Notebook Jupyter 2_geraWordclouds

Geração das WordClouds das descrições

Os dados gerados após o tratamento, objeto de estudo, estão contidos em um campo de descrição, onde o importador ou seu representante faz a entrada livre de dados, sem qualquer exigência ou padronização, nossa análise exploratória foi convincente no sentido de que deveria ser feito um tratamento com a utilização de Programação de Linguagem Natural – PLN. Dentro dessa análise exploratória de dados, observou-se a ocorrência de palavras que, embora frequentes, serão irrelevantes para a determinação da diferenciação entre os itens que estão classificados dentro do agrupamento representado pela NCM em estudo. Comum no estudo de categorização de textos, o uso de stopwords é recomendado nesse caso, pois da mesma forma que artigos, pronomes e outras palavras tradicionalmente identificadas como não relevantes para a distinção, termos como kit, transmissão, corrente, coroa e pinhão também são igualmente irrelevantes para distinguir um item de outro. Na biblioteca Natural Language Processing Toolkit - nltk já existe uma lista de stopwords para a língua portuguesa. Como essa biblioteca já possui a funcionalidade de complementação dessa lista, adicionou-se os itens kit, transmissão, corrente, coroa e pinhão; termos que pela sua característica não distinguem os itens dentro do dataset estudado.

Importa os dados já tratados

In [1]:

```
import pandas as pd
import time
```

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 17:14
```

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
16225	CHINA, REPUBLICA POP	10530041 IN KIT TRANSMISSAO P/MOTOCICLETAS(COR...	4.3750
10597	CHINA, REPUBLICA POP	KIT DE TRANSMISSAO (1045) COMPOSTO DE CORRENTE...	3.8710
10164	CHINA, REPUBLICA POP	KIT DE TRANSMISSAO , MARCA RIFFEL, TITANIUM (1...	4.5657
12269	CHINA, REPUBLICA POP	item 05;Partes e peças para Motocicletas,Desta...	4.7022
15584	CHINA, REPUBLICA POP	ENGRENAGENS PARA TRANSMISSÃO DE MOTOCICLETAS E...	6.3900

Importa as stopwords da língua portuguesa

In [6]:

```
# Importar lista de Stopwords
from nltk.corpus import stopwords
stopwords = set(stopwords.words('portuguese'))
```

In [7]:

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

Out[7]:

```
204
```

In [8]:

```
# 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']
```

Instala a biblioteca wordcloud e importa as bibliotecas necessárias

In [9]:

```
# Caso já esteja instalada atualiza
!pip install wordcloud -q
```

In [10]:

```
import matplotlib.pyplot as plt
from wordcloud import WordCloud
```

Cria a string contendo todas as descrições

In [11]:

```
# Mostra algumas descrições do dataset
df['DESCRICAO DO PRODUTO'].sample(5)
```

Out[11]:

```
3427     ITEM 100240WR - KIT TRANSMISSÃO PARA MOTOCICLE...
17013    KIT DE TRANSMISSAO, MARCA RIFFEL, TITANIUM (10...
13289    KIT DE TRANSMISSAO PARA MOTOCICLETAS MODELO: N...
5509     KIT DE TRANSMISSAO , MARCA RIFFEL, TITANIUM (1...
3212     71788 - KIT TRANSMISSAO TITANIUM PARA MOTOCICL...
Name: DESCRICAO DO PRODUTO, dtype: object
```

In [12]:

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

Gera a WodCloud aplicando o filtro das stopwords

In [13]:

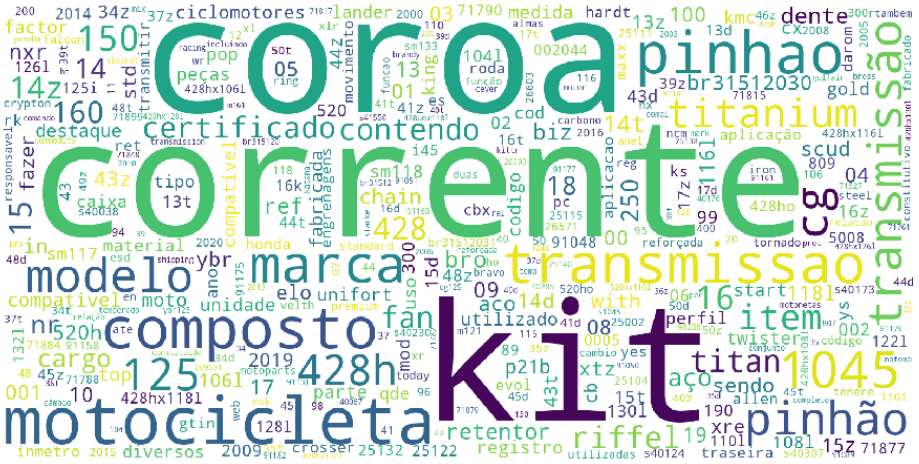
```
# Define e gera a wordcloud para um máximo de 400 palavras de tamanho mínim
o 2, sem termos duplos
wordcloud = WordCloud(stopwords=stopwords,
                      background_color="white",
                      width=1600, height=800,
                      max_words=400,
                      min_word_length=2,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

In [14]:

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

Out[14]:

```
<matplotlib.figure.AxesImage at 0x165c41c65f8>
```



In [15]:

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

Out[15]:

```
<wordcloud.wordcloud.WordCloud at 0x165c41ba748>
```

Atualiza as stopwords contendo as palavras irrelevantes

Palavras a adicionar, tais como: kit, transmissao, transmissão, coroa, pinhão, etc. que não agregam nenhuma diferença aos itens da lista

In [16]:

```
# 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 [17]:

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

In [18]:

```
# 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', 'tracão', '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']
```

Gera a WodCloud aplicando o filtro das stopwords atualizado

In [19]:

```
# 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 [20]:

```
# 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[20]:

```
<matplotlib.image.AxesImage at 0x165c42132b0>
```

In [21]:

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

Out[21]:

```
<wordcloud.wordcloud.WordCloud at 0x165c4208fd0>
```

In [22]:

```
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: 15.40 segundos.
```