Notebook Jupyter 3_classificarAplicação

Classificação dos modelos de motocicleta a partir da descrição

A grande dificuldade na tarefa de análise de valores compatíveis na importação de peças de motocicletas, em especial dos kits de transmissão, se dá no fato de que milhares de importadores adquirem essas peças no exterior e informam sua descrição em um campo texto livre.
Nem mesmo a utilização da classificação fiscal normatizada no Mercosul, chamada de Nomenclatura Comum do Mercosul – NCM, ajuda nesse caso específico, tendo em vista que grande parte das peças de motocicletas e todos os kits de transmissão são classificados em uma mesma posição na tabela da NCM.
Para que se possa tratar corretamente o dataset obtido na nossa etapa de processamento e tratamento de dados, e permitir o futuro aprendizado de máquina, com predições do modelo de motocicleta que aquele item se aplica, é preciso que primeiro se proceda a uma classificação de aplicações que futuramente será utilizado em um aprendizado supervisionado.
A ideia é se aplicar uma busca na descrição da mercadoria pelos termos conhecidos de aplicações e se buscar a qual aplicação aquela descrição se refere, fazendo desse modo a primeira classificação.
Posteriormente será utilizado um algorítimo de aprendizado de máquina para aprender com o próprio texto da descrição da aplicação e fazer a classificação utilizando a descrição já limpa de stopwords e outros termos desnecessários.
A interseção dos dois conjuntos de classificação será o dataset utilizado para fazer o treinamento do classificador, que será o primeiro passo antes da análise de valor do item importado.

Importa as bibliotecas necessárias

In [1]:

```
import pandas as pd, numpy as np
import re, time
# stopwords
from nltk.corpus import stopwords
# wordcloud
from wordcloud import WordCloud
# plotagem do gráfico
import matplotlib.pyplot as plt
```

In [2]:

```
# Data e hora da execução do script
initot=time.time()
print(f'Código executado em {time.strftime("%d/%m/%Y às %H:%M", time.localt
ime(time.time()))}')
```

```
Código executado em 20/01/2022 às 16:36
```

Importa os dados já tratados

In [3]:

```
# Importa base de dados para um dataframe
df = pd.read_excel(r'./bases/dataframe.xlsx')
```

In [4]:

```
# Verifica o tamnanho do dataframe
df.shape
```

Out[4]:

```
(18276, 3)
```

In [5]:

```
# Mostra linhas de exemplo do dataframe
df.sample(5)
```

Out[5]:

	PAIS DE ORIGEM	DESCRICAO DO PRODUTO	VALOR UN.PROD.DOLAR
5908	CHINA, REPUBLICA POP	10530002 IN KIT TRANSMISSAO P/MOTOCICLETAS(COR...	3.720000
12613	CHINA, REPUBLICA POP	KIT DE TRANSMISSÃO PARA MOTOCICLETAS, COMPOSTO...	3.360000
12836	CHINA, REPUBLICA POP	881606 - KIT DE TRANSMISSAO, COMPOSTO DE CORRE...	5.024464
14707	CHINA, REPUBLICA POP	152624 # KIT TRANSMISSAO STANDARD TEMP. COMPL....	3.197229
2151	CHINA, REPUBLICA POP	71848 - KIT NXR 150 BROS ESD (03-05) 50Z X 17Z...	5.757700

In [6]:

```
df['DESCRICAO DO PRODUTO'][5]
```

Out[6]:

```
'80348 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE, COROA E PINHÃ
O PARA MOTOCICLETA CBX 250 TWISTER, MARCA ALLEN.'
```

In [7]:

```
type(df['DESCRICAO DO PRODUTO'][1])
```

Out[7]:

```
str
```

Importa as stopwords da língua portuguesa

In [8]:

```
# Importar lista de Stopwords
stopwords = set(stopwords.words('portuguese'))
```

In [9]:

```
# Mostra tamanho da lista de stopwords
len(stopwords)
```

Out[9]:

```
204
```

In [10]:

```
# Mostra toda a lista de stopwords
swtemp = list(stopwords)
swtemp.sort()
print(swtemp)
```

```
['a', 'ao', 'aos', 'aquela', 'aquelas', 'aquele', 'aqueles',
'aquilo', 'as', 'até', 'com', 'como', 'da', 'das', 'de', 'del
a', 'delas', 'dele', 'deles', 'depois', 'do', 'dos', 'e', 'el
a', 'elas', 'ele', 'eles', 'em', 'entre', 'era', 'eram', 'ess
a', 'essas', 'esse', 'esses', 'esta', 'estamos', 'estas', 'est
ava', 'estavam', 'este', 'esteja', 'estejam', 'estejamos', 'es
tes', 'esteve', 'estive', 'estivemos', 'estiver', 'estivera',
'estiveram', 'estiverem', 'estivermos', 'estivesse', 'estivess
em', 'estivéramos', 'estivéssemos', 'estou', 'está', 'estávamo
s', 'estão', 'eu', 'foi', 'fomos', 'for', 'fora', 'foram', 'fo
rem', 'formos', 'fosse', 'fossem', 'fui', 'fôramos', 'fôssemo
s', 'haja', 'hajam', 'hajamos', 'havemos', 'hei', 'houve', 'ho
uvemos', 'houver', 'houvera', 'houveram', 'houverei', 'houvere
m', 'houveremos', 'houveria', 'houveriam', 'houvermos', 'houve
rá', 'houverão', 'houveríamos', 'houvesse', 'houvessem', 'houv
éramos', 'houvéssemos', 'há', 'hão', 'isso', 'isto', 'já', 'lh
e', 'lhes', 'mais', 'mas', 'me', 'mesmo', 'meu', 'meus', 'minh
a', 'minhas', 'muito', 'na', 'nas', 'nem', 'no', 'nos', 'noss
a', 'nossas', 'nosso', 'nossos', 'num', 'numa', 'não', 'nós',
'o', 'os', 'ou', 'para', 'pela', 'pelas', 'pelo', 'pelos', 'po
r', 'qual', 'quando', 'que', 'quem', 'se', 'seja', 'sejam', 's
ejamos', 'sem', 'serei', 'seremos', 'seria', 'seriam', 'será',
'serão', 'seríamos', 'seu', 'seus', 'somos', 'sou', 'sua', 'su
as', 'são', 'só', 'também', 'te', 'tem', 'temos', 'tenha', 'te
nham', 'tenhamos', 'tenho', 'terei', 'teremos', 'teria', 'teri
am', 'terá', 'terão', 'teríamos', 'teu', 'teus', 'teve', 'tinh
a', 'tinham', 'tive', 'tivemos', 'tiver', 'tivera', 'tiveram',
'tiverem', 'tivermos', 'tivesse', 'tivessem', 'tivéramos', 'ti
véssemos', 'tu', 'tua', 'tuas', 'tém', 'tínhamos', 'um', 'um
a', 'você', 'vocês', 'vos', 'à', 'às', 'é', 'éramos']
```

Atualiza a lista de stopwords

In [11]:

```
# Palavras a adicionar na lista de stopwords estão contidas em um arquivo c
sv externo
dfsw = pd.read_csv('./bases/stopwords.csv', encoding='ISO-8859-1')
stopwords_df=sorted(list(dfsw['stopword']))
swtemp = list(stopwords_df)
swtemp.sort()
print(swtemp)
```

