

# BIG DATA PROCESSING AND ANALYTICS

RECUPERAÇÃO DA INFORMAÇÃO NA WEB  
E EM REDES SOCIAIS



**Professor curador**  
Luciano Moreira Camilo e Silva



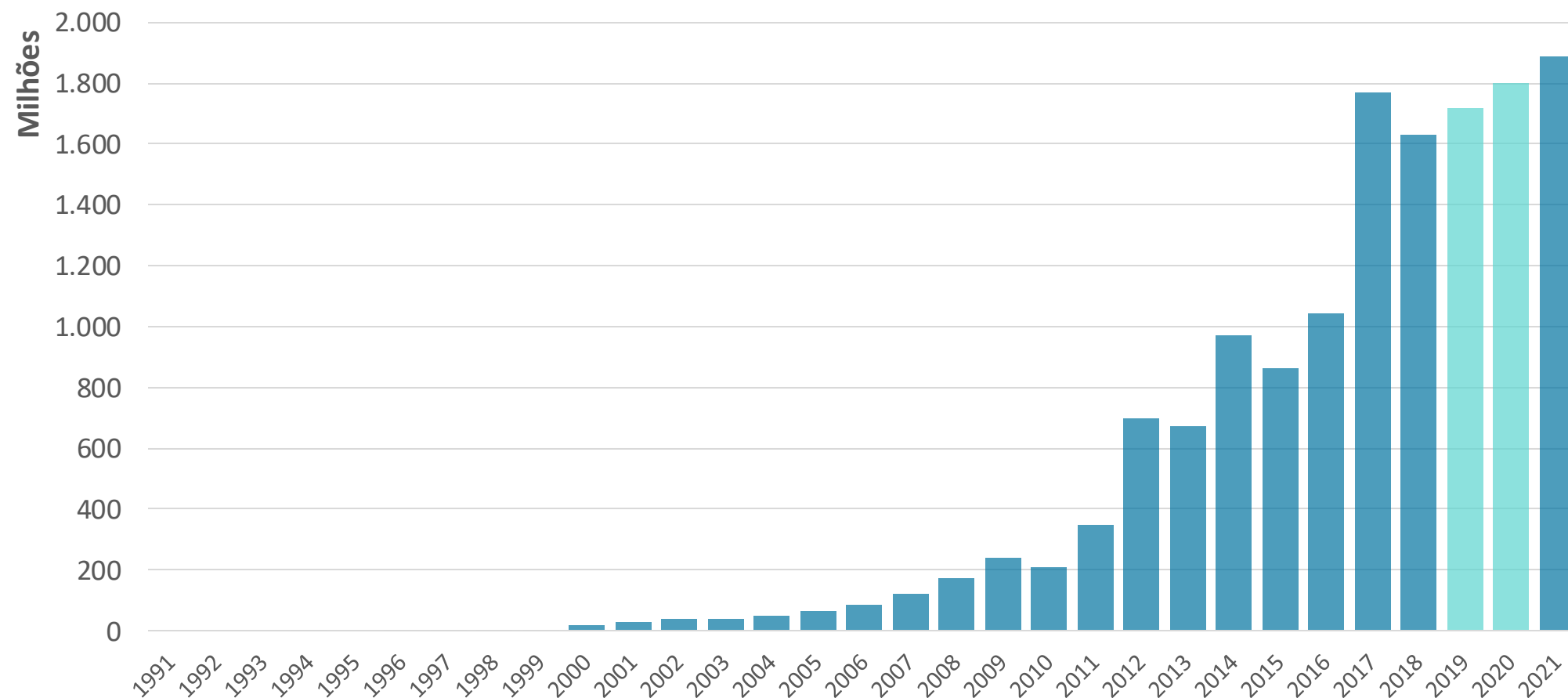


# **TRILHA 3**

## **RECUPERAÇÃO DE INFORMAÇÃO POR RASPAGEM – ROBÔS**

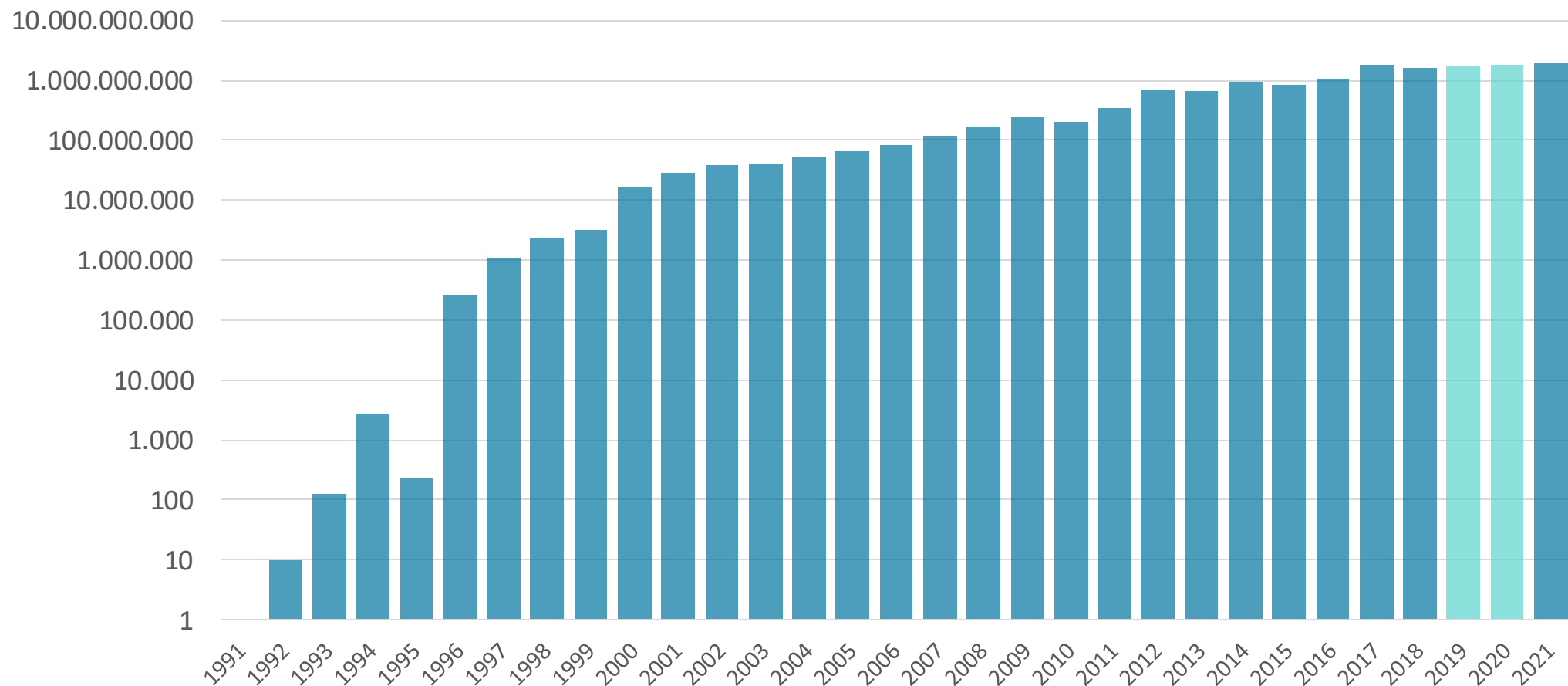
---

# Quantidade de sites na internet



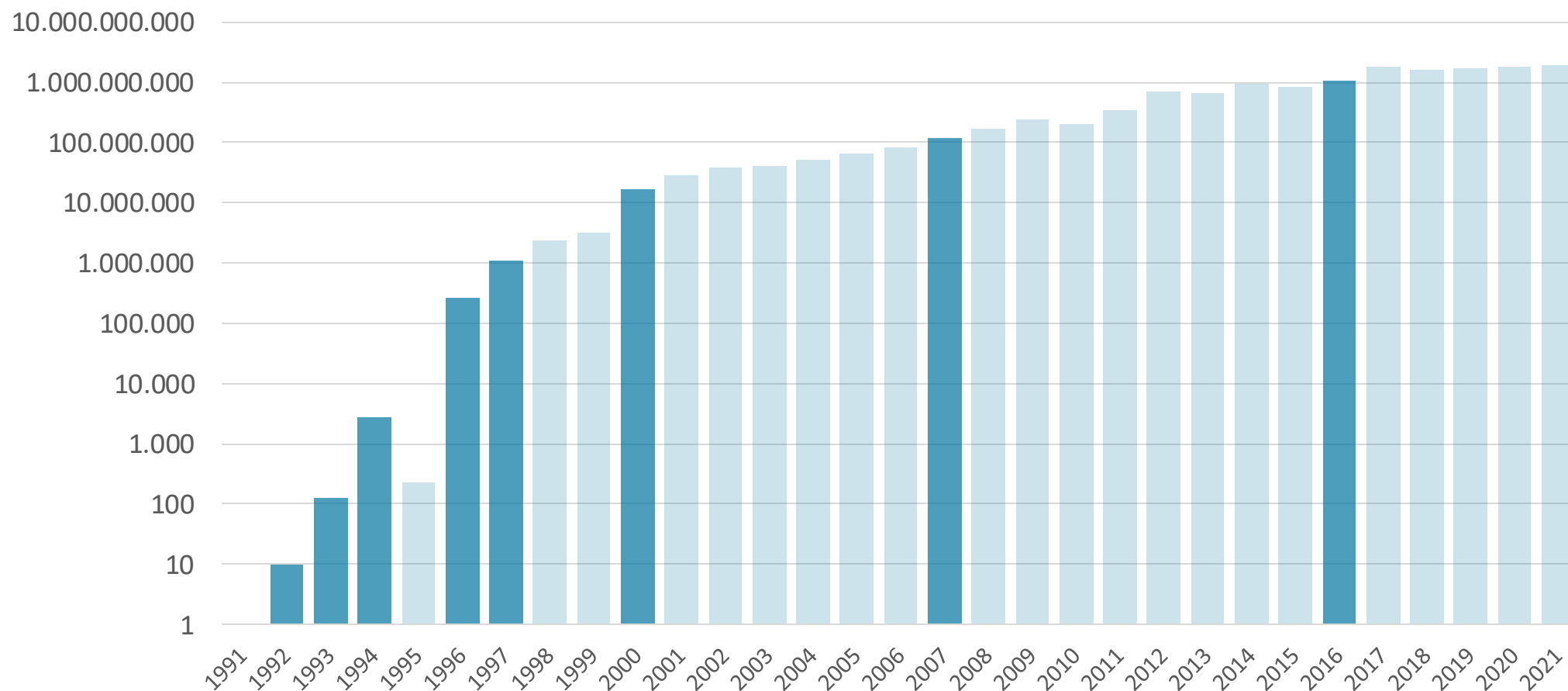
