

Ciência de Dados e Big Data

Recuperação da Informação na Web e em Redes Sociais

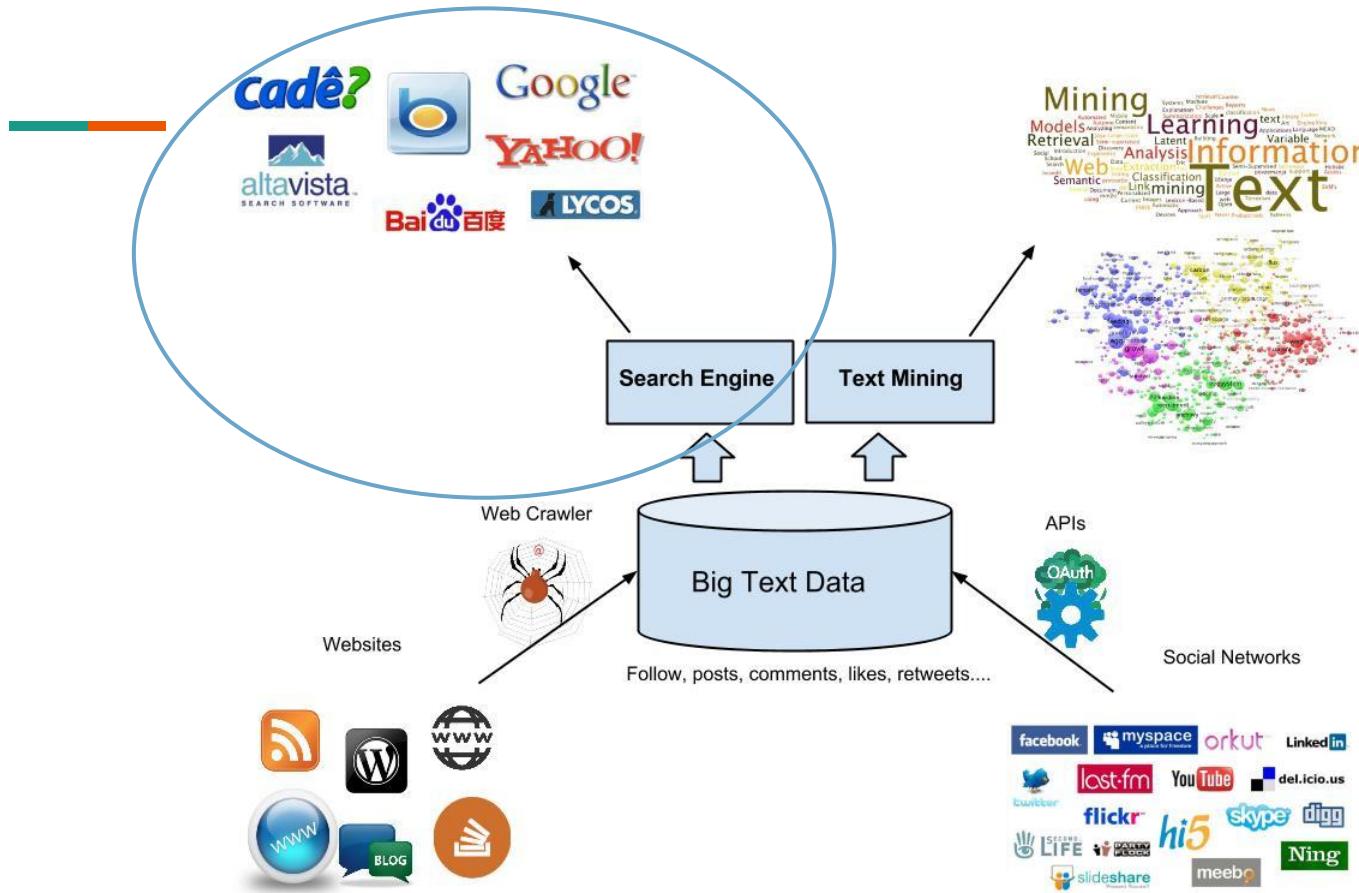
PUC-Minas IEC | Pós-Graduação Lato Sensu

Zilton Cordeiro Jr.

Projeto Final

- ❖ O **projeto final** consiste em realizar um estudo da Web para um assunto real e de livre escolha.
 - Exemplos: Automóveis, moda, música, imóveis...
- ❖ Será necessário
 - Coletar dados em texto de redes sociais e sites da Web
 - Analisar o conteúdo textual obtido
 - Analisar dados de relacionamentos entre usuários (i.e. nas redes)
 - **Relatório final**
- ❖ **Data de Entrega**
 - 15º dia após a última aula às 23:59hrs

