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

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.localtime(tim
e.time()))}')
```

```
Código executado em 25/01/2022 às 11:25
```

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
3109	CHINA, REPUBLICA POP	KIT DE TRANSMISSAO , MARCA RIFFEL, TITANIUM (1...	6.708000
10980	CHINA, REPUBLICA POP	Ý10821/i45¨ KIT DE TRANSMISSÃO, EM AÇO 1045, C...	2.950000
6336	CHINA, REPUBLICA POP	878193 - KIT DE TRANSMISSÃO, COMPOSTO DE CORRE...	23.945997
14741	CHINA, REPUBLICA POP	1104970 - DAROM / Kit de transmissao 428 - 110...	4.260000
1734	CHINA, REPUBLICA POP	80350 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENTE...	5.925957

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', 'at
é', 'com', 'como', 'da', 'das', 'de', 'dela', 'delas', 'dele', 'deles', 'depois',
'do', 'dos', 'e', 'ela', 'elas', 'ele', 'eles', 'em', 'entre', 'era', 'eram', 'ess
a', 'essas', 'esse', 'esses', 'esta', 'estamos', 'estas', 'estava', 'estavam', 'es
te', 'esteja', 'estejam', 'estejamos', 'estes', 'esteve', 'estive', 'estivemos',
'estiver', 'estivera', 'estiveram', 'estiverem', 'estivermos', 'estivesse', 'estiv
essem', 'estivéramos', 'estivéssemos', 'estou', 'está', 'estávamos', 'estão', 'e
u', 'foi', 'fomos', 'for', 'fora', 'foram', 'forem', 'formos', 'fosse', 'fossem',
'fui', 'fôramos', 'fôssemos', 'haja', 'hajam', 'hajamos', 'havemos', 'hei', 'houv
e', 'houvemos', 'houver', 'houvera', 'houveram', 'houverei', 'houverem', 'houverem
os', 'houveria', 'houveriam', 'houvermos', 'houverá', 'houverão', 'houveríamos',
'houvesse', 'houvessem', 'houvéramos', 'houvéssemos', 'há', 'hão', 'isso', 'isto',
'já', 'lhe', 'lhes', 'mais', 'mas', 'me', 'mesmo', 'meu', 'meus', 'minha', 'minha
s', 'muito', 'na', 'nas', 'nem', 'no', 'nos', 'nossa', 'nossas', 'nosso', 'nosso
s', 'num', 'numa', 'não', 'nós', 'o', 'os', 'ou', 'para', 'pela', 'pelas', 'pelo',
'pelos', 'por', 'qual', 'quando', 'que', 'quem', 'se', 'seja', 'sejam', 'sejamos',
'sem', 'serei', 'seremos', 'seria', 'seriam', 'será', 'serão', 'seríamos', 'seu',
'seus', 'somos', 'sou', 'sua', 'suas', 'são', 'só', 'também', 'te', 'tem', 'temo
s', 'tenha', 'tenham', 'tenhamos', 'tenho', 'terei', 'teremos', 'teria', 'teriam',
'terá', 'terão', 'teríamos', 'teu', 'teus', 'teve', 'tinha', 'tinham', 'tive', 'ti
vemos', 'tiver', 'tivera', 'tiveram', 'tiverem', 'tivermos', 'tivesse', 'tivesse
m', 'tivéramos', 'tivéssemos', 'tu', 'tua', 'tuas', 'tém', 'tínhamos', 'um', 'um
a', 'você', 'vocês', 'vos', 'à', 'às', 'é', 'éramos']
```

Cria a string contendo todas as descrições

In [9]:
```
# Mostra algumas descrições do dataset
df['DESCRICAO DO PRODUTO'].sample(5)
```

Out[9]:
```
12040    KIT TRANSMISSÃO COROA,CORRENTE E PINHAO PARA M...
16723    KIT DE TRANSMISSAO P/MOTOCICLETA MOD: XLR 125;...
16761    KIT DE TRANSMISSAO, MARCA RIFFEL, TITANIUM (10...
3674     item .12;Partes e peças para Motocicletas,Dest...