Fonte: NetCraft and Internet Live Stats (elaboration of data by Matthew Gray of MIT and Hobbes' Internet Timeline and Pingdom).

# Quantidade de sites na internet – Escala Logarítmica



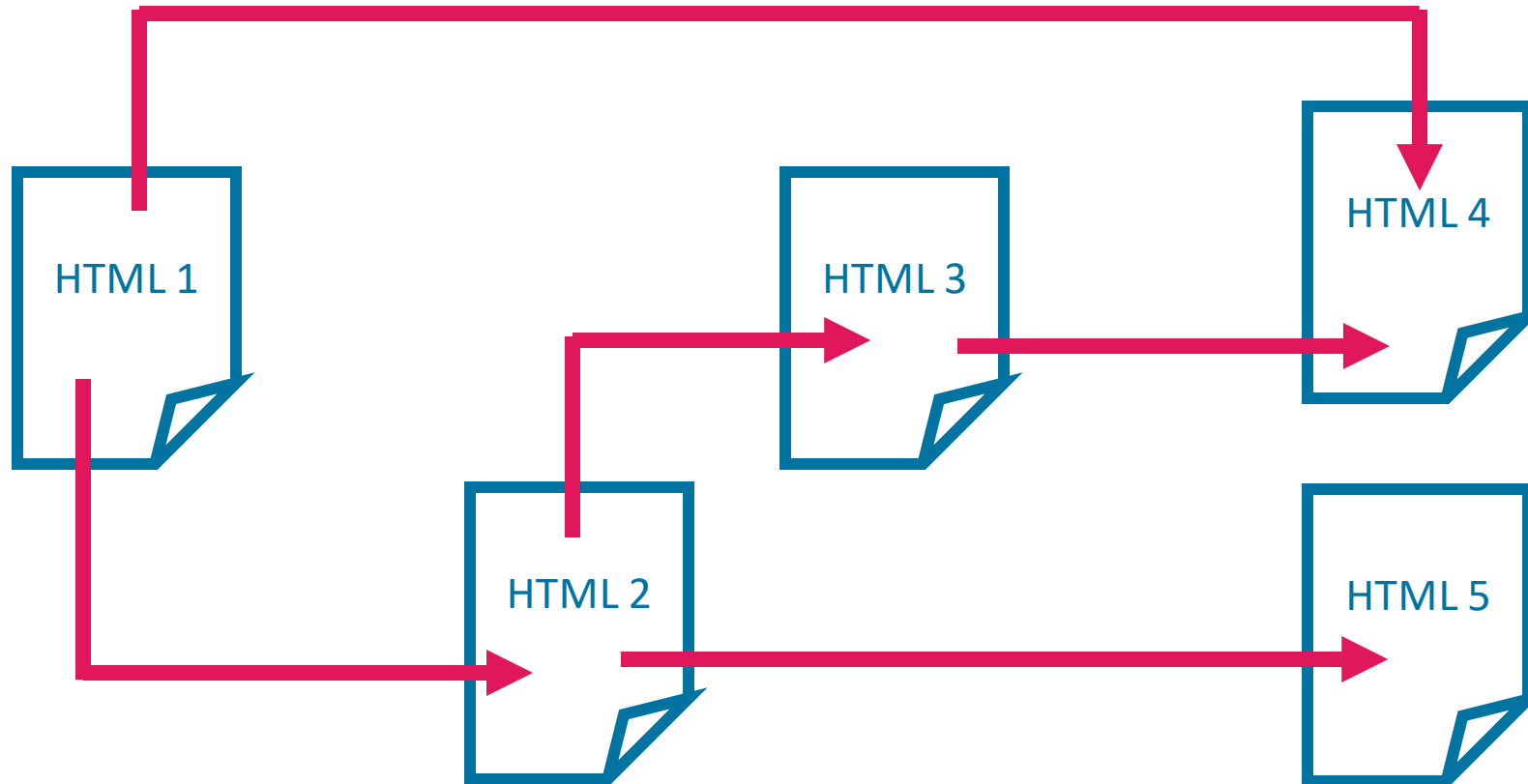
Fonte: NetCraft and Internet Live Stats (elaboration of data by Matthew Gray of MIT and Hobbes' Internet Timeline and Pingdom).