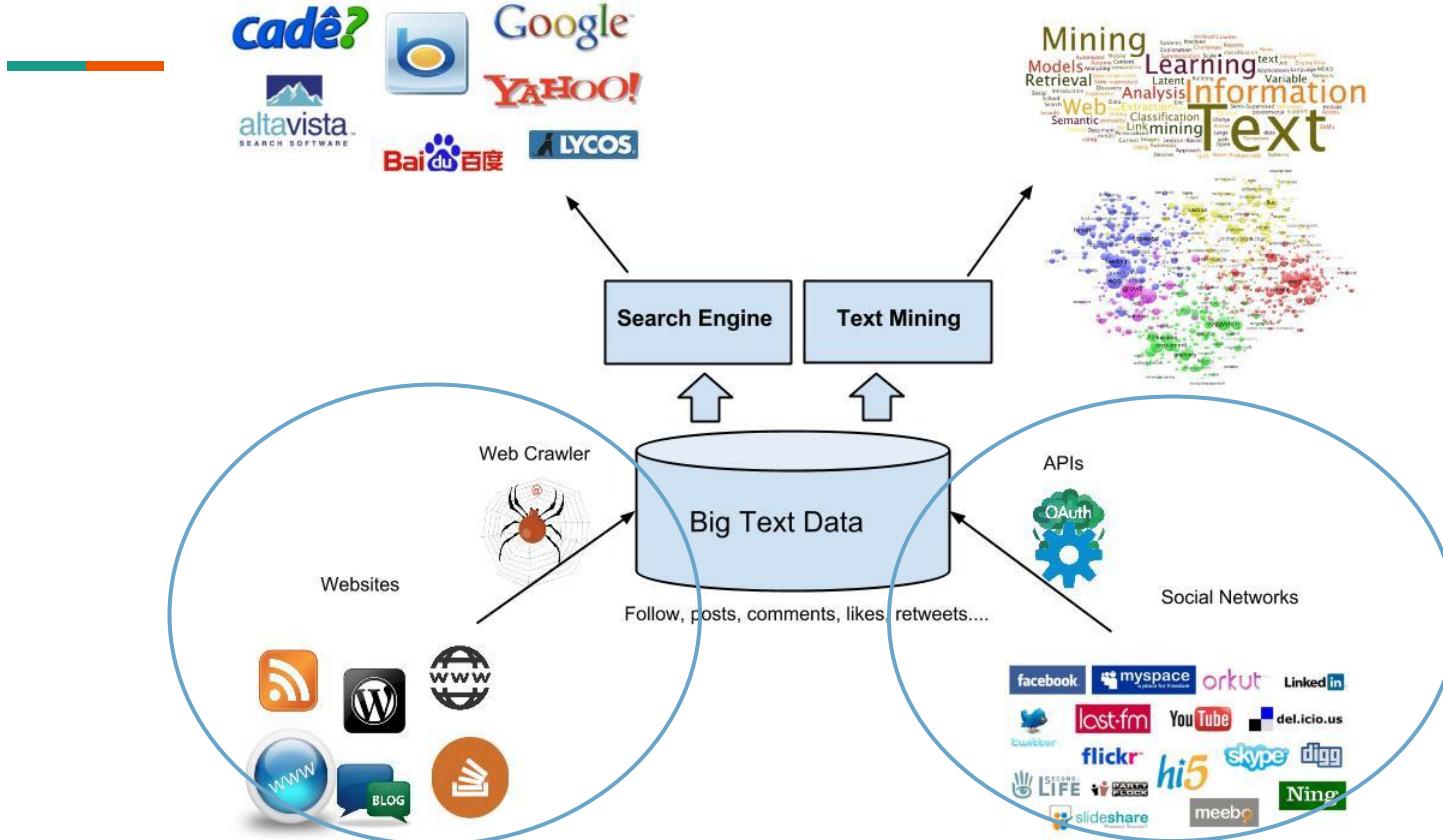
Mineração da Web e Redes Sociais



Coleta de Dados



Mineração da Web e Redes Sociais



Coleta de Dados

❖ Introdução

- São necessárias ferramentas e estratégias para o acesso aos dados na Web
- Existem diferentes formas de abordar uma tarefa de coleta
- Existem também uma série de desafios e problemas
- Atualmente muitas formas vêm sendo desenvolvidas para facilitar a disseminação de conteúdo (APIs, RSS, bases completas para download ...)

Técnicas de Extração

◆ Formatos de conteúdo na Web



Técnicas de Extração

❖ HTML

```
<title>
    atlético-mg | globoesporte.com
</title>
<li class="menu-item " id="menu-2-brasileirao-serie-b">
    <a href="http://globoesporte.globo.com/futebol/brasileirao-serie-b/">
        <span class="menu-item-link">
            <span class="menu-item-title">brasileirão série b</span>
        </span>
    </a>
</li>
```

Técnicas de Extração

❖ XML

- HTML e XML são primos. Ambos identificam elementos em uma página e utilizam sintaxes similares.
- A grande diferença entre HTML e XML é que o **HTML descreve a aparência** e as ações em uma página na rede, já o **XML** descreve o que **cada trecho de dados representa**
- Em outras palavras, o **XML descreve o conteúdo do documento**

Técnicas de Extração

❖ XML

```
<item>
<title>
Intestino gigante alerta população quanto ao câncer em Juiz de Fora
</title>
<link>
http://g1.globo.com/mg/zona-da-mata/noticia/2015/08/intestino-gigante-alerta-populacao-quanto-ao-cancer-em-juiz-de-f
ora.html
</link>
<description><a
href='http://g1.globo.com/mg/zona-da-mata/noticia/2015/08/intestino-gigante-alerta-populacao-quanto-ao-cancer-em-jui
z-de-fora.html' alt='Intestino gigante alerta população quanto ao câncer em Juiz de Fora'><img border='0'
src='http://s2.glbimg.com/beySFY5R3UYkuGuEsBXGksS4AWU=/90x68/s.glbimg.com/jo/g1/f/original/2015/08/07/jf_intesti
nogigante.mov_snapshot_01.43_2015.08.07_17.53.28.jpg' alt='Intestino gigante alerta população quanto ao câncer em
Juiz de Fora' title='Intestino gigante alerta população quanto ao câncer em Juiz de Fora' /></a><br />Objetivo também é
chamar da população para exames contra a doença. Evento ocorre na cidade até domingo (9) na UFJF.
</description>
<category>Zona da Mata</category>
<pubDate>Fri, 07 Aug 2015 19:19:21 -0300</pubDate>
</item>
```

Técnicas de Extração

◆ JSON

- O formato JSON é eficiente e simples. Sua estrutura é de fácil entendimento e utiliza convenções comuns em muitas linguagens de programação.
- Muitos desenvolvedores utilizam essa alternativa, porém o XML ainda é um bom candidato, especialmente quando se quer **validações mais robustas**, usando os chamados **XML-schemas**

Técnicas de Extração

◆ JSON

```
{  
    "name": "John",  
    "age": 31,  
    "city": "New York"  
}
```

Tipos de Coleta

❖ Web Crawling vs Web Scraping

➤ Web Scraping

Extração de conteúdo específico

Geralmente focam em **sites específicos** em busca de **dados específicos**.

Ex: comparação de preços, notícias sobre moda...