11420    980727 KIT DE TRANSMISSÃO, COMPOSTO DE CORRENT...
Name: DESCRICAO DO PRODUTO, dtype: object
```

In [10]:
```
# 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 [11]:
```
# Define e gera a wordcloud para um máximo de 400 palavras de tamanho mínimo 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 [12]:
```
# 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[12]:
```
<matplotlib.figure.AxesImage at 0x266cd0df3c8>
```

In [13]:
```
# Exporta para um aquivo
wordcloud.to_file(r"./imagens/wordcloud_descricoes_antes.png")
```

Out[13]:
```
<wordcloud.wordcloud.WordCloud at 0x266cd169eb8>
```

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 [14]:
```
# Palavras a adicionar na lista de stopwords estão contidas em um arquivo csv exter
no
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', 'acondicionados', 'adaptavel', 'adap
tável', 'allen', 'almas', 'alta', 'am', 'anel', 'ano', 'aplicacao', 'aplication',
'aplicavel', 'aplicação', 'aplicável', 'application', 'ate', 'atitanium', 'aç', 'a
ço', 'bicicleta', 'bicycle', 'bike', 'bravo', 'cada', 'caixa', 'caixas', 'cambio',
'carbono', 'certificado', 'cever', 'chain', 'chh', 'china', 'ciclomotor', 'ciclomo
tores', 'cilindrada', 'cilindradas', 'cod', 'code', 'codigo', 'comando', 'combustã
o', 'comercial', 'comercialmente', 'commodity', 'compativel', 'compatível', 'comp
l', 'completo', 'completos', 'composto', 'composto', 'compostopor', 'compostpo',
'comum', 'condicao', 'condicoes', 'condição', 'condições', 'confeccionado', 'confo
rmidade', 'conhecido', 'conj', 'conjunto', 'conjuntos', 'constituido', 'constituti
vo', 'constituído', 'contendo', 'coposto', 'coroa', 'coroaes', 'corr', 'corrent',
'corrente', 'correntee', 'correntes', 'corrnte', 'cx', 'câmbio', 'cãmbio', 'códig
o', 'decreto', 'denominada', 'dente', 'dentes', 'descricao', 'descrição', 'descriç
ao', 'descrição', 'destaque', 'destaque', 'destaques', 'detransmissão', 'diante',
'dimensao', 'dimensoes', 'dimensão', 'dimensões', 'diverso', 'diversos', 'dominad
o', 'durabilidade', 'elo', 'elos', 'embalagem', 'engine', 'engrenagem', 'engrenage
ns', 'epinhao', 'epinhão', 'especifico', 'específico', 'espessura', 'evol', 'exclu
sivo', 'fabr', 'fabri', 'fabricada', 'fabricado', 'final', 'foiproduzida', 'formad
o', 'funcao', 'funcao', 'funcão', 'funcão', 'funçao', 'função', 'gtin', 'hardt',
'ho', 'hp', 'imetro', 'in', 'incluindo', 'incluso', 'indicado', 'ingles', 'inmetr
o', 'inv', 'invoice', 'iron', 'item', 'jc', 'ki', 'kif', 'kit', 'kitr', 'kittr',
'kittr', 'kmc', 'ligacoes', 'ligaçoes', 'ligaçoes', 'ligações', 'marca', 'mark',