## Quantidade de sites na internet – Escala Logarítmica – 10x



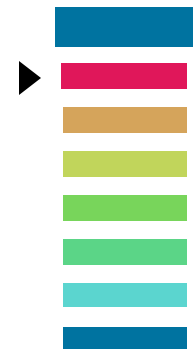
Fonte: NetCraft and Internet Live Stats (elaboration of data by Matthew Gray of MIT and Hobbes' Internet Timeline and Pingdom).

# Documentos de hipertexto



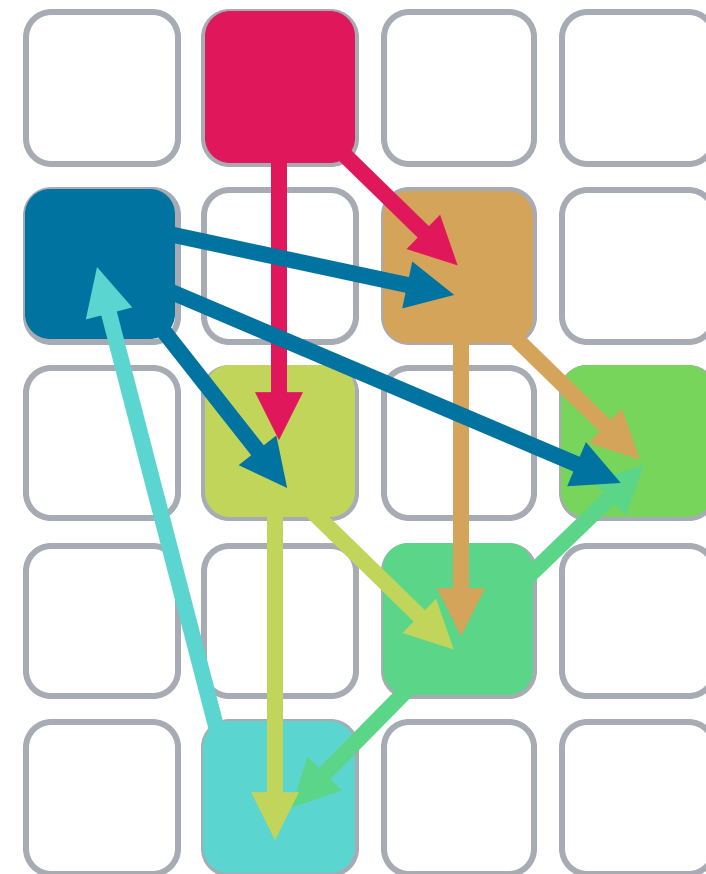
Fonte: Elaborada pelo autor.

# Como funciona um *Web Crawler*



**Web Crawler**

URL



# Uniform Resource Locator

subdomínio  
usuário domínio porta  
topo de domínio

`https://jose@www.site.com.br:123/caminho/para/documento.html?tag=price&order=new#top`