➤ Web Crawling

Cobertura em largura

Essencialmente o que Google, Yahoo, Bing, etc. fazem, procuram por **qualquer** informação

Tipos de Coleta

❖ APIs e Feeds RSS

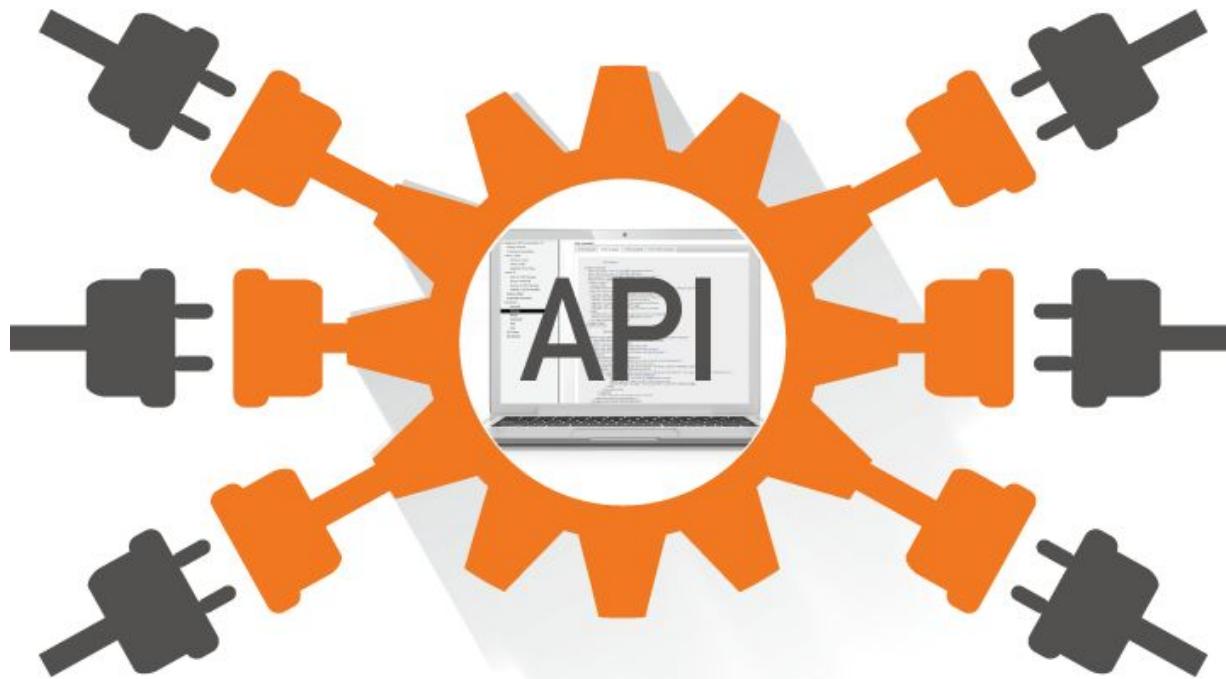
➤ API

Fornece acesso direto a dados de uma aplicação
Formatos bem definidos

➤ Feed RSS

Disponibiliza conteúdo de notícias e textos completos de sites e blogs
Atualização contínua do conteúdo
Formato de entrega preestabelecido

Coleta de Dados via



Interface de Programação de Aplicativos

❖ API - Application Programming Interface

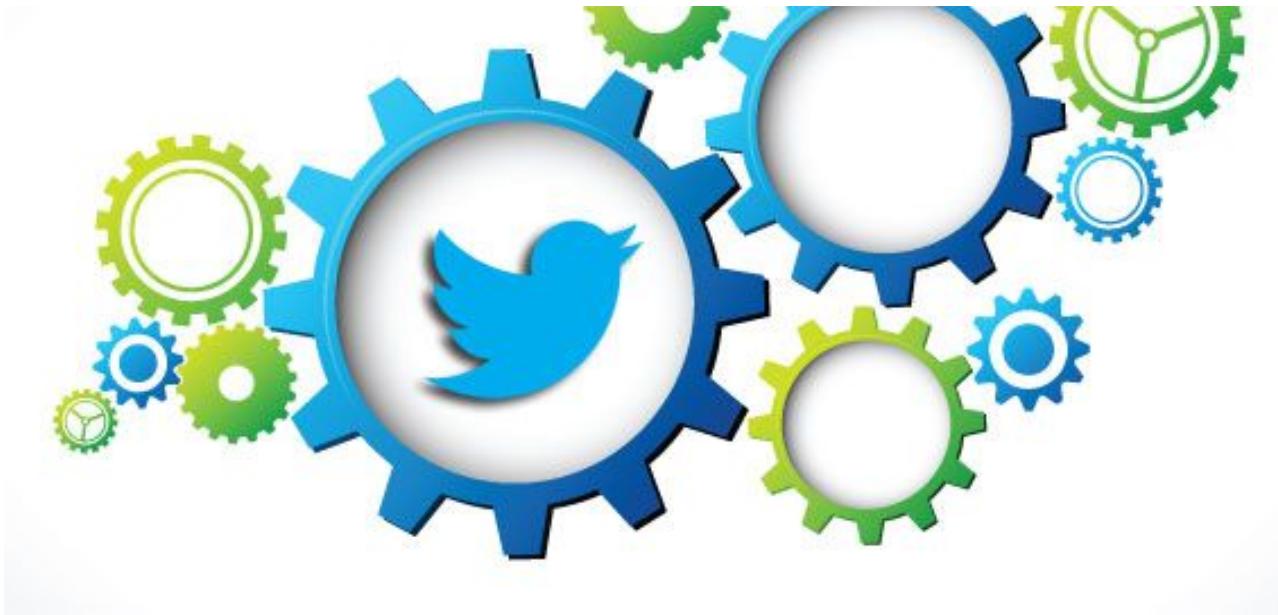
- É um conjunto definido de mensagens de **requisição e resposta** HTTP, geralmente expressado nos formatos XML ou JSON
- Um exemplo popular é a utilização para leitura e publicação de mensagens no Twitter
- No nosso caso queremos coletar dados