'match', 'material', 'materialdo', 'maxx', 'medida', 'medidas', 'medindo', 'meta
l', 'mini', 'mod', 'modelo', 'modelos', 'moto', 'motociclet', 'motocicleta', 'moto
cicletas', 'motoneta', 'motonetas', 'motoparts', 'motor', 'motorcicleta', 'motos',
'motos', 'movimento', 'nbsp', 'ncm', 'nome', 'nopinh', 'normais', 'nova', 'novo',
'nr', 'numero', 'número', 'onde', 'or', 'origem', 'oring', 'ox', 'papelao', 'papel
ão', 'part', 'parte', 'partes', 'parts', 'pc', 'pc-coroa', 'pc-corrente', 'pc-pinh
ao', 'pcs', 'pec', 'pecas', 'perfil', 'peç', 'peças', 'pinh', 'pinhao', 'pinhão',
'posição', 'premium', 'procedencia', 'procedência', 'prodepe', 'produto', 'própri
o', 'pç', 'qdes', 'qtd', 'qtds', 'qty', 'quadriciclo', 'quadriciclos', 'quantida
d', 'quantidade', 're', 'ref', 'reforcada', 'reforçada', 'registro', 'rel', 'relac
ao', 'relação', 'reposicao', 'resp', 'respo', 'respons', 'responsa', 'responsav',
'responsave', 'responsavel', 'responsáv', 'responsável', 'ret', 'retalho', 'retent
or', 'riffel', 'ring', 'roda', 'sae', 'scud', 'semi', 'semi-', 'semi-kit', 'send
o', 'serve', 'set', 'shipping', 'sistema', 'sm', 'sprocket', 'standard', 'standar
t', 'standartt', 'std', 'stdmodelo', 'steel', 'tambem', 'também', 'tec', 'temp',
'temperado', 'tipo', 'tipos', 'titanio', 'titaniu', 'titanium', 'titaniun', 'tr',
'tracao', 'tracão', 'trans', 'transmis', 'transmisao', 'transmiss', 'transmissa',
'transmissao', 'transmission', 'transmissão', 'transmitir', 'traseira', 'traçao',
'tração', 'und', 'unds', 'unid', 'unidade', 'unidade', 'unidades', 'unifort', 'u
o', 'uso', 'utilizada', 'utilizadas', 'utilizado', 'utilizados', 'utilização', 've
m', 'venda', 'with', 'xy', 'year']
```

In [15]:
```
# Atualizar stopwords
stopwords.update(stopwords_df)
```

In [16]:
```
# Mostra toda a lista de stopwords
swtemp = list(stopwords)
swtemp.sort()
print(swtemp)
```

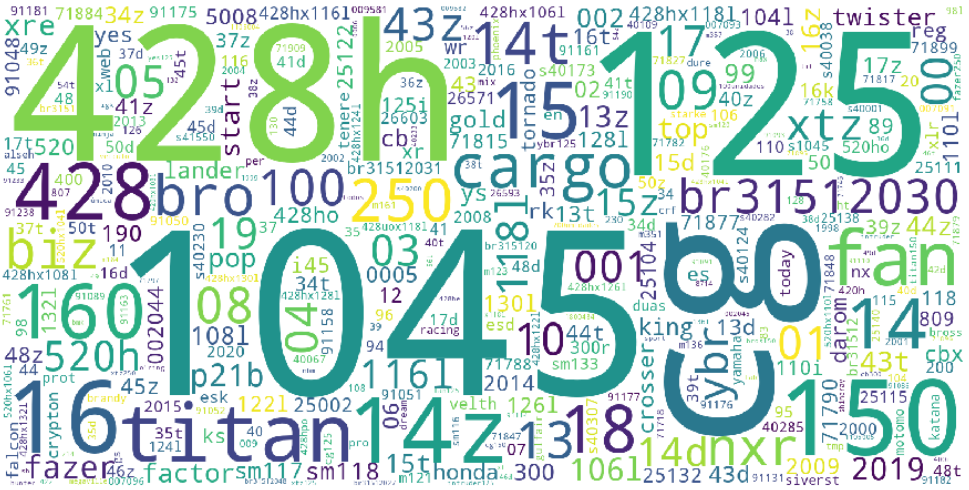
```
['a', 'abaixo', 'acessorios', 'acessórios', 'aco', 'acondicionados', 'adaptavel',
'adaptável', 'allen', 'almas', 'alta', 'am', 'anel', 'ano', 'ao', 'aos', 'aplicaca
o', 'aplication', 'aplicavel', 'aplicação', 'aplicável', 'application', 'aquela',
'aquelas', 'aquele', 'aqueles', 'aquilo', 'as', 'ate', 'atitanium', 'até', 'aç',
'aço', 'bicicleta', 'bicycle', 'bike', 'bravo', 'cada', 'caixa', 'caixas', 'cambi
o', 'carbono', 'certificado', 'cever', 'chain', 'chh', 'china', 'ciclomotor', 'cic
lomotores', 'cilindrada', 'cilindradas', 'cod', 'code', 'codigo', 'com', 'comand
o', 'combustão', 'comercial', 'comercialmente', 'commodity', 'como', 'compativel',
