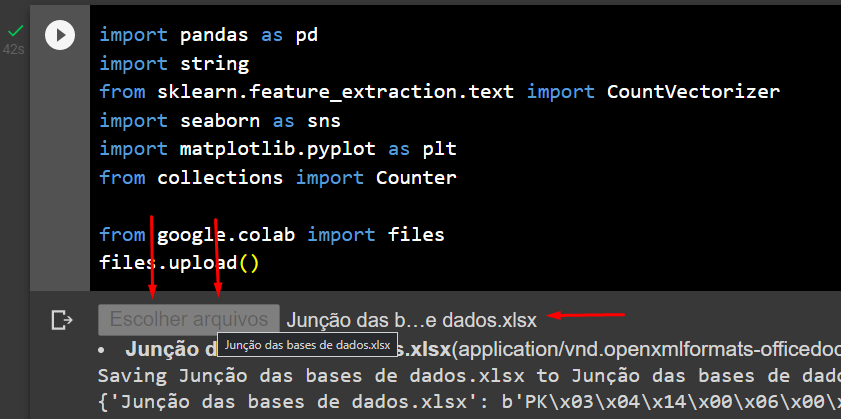
Após entrar no Colab, a primeira coisa que deverá ser feita é a importação do arquivo. Para isso, basta clicar em “Arquivo” e, depois, “Fazer upload de notebook”. Após aparecer o painel de seleção, basta selecionar o arquivo disponibilizado em .ipynb, chamado “Análise de Palavras - TOK&STOK”.



Dessa forma, aparecerá todos os códigos que foram utilizados ao longo das análise.

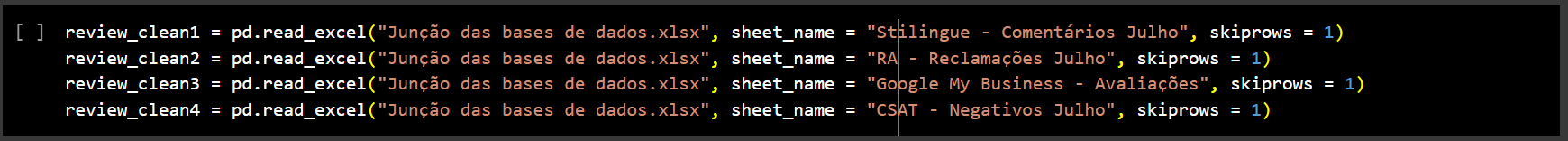
# **Importação da base de dados**

Dentro do arquivo no *Colab*, o primeiro código utilizado é destinado à importação da base de dados (no caso, Planilha de Excel). Caso desejem “importar” alguma planilha, basta executar o primeiro código e escolher o arquivo desejado logo em seguida. Aqui, foi utilizado o arquivo em Excel “Junção das bases de dados”, disponibilizado para a TOK&STOK, mas pode ser importado quaisquer bases desejadas. É evidente que, em caso de outras bases, alguns pequenos ajustes ao código deverão ser feitos, como a alteração do nome do arquivo que se deseja importar.



No segundo código, são importadas as abas dentro do arquivo em Excel, já que em uma mesma planilha podem conter mais de uma aba. Dessa forma, caso o nome do arquivo ou o nome das abas do Excel mudem, é necessário alterar seus nomes para conseguir importá-las corretamente. O nome do arquivo deve ficar no lugar de “Junção das bases de dados.xlsx”, mantendo as aspas e a expressão “.xlsx”. Ressalta-se que o nome das abas deve estar depois de “sheet\_name =”.

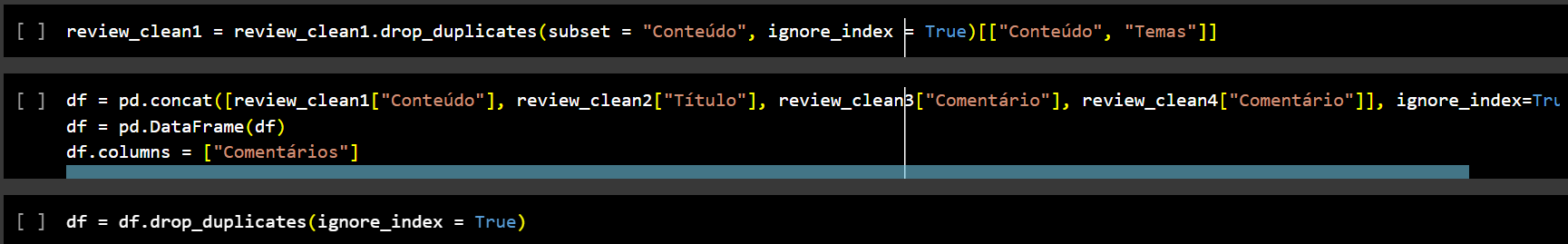
Caso desejem importar mais de uma planilha, é necessário criar um novo código e escrever “files.upload()”, igual foi feito na foto anterior. E, caso haja mais abas a serem importadas, deve-se copiar as linhas do código abaixo, chamando a variável (nome à esquerda da igualdade) de “review\_clean 5, 6, 7, …” e mudar o nome da aba em “sheet\_name”.