protocolo autoridade caminho documento parâmetros de consulta fragmento

Fonte: Ilustração do professor adaptada de wikipedia.org



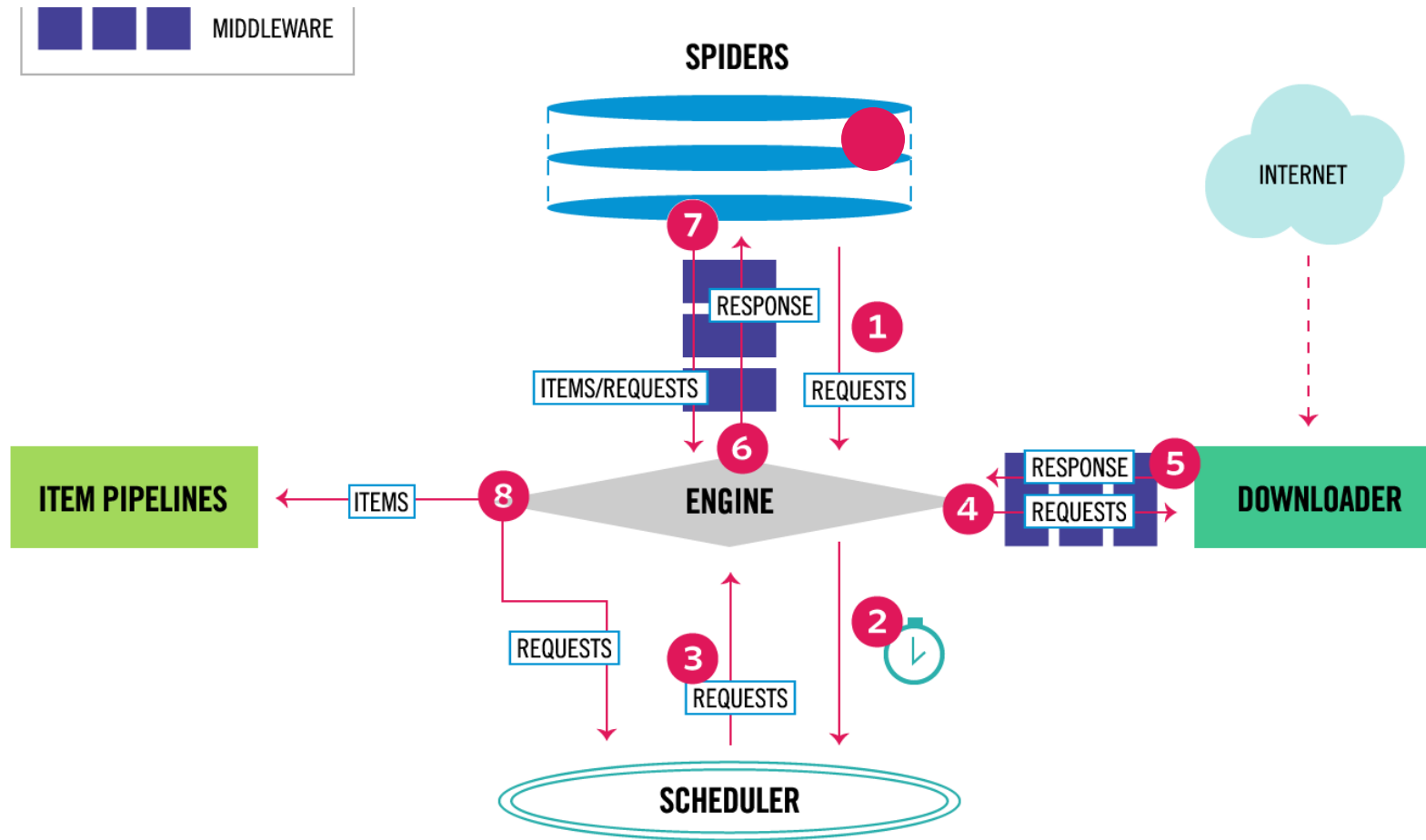
## *O framework Scrapy*

- É um framework!
- Possui todas as funções de um robô pré-programadas e arquitetadas.
- É fácil de expandir.
- É ideal para construção de *web crawlers*.

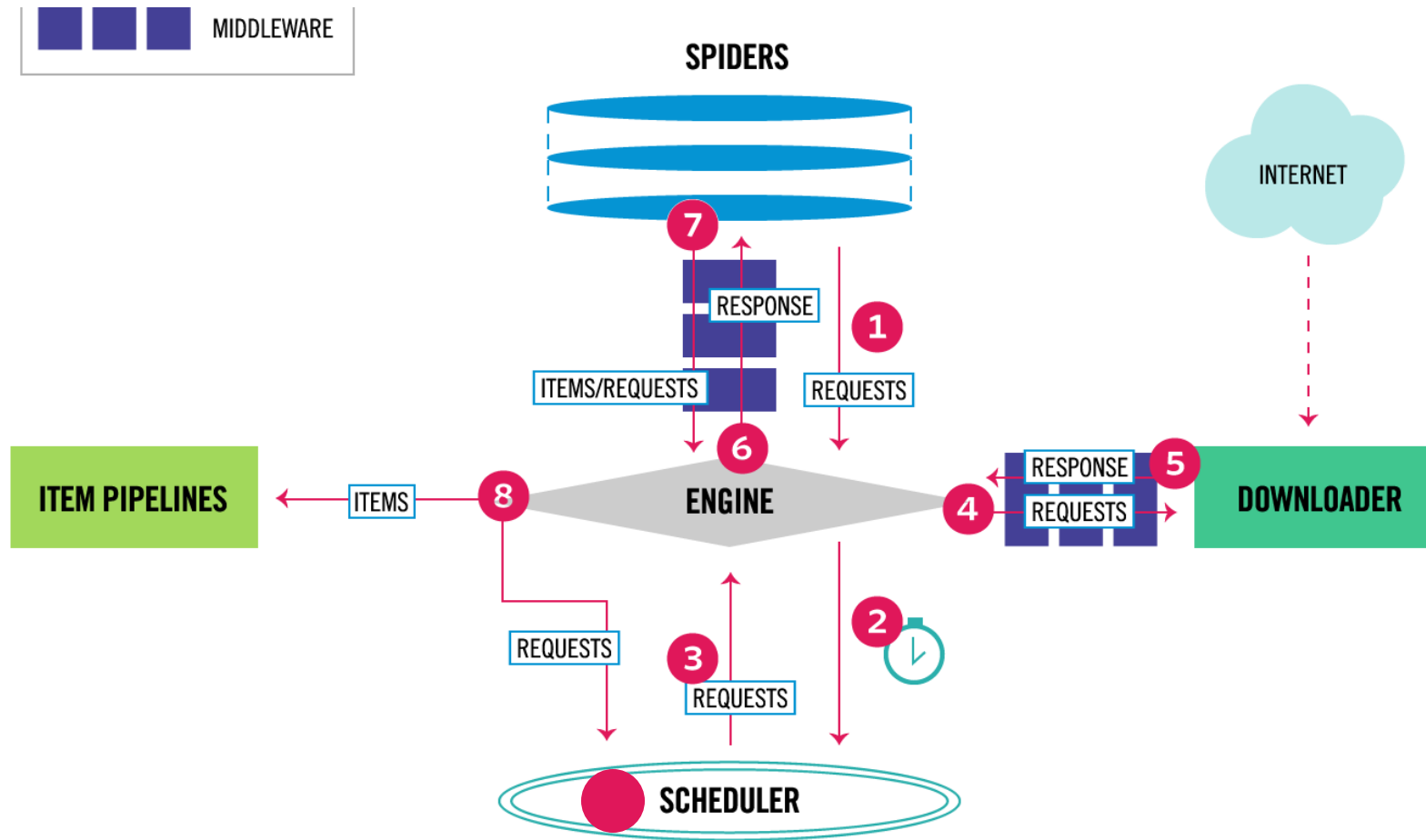


Fonte: <<https://scrapy.org/>>.

