Aula_1_DTS_PLN_Exercício_2_corrigido

June 30, 2025

#Exercícios - Aula 1

0.1 2) Utilizando o dataset de produtos [1]:

[1]:
```
import pandas as pd

df = pd.read_csv(
    "https://dados-ml-pln.s3-sa-east-1.amazonaws.com/produtos.csv",
    delimiter=";",
    encoding='utf-8' )
```

[2]:
```
df.head()
```

[2]:
```
                                                nome  \
0                            O Hobbit - 7ª Ed. 2013
1                  Livro - It A Coisa - Stephen King
2   Box  As Crônicas De Gelo E Fogo  Pocket  5 Li…
3                                  Box Harry Potter
4                          Livro Origem - Dan Brown

                                            descricao categoria
0  Produto NovoBilbo Bolseiro é um hobbit que lev…     livro
1  Produto NovoDurante as férias escolares de 195…     livro
2  Produto NovoTodo o reino de Westeros ao alcanc…     livro
3  Produto Novo e Físico  A série Harry Potter ch…     livro
4  Produto NovoDe Onde Viemos? Para Onde Vamos? R…     livro
```

2.1. Elimine linhas com valores nulos

[3]:
```
df.dropna(inplace=True)
```

2.2. Adicione uma nova coluna chamada texto, formada pela composição das colunas nome e descrição

[4]:
```
df["texto"] = df['nome'] + " " + df['descricao'] # cria uma nova culuna com os␣
↪valores concatenados
```

[5]:
```
df.head()
```

[5]:
```
                                                nome  \
0                                O Hobbit - 7ª Ed. 2013
1                      Livro - It A Coisa - Stephen King
2   Box  As Crônicas De Gelo E Fogo  Pocket  5 Li…
3                                    Box Harry Potter
4                             Livro Origem - Dan Brown

                                              descricao categoria  \
0  Produto NovoBilbo Bolseiro é um hobbit que lev…     livro
1  Produto NovoDurante as férias escolares de 195…     livro
2  Produto NovoTodo o reino de Westeros ao alcanc…     livro
3  Produto Novo e Físico  A série Harry Potter ch…     livro
4  Produto NovoDe Onde Viemos? Para Onde Vamos? R…     livro

                                               texto
0   O Hobbit - 7ª Ed. 2013  Produto NovoBilbo Bol…
1   Livro - It A Coisa - Stephen King  Produto No…
2   Box  As Crônicas De Gelo E Fogo  Pocket  5 Li…
3   Box Harry Potter  Produto Novo e Físico  A sé…
4   Livro Origem - Dan Brown  Produto NovoDe Onde…
```

2.3. Quantos Unigramas existem antes e depois de remover stopwords (use a coluna texto)

[6]:
```
from sklearn.feature_extraction.text import CountVectorizer # Converte uma␣
 ↪coleção de documentos de texto em uma matriz de contagens de tokens

vect = CountVectorizer(ngram_range=(1,1))
vect.fit(df.texto)
text_vect = vect.transform(df.texto)

print('UNIGRAMAS com STOPWORDS', text_vect.shape[1])
```

```
UNIGRAMAS com STOPWORDS 35466
```

[7]:
```
len(vect.get_feature_names_out())
```

[7]: 35466

[8]:
```
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data…
[nltk_data]   Unzipping corpora/stopwords.zip.
```

[8]: True

[9]:
```
from sklearn.feature_extraction.text import CountVectorizer
```

```
stopwords = nltk.corpus.stopwords.words('portuguese')

vect = CountVectorizer(ngram_range=(1,1), stop_words=stopwords)
vect.fit(df.texto)
text_vect = vect.transform(df.texto)

print('UNIGRAMAS sem STOPWORDS', text_vect.shape[1])
```

```
UNIGRAMAS sem STOPWORDS 35307
```

2.4. Quantos Bigramas existem antes e depois de remover stopwords (use a coluna texto)

[10]:
```
from sklearn.feature_extraction.text import CountVectorizer

vect = CountVectorizer(ngram_range=(2,2))
vect.fit(df.texto)
text_vect = vect.transform(df.texto)

print('BIGRAMAS com STOPWORDS', text_vect.shape[1])
```

```
BIGRAMAS com STOPWORDS 159553
```

[11]:
```
from sklearn.feature_extraction.text import CountVectorizer

stopwords = nltk.corpus.stopwords.words('portuguese')

vect = CountVectorizer(ngram_range=(2,2), stop_words=stopwords)
vect.fit(df.texto)
text_vect = vect.transform(df.texto)

print('BIGRAMAS sem STOPWORDS', text_vect.shape[1])
```

```
BIGRAMAS sem STOPWORDS 145224
```

2.5. Quantos Trigramas existem antes e depois de remover stopwords (use a coluna texto)

[12]:
```
from sklearn.feature_extraction.text import CountVectorizer

vect = CountVectorizer(ngram_range=(3,3))
vect.fit(df.texto)
text_vect = vect.transform(df.texto)

print('TRIGRAMAS com STOPWORDS', text_vect.shape[1])
```

```
TRIGRAMAS com STOPWORDS 228162
```