Interface de Programação de Aplicativos

❖ Coleta de dados via APIs

- Não é necessária a implementação de scripts externos de coleta, uma vez que os dados podem ser obtidos por simples requisições estabelecidas pela aplicação
- É preciso criar e registrar uma aplicação para obter credenciais de acesso
- Fácil extração dos dados disponibilizados em formatos semi-estruturados

API - Twitter



Cadastro de desenvolvedor de aplicações



1. Caso não tenha conta, criar um usuário em <https://twitter.com/>

1. Acessar área de desenvolvedor <https://developer.twitter.com/en/apply/user>

1. Preencher o formulário solicitando acesso como desenvolvedor
[Exemplos de Alunos e dados preenchidos](#)

Cadastro de desenvolvedor de aplicações

- Após o preenchimento pode ser que o Twitter te mande um e-mail para saber mais informações



- Responda (em inglês) que suas intenções são educativas e que você vai realizar tarefas como text processing, count terms frequency, text similarities matching...
- E que além disso não pretende postar na conta de outros usuários e nem divulgar esses resultados

Após aprovação do Twitter

<https://developer.twitter.com/en/account/get-started>

The screenshot shows the Twitter Developer website's "Get started" section. At the top, there's a purple navigation bar with links for "Developer", "Use cases", "Products", "Docs", and "More". On the right side of the bar are "Dashboard" and a user profile icon for "ziltonjunior". Below the bar, the page title "Account / Get started" is displayed. The main content area has a purple background and features a large "Welcome!" heading. It congratulates the user on creating a new account and provides steps to get started with creating an app and setting up a development environment. A sidebar on the left is titled "Helpful tools" and includes links for "Dive into the docs", "View API usage", "Have a question?", and "Looking for something else?".

Welcome!

Congratulations! You have successfully created a new Twitter developer account. With this account, you now have access to new APIs, app management, and tools to facilitate and support development.

Below are a few steps to help you create an app and to get up and running with the new premium APIs.

If you're planning to use our standard APIs instead of our premium APIs, simply follow the steps below to create an app, then refer to the "[Getting started](#)" guide in our documentation for next steps.

The screenshot shows the "Get started" section of the Twitter Developer site. It features three main steps: "Create an app", "Set up a dev environment", and "Start using the endpoints!". Each step is accompanied by a brief description and a link to further documentation. The sidebar on the left remains the same as in the previous screenshot.

Get started

- Create an app**
To use an API, we require you create an app as part of our OAuth authorization scheme. Visit the [Apps](#) page of this developer portal to create one. Then, return to this page to complete the next step.
- Set up a dev environment**
To begin using the new Premium APIs, you need to [set up one or more dev environments](#) for the endpoint — and connect it to an app. Dev environments can be used to isolate usage, rules, rate limits, and more. If you are planning to use our standard APIs, you can skip this step.
- Start using the endpoints!**
Once you've set up your account, accessing the endpoint is super simple. Check out our [documentation](#) and [API reference](#) for additional details about each endpoint.

Uma vez aprovados, criamos um App

<https://developer.twitter.com/en/apps>

The screenshot shows the Twitter Developer Apps page. At the top, there's a purple navigation bar with links for Developer, Use cases, Products, Docs, More, Dashboard, ziltonjunior, and a user icon. Below the bar, the word "Apps" is displayed in blue, and a "Create an app" button is in a blue box. A single app entry is shown in a card: it has a logo with a gear and a bird, the name "coleta_ri", the App ID "16084167", and a "Details" button.

Uma vez aprovados, criamos um App

<https://developer.twitter.com/en/apps>

The screenshot shows the Twitter Developer website's 'Create an app' interface. On the left, there's a sidebar with links like 'Understanding apps', 'What is an app?', 'Why register an app?', and 'Which products require an API key?'. The main area is titled 'App details' and contains fields for 'App name' (with a character limit of 32) and 'Application description'. A large text area for the description starts with 'Please be detailed.' An arrow points from the text 'Preenchemos as informações necessárias' to the 'App name' field.

Apps / Create an app

Create an app

Developer Use cases Products Docs More

Dashboard ziltonjunior

Twitter Understanding apps

What is an app?

Why register an app?

Which products require an API key?

App details

The following app details will be visible to app users and are required to generate the API keys needed to authenticate Twitter developer products.

App name (required) ?

Maximum characters: 32

Application description (required)

Share a description of your app. This description will be visible to users so this is a good place to tell them what your app does.

Please be detailed.

Between 10 and 200 characters

Preenchemos as informações necessárias

Uma vez aprovados, criamos um App

<https://developer.twitter.com/en/apps>



- Se não tiver uma url própria para usar, pode utilizar a minha:

<https://homepages.dcc.ufmg.br/~zilton/>

Com o App criado, temos as credenciais

<https://developer.twitter.com/en/apps>



Screenshot of the Twitter Developer Dashboard showing the 'Apps' section.

Header navigation: Twitter icon, Developer, Use cases, Products, Docs, More, Dashboard, ziltonjunior, profile icon.

Apps

Create an app

| | | |
|---|--------------------|-------------------------|
|  coleta_ri | App ID 16084167 | Details |
|---|--------------------|-------------------------|

Com o App criado, temos as credenciais

<https://developer.twitter.com/en/apps>

The screenshot shows the Twitter Developer API Keys and tokens page. At the top, there's a purple navigation bar with links for Developer, Use cases, Products, Docs, More, Dashboard, and a user profile for ziltonjunior. Below the navigation, the URL https://developer.twitter.com/en/apps is visible. The main content area has tabs for App details, Keys and tokens (which is selected and highlighted with a red box), and Permissions. The Keys and tokens section contains fields for Consumer API keys (API key: 5KkOejODws, API secret key: Jm2r9zE8nGVffMUF) and Access token & access token secret (Access token: 68135902-HzmDw, Access token secret: m4Zz1xBm5n). It also includes a 'Read and write' access level and buttons for Revoking or Regenerating tokens.