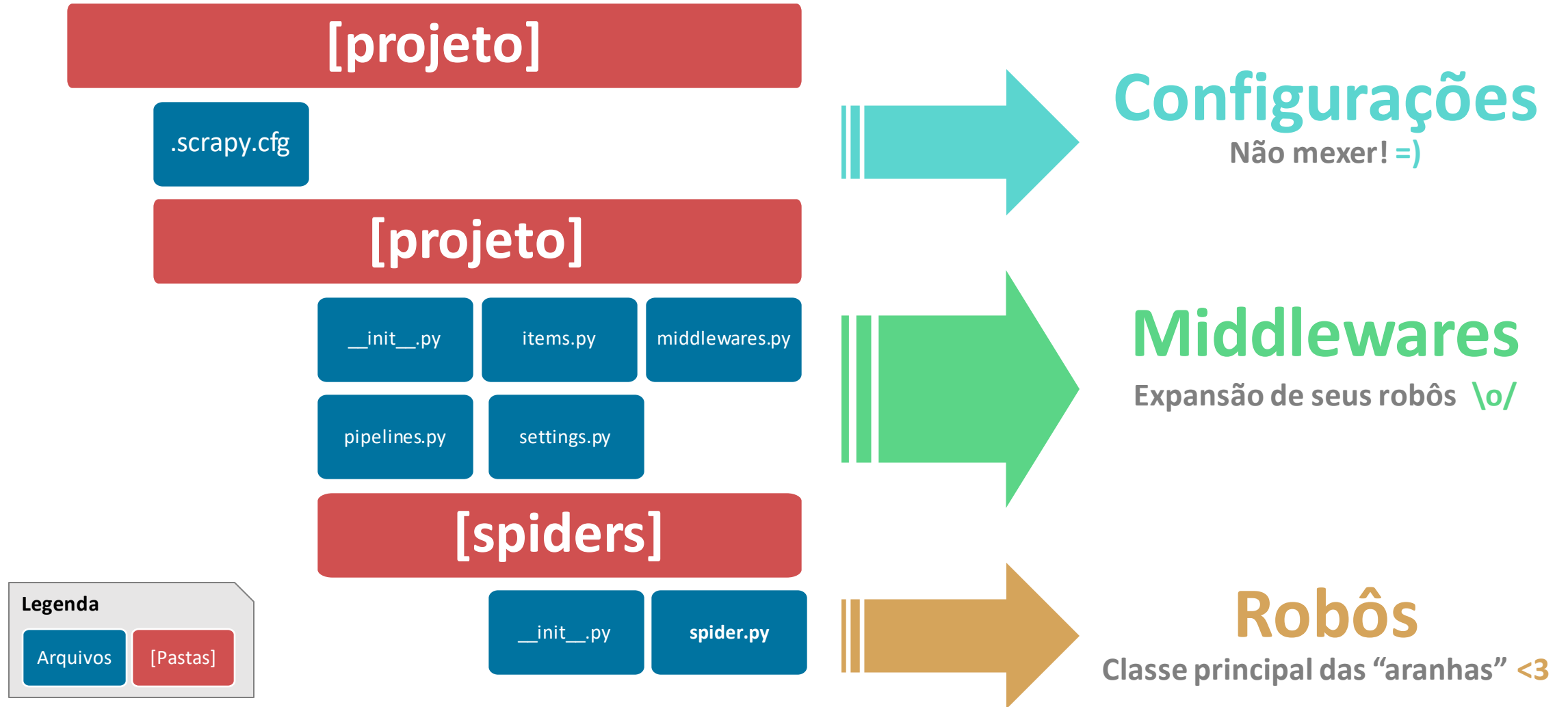
# Arquitetura do Scrapy



# Arquitetura do Scrapy



# Estrutura de diretórios e arquivos do Scrapy



# Exercício prático

Arnaldo Jabor é cineasta, roteirista, diretor de cinema e TV, produtor cinematográfico, dramaturgo, crítico, jornalista e escritor brasileiro.

Durante 26 anos, escreveu mais de 1500 artigos para jornais e revistas. E teve seu nome associado a muito outros textos vistos na internet!

O exercício consiste em puxar e formatar todos os artigos de Arnaldo Jabor disponíveis no site do jornal *O Tempo*.

The screenshot shows the website of the newspaper 'O TEMPO'. The header is green with the newspaper's name in white. Below the header is a navigation bar with links: SUPER NOTÍCIA, RÁDIO SUPER, SUPER.FC, TEMPO TV, O TEMPO BETIM, CLUBE O TEMPO, TEMPOSTORE, and VERSÃO DIGITAL. On the right, there's a search bar and a weather widget showing '26°C | Belo Horizonte 07/09/2021'. The main content area is titled 'ÚLTIMAS COLUNAS' and displays a grid of 12 opinion columns, each with a small image and a title. The columns are: 'Adeus' (image of a man's face), 'Nada vem do nada' (image of a nuclear explosion), 'Direita, esquerda e realidade' (image of a person holding a red flag), 'A barbárie dos fatos' (image of a man's face), 'Amor ao fracasso' (image of a globe), 'A tragicomédia' (image of a man's face with stars), 'O tríduo de Momo' (image of a hand holding a bowl), 'Os burros n'água' (image of a person in water), 'A "pós-mentira"' (image of a person in a doorway), and three more at the bottom. On the right side, there's a sidebar with a profile picture of Arnaldo Jabor, the text 'AUTOR Arnaldo Jabor', and a 'LEIA MAIS' link at the bottom.

MENU

O TEMPO

26°C | Belo Horizonte 07/09/2021

SUPER NOTÍCIA RÁDIO SUPER SUPER.FC TEMPO TV O TEMPO BETIM CLUBE O TEMPO TEMPOSTORE VERSÃO DIGITAL

BUSCA

ÚLTIMAS COLUNAS

Adeus

Nada vem do nada

Direita, esquerda e realidade

A barbárie dos fatos

Amor ao fracasso

A tragicomédia

O tríduo de Momo

Os burros n'água

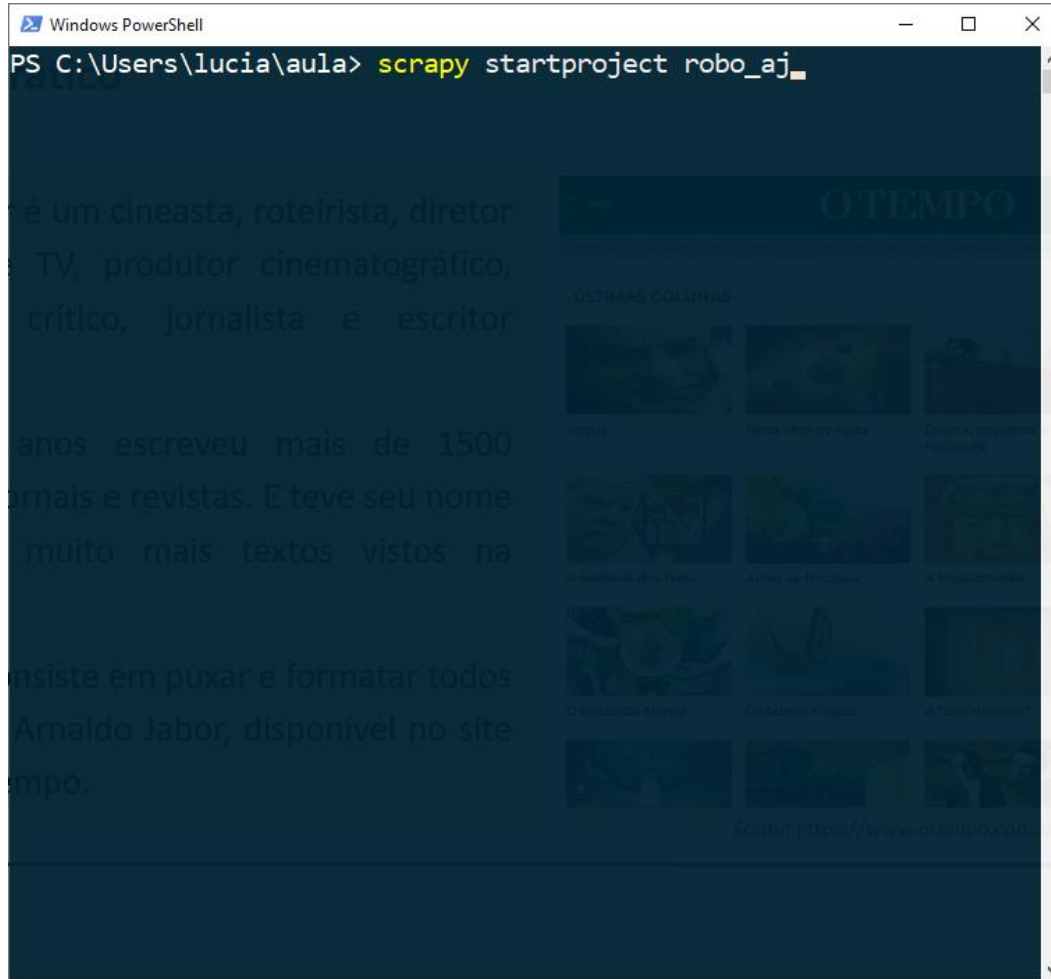
A "pós-mentira"

AUTOR Arnaldo Jabor

LEIA MAIS

Fonte: <<https://www.otempo.com.br/opiniao/arnaldo-jabor>>.