['abaixo', 'acessorios', 'acessórios', 'aco', 'acondicionado
s', 'adaptavel', 'adaptável', 'allen', 'almas', 'alta', 'am',
'anel', 'ano', 'aplicacao', 'aplication', 'aplicavel', 'aplica
ção', 'aplicável', 'application', 'ate', 'atitanium', 'aç', 'a
ço', 'bicicleta', 'bicycle', 'bike', 'bravo', 'cada', 'caixa',
'caixas', 'cambio', 'carbono', 'certificado', 'cever', 'chai
n', 'chh', 'china', 'ciclomotor', 'ciclomotores', 'cilindrad
a', 'cilindradas', 'cod', 'code', 'codigo', 'comando', 'combus
tão', 'comercial', 'comercialmente', 'commodity', 'compative
l', 'compatível', 'compl', 'completo', 'completos', 'compost
o', 'composto', 'compostopor', 'compostpo', 'comum', 'condica
o', 'condicoes', 'condição', 'condições', 'confeccionado', 'co
nformidade', 'conhecido', 'conj', 'conjunto', 'conjuntos', 'co
nstituido', 'constitutivo', 'constituído', 'contendo', 'copost
o', 'coroa', 'coroaes', 'corr', 'corrent', 'corrente', 'corren
tee', 'correntes', 'corrnte', 'cx', 'câmbio', 'cãmbio', 'códig
o', 'decreto', 'denominada', 'dente', 'dentes', 'descricao',
'descrição', 'descriçao', 'descrição', 'destaque', 'destaque',
'destaques', 'detransmissão', 'diante', 'dimensao', 'dimensoe
s', 'dimensão', 'dimensões', 'diverso', 'diversos', 'dominad
o', 'durabilidade', 'elo', 'elos', 'embalagem', 'engine', 'eng
renagem', 'engrenagens', 'epinhao', 'epinhão', 'especifico',
'específico', 'espessura', 'evol', 'exclusivo', 'fabr', 'fabr
i', 'fabricada', 'fabricado', 'final', 'foiproduzida', 'formad
o', 'funcao', 'funcao', 'funcão', 'funcão', 'funçao', 'funçã
o', 'gtin', 'hardt', 'ho', 'hp', 'imetro', 'in', 'incluindo',
'incluso', 'indicado', 'ingles', 'inmetro', 'inv', 'invoice',
'iron', 'item', 'jc', 'ki', 'kif', 'kit', 'kitr', 'kittr', 'ki
ttr', 'kmc', 'ligacoes', 'ligaçoes', 'ligaçoes', 'ligações',
'marca', 'mark', 'match', 'material', 'materialdo', 'maxx', 'm
edida', 'medidas', 'medindo', 'metal', 'mini', 'mod', 'model
o', 'modelos', 'moto', 'motociclet', 'motocicleta', 'motocicle
tas', 'motoneta', 'motonetas', 'motoparts', 'motor', 'motorcic
leta', 'motos', 'motos', 'movimento', 'nbsp', 'ncm', 'nome',
'nopinh', 'normais', 'nova', 'novo', 'nr', 'numero', 'número',
'onde', 'or', 'origem', 'oring', 'ox', 'papelao', 'papelão',
'part', 'parte', 'partes', 'parts', 'pc', 'pc-coroa', 'pc-corr
ente', 'pc-pinhao', 'pcs', 'pec', 'pecas', 'perfil', 'peç', 'p
eças', 'pinh', 'pinhao', 'pinhão', 'posição', 'premium', 'proc
edencia', 'procedência', 'prodepe', 'produto', 'próprio', 'p
ç', 'qdes', 'qtd', 'qtds', 'qty', 'quadriciclo', 'quadriciclo
s', 'quantidad', 'quantidade', 're', 'ref', 'reforcada', 'refo
rçada', 'registro', 'rel', 'relacao', 'relação', 'reposicao',
'resp', 'respo', 'respons', 'responsa', 'responsav', 'responsa
ve', 'responsavel', 'responsáv', 'responsável', 'ret', 'retalh
o', 'retentor', 'riffel', 'ring', 'roda', 'sae', 'scud', 'sem
i', 'semi-', 'semi-kit', 'sendo', 'serve', 'set', 'shipping',
'sistema', 'sm', 'sprocket', 'standard', 'standart', 'standart
t', 'std', 'stdmodelo', 'steel', 'tambem', 'também', 'tec', 't
emp', 'temperado', 'tipo', 'tipos', 'titanio', 'titaniu', 'tit
anium', 'titaniun', 'tr', 'tracao', 'tracão', 'trans', 'transm
is', 'transmisao', 'transmiss', 'transmissa', 'transmissao',
'transmission', 'transmissão', 'transmitir', 'traseira', 'traç
ao', 'tração', 'und', 'unds', 'unid', 'unidade', 'unidade', 'u
nidades', 'unifort', 'uo', 'uso', 'utilizada', 'utilizadas',

```
'utilizado', 'utilizados', 'utilização', 'vem', 'venda', 'wit
h', 'xy', 'year']
```

In [12]:

```
# Atualizar stopwords
stopwords.update(stopwords_df)
```

In [13]:

```
# Mostra toda a lista de stopwords
swtemp = list(stopwords)
swtemp.sort()
print(swtemp)
```