Keys and tokens

Keys, secret keys and access tokens management.

Consumer API keys

| | |
|------------------|------------------|
| 5KkOejODws | (API key) |
| Jm2r9zE8nGVffMUF | (API secret key) |

[Regenerate](#)

Access token & access token secret

| | |
|----------------|-----------------------|
| 68135902-HzmDw | (Access token) |
| m4Zz1xBm5n | (Access token secret) |

Read and write (Access level)

[Revoke](#) [Regenerate](#)



KNIME - Coleta de Dados via APIs

KNIME - Coleta via API

❖ Coleta no Twitter

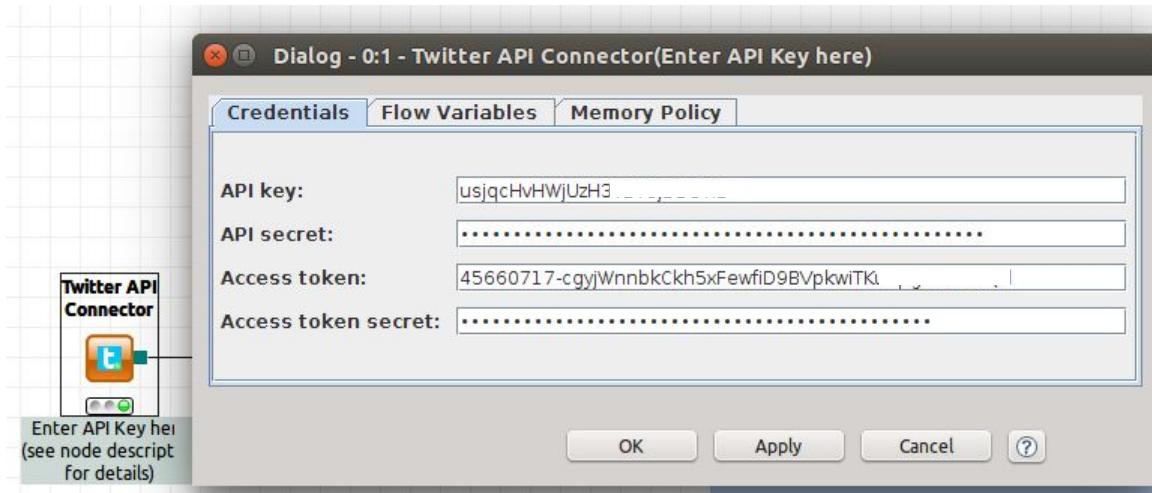
- Fluxo: DataCollection-pratica
- Prática em coleta e armazenamento
- Exercício de criação de aplicação e requisição de consultas

Coleta em Redes Sociais

❖ Conexão e Autorização

via node

*Twitter API
Connector*



- ❖ Para obter as credenciais de acesso é preciso ter uma aplicação registrada no Twitter.

Coleta em Redes Sociais

❖ Consulta

Com o node *Twitter Search* é possível realizar buscas e os dados já vêm formatados em uma tabela. Não é necessário nenhum processo de extração.

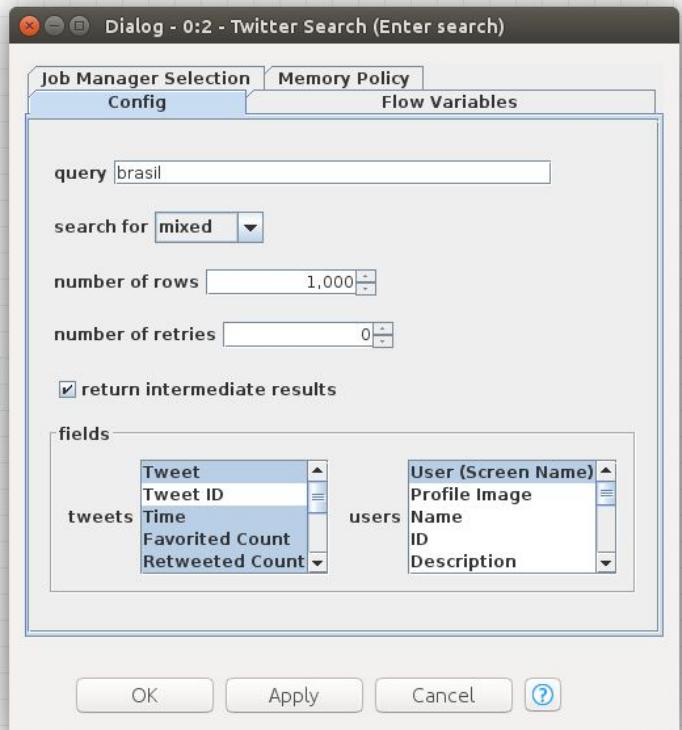
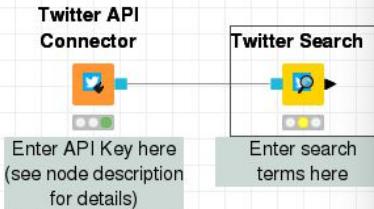