# Passo 1 – Criando o projeto



```
Windows PowerShell
PS C:\Users\lucia\aula> scrapy startproject robo_aj_
```

## Legenda

Arquivos

[Pastas]

# Passo 1 – Criando o projeto

```
Windows PowerShell
PS C:\Users\lucia\aula> scrapy startproject robo_aj
New Scrapy project 'robo_aj', using template directory 'c:\python39\lib\site-packages\scrapy\templates\project', created in:
  C:\Users\lucia\aula\robo_aj

You can start your first spider with:
  cd robo_aj
  scrapy genspider example example.com
PS C:\Users\lucia\aula>
```

[projeto]

.scrapy.cfg

[projeto]

\_\_init\_\_.py

items.py

middlewares.py

pipelines.py

settings.py

[spiders]

\_\_init\_\_.py

Legenda

Arquivos

[Pastas]

# Passo 1 – Criando o projeto

```
Windows PowerShell
PS C:\Users\lucia\aula> scrapy startproject robo_aj
New Scrapy project 'robo_aj', using template directory 'c:\python39\lib\site-packages\scrapy\templates\project', created in:
  C:\Users\lucia\aula\robo_aj

You can start your first spider with:
  cd robo_aj
  scrapy genspider example example.com
PS C:\Users\lucia\aula> cd .\robo_aj\
PS C:\Users\lucia\aula\robo_aj> █
```

[projeto]

.scrapy.cfg

[projeto]

\_\_init\_\_.py

items.py

middlewares.py

pipelines.py

settings.py

[spiders]

\_\_init\_\_.py

Legenda

Arquivos

[Pastas]



# Passo 1 – Criando o projeto

```
Windows PowerShell
PS C:\Users\lucia\aula> scrapy startproject robo_aj
New Scrapy project 'robo_aj', using template directory 'c:\python39\lib\site-packages\scrapy\templates\project', created in:
  C:\Users\lucia\aula\robo_aj
You can start your first spider with:
  cd robo_aj
  scrapy genspider example example.com
PS C:\Users\lucia\aula> cd .\robo_aj\
PS C:\Users\lucia\aula\robo_aj> scrapy genspider spider_aj del.com
```

[projeto]

.scrapy.cfg

[projeto]

\_\_init\_\_.py

items.py

middlewares.py

pipelines.py

settings.py

[spiders]

\_\_init\_\_.py

## Legenda

Arquivos

[Pastas]

# Passo 1 – Criando o projeto

```
Windows PowerShell
PS C:\Users\lucia\aula> scrapy startproject robo_aj
New Scrapy project 'robo_aj', using template directory 'c:\python39\lib\site-packages\scrapy\templates\project', created in:
  C:\Users\lucia\aula\robo_aj
You can start your first spider with:
  cd robo_aj
  scrapy genspider example example.com
PS C:\Users\lucia\aula> cd .\robo_aj\
PS C:\Users\lucia\aula\robo_aj> scrapy genspider spider_aj del.com
Created spider 'spider_aj' using template 'basic' in module:
  robo_aj.spiders.spider_aj
PS C:\Users\lucia\aula\robo_aj>
```

[projeto]

.scrapy.cfg

[projeto]

\_\_init\_\_.py

items.py

middlewares.py

pipelines.py

settings.py

[spiders]

\_\_init\_\_.py

Spider\_aj.py

Legenda

Arquivos

[Pastas]

# Escrevendo o Robô

---

## Passo 2 – Importar bibliotecas

```
1. import scrapy
2. import os
3. import pathlib
4. import csv
5.
6. class SpiderAj2Spider(scrapy.Spider):
7.     name = "spider_aj"
8.
9.     # abre o arquivo local, com todos os links
10.    start_urls = [f"{pathlib.Path(os.path.abspath('arnaldo_jabor.html')).as_uri()}"]
11.
12.    # função que interpreta o documento que lista os artigos
13.    def parse(self, response):
14.        for link in response.css(".item-ultimas").css("h2").css("a::attr(href)").getall():
15.            text_page = f"https://www.otempo.com.br/{link}"
16.            yield scrapy.Request(text_page, callback=self.parse_text)
17.
18.    # função que interpreta os documentos com os textos
19.    def parse_text(self, response):
20.        content = ""
21.        for line in response.css('div.texto-artigo p::text').getall() :
22.            content = content + line + "\n"
23.
24.        post = {
25.            'author': 'Arnaldo Jabor',
26.            'title': response.css('h1::text').get(),
27.            'content': content.encode('utf-8')
28.        }
29.
30.        with open('artigos_aj.csv', 'a', newline='', encoding="utf-8") as output_file:
31.            dict_writer = csv.DictWriter(output_file, post.keys())
```

## Passo 3 – Indicar o arquivo com todos as ligações do texto

```
1. import scrapy
2. import os
3. import pathlib
4. import csv
5.
6. class SpiderAj2Spider(scrapy.Spider):
7.     name = "spider_aj"
8.
9.     # abre o arquivo local, com todos os links
10.    start_urls = [f"{pathlib.Path(os.path.abspath('arnaldo_jabor.html')).as_uri()}"]
11.
12.    # função que interpreta o documento que lista os artigos
13.    def parse(self, response):
14.        for link in response.css(".item-ultimas").css("h2").css("a::attr(href)").getall():
15.            text_page = f"https://www.otempo.com.br/{link}"
16.            yield scrapy.Request(text_page, callback=self.parse_text)
17.
18.    # função que interpreta os documentos com os textos
19.    def parse_text(self, response):
20.        content = ""
21.        for line in response.css('div.texto-artigo p::text').getall() :
22.            content = content + line + "\n"
23.
24.        post = {
25.            'author': 'Arnaldo Jabor',
26.            'title': response.css('h1::text').get(),
27.            'content': content.encode('utf-8')
28.        }
29.
30.        with open('artigos_aj.csv', 'a', newline='', encoding="utf-8") as output_file:
31.            dict_writer = csv.DictWriter(output_file, post.keys())
```