['a', 'abaixo', 'acessorios', 'acessórios', 'aco', 'acondicion
ados', 'adaptavel', 'adaptável', 'allen', 'almas', 'alta', 'a
m', 'anel', 'ano', 'ao', 'aos', 'aplicacao', 'aplication', 'ap
licavel', 'aplicação', 'aplicável', 'application', 'aquela',
'aquelas', 'aquele', 'aqueles', 'aquilo', 'as', 'ate', 'atitan
ium', 'até', 'aç', 'aço', 'bicicleta', 'bicycle', 'bike', 'bra
vo', 'cada', 'caixa', 'caixas', 'cambio', 'carbono', 'certific
ado', 'cever', 'chain', 'chh', 'china', 'ciclomotor', 'ciclomo
tores', 'cilindrada', 'cilindradas', 'cod', 'code', 'codigo',
'com', 'comando', 'combustão', 'comercial', 'comercialmente',
'commodity', 'como', 'compativel', 'compatível', 'compl', 'com
pleto', 'completos', 'composto', 'compostopor', 'compostpo',
'comum', 'condicao', 'condicoes', 'condição', 'condições', 'co
nfeccionado', 'conformidade', 'conhecido', 'conj', 'conjunto',
'conjuntos', 'constituido', 'constitutivo', 'constituído', 'co
ntendo', 'coposto', 'coroa', 'coroaes', 'corr', 'corrent', 'co
rrente', 'correntee', 'correntes', 'corrnte', 'cx', 'câmbio',
'cãmbio', 'código', 'da', 'das', 'de', 'decreto', 'dela', 'del
as', 'dele', 'deles', 'denominada', 'dente', 'dentes', 'depoi
s', 'descricao', 'descrição', 'descriçao', 'descrição', 'desta
que', 'destaques', 'detransmissão', 'diante', 'dimensao', 'dim
ensoes', 'dimensão', 'dimensões', 'diverso', 'diversos', 'do',
'dominado', 'dos', 'durabilidade', 'e', 'ela', 'elas', 'ele',
'eles', 'elo', 'elos', 'em', 'embalagem', 'engine', 'engrenage
m', 'engrenagens', 'entre', 'epinhao', 'epinhão', 'era', 'era
m', 'especifico', 'específico', 'espessura', 'essa', 'essas',
'esse', 'esses', 'esta', 'estamos', 'estas', 'estava', 'estava
m', 'este', 'esteja', 'estejam', 'estejamos', 'estes', 'estev
e', 'estive', 'estivemos', 'estiver', 'estivera', 'estiveram',
'estiverem', 'estivermos', 'estivesse', 'estivessem', 'estivér
amos', 'estivéssemos', 'estou', 'está', 'estávamos', 'estão',
'eu', 'evol', 'exclusivo', 'fabr', 'fabri', 'fabricada', 'fabr
icado', 'final', 'foi', 'foiproduzida', 'fomos', 'for', 'for
a', 'foram', 'forem', 'formado', 'formos', 'fosse', 'fossem',
'fui', 'funcao', 'funcão', 'funçao', 'função', 'fôramos', 'fôs
semos', 'gtin', 'haja', 'hajam', 'hajamos', 'hardt', 'havemo
s', 'hei', 'ho', 'houve', 'houvemos', 'houver', 'houvera', 'ho
uveram', 'houverei', 'houverem', 'houveremos', 'houveria', 'ho
uveriam', 'houvermos', 'houverá', 'houverão', 'houveríamos',
'houvesse', 'houvessem', 'houvéramos', 'houvéssemos', 'hp', 'h
á', 'hão', 'imetro', 'in', 'incluindo', 'incluso', 'indicado',
'ingles', 'inmetro', 'inv', 'invoice', 'iron', 'isso', 'isto',
'item', 'jc', 'já', 'ki', 'kif', 'kit', 'kitr', 'kittr', 'km
c', 'lhe', 'lhes', 'ligacoes', 'ligaçoes', 'ligações', 'mais',
'marca', 'mark', 'mas', 'match', 'material', 'materialdo', 'ma
xx', 'me', 'medida', 'medidas', 'medindo', 'mesmo', 'metal',
'meu', 'meus', 'minha', 'minhas', 'mini', 'mod', 'modelo', 'mo
delos', 'moto', 'motociclet', 'motocicleta', 'motocicletas',
'motoneta', 'motonetas', 'motoparts', 'motor', 'motorcicleta',
'motos', 'movimento', 'muito', 'na', 'nas', 'nbsp', 'ncm', 'ne
m', 'no', 'nome', 'nopinh', 'normais', 'nos', 'nossa', 'nossa
s', 'nosso', 'nossos', 'nova', 'novo', 'nr', 'num', 'numa', 'n
umero', 'não', 'nós', 'número', 'o', 'onde', 'or', 'origem',
'oring', 'os', 'ou', 'ox', 'papelao', 'papelão', 'para', 'par
t', 'parte', 'partes', 'parts', 'pc', 'pc-coroa', 'pc-corrent
e', 'pc-pinhao', 'pcs', 'pec', 'pecas', 'pela', 'pelas', 'pel

```
o', 'pelos', 'perfil', 'peç', 'peças', 'pinh', 'pinhao', 'pinh
ão', 'por', 'posição', 'premium', 'procedencia', 'procedênci
a', 'prodepe', 'produto', 'próprio', 'pç', 'qdes', 'qtd', 'qtd
s', 'qty', 'quadriciclo', 'quadriciclos', 'qual', 'quando', 'q
uantidad', 'quantidade', 'que', 'quem', 're', 'ref', 'reforcad
a', 'reforçada', 'registro', 'rel', 'relacao', 'relação', 'rep
osicao', 'resp', 'respo', 'respons', 'responsa', 'responsav',
'responsave', 'responsavel', 'responsáv', 'responsável', 're
t', 'retalho', 'retentor', 'riffel', 'ring', 'roda', 'sae', 's
cud', 'se', 'seja', 'sejam', 'sejamos', 'sem', 'semi', 'semi-
', 'semi-kit', 'sendo', 'serei', 'seremos', 'seria', 'seriam',
'serve', 'será', 'serão', 'seríamos', 'set', 'seu', 'seus', 's
hipping', 'sistema', 'sm', 'somos', 'sou', 'sprocket', 'standa
rd', 'standart', 'standartt', 'std', 'stdmodelo', 'steel', 'su
a', 'suas', 'são', 'só', 'tambem', 'também', 'te', 'tec', 'te
m', 'temos', 'temp', 'temperado', 'tenha', 'tenham', 'tenhamo
s', 'tenho', 'terei', 'teremos', 'teria', 'teriam', 'terá', 't
erão', 'teríamos', 'teu', 'teus', 'teve', 'tinha', 'tinham',
'tipo', 'tipos', 'titanio', 'titaniu', 'titanium', 'titaniun',
'tive', 'tivemos', 'tiver', 'tivera', 'tiveram', 'tiverem', 't
ivermos', 'tivesse', 'tivessem', 'tivéramos', 'tivéssemos', 't
r', 'tracao', 'tração', 'trans', 'transmis', 'transmisao', 'tr
ansmiss', 'transmissa', 'transmissao', 'transmission', 'transm
issão', 'transmitir', 'traseira', 'traçao', 'tração', 'tu', 't
ua', 'tuas', 'tém', 'tínhamos', 'um', 'uma', 'und', 'unds', 'u
nid', 'unidade', 'unidades', 'unifort', 'uo', 'uso', 'utilizad
a', 'utilizadas', 'utilizado', 'utilizados', 'utilização', 've
m', 'venda', 'você', 'vocês', 'vos', 'with', 'xy', 'year',
'à', 'às', 'é', 'éramos']
```

In [14]:

```
len(stopwords)
```

Out[14]:

518

Carrega lista de aplicações

In [15]:

```
# carrega a lista de marcas de motos do arquivo
dftemp=pd.read_csv('./bases/Aplicacoes.csv')
```

In [16]:

```
dftemp.head()
```

Out[16]:

	APLICACOES
0	ACELLERA ACX 250F 250
1	ACELLERA FRONTLANDER 500
2	ACELLERA FRONTLANDER 800 EFI
3	ACELLERA HOTZOO SPORT 90
4	ACELLERA QUADRILANDER 300

Cria a lista de Palavras Chave das Aplicações

In [17]:

```
# remove caracteres especiais ou soltos e termos duplicados, salvando na li
sta
palavrasChave=sorted(set(re.sub(r"\b \w \b",
                                ' ',
                                re.sub(r"[/<>()|\+\$%&#@\'\"]+",
                                       " ",
                                       " ".join(dftemp['APLICACOES'].tolist
()))).split()))
```