Table "default" - Rows: 1000 Spec - Columns: 6 Properties Flow Variables

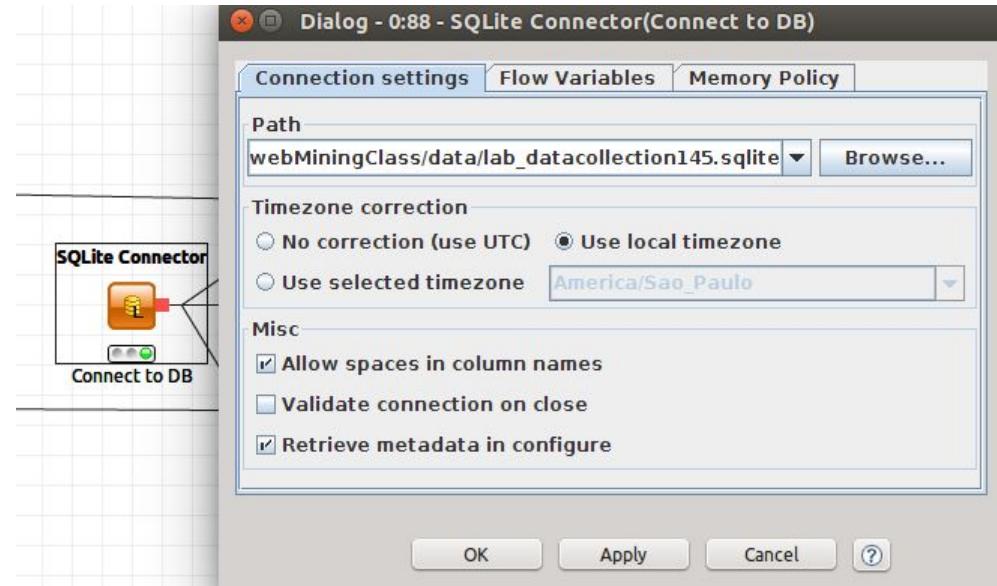
| Row ID | S Tweet |
|--------|---|
| Row0 | Nenhuma outra proposta de reforma foi tão firme contra privilégios. Está claro no texto da Nova Previdência: quem tem menos, paga menos. É preciso consciência. A mudança é dura, |
| Row1 | De laranjal o Brasil entende https://t.co/Nj0JSmmj6K |
| Row2 | BRASIL ACIMA DE TUDO! [REDACTED] |
| Row3 | @raquelgotthilf @manujpeg É cultura do CN no Brasil inteiro mo |
| Row4 | #EuSouUnimigodaGlobo A Rede Globo é o câncer do Brasil. @RedeGlobo https://t.co/6KKv1X3alm |
| Row5 | RT @RealitySocial: Gabi dizendo o que estava pensando em falar ontem durante o paredão pra Rodrigo: |
| Row6 | RT @Manu_Bahia : Eu sou a única pessoa do Brasil que não tem esse tênis https://t.co/NTGc7Hyclo |

Coleta em Redes Sociais

❖ Salvando em Banco - Conecta

O *SQLite* é um pequeno banco de dados que pode ser disponibilizado junto com uma aplicação.

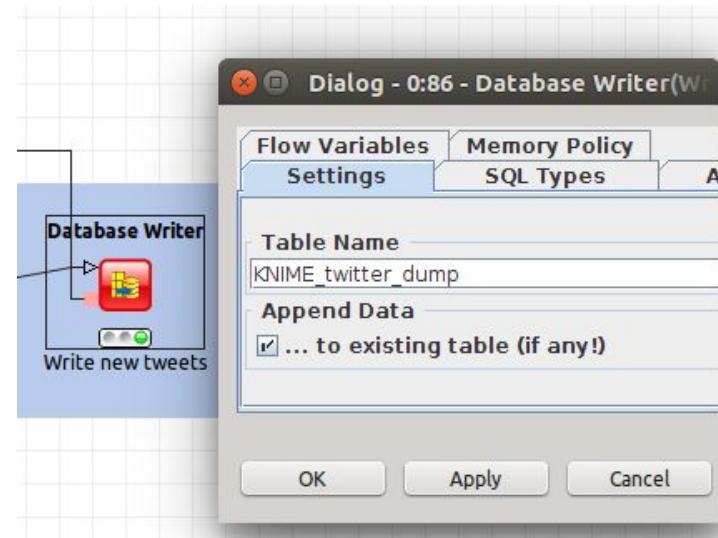
Utilizando o node *SQLite* defina o nome do banco e local onde será salvo no disco (em seu workspace).



Coleta em Redes Sociais

❖ Salvando em Banco - Escreve

Com o node *Database Writer* defina o nome da tabela e execute o node. Assim o banco será preenchido e salvo com os dados da coleta



Coleta em Redes Sociais

❖ Salvando em Banco - Atualiza (SET)

The image shows a KNIME workflow on the left and a 'Select SET Columns' dialog on the right.

Workflow Node: A 'Database Update' node is highlighted with a red box. It has inputs from a 'Favoriate/RetweetsRe...' node and outputs to a 'User' node.

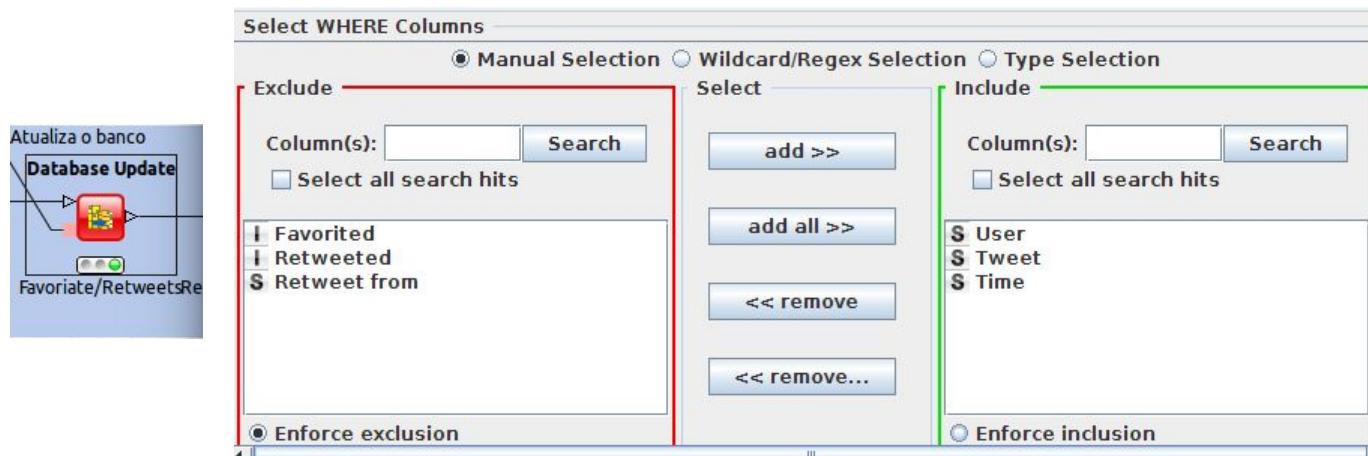
Select SET Columns Dialog:

