
**Pontifícia Universidade Católica - PUCRS**

**Disciplina: Coleta, Preparação e Análise de Dados**
**Prof. Lucas Rafael Costella Pessutto**

**Exercício: Coleta de Dados**

---

# Coleta de Dados na Wikipedia

*Administrativo*
Este exercício pode ser feito em **trios ou quartetos**, não sendo possível realizar individualmente. Apenas um dos integrantes deve submeter os arquivos do trabalho no moodle.
Data de Entrega: 07/04/2024 (Turma 10) – 09/04/2024 (Turma 30)
Forma de Entrega: O trabalho deve ser apresentado durante o período de aula. Não haverá uma aula de apresentação para este trabalho. O grupo deve marcar com o professor um horário para realizar a apresentação. A apresentação consiste em mostrar o código e o programa em funcionamento para o professor. Poderão ser feitas perguntas sobre o funcionamento do trabalho para o grupo.
Além disso, deve ser entregue no moodle:

- Scripts Python ou Jupyter notebooks com o código que realiza as tarefas solicitadas.
- Link para um repositório com os dados obtidos via scraping (páginas html) e com os dados extraídos (arquivos json)
- Orientações sobre como executar os scripts (como comentário no código, arquivo README.txt ou células de texto em jupyter notebook). Incluir comentários sobre a configuração do ambiente de desenvolvimento necessário para rodar os scripts.

*Tarefa 1 – Desenvolvendo um crawler*
Escreva um crawler para descobrir e coletar páginas da Wikipedia em Português. Seu programa deve coletar 5.000 páginas de verbetes diferentes da Wikipedia, à partir da página inicial: `https://pt.wikipedia.org`. Não coletar páginas de outros sites e também não deve ser coletado páginas internas da Wikipedia que não são verbetes.
O crawler deve funcionar da seguinte maneira:

1. Obter uma página.
2. Salvar a página como um arquivo html, chamado `<titulo_verbete>.html`
3. Extrair todos os links que se encontram nessa página
4. Filtrar os links, removendo os que não se referem à verbetes e os verbetes que já foram visitados.

5. Guardar esses links em uma lista

6. Escolher um link não visitado para ser a próxima página.

7. Voltar ao passo inicial

As páginas de verbetes coletadas deverão ser salvas como arquivos com a extensão `.html`. Lembre-se de tomar cuidado para não estressar o servidor com requisições em excesso

*Tarefa 2 – Extraindo informações de Infoboxes*
A segunda tarefa consiste em identificar as páginas que possuem infoboxes, que são usadas para resumir as informações de um artigo na Wikipédia. Veja na figura a seguir um exemplo de página que contém um infobox (ele está destacado em vermelho na imagem).
Infoboxes estruturam informações de diversas maneiras, por isso é difícil conseguir extrair todos os seus elementos de forma fácil. Portanto, focaremos nossos esforços em extrair apenas alguns elementos. São eles:

1. **Título**: toda infobox possui um título que fica no topo da caixa. No exemplo da figura, o título é Alan Turing

2. **Pares chave – valor**: esses pares são identificados por uma chave que está associada a um único valor. Por exemplo, na figura temos a chave "Nome completo" e o valor "Alan Mathison Turing".

3. **Pares chave – lista**: nesse tipo de item, uma chave está associada a uma lista de valores. Na figura de exemplo são pares de chave – lista os campos: Conhecido(a) por, Alma mater, Orientado(a)(s), Instituições e Campos.

Sua tarefa consiste em extrair o conteúdo das infoboxes de todas as páginas que foram extraídas no exercício anterior. Seu programa deve identificar quando uma página possui uma infobox, realizar a extração das informações e salvá-las em um arquivo `.json` cujo nome é o título da infobox.
Para fins de teste, será fornecido um conjunto de páginas juntamente com a saída esperada para cada uma delas.

# Alan Turing

157 línguas

Artigo Discussão Ler Editar Ver histórico Ferramentas

**Alan Mathison Turing** (Londres, 23 de junho de 1912 — Wilmslow, Cheshire, 7 de junho de 1954) foi um matemático,[1] cientista da computação, lógico, criptoanalista, filósofo e biólogo teórico britânico. Turing foi altamente influente no desenvolvimento da moderna ciência da computação teórica, proporcionando uma formalização dos conceitos de algoritmo e computação com a máquina de Turing, que pode ser considerada um modelo de um computador de uso geral.[2][3][4] Ele é amplamente considerado o pai da ciência da computação teórica e da inteligência artificial.[5] Apesar dessas realizações ele nunca foi totalmente reconhecido em seu país de origem durante sua vida por ser homossexual e porque grande parte de seu trabalho foi coberto pela Lei de Segredos Oficiais.