In [18]:

```
len(palavrasChave)
```

Out[18]:

```
894
```

In [19]:

```
# amostra de palavrasChave
print(palavrasChave[:20],' ... ', palavrasChave[-20:])
```

```
['1000', '1000F', '1000R', '1000V', '1098', '1100', '1100XX',
'110S', '1125', '1190', '1198', '1200', '1200Z', '125', '125
R', '130', '1300', '135', '1400', '150']  ...  ['YFS', 'YS15
0', 'YS250', 'YZ', 'YZF', 'YZR', 'Z1000', 'Z750', 'Z800', 'ZAC
H', 'ZANELLA', 'ZENITH', 'ZIG', 'ZING', 'ZIP', 'ZONGSHEN', 'Z
R', 'ZRX', 'ZS', 'ZZR']
```

In [20]:

```
'RT' in palavrasChave
```

Out[20]:

```
True
```

Limpeza e criação da coluna DESCRICAO

Função de limpeza de dados irrelevantes para a classificação e remoção de stopwords

In [21]:

```
def limpaDescricao(descricao): #
    descricao=descricao.lower() #transformar em minúsculas
    # remove top (1045) e variantes
    descricao=re.sub(r'\b[ (-]*top \( *1045 *\)[-) ]*\b',' ',descricao)
    # remove códigos numéricos entre parênteses com -*/
    descricao=re.sub(r"\( *\d*[\/\*\-\d]*\d* *\)", ' ', descricao)
    # remove a ocorrência de "código e etc." e o termo seguinte começado co
m número
    # att: (alguns tem hífen ou asterisco) (colocar antes de remover pontua
ção)
    descricao=re.sub(r"\b(invoice|código|codigo|cod|cód|(certificado|cert)(
no|nr|)|ref)[0-9a-z/\-\*\.\:]* *\d[^ ]+", ' ', descricao)
    # remove identificação de referência de engrenagens dos kits (antes da
 pontuação)
    #descricao=re.sub(r"([^a-z]|)(ho|uo|h|l|t|ktd|sm|m|d| x|elos )\d{1,}[ x
\-\/,);.]|[ x\-\/(]*\d{1,}(ho|uo|h|l|z|t|ktd|m|d|x| dentes| elos)[ \-
\/,;)]", ' ', descricao) # 00h
    descricao=re.sub(r"\d*(ho|uo|h|l|t|ktd|sm|m|d|elos )\d{1,}[ \-\/,);.]|[
\-\/(]*\d{1,}(ho|uo|h|l|z|t|ktd|m|d| dentes| elos)", ' ', descricao) # 00h
    # substitui os termos "s/re" e "s/ret" por "sem retentor"
    descricao=re.sub(r"\b(s\/re|s\/ret)\b", 'sem retentor', descricao)
    # substitui os termos "c/re" ou "c/ret" por "com retentor"
    descricao=re.sub(r"\b(c\/re|c\/ret)\b", 'com retentor', descricao)
    # substitui o termo "aplicação" e "modelo" emendado com outro
    descricao=re.sub(r"aplicacao", "aplicacao ", descricao)
    descricao=re.sub(r"modelo", "modelo ", descricao)
    # remove códigos no início da descrição
    descricao=re.sub(r"^\b\d{2,}[^ ]*\b", ' ', descricao)
    descricao=re.sub(r"^k[^ ]+", ' ', descricao)
    descricao=re.sub(r"- | -|[\\\+,.:;!?/]+", ' ', descricao) #remover pont
uação (att: "- " ou " -")
    #correção de erros de digitação comuns
    termos={'titan': ['titian','tita','tintan','tit'],
            'honda': ['hond','hnda','hon'],
            'twister': ['twist', 'twiste'],
            'dafra kansas': ['dafra kan'],
            'tenere': ['tener','tenerre'],
            'broz': ['bros','bross'],
            'titan 150': ['titan150'],
            'broz 150': ['bross125.','bros125.','broz125','bross150.','bros
150.','broz150'],
            'pop 100':['pop100'],
            'phoenix':['phoeni','phenix'],
            'c100': ['c 100']}
    for termo in termos:
        for termoerrado in termos[termo]:
            descricao=re.sub(r"\b"+termoerrado+r"\b", termo, descricao)
    descricao=re.sub(r"[/<>()|\+\\$%&#@\'\"]+", ' ', descricao) #remover ca
rcteres especiais
    # remove a ocorrência de medidas tipo 00x000x00 ou 000x0000
    descricao=re.sub(r"\b\d{1,}(x|\*)\d{1,}(x|\*)\d{1,}|\d{1,}(x|\*)\d{1,}
\b", ' ', descricao)
    # remove identificação de quantidades, unidades, peças e conjuntos
    descricao=re.sub(r"\b(\d* *(conj|und|uni|pc|pç|pec|peç)( \w|\w)+?)\b",
```

```
' ', descricao)
    # remove identificação de mais de 4 dígitos com ou sem letras no início
e no final
    descricao=re.sub(r"\w+\d{4,}\w+", ' ', descricao)
    # remove números de 4 dígitos ou mais começados de 2 a 9
    descricao=re.sub(r"\b[02-9]\d{3,}\b", ' ', descricao)
    # remove identificação de termos começados por zero
    descricao=re.sub(r"\b0\d*\w+?(?=\b)", ' ', descricao)
    # remove a ocorrência de "marca " e o termo na lista até o próximo espa
ço
    for marca in ['kmc *gold','am *gold','king','bravo *racing','riffel *to
p']:
        descricao=re.sub(r"\bmarca[ :\./]*"+str(marca)+r"[^ ]*", ' ', descr
icao) # colocar antes das stopwords
    descricao=re.sub(r"marca[ :\./]*\w+", ' ', descricao)
    descricao=re.sub(r"(^-| -|- )", ' ', descricao)
    # remove stopwords mantendo a ordem original da descrição
    descricao=list(dict.fromkeys(descricao.split())) # cria lista com termo
s únicos
    descricao=" ".join([x for x in descricao if x not in set(stopwords)]) #
exclui stopwords
    # limpa os número que não estão na lista de aplicações (colocar depois
 das stopwords)
    desc=descricao.upper().split() # quebra a descrição
    dif=list(set(descricao.upper().split()).difference(palavrasChave)) # pe
ga os termos diferentes de palavrasChave
    [desc.remove(x) for x in desc if (x in dif and x.isnumeric())] # exclui
de desc os termos numéricos diferentes
    descricao=" ".join(desc).lower() # volta para texto
    #remover hífen, letras ou números soltos (deixar duplicado mesmo)
    descricao=re.sub(r"(^-| -|- |\b\w\b)", ' ', descricao)
    descricao=re.sub(r"(^-| -|- |\b\w\b)", ' ', descricao)
    #substitui remove o i das cilindradas: ex.: 125i por 125
    termos=re.findall(r"\d{1,}i\b",descricao)
    if termos:
        for termo in termos:descricao=descricao.replace(termo,termo[:-1])
    # remove espaços em excesso (colocar no final)
    descricao=re.sub(r" {2,}", ' ', descricao)
    descricao=descricao.strip()
    # retorna a descricao como saída da função
    return descricao # retorna a descrição
```

