MINERAÇÃO DE TEXTO (Parte 1) - WORDCLOUD

Livro: The Amateur Emigrant de Robert Louis Stevenson

Resumo: Em 1879, o escritor escocês Robert Louis Stevenson (1850-1894) viajou da Grã-Bretanha para os Estados Unidos, para encontrar-se com a mulher que amava, a americana Funny Vandegrift Osbourne. Ele embarcou num navio, em Glasgow, rumo a Nova York e, de lá, pegou um trem para São Francisco. "The Amateur Emigrant" cobre a primeira parte dessa longa viagem e foi publicada em 1895, um ano depois da morte do escritor. Em sua narrativa, Stevenson conta como eram as condições dos viajantes de todas as classes do navio, dando destaque para os que embarcaram na classe inferior; a maioria, composta de trabalhadores desempregados que procuravam por melhores condições de vida no Novo Mundo.

Importando as bibliotecas:

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
from PIL import Image
```

Importando o texto (dados não estruturados):

Domínio público e disponível no site Project Gutenberg.

In [2]:
```
df = pd.read_csv('data/TheAmateurEmigrant.txt',sep='\t')\
    .dropna()

df.head(10)
```

Out[2]:

	text
0	THE AMATEUR EMIGRANT
1	THE SECOND CABIN
2	I first encountered my fellow-passengers on th...
3	Thence we descended the Clyde in no familiar s...
4	on each other as on possible enemies. A few S...
5	already grown acquainted on the North Sea, wer...
6	their long pipes; but among English speakers d...
7	reigned supreme. The sun was soon overclouded...
8	grew sharp as we continued to descend the wide...
9	falling temperature the gloom among the passen...

Preparando o texto:

In [3]:
```
textoRLS = df['text']
```

In [4]:
```
texto_taem = " ".join(w for w in textoRLS)
```

In [5]:
```
stopwords = set(STOPWORDS)
```

Criando a wordcloud:

In [6]:
```
def plot_wordcloud(wc):
    fig, ax = plt.subplots(figsize=(14,7))
    ax.imshow(wc, interpolation='bilinear')
    ax.set_axis_off()
    plt.imshow(wc)
```

In [7]:
```
colors = ['darkblue', 'gray', 'lightblue']
meu_cmap = ListedColormap(sns.color_palette(colors).as_hex())
```

In [8]:
```
wc = WordCloud(stopwords=stopwords,
               colormap=meu_cmap,
               background_color='white',
               width=1600,
               height=800).generate(texto_taem)
```

In [9]:
```
plot_wordcloud(wc)
```

Se o editor não conhecer a história de um livro com antecedência, ele pode gerar uma wordcloud para ter uma ideia do que o seu tema principal se trata.