Caso o código acima não funcione, deve-se rodar o código abaixo dele, porém retirando as hashtags (#). Nas linhas abaixo, não há muito o que se preocupar, já que a única coisa importante é acrescentar dentro de pd.concat([]) as abas que foram acrescentadas. Por exemplo, caso haja um “review\_clean5”:

pd.concat([review\_clean1["Conteúdo"], review\_clean2["Título"], review\_clean3["Comentário"], review\_clean4["Comentário"], review\_clean5[“Conteúdo”]], ignore\_index=True)

Na hipótese do nome das variáveis em cada aba do Excel seja diferente, também é necessário mudar o nome da respectiva aba dentro das aspas de cada “review\_clean”.



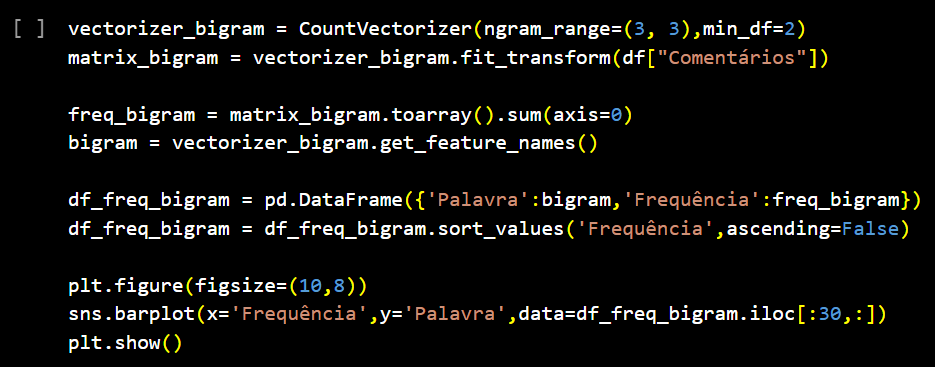
Por fim, vale ressaltar que o comando de df.drop\_duplicates() remove todos os comentários repetidos da base de dados. Se a TOK&STOK não quiser remover a repetição, basta deletar o que contém nele (terceira linha da foto anterior).

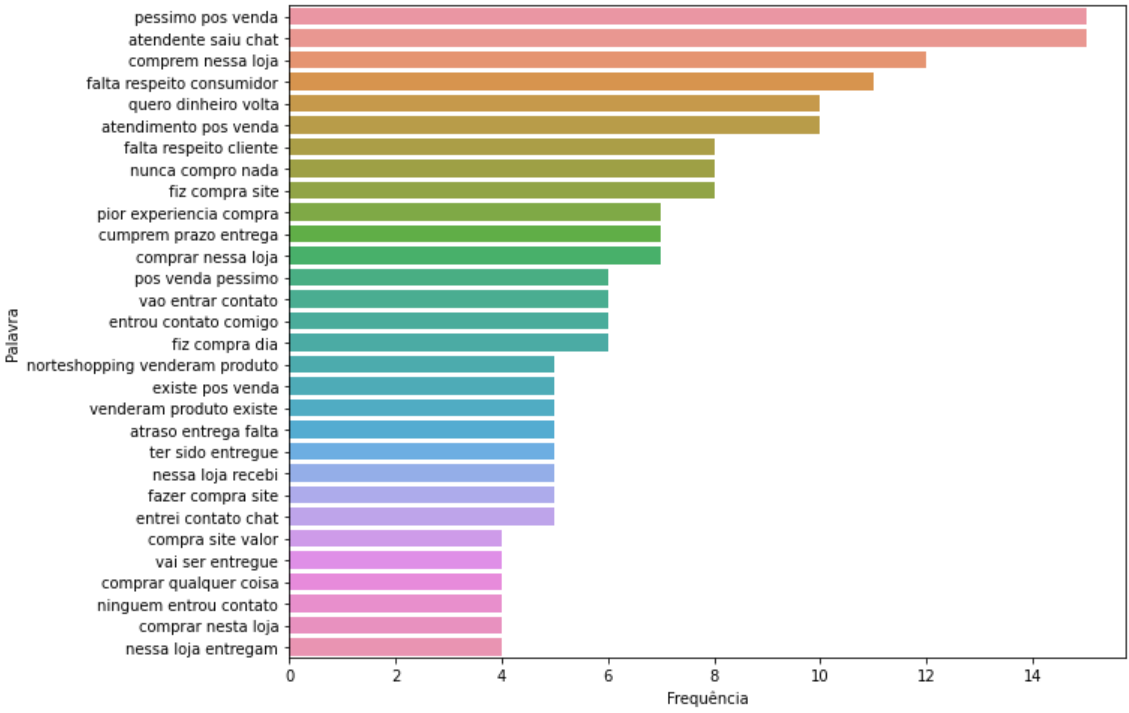
**Análise de Palavras**

Na análise de palavras, o código é mais intuitivo e não são necessárias mudanças nas variáveis, pois seus nomes já foram ajustados anteriormente. Porém, a fim de mostrar melhor o que está acontecendo, esse manual também busca evidenciar o que o código está extraindo para a análise.