Exemplo de execução da função

In [22]:

```
linha=745
```

In [23]:

```
df.iloc[linha]['DESCRICAO DO PRODUTO']
```

Out[23]:

```
'KIT TRANSMISSÃO PARA MOTOCICLETAS,COMPOSTO DE CORRENTE,PINHAO
E COROA,PARA MODELOS DIVERSOS DE MOTOCICLETAS (KIT C 100 BIZ
(13-15) 34Z X 14Z WITH CHAIN 428H X 108L - TITANIUM (1045)) MO
DELO KIT C 100 BIZ (13-15) 34Z X 14Z WITH CHAIN 428H X 108L -
TIT'
```

In [24]:

```
limpaDescricao(df.iloc[linha]['DESCRICAO DO PRODUTO'])
```

Out[24]:

```
'c100 biz titan'
```

Execução da função para criação da coluna DESCRICAO limpa

In [25]:

```
ini=time.time()
df['DESCRICAO']=df['DESCRICAO DO PRODUTO'].apply(limpaDescricao)
fim=time.time()
print(f'Tempo de execução: {fim-ini:.2} segundos.')
```

```
Tempo de execução: 7.8 segundos.
```

In [26]:

```
df.sample(5)
```

Out[26]:

	PAIS DE ORIGEM	DESCRICAO DO PRODUTO	VALOR UN.PROD.DOLAR	DESCRICAO
10292	CHINA, REPUBLICA POP	-54T17T/428H132L - KIT TRANSMISSAO EM ACO COMP...	3.600100	
1897	CHINA, REPUBLICA POP	ENGRENAGENS PARA TRANSMISSÃO DE MOTOCICLETAS E...	3.713600	c100 biz
12069	CHINA, REPUBLICA POP	10530031 IN KIT TRANSMISSAO P/MOTOCICLETAS(COR...	3.744000	spee
2927	CHINA, REPUBLICA POP	007308# KIT TRANSMISSAO STANDARD TEMP. COMPL. ...	3.221500	suzuki yes intruder katana 125
12202	CHINA, REPUBLICA POP	91256 KIT CG 160 TITAN (16-19) / CG 160 FAN (1...	9.711875	cg 160 titan fan start cargo

In [27]:

```
df['DESCRICAO'].iloc[linha]
```

Out[27]:

```
'c100 biz titan'
```

Criação de colunas Modelo

Função de determinação de palavras chave na coluna Modelo

In [28]:

```
def achaPalavraChave(descricao):
    palavras=[]
    descricao=descricao.upper()
    desc=descricao.split()
    for palavra in palavrasChave:
        if palavra in desc:
            palavras.append(palavra)
        else:
            if palavra.isnumeric():
                pat=r"[0-9]*"+str(palavra)+r"[0-9]*"
            elif palavra.isalpha():
                pat=r"[A-Z]*"+str(palavra)+r"[A-Z]*"
            else:
                pat=r"\b"+palavra+r"\b"
            a = re.findall(pat,descricao)
            if len(a)>0:
                 # adiciona resultado nas palavras se o resultado estiver e
m palavrasChave
                palavras+=[a[i] for i in range(len(a)) if a[i] in palavrasC
have]
    palavras=list(set(palavras)) # remove duplicados
    palavras=" ".join(palavras) # converte para string
    return palavras.lower()
```

In [29]:

```
achaPalavraChave(limpaDescricao(df['DESCRICAO DO PRODUTO'].iloc[linha]))
```

Out[29]:

```
'c100 titan biz'
```

Função para acrescentar a marca da motocicleta

In [30]:

```
# termos que iniciam item da descrição correspondem a marca
# As que começam com espaço devem permanecer assim, pois há outros modelos
 com o mesmo final
Marcas = {'HONDA': ['CG','CD','CBX','CB','CBR','CRF','BIZ','BROS','BROZ','X
L',' FAN','XR','XRE'
                    'DREAM','TITAN','TODAY','TWIN','POP','NX','NXR','TWISTE
R', 'HORNET',
                    'AMERICA','BOLDOR','DUTY','FIREBLADE','FURY','WING','LE
AD','MAGNA','NL',
                    ' NC','NSR','NC','NXR','PACIFIC','COAST','SHADOW',' STR
ADA','STUNNER','HAWK',
                    'SUPERBLACKBIRD','TORNADO','TURUNA','XRV','AFRICA','VAL
KYRIE','VARADERO',
                    'VFR','VLR','VTR','VTX','TRANSALP'],
          'YAMAHA': ['AEROX','ALBA','AXIS','BWS','DRAG ','DT','FZ','FJ',' R
D','TENERE',
                     'MT','XF','XJ','XS','XT','XZ','YF','YZ','LANDER','GLAD
IATOR','GRIZZLY',
                     'YBR','YZ','VIRAGO','FACTOR','EC','CRYPTON','FAZER','J
OG',' LANDER',
                     'FROG','LIBERO','MAJESTY','MEST','MIDNIGHT','MORPH','N
EO','PASSOL'],
          'DAFRA': ['APACHE','CITYCOM','KANSAS','LASER','NEXT','RIVA','ROAD
WIN','ZIG','SPEED'],
          'SUZUKI': ['KATANA','YES','INTRUDER'],
          'ZONGSHEN': ['ZS'],
          'KASINSKI': ['COMET','MIRAGE'],
          'POLARIS': ['SPORTSMAN','RZR','RANGER'],
          'KAWASAKI': ['NINJA','VERSYS','VOYAGER','GTR','KDX','KL','KX','K
Z','ZR','ZZ','ER6N','ER6F'],
          'DAYANG': ['DY1','DY2','DY5'],
          'SUNDOWN': ['WEB','FIFITY','PALIO','PGO','STX','VBLADE','EVO', 'H
UNTER MAX'],
          'SHINERAY': ['BIKE','BRAVO','DISCOVER','EAGLE','INDIANAPOLIS','JE
T','NEW','WAVE',
                       'STRONG','SUPER SMART','VENICE',' XY']}
```

In [31]:

```
# Função para pegar a chave pelo valor, dado que valor é único.
def pegaChave(v, dict):
    for chave, valores in dict.items():
        if type(valores)!=type([1,2]):
            valores=[valores]
        for valor in valores:
            if v == valor:
                return chave
    return "Não existe chave para esse valor."
```

In [32]:

```
def acrescentaMarca(descricao):
    for marca in Marcas:
        if re.search(marca,descricao.upper()):
            descricao += " "+marca
        for termo in Marcas[marca]:
            t1=termo.split()
            if len(t1)>1:
                pat=r"(?:"+t1[0]+r"|"+t1[1]+r").*(?:"+t1[0]+r"|"+t1[1]+r")"
            elif len(termo)<3:
                pat=termo+r"([0-9]{1,}|\b)"
            else:
                pat=termo
            resultados = re.findall(pat,descricao.upper())
            if resultados:
                descricao += " "+marca
                descricao += " "+" ".join(resultados)
                descricao += " "+termo
    descricao=" ".join(sorted(set(descricao.lower().split())))
    return descricao
```

In [33]:

```
acrescentaMarca(achaPalavraChave(limpaDescricao(df['DESCRICAO DO PRODUTO'].
iloc[linha])))
```

Out[33]:

```
'biz c100 honda titan'
```

Aplica as funções

Tenha paciência, demora cerca de 1 minuto para cada mil registros.

