



**FACULDADE DE TECNOLOGIA BAIXADA SANTISTA**  
**CIÊNCIA DE DADOS**

**Nícolas de Almeida Lopes**

**RELATÓRIO – SIMILARIDADE DE EVENTOS HISTÓRICOS**  
**ÁLGEBRA LINEAR**

**SANTOS – SP**  
**2025**

## 1 INTRODUÇÃO

Para melhorar a experiência dos seus usuários, diversas empresas utilizam sistemas de recomendação. Esses sistemas são baseados em algoritmos de aprendizado de máquina a partir de cálculos matemáticos que analisam o comportamento do usuário e para sugerir produtos, serviços ou conteúdo personalizados.

Esses sistemas ajudam a reduzir a sobrecarga de informações, tornando a escolha de produtos ou conteúdos mais rápida e intuitiva. Além disso, com o uso desse sistema, é possível aumentar a retenção de usuários e as vendas.

Para medir a similaridade, um dos métodos mais utilizados é a similaridade por cosseno, uma medida que quantifica a semelhança entre dois vetores, avaliando o cosseno do ângulo entre eles. Um valor próximo de 1 indica alta similaridade, enquanto valores próximos de -1 indicam alta dissimilaridade.

Por exemplo, na Netflix, um usuário pode ser representado por um vetor cujas dimensões correspondem a diferentes filmes ou séries e cujos valores refletem suas preferências.

## 2 OBJETIVO

A partir desse método, o trabalho tem como objetivo desenvolver algoritmos na linguagem de programação Python, para implementar um sistema de similaridade sobre eventos históricos onde o usuário seleciona um evento de um ano específico e, com base nessa escolha, o sistema retorna diversos eventos históricos semelhantes ao que o usuário escolheu.

## 3 EXTRAÇÃO DE DADOS

Para a obtenção dos dados, foi desenvolvido um algoritmo de *web scraping* para extrair dados de eventos históricos que foram obtidos a partir do site 'On This Day', no qual contém eventos do ano 1 até o ano de 2024.

### 3.1 WEB SCRAPING

*Web scraping* é uma técnica utilizada para extrair dados de sites de forma automatizada, permitindo coletar dados de páginas da web e armazená-las para análise ou utilização posterior.

Para desenvolver a técnica em Python, utilizou-se duas bibliotecas: o BeautifulSoup, uma biblioteca que permite navegar e extrair informações estruturadas a partir do HTML de uma página e a biblioteca 'subprocess', utilizada para executar comandos no *Command Prompt*.

**Figura 1 - Algoritmo de Web scraping em Python para extração de dados**

```
def extrair_eventos(ano_inicio, ano_fim):
    data = []
    for ano in range(ano_inicio, ano_fim+1):
        url = "https://www.ontthisday.com/date" if ano < 1492 else "https://www.ontthisday.com/events/date"
        for pagina in range(1, 6):
            command = f"{url}/{ano}?p={pagina}" if pagina >= 2 else f"{url}/{ano}"
            result = subprocess.run(["curl", command], capture_output=True, text=True, encoding='utf-8')
            soup = BeautifulSoup(result.stdout, 'html.parser')
            try:
                year = soup.find('span', class_='year').get_text()
                for event_el in (soup.find_all('li', class_=['event', 'person']) +
                                soup.find_all('div', class_='section--highlight')):
                    if event_el.b:
                        date = event_el.b.get_text()
                        for cls in ['date', 'deathDate', 'birthDate']:
                            date_el = event_el.find('a', class_=cls)
                            if date_el:
                                date = date_el.get_text()
                        event = event_el.get_text().replace(date, '', 1).strip()
                        data.append({"Year": year, "Date": date, "Event": event})
            except:
                break
    return data
```

**Fonte: Elaborado pelo autor**

Conforme é mostrado na Figura 1, após ser fornecido o intervalo de ano desejado para extração, o algoritmo executa *prompt* de comando 'curl' que retorna todo o HTML da URL atribuída e armazena em uma variável. Em seguida, com a utilização do BeautifulSoup, obtém-se os dados de um evento a partir de suas classes HTML.

O website conta com duas seções de eventos: uma dedicada a eventos históricos comuns e outra a eventos em destaque (*highlights*): aqueles eventos que ficaram marcados na história. O algoritmo combina as duas categorias, somando primeiramente os eventos em destaque com os eventos comuns. Dessa forma, os primeiros eventos exibidos para cada ano serão os de destaque.

A Figura 2 mostra que, de acordo com o nome da classe de um elemento, é possível extrair seus dados. No elemento onde é mostrado a data, como por exemplo 'Feb 17', o nome da classe pode variar de acordo com o contexto do evento.

Figura 2 – Classes dos elementos do website 'On This Day'

