MINERAÇÃO DE TEXTO

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 americana Funny Vandegrift Osbourne. Em Glasgow, ele embarcou num navio rumo a Nova York e de lá, ele pegou um trem para São Francisco, onde a sua amada morava. "The Amateur Emigrant" cobre a parte da viagem do escritor feita no navio e foi publicada em 1895, um ano depois de sua morte. Em sua narrativa, Stevenson conta como eram as condições dos viajantes de todas as classes, destacando os que embarcaram na classe inferior; grupo composto, em sua maioria, por europeus desempregados que procuravam por melhores condições de vida no Novo Mundo. Essa história faz parte do livro "Essays of Travel", um domínio público que encontra-se disponível no site do Project Gutenberg.

Objetivo do projeto:

Identificar quais palavras são mais recorrentes ao longo do livro e se o tom da narrativa tende mais para o "positivo" ou para o "negativo". Para tanto, serão aplicados quatro processos de mineração de textos:

- a criação de uma **Wordcloud**, para obtermos uma visão geral das palavras mais frequentes;
- o uso de algoritmo de **Word Pairs**, para entendermos quais palavras foram mais usadas em duplas;
- a **Classificação Binária de Sentimentos**, para separarmos as palavras entre "positivas" e "negativas"; e
- o uso de um **Escore de Sentimentos**, para sabermos com que intensidade as palavras são positivas ou negativas e para verificarmos como elas se distribuem ao longo do arco narrativo.

Parte 1 - CRIAÇÃO DE WORDCLOUD

Importando as bibliotecas necessárias:

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):

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

df.head()
```

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...

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', 'orange', 'gray']
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)
```

O texto de Stevenson mostra a predominância de palavras como "man" e "men" (homem e homens), "ship" (navio), "sea" (mar), "steerage" (nesse contexto, a classe inferior do navio) e "passenger" (passageiro). Combinadas, elas indicam que essa história acontece num navio, com homens que viajam na classe mais barata de todas.

Importante: a criação de uma Wordcloud pode ser a primeira forma de um(a) editor(a) de livros ou de conteúdo ter uma visão prévia sobre o tema do material que ele(a) tem em mãos para ler e avaliar.