In [34]:

```
# cria as colunas
df=df.assign(Modelo=df['DESCRICAO'])
df.iloc[linha]['DESCRICAO']
```

Out[34]:

```
'c100 biz titan'
```

In [35]:

```
df.iloc[:,-2:].sample(5)
```

Out[35]:

	DESCRICAO	Modelo
7430	nxr-125 broz	nxr-125 broz
7152	crf 230	crf 230
4746	titan 160	titan 160
15923	xls 125 96	xls 125 96
4524	xls-125	xls-125

In [36]:

```
# aplica as funções a cada coluna
ini=time.time()
now = time.strftime("%H:%M", time.localtime(time.time()))
print("Hora de início:" + now)
print(f"Tempo estimado de execução: {df.shape[0]//1000} minutos.") # 1000 r
egistros por minuto

print('\nBuscando palavras chave... Aguarde...')
df['Modelo']=df['Modelo'].apply(achaPalavraChave)

print('\nBuscando marcas... Aguarde...')
df['Modelo']=df['Modelo'].apply(acrescentaMarca)

now = time.strftime("%H:%M", time.localtime(time.time()))
fim=time.time()
print("\nHora de término:" + str(now))
print("Tempo decorrido: " + str(round((fim-ini)/60,2)) + " minutos.")
```

```
Hora de início:16:36
Tempo estimado de execução: 18 minutos.

Buscando palavras chave... Aguarde...

Buscando marcas... Aguarde...

Hora de término:16:52
Tempo decorrido: 16.57 minutos.
```

In [37]:

```
df['DESCRICAO DO PRODUTO'].iloc[linha]
```

Out[37]:

```
'KIT TRANSMISSÃO PARA MOTOCICLETAS,COMPOSTO DE CORRENTE,PINHAO
E COROA,PARA MODELOS DIVERSOS DE MOTOCICLETAS (KIT C 100 BIZ
(13-15) 34Z X 14Z WITH CHAIN 428H X 108L - TITANIUM (1045)) MO
DELO KIT C 100 BIZ (13-15) 34Z X 14Z WITH CHAIN 428H X 108L -
TIT'
```

In [38]:

```
df[['DESCRICAO DO PRODUTO','DESCRICAO','Modelo']].sample(5)
```

Out[38]:

	DESCRICAO DO PRODUTO	DESCRICAO	Modelo
3123	21989 - 71815 - KIT DE TRANSMISSAO PARA MOTOCI...	ybr 125 factor	125 factor yamaha ybr
13490	20257/i45 - KIT DE TRANSMISSÃO EM AÇO 1045, MA...	xtz150 crosser	150 crosser xtz
9440	TM10300 - KIT DE TRANSMISSÃO COMPOSTO DE COROA...	yamaha yes 125 14	125 suzuki yamaha yes
11269	KIT DE TRANSMISSÃO EM AÇO 1045, PARA USO EM MO...	titan	honda titan
6473	KIT DE TRANSMISSAO PARA MOTOCICLETA, CONTENDO ...	titan fan *	fan honda titan

In [39]:

```
df_sem_modelo = df[df['Modelo']=='']
df_sem_modelo['DESCRICAO'].to_excel("./bases/sem_modelo.xlsx")
```

In [40]:

```
df_sem_modelo[['DESCRICAO DO PRODUTO','DESCRICAO','Modelo']].sample(10)
```

Out[40]:

	DESCRICAO DO PRODUTO	DESCRICAO	Modelo
9362	3 - PARTES E ACESSORIOS DE MOTOCICLETA, SENDO ...	49	
5586	TRANSMISSAO PARA USO EM MOTOCICLETA COMPOSTO D...		
8166	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...		
3249	45T14T/428H118L- KIT TRANSMISSAO EM ACO COMPOS...		
9048	2 - PARTES E ACESSORIOS DE MOTOCICLETA, SENDO ...	34 100	
3482	REF: 428HX118LX43TX14T (1104983) -KIT TRANSMIS...		
10253	TRANSMISSAO PARA USO EM MOTOCICLETA COMPOSTO D...		
9365	6 - PARTES E ACESSORIOS DE MOTOCICLETA, SENDO ...	45	
13083	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...		
17756	001-P21B-00600 - Kit transmissao Titanium para...		

In [41]:

```
df_sem_modelo.shape
```

Out[41]:

```
(792, 5)
```

In [42]:

```
df_sem_modelo.reset_index(drop=True)
```

Out[42]:

	PAIS DE ORIGEM	DESCRICAO DO PRODUTO	VALOR UN.PROD.DOLAR	DESCRICAO	Modelo
0	CHINA, REPUBLICA POP	KIT DE TRANSMISSÃO, COMPOSTO DE COROA, CORRENT...	3.900000	f80 kits	
1	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.809000		
2	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	4.052000		
3	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.876000		
4	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.877000		
5	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.294000		
6	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.871000		
7	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	4.173000		
8	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.673000		
9	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.877000		

	PAIS DE ORIGEM	DESCRICAO DO PRODUTO	VALOR UN.PROD.DOLAR	DESCRICAO	Modelo
10	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	4.166000		
11	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.849000		
12	CHINA, REPUBLICA POP	880375 - KIT DE TRANSMISSÃO, COMPOSTO DE CORRE...	3.329221	spee	
13	CHINA, REPUBLICA POP	880349 - KIT DE TRANSMISSÃO, COMPOSTO DE CORRE...	3.813052	come	
14	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	4.166000		
15	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.809000		
16	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.213000		
17	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.279000		
18	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.871000		
19	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.871000		
20	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.909000		

	PAIS DE ORIGEM	DESCRICAO DO PRODUTO	VALOR UN.PROD.DOLAR	DESCRICAO	Modelo
21	CHINA, REPUBLICA POP	980730 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENT...	4.701278	56	
22	CHINA, REPUBLICA POP	070593 - KIT TRANSMISSÃO COMPOSTO DE CORRENTE,...	3.730000	ti x118 aço1045	
23	CHINA, REPUBLICA POP	072165 - KIT TRANSMISSÃO COMPOSTO DE CORRENTE,...	4.846000	15	
24	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	4.218000		
25	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.918000		
26	CHINA, REPUBLICA POP	KIT TRANSMISSÃO AÇO (1045), COMPOSTO DE CORREN...	3.984000		
27	CHINA, REPUBLICA POP	43T14T/428H116L - KIT TRANSMISSAO EM ACO COMPO...	3.024853		
28	CHINA, REPUBLICA POP	43T16T/428H118L - KIT TRANSMISSAO EM ACO COMPO...	2.975183		
29	CHINA, REPUBLICA POP	80422 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	9.480000	mt r3 43	
...	...	...	...	...	...
762	CHINA, REPUBLICA POP	001-P21B-03700 - Kit transmissao Titanium para...	4.100000		
763	CHINA, REPUBLICA POP	001-P21B-04900 - Kit transmissao Titanium para...	4.550000		
764	CHINA, REPUBLICA POP	001-P21B-06100 - Kit transmissao Titanium para...	6.240000		
765	CHINA, REPUBLICA POP	001-P21B-06500 - Kit transmissao Titanium para...	3.460000		