'compatível', 'compl', 'completo', 'completos', 'composto', 'compostopor', 'compos
tpo', 'comum', 'condicao', 'condicoes', 'condição', 'condições', 'confeccionado',
'conformidade', 'conhecido', 'conj', 'conjunto', 'conjuntos', 'constituido', 'cons
titutivo', 'constituído', 'contendo', 'coposto', 'coroa', 'coroaes', 'corr', 'corr
ent', 'corrente', 'correntee', 'correntes', 'corrnte', 'cx', 'câmbio', 'cãmbio',
'código', 'da', 'das', 'de', 'decreto', 'dela', 'delas', 'dele', 'deles', 'denomin
ada', 'dente', 'dentes', 'depois', 'descricao', 'descrição', 'descriçao', 'descriç
ão', 'destaque', 'destaques', 'detransmissão', 'diante', 'dimensao', 'dimensoes',
'dimensão', 'dimensões', 'diverso', 'diversos', 'do', 'dominado', 'dos', 'durabili
dade', 'e', 'ela', 'elas', 'ele', 'eles', 'elo', 'elos', 'em', 'embalagem', 'engin
e', 'engrenagem', 'engrenagens', 'entre', 'epinhao', 'epinhão', 'era', 'eram', 'es
pecifico', 'específico', 'espessura', 'essa', 'essas', 'esse', 'esses', 'esta', 'e
stamos', 'estas', 'estava', 'estavam', 'este', 'esteja', 'estejam', 'estejamos',
'estes', 'esteve', 'estive', 'estivemos', 'estiver', 'estivera', 'estiveram', 'est
iverem', 'estivermos', 'estivesse', 'estivessem', 'estivéramos', 'estivéssemos',
'estou', 'está', 'estávamos', 'estão', 'eu', 'evol', 'exclusivo', 'fabr', 'fabri',
'fabricada', 'fabricado', 'final', 'foi', 'foiproduzida', 'fomos', 'for', 'fora',
'foram', 'forem', 'formado', 'formos', 'fosse', 'fossem', 'fui', 'funcao', 'funcã
o', 'funçao', 'função', 'fôramos', 'fôssemos', 'gtin', 'haja', 'hajam', 'hajamos',
'hardt', 'havemos', 'hei', 'ho', 'houve', 'houvemos', 'houver', 'houvera', 'houver
am', 'houverei', 'houverem', 'houveremos', 'houveria', 'houveriam', 'houvermos',
'houverá', 'houverão', 'houveríamos', 'houvesse', 'houvessem', 'houvéramos', 'houv
éssemos', 'hp', 'há', 'hão', 'imetro', 'in', 'incluindo', 'incluso', 'indicado',
'ingles', 'inmetro', 'inv', 'invoice', 'iron', 'isso', 'isto', 'item', 'jc', 'já',
'ki', 'kif', 'kit', 'kitr', 'kittr', 'kmc', 'lhe', 'lhes', 'ligacoes', 'ligaçoes',
'ligações', 'mais', 'marca', 'mark', 'mas', 'match', 'material', 'materialdo', 'ma
xx', 'me', 'medida', 'medidas', 'medindo', 'mesmo', 'metal', 'meu', 'meus', 'minh
a', 'minhas', 'mini', 'mod', 'modelo', 'modelos', 'moto', 'motociclet', 'motocicle
ta', 'motocicletas', 'motoneta', 'motonetas', 'motoparts', 'motor', 'motorciclet
a', 'motos', 'movimento', 'muito', 'na', 'nas', 'nbsp', 'ncm', 'nem', 'no', 'nom
e', 'nopinh', 'normais', 'nos', 'nossa', 'nossas', 'nosso', 'nossos', 'nova', 'nov
o', 'nr', 'num', 'numa', 'numero', 'não', 'nós', 'número', 'o', 'onde', 'or', 'ori