Nos primeiros códigos, todas as palavras são colocadas em minúsculo, e são removidos os emojis, as pontuações e as acentuações. Depois disso, também são retiradas as chamadas “stopwords”, que são palavras da língua portuguesa que não têm significado intrínseco, como preposições, artigos, conjunções, etc. Alguns exemplos são: “que”, “é”, “e”. Também foram removidas as palavras “tok”, “stok”, “tokstock” e “tokstok”.

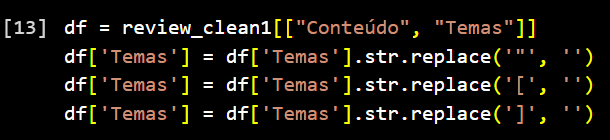
Além disso, no código anterior aos gráficos, há uma pequena diferença usada para analisar uma, duas ou três palavras juntas. Essa distinção pode ser visualizada no código "n\_gram\_range = (2,2)" e "n\_gram\_range = (3,3"). Quando utilizados o primeiro, a análise é de duas palavras, enquanto quando aplicamos o segundo, a análise é de três palavras juntas. Cabe ressaltar, ainda, que se pode analisar quantas palavras juntas quiser, apenas substituindo o "2,2" pelo número desejado. Ou seja, se almejar analisar 5 palavras juntas, basta alterar o código para "n\_gram\_range = (5,5)".



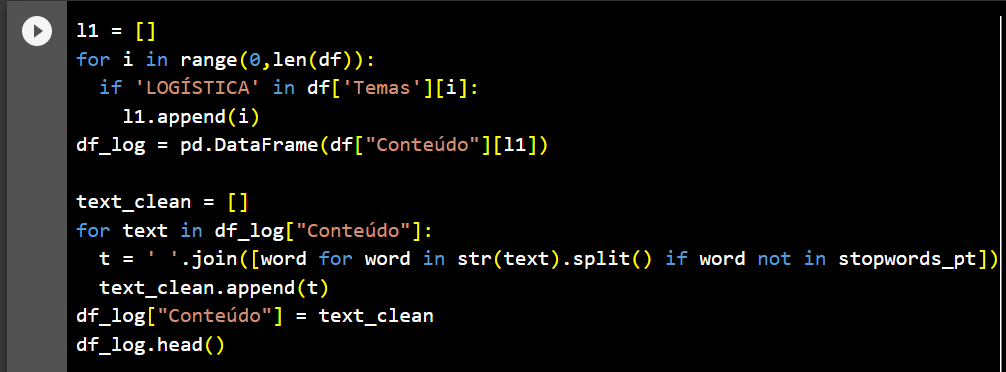


**Análise por setores**

Os códigos para análise por setores são bem parecidos com aqueles usados na análise das palavras. A principal diferença reside no fato de ser necessário realizar uma “limpeza” dos dados iniciais. Para isso, foi utilizado o “review\_clean1”, que corresponde à base Stilingue, extraídas as colunas de “Conteúdo” e “Temas” (esses nomes podem mudar se as bases tiverem outros nomes) e depois retiradas as expressões dispensáveis das células de “Temas” (como [, ] e “).



Depois disso, também são retiradas as stopwords do texto, conforme código abaixo:



Porém, vale ressaltar que essa análise pode ser feita para outros setores da TOK&STOK (basta apenas mudar no código acima a expressão ‘LOGÍSTICA’ pelo setor a ser analisado, em maiúsculo e entre aspas). Essa mesma análise pode ser feita para outras variáveis que não são “Temas”, já que bastaria apenas mudar todos os códigos que têm “Temas” pela variável desejada e depois mudar o código acima para “if ‘NOME’ in df[‘Variável’][i]:” na linha correspondente, sendo “NOME” um dos valores presentes na nova variáveis e “Variável” o nome da variável desejada.

**Conclusão**

Desse modo, o código utilizado pode ser reciclado pela TOK&STOK e, caso deseje, também é possível alterar alguns detalhes de forma que a análise permita novas mudanças. Vale ressaltar que os dados utilizados na análise dos consultores foram aqueles presentes no PowerPoint enviado.