	PAIS DE ORIGEM	DESCRICAO DO PRODUTO	VALOR UN.PROD.DOLAR	DESCRICAO	Modelo
766	CHINA, REPUBLICA POP	001-P21B-06600 - Kit transmissao Titanium para...	3.440000		
767	CHINA, REPUBLICA POP	TRANSMISSAO PARA USO EM MOTOCICLETA COMPOSTO D...	3.990000	ft2368	
768	CHINA, REPUBLICA POP	TRANSMISSAO PARA USO EM MOTOCICLETA COMPOSTO D...	3.990000	ft2388	
769	CHINA, REPUBLICA POP	12565-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	2.341216		
770	CHINA, REPUBLICA POP	12570-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	2.341216		
771	CHINA, REPUBLICA POP	12588-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	2.200000		
772	CHINA, REPUBLICA POP	12591-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	2.359939		
773	CHINA, REPUBLICA POP	12603-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	3.200000		
774	CHINA, REPUBLICA POP	12604-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	3.200000		
775	CHINA, REPUBLICA POP	12578-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	3.200000		
776	CHINA, REPUBLICA POP	12580-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	2.600000		
777	CHINA, REPUBLICA POP	12606-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	3.200000		
778	CHINA, REPUBLICA POP	12566-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	2.600000		

	PAIS DE ORIGEM	DESCRICAO DO PRODUTO	VALOR UN.PROD.DOLAR	DESCRICAO	Modelo
779	CHINA, REPUBLICA POP	12562-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	3.200000		
780	CHINA, REPUBLICA POP	12582-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	3.200000		
781	CHINA, REPUBLICA POP	12583-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	2.512246		
782	CHINA, REPUBLICA POP	12584-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	2.512246		
783	CHINA, REPUBLICA POP	12585-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	3.200000		
784	CHINA, REPUBLICA POP	12586-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	3.200000		
785	CHINA, REPUBLICA POP	12567-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	1.946308		
786	CHINA, REPUBLICA POP	12592-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	3.200000		
787	CHINA, REPUBLICA POP	12593-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	3.200000		
788	CHINA, REPUBLICA POP	12594-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	3.200000		
789	CHINA, REPUBLICA POP	12595-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	2.260000		
790	CHINA, REPUBLICA POP	12596-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	2.260000		
791	CHINA, REPUBLICA POP	12600-E - PARTES E PEÇAS DE MOTOCICLETAS, SEND...	2.260000		

792 rows × 5 columns

In [43]:

```
print(f'Número de registros sem aplicação contida na descrição: {df_sem_modelo.shape[0]}')
```

```
Número de registros sem aplicação contida na descrição: 792
```

Exclusão dos registros sem aplicação contida na descrição

Neste momento é necessário tomar uma decisão sobre o que fazer com os registros que permaneceram sem nenhuma extração na coluna **Modelo**.
Para tal decisão foi necessário observar cada um desses registros no arquivo "sem_modelo.xls" exportado e constatar que nenhum dos registros possui realmente qualquer alusão à aplicação do item descrito.

In [44]:

```
df=df[df['Modelo']!='']
```

In [45]:

```
df.reset_index(drop=True)
```

Out[45]:

	PAIS DE ORIGEM	DESCRICAO DO PRODUTO	VALOR UN.PROD.DOLAR	DESCRICAO	M
0	CHINA, REPUBLICA POP	007293# KIT TRANSMISSAO STANDARD TEMPERADO COM...	3.728000	honda cg 150 titan ks es mix fan	
1	CHINA, REPUBLICA POP	007295# KIT TRANSMISSAO STANDARD TEMPERADO COM...	3.700000	honda cg 125 titan ks es cargo	ca
2	CHINA, REPUBLICA POP	007296# KIT TRANSMISSAO STANDARD TEMPERADO COM...	3.686000	honda cg 125 fan	
3	CHINA, REPUBLICA POP	80341 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.343258	c100 biz	biz
4	CHINA, REPUBLICA POP	80364 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.826396	mirage 150	ka r
5	CHINA, REPUBLICA POP	80348 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	5.232801	cbx 250 twister	25 t
6	CHINA, REPUBLICA POP	80350 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	5.969694	crf 230	2
7	CHINA, REPUBLICA POP	80373 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.245770	shineray phoenix 50cc	ph sh
8	CHINA, REPUBLICA POP	80371 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.255806	pop 110	
9	CHINA, REPUBLICA POP	80370 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.245770	pop	
10	CHINA, REPUBLICA POP	80344 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	2.950439	c125 biz	1
11	CHINA, REPUBLICA POP	80360 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.284478	hunter max 125	sur
12	CHINA, REPUBLICA POP	80342 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.255806	c100 biz	biz

	PAIS DE ORIGEM	DESCRICAO DO PRODUTO	VALOR UN.PROD.DOLAR	DESCRICAO	M
13	CHINA, REPUBLICA POP	80385 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.265841	web	sur
14	CHINA, REPUBLICA POP	80345 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	6.756764	cb 250f	25 cb
15	CHINA, REPUBLICA POP	80372 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	4.022805	riva150 dafra	
16	CHINA, REPUBLICA POP	80392 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	4.907363	xt 250 tenere	ter ya
17	CHINA, REPUBLICA POP	80397 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.136813	yes intruder 125	in
18	CHINA, REPUBLICA POP	80395 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.136813	ybr	ya
19	CHINA, REPUBLICA POP	80346 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	4.828513	cb 300r	30 cb
20	CHINA, REPUBLICA POP	80367 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	5.034958	next 250 dafra	ya
21	CHINA, REPUBLICA POP	80363 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.668695	kansas 150	k
22	CHINA, REPUBLICA POP	80362 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.265841	jet 49cc	sh
23	CHINA, REPUBLICA POP	80356 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.314585	fan 125	1
24	CHINA, REPUBLICA POP	80393 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.992699	xtz 125	1
25	CHINA, REPUBLICA POP	80355 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.324620	fan 125	1

	PAIS DE ORIGEM	DESCRICAO DO PRODUTO	VALOR UN.PROD.DOLAR	DESCRICAO	M
26	CHINA, REPUBLICA POP	80391 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	5.103773	xre 300	3
27	CHINA, REPUBLICA POP	80379 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.314585	titan	
28	CHINA, REPUBLICA POP	80378 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.314585	titan fan 150	1
29	CHINA, REPUBLICA POP	80381 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	3.314585	titan 160	
...	...	...	...	...	
17454	CHINA, REPUBLICA POP	10530037 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	4.865000	nxr150bros	
17455	CHINA, REPUBLICA POP	10530040 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	4.170000	crosser 150	c
17456	CHINA, REPUBLICA POP	10530041 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	4.598000	fazer 150	ya
17457	CHINA, REPUBLICA POP	10530043 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	3.751000	crypton	c ya
17458	CHINA, REPUBLICA POP	10530046 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	4.440000	fan 125	1
17459	CHINA, REPUBLICA POP	10530048 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	4.792000	broz 160	16
17460	CHINA, REPUBLICA POP	10530049 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	6.489000	ninja300 chh	kav
17461	CHINA, REPUBLICA POP	10530050 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	3.634000	pop	
17462	CHINA, REPUBLICA POP	10530051 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	4.461000	titan 160	
17463	CHINA, REPUBLICA POP	10530054 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	4.950000	fazer250 chh	ya
17464	CHINA, REPUBLICA POP	10540002 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	8.252000	nxr150bros	