## Passo 4 – Escrever o 1º interpretador (extração de links)

```
1. import scrapy
2. import os
3. import pathlib
4. import csv
5.
6. class SpiderAj2Spider(scrapy.Spider):
7.     name = "spider_aj"
8.
9.     # abre o arquivo local, com todos os links
10.    start_urls = [f"{pathlib.Path(os.path.abspath('arnaldo_jabor.html')).as_uri()}"]
11.
12.    # função que interpreta o documento que lista os artigos
13.    def parse(self, response):
14.        for link in response.css(".item-ultimas").css("h2").css("a::attr(href)").getall():
15.            text_page = f"https://www.otempo.com.br/{link}"
16.            yield scrapy.Request(text_page, callback=self.parse_text)
17.
18.    # função que interpreta os documentos com os textos
19.    def parse_text(self, response):
20.        content = ""
21.        for line in response.css('div.texto-artigo p::text').getall() :
22.            content = content + line + "\n"
23.
24.        post = {
25.            'author': 'Arnaldo Jabor',
26.            'title': response.css('h1::text').get(),
27.            'content': content.encode('utf-8')
28.        }
29.
30.        with open('artigos_aj.csv', 'a', newline='', encoding="utf-8") as output_file:
31.            dict_writer = csv.DictWriter(output_file, post.keys())
```

## Passo 5 – Escrever o 2º interpretador (extração do conteúdo)

```
1. import scrapy
2. import os
3. import pathlib
4. import csv
5.
6. class SpiderAj2Spider(scrapy.Spider):
7.     name = "spider_aj"
8.
9.     # abre o arquivo local, com todos os links
10.    start_urls = [f"{pathlib.Path(os.path.abspath('arnaldo_jabor.html')).as_uri()}"]
11.
12.    # função que interpreta o documento que lista os artigos
13.    def parse(self, response):
14.        for link in response.css(".item-ultimas").css("h2").css("a::attr(href)").getall():
15.            text_page = f"https://www.otempo.com.br/{link}"
16.            yield scrapy.Request(text_page, callback=self.parse_text)
17.
18.    # função que interpreta os documentos com os textos
19.    def parse_text(self, response):
20.        content = ""
21.        for line in response.css('div.texto-artigo p::text').getall() :
22.            content = content + "".join(line) + "\n"
23.
24.        post = {
25.            'author': 'Arnaldo Jabor',
26.            'title': response.css('h1::text').get(),
27.            'content': content.encode('utf-8')
28.        }
29.
30.        with open('artigos_aj.csv', 'a', newline='', encoding="utf-8") as output_file:
31.            dict_writer = csv.DictWriter(output_file, post.keys())
```

## Passo 6 – Salvar em CSV para leitura em MS Excel

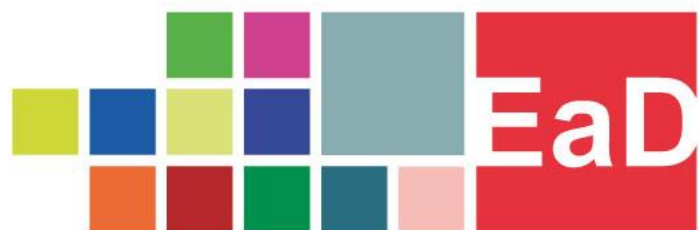
```
1. import scrapy
2. import os
3. import pathlib
4. import csv
5.
6. class SpiderAj2Spider(scrapy.Spider):
7.     name = "spider_aj"
8.
9.     # abre o arquivo local, com todos os links
10.    start_urls = [f"{pathlib.Path(os.path.abspath('arnaldo_jabor.html')).as_uri()}"]
11.
12.    # função que interpreta o documento que lista os artigos
13.    def parse(self, response):
14.        for link in response.css(".item-ultimas").css("h2").css("a::attr(href)").getall():
15.            text_page = f"https://www.otempo.com.br/{link}"
16.            yield scrapy.Request(text_page, callback=self.parse_text)
17.
18.    # função que interpreta os documentos com os textos
19.    def parse_text(self, response):
20.        content = ""
21.        for line in response.css('div.texto-artigo p::text').getall() :
22.            content = content + line + "\n"
23.
24.        post = {
25.            'author': 'Arnaldo Jabor',
26.            'title': response.css('h1::text').get(),
27.            'content': content.encode('utf-8')
28.        }
29.
30.        with open('artigos_aj.csv', 'a', newline='', encoding="utf-8") as output_file:
31.            dict_writer = csv.DictWriter(output_file, post.keys())
```