Durante a Segunda Guerra Mundial, Turing trabalhou para a Escola de Código e Cifras do Governo (GC&CS) em Bletchley Park, o centro britânico de criptoanálise que produzia ultra inteligência. Por um tempo ele liderou a Hut 8, a seção responsável pela análise criptográfica naval alemã. Lá ele desenvolveu várias técnicas para acelerar a quebra das cifras alemãs, incluindo melhorias no método de bombardeio polonês antes da guerra, bem como uma máquina eletromecânica que poderia encontrar configurações para a máquina Enigma. Turing desempenhou um papel crucial na quebra de mensagens codificadas interceptadas que permitiram aos Aliados derrotar os nazistas em muitos compromissos cruciais, incluindo a Batalha do Atlântico, e ao fazê-lo os ajudou a vencer a guerra. Devido aos problemas da história contrafactual, é difícil estimar o efeito preciso que a inteligência ultra teve na guerra[6] mas foi estimado que este trabalho encurtou a guerra na Europa em mais de dois anos e salvou mais de 14 milhões de vidas.[7]

Após a guerra Turing trabalhou no Laboratório Nacional de Física, onde projetou o Mecanismo de Computação Automática, um dos primeiros projetos para um computador de programa armazenado. Em 1948 Turing ingressou no Laboratório de Máquinas de Computação de Max Newman, na Victoria University de Manchester, onde ajudou a desenvolver os computadores de Manchester[8] e se interessou por biologia matemática. Ele escreveu um artigo sobre as bases químicas da morfogênese e previu reações químicas oscilantes, como a reação de Belousov – Zhabotinsky, observada pela primeira vez na década de 1960.

Turing foi processado judicialmente em 1952 por atos homossexuais: a Emenda Labouchere de 1885 determinara que "indecência grosseira" era uma ofensa criminal no Reino Unido. Ele aceitou o tratamento de castração química, com dietilestilbestrol, como alternativa à prisão. Turing morreu em 1954, 16 dias antes de seu 42º aniversário, por envenenamento por cianeto. Um inquérito determinou sua morte como suicídio, mas se observou que a evidência conhecida também é consistente com envenenamento acidental. Em 2009, após uma campanha na Internet, o primeiro-ministro britânico Gordon Brown fez um pedido de desculpas público e oficial a Turing em nome do governo britânico pela "maneira terrível como foi tratado". A rainha Elizabeth II concedeu a Turing um perdão póstumo em 2013. A "lei Alan Turing" é agora um termo informal para uma lei britânica de 2017 que retroativamente perdoou homens advertidos ou condenados sob a legislação histórica que proibia atos homossexuais.[9]

## Infância e educação

### Família

Turing nasceu em Maida Vale, Londres, enquanto seu pai, Julius Mathison Turing (1873-1947), estava de licença de seu cargo no Serviço Civil Indiano (ICS) em Chatrapur, atual estado de Odisha, na Índia.[10][11] O pai de Turing era filho de um clérigo, o Rev. John Robert Turing, de uma família escocesa de comerciantes sediada nos Países Baixos e que incluía um baronete. A mãe de Turing, esposa de Julius, era Ethel Sara Turing (1881–1976), filha de Edward Waller Stoney, engenheiro chefe das Ferrovias Madras. Os Stoneys eram uma família de nobres protestantes anglo-

**Alan Turing**

Turing em 1928, aos dezesseis anos de idade

| | |
|---|---|
| **Nome completo** | Alan Mathison Turing |
| **Conhecido(a) por** | Lista [Expandir] |
| **Nascimento** | 23 de junho de 1912<br>Maida Vale, Londres |
| **Morte** | 7 de junho de 1954 (41 anos)<br>Wilmslow, Cheshire |
| **Causa da morte** | Suicídio por ingestão de cianeto |
| **Nacionalidade** | britânico |
| **Educação** | Sherborne School |
| ***Alma mater*** | Universidade de Cambridge (BA, MA)<br>Universidade de Princeton (PhD) |
| **Prêmios** | Prêmio Smith (1936) |
| **Religião** | Ateísmo |
| **Orientador(es) (as)** | Alonzo Church |
| **Orientado(a) (s)** | Robin Gandy<br>Beatrice Worsley |
| **Instituições** | Universidade de Manchester<br>Escola Governamental de Código e Cifra<br>Laboratório Nacional de Física |
| **Campo(s)** | Lista [Expandir] |
| **Tese** | *Sistemas de Lógica Baseada em Ordinais* (1938) |

**Assinatura**



**Máquina de Turing**

**Máquina**

Máquina de Turing universal
Máquina de Turing alternada
Máquina de Turing quântica
Máquina de Turing não determinística
Máquina de Turing somente de leitura
Máquina de Turing movimentação à direita
Máquina de Turing probabilística
Máquina de Turing multifita
Máquina de Turing de Várias Faixas
Máquinas de Turing equivalentes
Exemplos de Máquinas de Turing

**Ciência**

**Alan Turing**
Categoria:Máquina de Turing

V · D · E