	PAIS DE ORIGEM	DESCRICAO DO PRODUTO	VALOR UN.PROD.DOLAR	DESCRICAO	M
17465	CHINA, REPUBLICA POP	10540012 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	8.177000	fazer250	ya
17466	CHINA, REPUBLICA POP	10540024 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	9.287000	xre 300	3
17467	CHINA, REPUBLICA POP	10540025 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	7.273000	crosser 150	c
17468	CHINA, REPUBLICA POP	10540026 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	7.800000	fazer 150	ya
17469	CHINA, REPUBLICA POP	10540029 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	8.044000	broz 160	16
17470	CHINA, REPUBLICA POP	10540031 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	7.463000	titan 160	
17471	CHINA, REPUBLICA POP	10540034 IN - KIT TRANSMISSAO P/MOTOCICLETAS(C...	8.400000	fazer250 chh-uo	ya
17472	CHINA, REPUBLICA POP	KIT DE TRANSMISSÃO AÇO BIZ 100 1045 COMPOSTO D...	3.258000	biz 1045 pro honda	
17473	CHINA, REPUBLICA POP	KIT DE TRANSMISSÃO AÇO BIZ 125/POP 100 1045, C...	3.019000	biz 125 pop 1045 pro honda 14	1
17474	CHINA, REPUBLICA POP	KIT DE TRANSMISSÃO AÇO BROS 150 1045, COMPOSTO...	4.159000	broz 150 pro honda nxr corren	15
17475	CHINA, REPUBLICA POP	KIT DE TRANSMISSÃO AÇO BROS 160 1045, COMPOSTO...	4.012000	broz 160 pro honda es ks	16
17476	CHINA, REPUBLICA POP	KIT DE TRANSMISSÃO AÇO FAN 125 2009 1045, COMP...	3.607000	fan 125 pro honda cg	
17477	CHINA, REPUBLICA POP	KIT DE TRANSMISSÃO AÇO SHIN/WEB 100 36DTS 1045...	2.980000	shin web ts pro sundown	su
17478	CHINA, REPUBLICA POP	KIT DE TRANSMISSÃO AÇO TITAN 150 1045, COMPOST...	3.648000	titan 150 pro honda cg ks es mix den	
17479	CHINA, REPUBLICA POP	KIT DE TRANSMISSÃO AÇO TITAN 160 1045, COMPOST...	3.747000	titan 160 pro honda cg	

	PAIS DE ORIGEM	DESCRICAO DO PRODUTO	VALOR UN.PROD.DOLAR	DESCRICAO	M
17480	CHINA, REPUBLICA POP	KIT DE TRANSMISSÃO AÇO TITAN 2000 1045, COMPOS...	3.620000	titan pro honda cg 125 ks es cargo	ca
17481	CHINA, REPUBLICA POP	KIT DE TRANSMISSÃO AÇO TITAN 99 1045 COMPOSTO ...	3.581000	titan 1045 pro honda cg 125	
17482	CHINA, REPUBLICA POP	KIT DE TRANSMISSÃO AÇO YBR 125 00/02 1045, COM...	3.450000	ybr 125 pro yamaha corre	ya
17483	CHINA, REPUBLICA POP	KIT DE TRANSMISSÃO AÇO YBR 125 03/05, COMPOSTO...	3.503000	ybr 125 pro yamaha factor	ya

17484 rows × 5 columns

In [46]:

```
df.shape
```

Out[46]:

```
(17484, 5)
```

Funcão final que transforma a DESCRICAO DO PRODUTO em Modelo para classificar

In [47]:

```
def criaModelo(descricao):
    descricao=limpaDescricao(descricao)
    descricao=achaPalavraChave(descricao)
    descricao=acrescentaMarca(descricao)
    return descricao
```

In [48]:

```
criaModelo(df.iloc[linha]['DESCRICAO DO PRODUTO'])
```

Out[48]:

```
'250 cb honda twister'
```

Exportando DataFrame com Modelos de aplicação

In [49]:

```
df.to_excel(r'./bases/dataframe_modelos.xlsx', index = False, header = True
)
```

Gerando a WordCloud com o campo Modelo

In [50]:

```
# Mescla todas as descrições como uma string usado espaço como separador
descricoes = " ".join(df['Modelo']).lower()
```

In [51]:

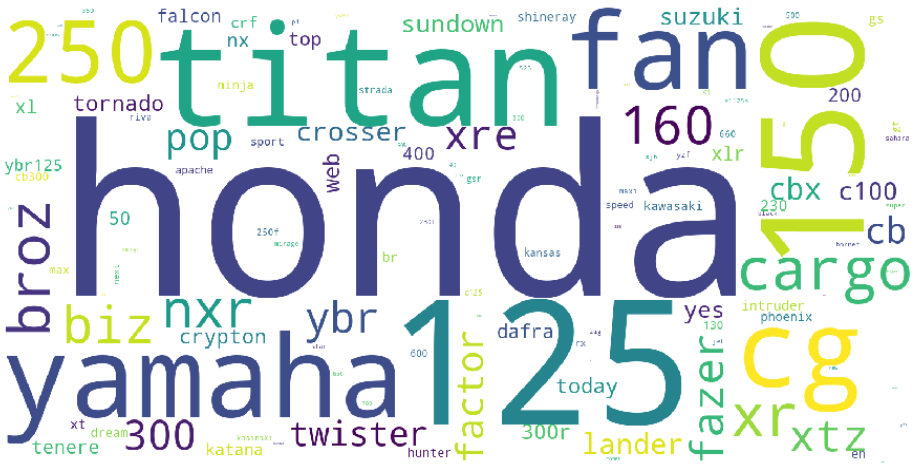
```
# Define e gera a wordcloud para um máximo de 400 palavras de tamanho mínim
o 2, sem termos duplos
wordcloud = WordCloud(stopwords=stopwords,
                      background_color="white",
                      width=1600, height=800,
                      max_words=400,
                      min_word_length=2,
                      collocations=False,
                      include_numbers=True).generate(descricoes)
```

In [52]:

```
# Exibe a imagem da nova WordCloud gerada
fig, ax = plt.subplots(figsize=(20,8))
ax.imshow(wordcloud, interpolation='bilinear')
ax.set_axis_off()
plt.imshow(wordcloud)
```

Out[52]:

```
<matplotlib.image.AxesImage at 0x257b7ae46a0>
```



In [53]:

```
# Exporta para um aquivo
wordcloud.to_file("./imagens/wordcloud_descricoes_final.png")
```

Out[53]:

```
<wordcloud.wordcloud.WordCloud at 0x257b5059470>
```

In [54]:

```
tempotot=time.time()-initot
if tempotot>60:
    print(f'Tempo total de execução: {tempotot/60:.2f} minutos.')
else:
    print(f'Tempo total de execução: {tempotot:.2f} segundos.')
```

```
Tempo total de execução: 16.82 minutos.
```