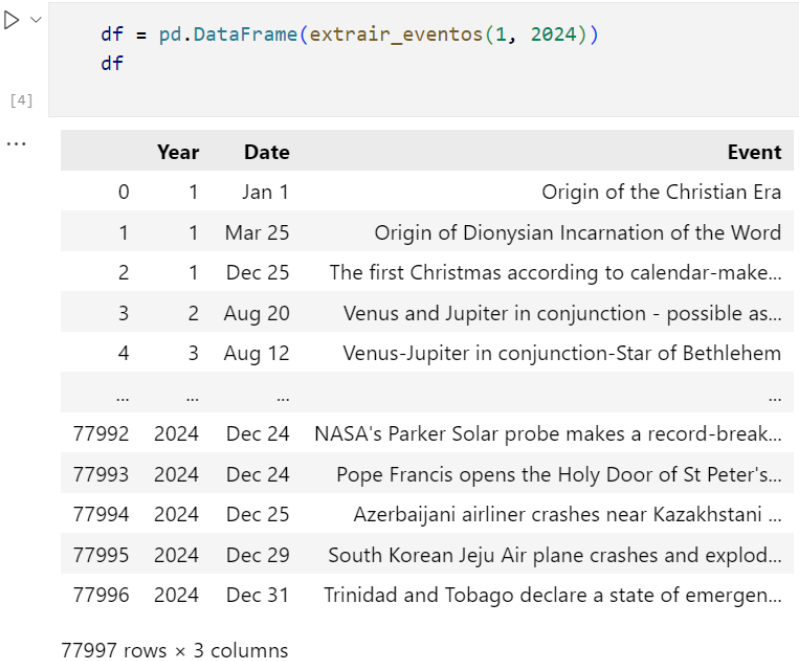


Fonte: Elaborado pelo autor a partir do site 'On This Day'

4 TRATAMENTO DOS DADOS

A partir dos dados dos eventos extraídos, foi criado um *DataFrame*, uma tabela gerada pela biblioteca Pandas que contém todos os dados dos eventos históricos entre o ano 1 até 2024. No total, foram extraídos quase 80 mil eventos históricos, a conforme mostra a Figura 3.

Figura 3 – Preview da tabela dos dados extraídos a partir do *DataFrame*



Fonte: Elaborado pelo autor

## 4.1 TRATAMENTO DE TEXTO

Além das três colunas que serão exibidas futuramente para o usuário, foi criada mais uma coluna que será utilizada para o algoritmo da similaridade. A nova coluna conterá o texto tratado do evento histórico, o tratamento também será feito em Python através das bibliotecas 'string' e 'nltk', como é mostrado na Figura 4.

**Figura 4 – Algoritmo para tratamento de texto.**

```
def tratamento_texto(texto):
    texto = texto.lower()
    texto = texto.translate(str.maketrans('', '', string.punctuation))
    tokens = word_tokenize(texto)
    stop_words = set(stopwords.words("english"))
    tokens = [palavra for palavra in tokens if palavra not in stop_words]
    return ' '.join(tokens)
```

**Fonte:** Elaborado pelo autor

Primeiramente, o algoritmo converte todas as letras para minúsculas do texto fornecido, em seguida, remove todas as pontuações. Posteriormente, as palavras são guardadas em uma lista através do comando *word\_tokenize*.

Para remover as *stopwords*, isto é, remover palavras muito comuns em um idioma (como 'the', 'of', 'a', 'for', 'to' em inglês) que não carregam muito significado por si só, o algoritmo verifica, para cada palavra *tokenizada*, se a palavra não é uma *stopword*.

Por último, o programa retorna o texto após o tratamento, separado normalmente por espaço. Esse texto tratado, será adicionada à nova coluna que foi criada, conforme mostra a Figura 5.

**Figura 5 – Preview da tabela com a nova coluna 'texto\_tratado'**

df["texto\_tratado"] = df["Event"].apply(tratamento\_texto)

df

	Year	Date	Event	texto_tratado
0	1	Jan 1	Origin of the Christian Era	origin christian era
1	1	Mar 25	Origin of Dionysian Incarnation of the Word	origin dionysian incarnation word
2	1	Dec 25	The first Christmas according to calendar-make...	first christmas according calendarmaker easter...
3	2	Aug 20	Venus and Jupiter in conjunction - possible as...	venus jupiter conjunction possible astrologica...
4	3	Aug 12	Venus-Jupiter in conjunction-Star of Bethlehem	venusjupiter conjunctionstar bethlehem
...	...	...	...	...
77992	2024	Dec 24	NASA's Parker Solar probe makes a record-break...	nasas parker solar probe makes recordbreaking ...
77993	2024	Dec 24	Pope Francis opens the Holy Door of St Peter's...	pope francis opens holy door st peters basilic...
77994	2024	Dec 25	Azerbaijani airliner crashes near Kazakhstani ...	azerbaijani airliner crashes near kazakhstani ...
77995	2024	Dec 29	South Korean Jeju Air plane crashes and explod...	south korean jeju air plane crashes explodes t...
77996	2024	Dec 31	Trinidad and Tobago declare a state of emergen...	trinidad tobago declare state emergency gang v...

77997 rows x 4 columns

## 4.2 DATASET

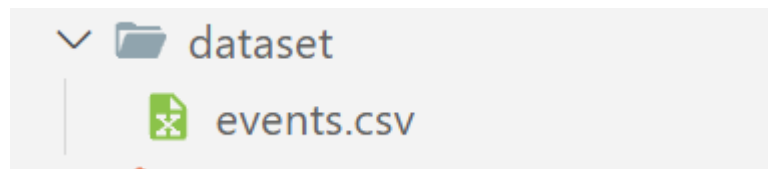
Com os dados carregados em um *DataFrame* e o texto dos eventos tratados adequadamente, é possível converter para um arquivo CSV, que conterà o conjunto de dados extraídos e tratados, como é mostrado na Figura 6 e 7. O *dataset* será utilizado no sistema final para encontrar as similaridades.

**Figura 6 – Conversão do *DataFrame* para um arquivo CSV.**

```
df.to_csv("dataset/events.csv", sep=";", index=False)
```

Fonte: Elaborado pelo autor

**Figura 7 – O arquivo CSV 'events.csv' na pasta 'dataset'.**



Fonte: Elaborado pelo autor

## 5 O SISTEMA