# Arquivo com o conteúdo

Salvamento Automático			Pasta2 - Excel	Pesquisar	Luciano McSilva	Compartilhar	Comentários			
Arquivo	Página Inicial	Inserir	Desenhar	Layout da Página	Fórmulas	Dados	Revisão	Exibir	Desenvolvedor	Ajuda
B4	Você está andando pela rua, e, de repente, uma imensa tempestade de luz cai sobre sua cabeça, como o sol despencando do céu. Você não sabe o que é nem vai saber nunca porque você derreteu como um sorvete em dois segundos. Fica um lago de seu									
	A	B								C
1	author	content	title							
2	Arnaldo Jabor	Caros leitores, meus semelhantes e irmãos, vou abandoná-los. Isso. Correndo o risco de "lugares-comuns" ou lamentos narcisistas, vou dizer por quê. Foram 26 anos escrevendo sem parar para vários jornais do país. E aqui já vai meu primeiro lugar-comum: "Como o tempo voa... Foi outro dia mesmo" que estreei na "Folha de S.Paulo", onde fiquei por dez anos. Depois, fui para outros jornais, incluindo o "Estadoão" e "de Belo Horizonte. Fiz as contas e, entre o espanto e o orgulho (outra obviedade), verifiquei que, nessas duas décadas e meia, escrevi cerca de 1.500 artigos para jornais. Mil e quinhentos? É. Logo depois, me meti na TV e no rádio, onde também estou há 20 anos mais ou menos. Rádio e TV juntos somam cerca de 3.000 comentários sobre a vida do país até hoje. Como ousei? Com que cara me meti nisso, deitando regra sobre tudo? Bem, foi por fome, e não por vaidade. Eu fiz cinema por 30 anos e, como todo cineasta, sofria de duas angústias básicas: ansiedade e frustração. Fiz nove filmes e, mesmo assim, passava necessidade para sustentar minhas filhas. Um dia falei: "Enchi. Chega de sofrer". Encontrei Fernando Gabeira num avião e pedi que ele me recomendasse à "Folha", para o qual ele escrevia. Pois							Adeus	
3	Arnaldo Jabor	"Como podem 60 milhões de pessoas serem tão estupidas?" Essa foi a manchete de capa do jornal inglês "The Guardian", quando Bush foi reeleito. E hoje? 52 milhões de imbecis jogaram fora a Grã Bretanha por ignorância e velhice (a maioria era de velhos burros). Como sentenciou o "The Economist", "foi um gesto de automutilação", "impensado, preconceituoso". Vocês viram aquele sósia do Trump, o Boris Johnson, ex-prefeito de Londres? Pois é, na ultima hora ele traliu o babaca do Cameron, que convocou aquele plebiscito desnecessário e imprudente, e liderou o "leave". Esse Boris é um rato igual ao Trump: o mesmo cabelinho louro, mesmas fuças boçais, mesmas frases agressivas e populistas para o povo entender, ou melhor, "não entender" a complexa situação econômica e política de hoje. O Reino Unido tem uma eterna saudade do império que se estendeu ao mundo todo. Ainda se sentem donos de um passado glorioso. Usando essa estupidez, Boris arrasou o Reino Unido. O triunfo da barbárie, da estupidez está no mundo todo. A Síria agoniza nas mãos daquele assassino Assad, que destrói o próprio país, envia milhões de desgraçados para a Europa e não pode ser destruído porque o outro assassino,							O triunfo da estupidez	
4	Arnaldo Jabor	Você está andando pela rua, e, de repente, uma imensa tempestade de luz cai sobre sua cabeça, como o sol despencando do céu. Você não sabe o que é nem vai saber nunca porque você derreteu como um sorvete em dois segundos. Fica um lago de seu corpo em volta de seus sapatos, enquanto a cidade inteira vira um deserto fervente povoado por cadáveres que vagam como zumbis pelas ruas em fogo. Falo assim para ver se sentimos no corpo o intenso horror do "segundo holocausto" da Guerra: as bombas atômicas no Japão. Há 71 anos, em 6 e 9 de agosto de 1945, os norte-americanos destruíram Hiroshima e Nagasaki. Todo ano me repito e escrevo artigos sobre a bomba nesta data, não apenas para condenar um dos maiores crimes da humanidade, mas para lembrar que o impensável pode acontecer a qualquer momento. Tudo pode acontecer hoje em um mundo onde um psicótico como Trump, um hitlerzinho repulsivo, pode ser candidato a presidente dos Estados Unidos. Isso não podia acontecer e, no entanto, acontece. A destruição de Hiroshima e Nagasaki, três dias depois, inaugurou a "guerra preventiva" de hoje. O holocausto dos judeus na Segunda Guerra fecha o século XX, ainda no contexto de							Hiroshima mon amour	
5	Arnaldo Jabor	Vivemos um suspense histórico, uma situação de trágicos conflitos descentralizados no mundo todo, principalmente no Oriente Médio. Como isso começou? Alguma coisa ou alguém deflagrou este tempo. Na minha opinião, foi o George W. Bush, nossa besta do apocalipse. Ele é culpado por tudo que acontece no mundo atual, e ninguém fala nele. Bush está pintando quadros em sua fazenda no Texas enquanto o mundo que ele armou se destrói. Finalmente, depois de 13 anos dessa vergonha, a Comissão de Crimes de Guerra de Kuala Lumpur, na Malásia, julgou e condenou Bush e Cheney por crimes de guerra. Isso. Claro que não há quem prenda o nefasto elemento. Mas já é um consolo. Tudo começou com a absurda invasão do Iraque em 2003. A invasão do Iraque foi um erro tão grave quanto, digamos, atacar o México por causa do bombardeio a Pearl Harbour em 1941. Aconselhado por seu vice-papai Dick Cheney – um dos piores ratos da América –, Bush mentiu que o Iraque tinha armas de destruição em massa. partir daí, Bush continuou a construir nosso futuro apavorante. Ele não era um Hitler nem um Mussolini, com seus dogmas psicóticos. Ele era a estupidez destrutiva, com trapalhadas							A boca, Bush e Trump	
6	Arnaldo Jabor	Dilma e o PT continuam a bradar que está em curso um golpe contra eles. Vão berrar isso nas Olimpíadas, vão continuar até 2018, quando esperam eleger o Lula. Mas creio que esse demagogo narcisista encontrará seu destino antes disso. É espantoso ver o ardor com que a "Barbie" de esquerda Gleisi Hoffmann e o Lindberg Farias, bem conhecido em Nova Iguaçu, defendem Dilma. Por que será? Para mostrar força, já que ambos são investigados na Lava Jato? E o José Eduardo Cardozo? Ele parece estar lutando pela própria vida. Sua fidelidade canina é emocionante. O que será que ele quer? Algum sonho de poder ou é só amor? Todos se afeiram à tecnicidade das chamadas "pedaladas fiscais", questionando-as, periciando-as, como se esse detalhe fosse a única razão para o impedimento. Sem dúvida, foram o irrefutável crime contábil de seu governo. Mas não só as malandragens da administração são crimes; também foram espantosos os desastres econômicos e políticos que essas práticas provocaram. Foi golpe, sim, quando deram as pedaladas, desrespeitando a Lei de Responsabilidade Fiscal, para fingir que as contas estavam sob controle. Mais do que aumentar o endividamento, o governo recorreu a							O golpe, o golpe, o golpe	
7	Arnaldo Jabor	A realidade está mais louca do que a ficção. Assim sendo, a ficção tem de ser muito mais louca do que a realidade. A destruição ambiental, a sordidez mercantil, a estupidez no poder, o fanatismo do terror, em suma, toda a tempestade de bosta que nos ronda está muito além de qualquer crítica. O mal é tão profundo que denunciá-lo ficou inútil. Essa anomalia da vida atual aumenta a tradicional paranoia ocidental, principalmente nos Estados Unidos. E aí surge um estranho fenômeno que tento entender: a vontade de salvar o país e um desejo simultâneo de destruí-lo. A América parece querer suicidar-se. Por exemplo, a possibilidade de Trump ser presidente já é um filme de horror. Se esse rato for eleito, aí, sim, o mundo pode acabar. Também na cultura norte-americana, são impressionantes os filmes de ação e catástrofe que destroem o país ou o mundo, produzidos por Hollywood. É estranho; imaginem o cinema francês destruindo Paris sem parar, invadido por alienigenas (aliás, como os terroristas), ou o cinema brasileiro arrebitando o Pão de Açúcar e o Corcovado! Eles acham isso normal. E lucrativo. Vejam os filmes dos últimos anos: "Independence Day", "Godzilla", "Armagedon", "Terremoto – A Falha							O "suicídio" da América	
	Arnaldo Jabor	Acho muito boas as denúncias recentes. Elas nos fazem avançar, mesmo de lado, como síris do mangue. O Brasil evolui pelo que perde e não pelo que ganha. Sempre houve no país uma desmontagem contínua de ilusões históricas. Esse é nosso							Nosso atraso ficou atrasado	
	teste	Planilha1								

Fonte: Elaborada pelo autor.



Universidade Presbiteriana  
**Mackenzie**