- Table Name:** KNIME_twitter_dump
- Select SET Columns:** Radio buttons for Manual Selection (selected), Wildcard/Regex Selection, and Type Selection.
- Exclude:** A red-bordered section containing:
 - Column(s): Search
 - Select all search hits
 - Items: User, Tweet, Time, Retweet from
- Select:** Buttons: add >>, add all >>, << remove, << remove...
- Include:** A green-bordered section containing:
 - Column(s): Search
 - Select all search hits
 - Items: Favorited, Retweeted
- Buttons:** Enforce exclusion (in Exclude) and Enforce inclusion (in Include).

Com o node Database Update defina os campos a serem atualizados (SET)

Coleta em Redes Sociais

❖ Salvando em Banco - Atualiza (WHERE)



Neste caso, sempre que houver novas interações para um mesmo tweet previamente coletado (mesmo usuário, tweet e hora de publicação)

Coleta em Redes Sociais

❖ Salvando em Banco - Filtrando dados atualizados

Dados novos possuem o valor 0 (zero) na coluna updateStatus, criada pelo node *Update*

The screenshot shows the KNIME workflow interface. At the top, there is a table view titled "Table 'default' - Rows: 34 Spec - Columns: 7 Properties Flow Variables". The table has columns: Row ID, User, Tweet, Time, Fa..., R..., Re..., Updat... . Below the table is a configuration dialog for a "Row Filter" node, titled "Dialog - 0:76 - Row Filter(Remove Updated)". The dialog has three tabs: Filter Criteria, Flow Variables, and Memory Policy. The "Filter Criteria" tab is active, showing the following configuration:

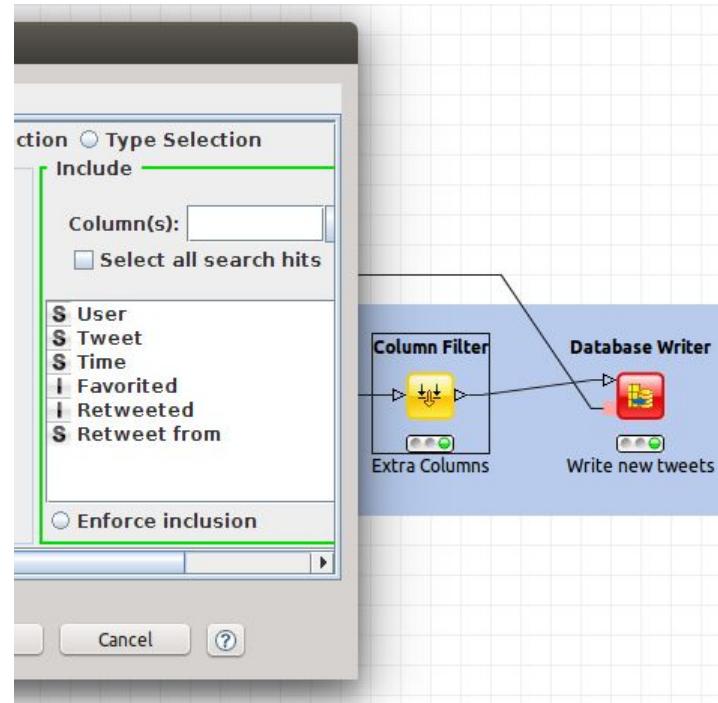
- Column value matching:
 - select the column to test: **UpdateStatus**
 - filter based on collection elements
- matching criteria:
 - use pattern matching
 - pattern: **0** **v=?**
 - cont.
 - case sensitive match
 - regu
 - use range checking

To the right of the dialog, a small preview window shows the "Row Filter" node with the label "Remove Updated".

Coleta em Redes Sociais

- ❖ Salvando em Banco - Salvando novos dados

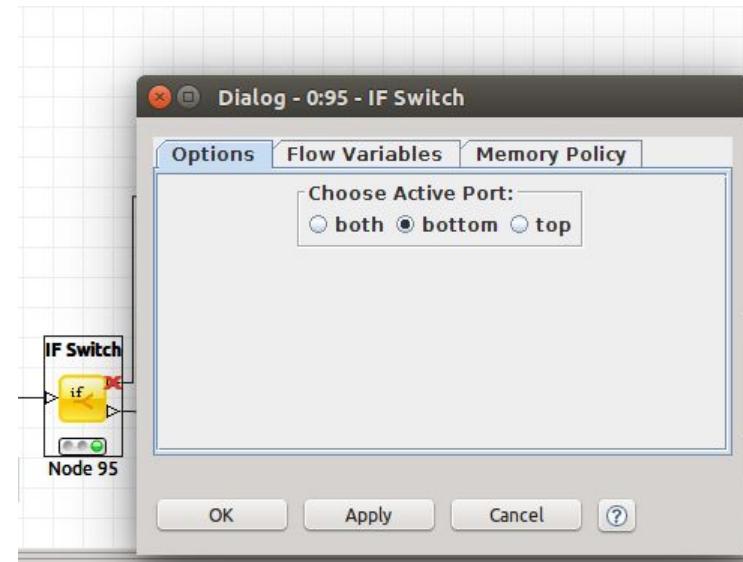
Selecionando apenas colunas relativas aos tweets