[13]:
```
from sklearn.feature_extraction.text import CountVectorizer
```

```
stopwords = nltk.corpus.stopwords.words('portuguese')

vect = CountVectorizer(ngram_range=(3,3), stop_words=stopwords)
vect.fit(df.texto)
text_vect = vect.transform(df.texto)

print('TRIGRAMAS sem STOPWORDS', text_vect.shape[1])
```

```
TRIGRAMAS sem STOPWORDS 177377
```

2.6. Quantos unigramas existem na coluna texto após aplicar Stemmer (utilize rslp)

[14]:
```
from nltk.tokenize import word_tokenize
import nltk
nltk.download('punkt')
nltk.download('punkt_tab')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data…
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package punkt_tab to /root/nltk_data…
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
```

[14]:
```
True
```

[15]:
```
df['tokens'] = df.texto.apply(word_tokenize)
```

[16]:
```
df.head()
```

[16]:

	nome	descricao	categoria	texto
0	O Hobbit - 7ª Ed. 2013	Produto NovoBilbo Bolseiro é um hobbit que lev…	livro	O Hobbit - 7ª Ed. 2013 Produto NovoBilbo Bol…
1	Livro - It A Coisa - Stephen King	Produto NovoDurante as férias escolares de 195…	livro	Livro - It A Coisa - Stephen King Produto No…
2	Box As Crônicas De Gelo E Fogo Pocket 5 Li…	Produto NovoTodo o reino de Westeros ao alcanc…	livro	Box As Crônicas De Gelo E Fogo Pocket 5 Li…
3	Box Harry Potter	Produto Novo e Físico A série Harry Potter ch…	livro	Box Harry Potter Produto Novo e Físico A sé…
4	Livro Origem - Dan Brown	Produto NovoDe Onde Viemos? Para Onde Vamos? R…	livro	Livro Origem - Dan Brown Produto NovoDe Onde…

```
                                                tokens
0  [O, Hobbit, -, 7ª, Ed, ., 2013, Produto, NovoB…
1  [Livro, -, It, A, Coisa, -, Stephen, King, Pro…
2  [Box, As, Crônicas, De, Gelo, E, Fogo, Pocket,…
3  [Box, Harry, Potter, Produto, Novo, e, Físico,…
4  [Livro, Origem, -, Dan, Brown, Produto, NovoDe…
```

[17]:
```
from nltk.stem.rslp import RSLPStemmer
import nltk
nltk.download('rslp')

rslp = RSLPStemmer()

def stem_pandas(doc):
  return ' '.join([rslp.stem(token) for token in doc])

df['stemmer'] = df.tokens.apply(stem_pandas)

df.stemmer.head()
```

```
[nltk_data] Downloading package rslp to /root/nltk_data…
[nltk_data]   Unzipping stemmers/rslp.zip.
```

[17]:
```
0    o hobbit - 7ª ed . 2013 produt novobilb bols é…
1    livr - it a cois - stephen king produt novodur…
2    box as crôn de gel e fog pocket 5 livr produt …
3    box harry pott produt nov e físic a séri harry…
4    livr orig - dan brown produt novod ond vi ? pa…
Name: stemmer, dtype: object
```

[18]:
```
from sklearn.feature_extraction.text import CountVectorizer

vect = CountVectorizer(ngram_range=(1,1))
vect.fit(df.stemmer)
text_vect = vect.transform(df.stemmer)

print('UNIGRAMAS com STOPWORDS', text_vect.shape[1])
```

```
UNIGRAMAS com STOPWORDS 26532
```

[20]:
```
nltk.download('stopwords')
stopwords = nltk.corpus.stopwords.words('portuguese')

vect = CountVectorizer(ngram_range=(1,1), stop_words=stopwords)
vect.fit(df.stemmer)

text_vect = vect.transform(df.stemmer)
```

```
print('UNIGRAMAS sem STOPWORDS', text_vect.shape[1])
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
```

```
UNIGRAMAS sem STOPWORDS 26465
```

Documento/Texto: “um dois três quatro”

- Unigrama ["um","dois","três","quatro"], temos 4 unigramas.
- Bigrama ["um dois","dois três","três quatro"], temos 3 bigramas.
- Trigrama ["um dois três","dois três quatro"], temos 2 trigramas.
- 4-grama ["um dois três quatro"], temos um 4-grama.

Documento/Texto: “um dois três quatro um três um”

- Unigrama ["um","dois","três","quatro"], temos 4 unigramas.
- Bigrama ["um dois","dois três","três quatro","quatro um","um três","três um"], temos 6 bigr
- Trigrama ["um dois três","dois três quatro","três quatro um","quatro um três","um três um"]
- 4-grama ["um dois três quatro","dois três quatro um","três quatro um três","quatro um três