gem', 'oring', 'os', 'ou', 'ox', 'papelao', 'papelão', 'para', 'part', 'parte', 'p
artes', 'parts', 'pc', 'pc-coroa', 'pc-corrente', 'pc-pinhao', 'pcs', 'pec', 'peca
s', 'pela', 'pelas', 'pelo', 'pelos', 'perfil', 'peç', 'peças', 'pinh', 'pinhao',
'pinhão', 'por', 'posição', 'premium', 'procedencia', 'procedência', 'prodepe', 'p
roduto', 'próprio', 'pç', 'qdes', 'qtd', 'qtds', 'qty', 'quadriciclo', 'quadricicl
os', 'qual', 'quando', 'quantidad', 'quantidade', 'que', 'quem', 're', 'ref', 'ref
orcada', 'reforçada', 'registro', 'rel', 'relacao', 'relação', 'reposicao', 'res
p', 'respo', 'respons', 'responsa', 'responsav', 'responsave', 'responsavel', 'res
ponsáv', 'responsável', 'ret', 'retalho', 'retentor', 'riffel', 'ring', 'roda', 's
ae', 'scud', 'se', 'seja', 'sejam', 'sejamos', 'sem', 'semi', 'semi-', 'semi-kit',
'sendo', 'serei', 'seremos', 'seria', 'seriam', 'serve', 'será', 'serão', 'seríamo
s', 'set', 'seu', 'seus', 'shipping', 'sistema', 'sm', 'somos', 'sou', 'sprocket',
'standard', 'standart', 'standartt', 'std', 'stdmodelo', 'steel', 'sua', 'suas',
'são', 'só', 'tambem', 'também', 'te', 'tec', 'tem', 'temos', 'temp', 'temperado',
'tenha', 'tenham', 'tenhamos', 'tenho', 'terei', 'teremos', 'teria', 'teriam', 'te
rá', 'terão', 'teríamos', 'teu', 'teus', 'teve', 'tinha', 'tinham', 'tipo', 'tipo
s', 'titanio', 'titaniu', 'titanium', 'titaniun', 'tive', 'tivemos', 'tiver', 'tiv
era', 'tiveram', 'tiverem', 'tivermos', 'tivesse', 'tivessem', 'tivéramos', 'tivés
semos', 'tr', 'tracao', 'tracão', 'trans', 'transmis', 'transmisao', 'transmiss',
'transmissa', 'transmissao', 'transmission', 'transmissão', 'transmitir', 'traseir
a', 'traçao', 'tração', 'tu', 'tua', 'tuas', 'tém', 'tínhamos', 'um', 'uma', 'un
d', 'unds', 'unid', 'unidade', 'unidades', 'unifort', 'uo', 'uso', 'utilizada', 'u
tilizadas', 'utilizado', 'utilizados', 'utilização', 'vem', 'venda', 'você', 'você
s', 'vos', 'with', 'xy', 'year', 'à', 'às', 'é', 'éramos']
```

Gera a WodCloud aplicando o filtro das stopwords atualizado

In [17]:
```
# Define e gera a wordcloud para um máximo de 400 palavras de tamanho mínimo 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 [18]:
```
# 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[18]: <matplotlib.image.AxesImage at 0x266ccdf9080>



In [19]:
```
# Exporta para um aquivo
wordcloud.to_file("./imagens/wordcloud_descricoes_depois.png")
```

Out[19]: <wordcloud.wordcloud.WordCloud at 0x266ccdd27b8>

In [20]:
```
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: 13.05 segundos.
```