Coleta em Redes Sociais

❖ Opção de caminhos em um fluxo

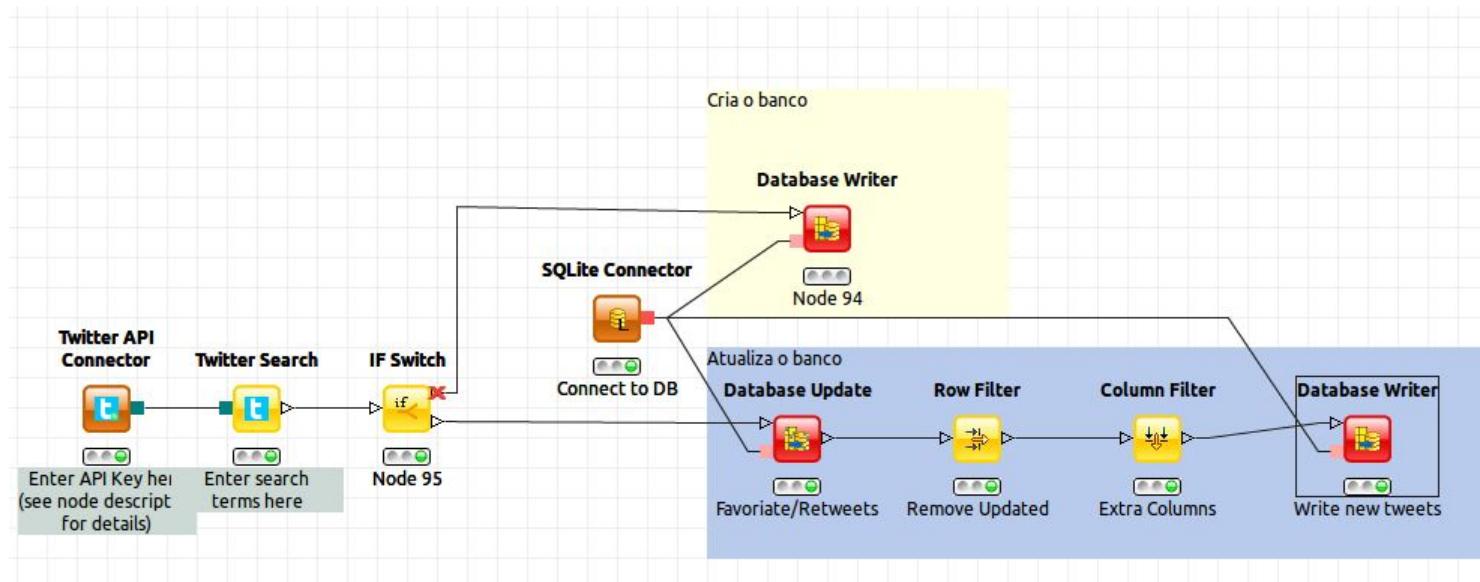
O node *IF Switch* permite configurar e trocar qual caminho do fluxo será executado em cada momento



Coleta em Redes Sociais

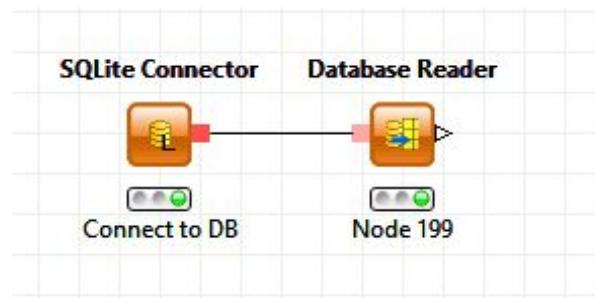
❖ Opção de caminhos em um fluxo

Podemos definir o caminho de execução onde o banco será sempre atualizado



Coleta em Redes Sociais

- ❖ Lendo dados com o node *Database Reader*



Exercício

- ❖ Criar uma aplicação no twitter (**tarefa**)

-
- Obter credenciais
 - Coletar e armazenar uma nova consulta (criar nova tabela/db)
 - Pré-processar o texto e verificar frequência de termos

- ❖ Utilidade no projeto final

- Coletar e extrair conteúdo de interesse no twitter
 - Escolha termos relevantes para o assunto

Entrega via formulário

Coleta de Dados via



Feeds RSS

◆ Feeds RSS

- Usado principalmente em sites de notícias e blogs para compartilhamento de notícias com textos completos e até mesmo arquivos multimídia.
- Um tipo de programa conhecido como "feed reader" ou agregador pode ler diversas páginas habilitadas para RSS.
Ex: agregar diversos itens relativos a futebol de diversos feeds de esporte criando então um novo feed de futebol.
- Existem tanto sites que funcionam como feed readers como extensões para navegadores web.



Feeds RSS

◆ Feeds RSS em XML

```
<item>
<title>
Intestino gigante alerta população quanto ao câncer em Juiz de Fora
</title>
<link>
http://g1.globo.com/mg/zona-da-mata/noticia/2015/08/intestino-gigante-alerta-populacao-quanto-ao-cancer-em-juiz-de-f
ora.html
</link>
<description><a
href='http://g1.globo.com/mg/zona-da-mata/noticia/2015/08/intestino-gigante-alerta-populacao-quanto-ao-cancer-em-jui
z-de-fora.html' alt='Intestino gigante alerta população quanto ao câncer em Juiz de Fora'><img border='0'
src='http://s2.glbimg.com/beySFY5R3UYkuGuEsBXGksS4AWU=/90x68/s.glbimg.com/jo/g1/f/original/2015/08/07/jf_intesti
nogigante.mov_snapshot_01.43_2015.08.07_17.53.28.jpg' alt='Intestino gigante alerta população quanto ao câncer em
Juiz de Fora' title='Intestino gigante alerta população quanto ao câncer em Juiz de Fora' /></a><br />Objetivo também é
chamar da população para exames contra a doença. Evento ocorre na cidade até domingo (9) na UFJF.
</description>
<category>Zona da Mata</category>
<pubDate>Fri, 07 Aug 2015 19:19:21 -0300</pubDate>
</item>
```

Feeds RSS

❖ Coleta de Feeds RSS

- A coleta de conteúdo a partir de feeds rss é feita através de requisições HTTP
- Facilita o processo de extração dos dados, por esses já virem em um padrão estabelecido (XML) e periodicidade automaticamente controlada pelo site.
- Não exige autenticação para obter os dados





KNIME - Coleta de Dados via RSS

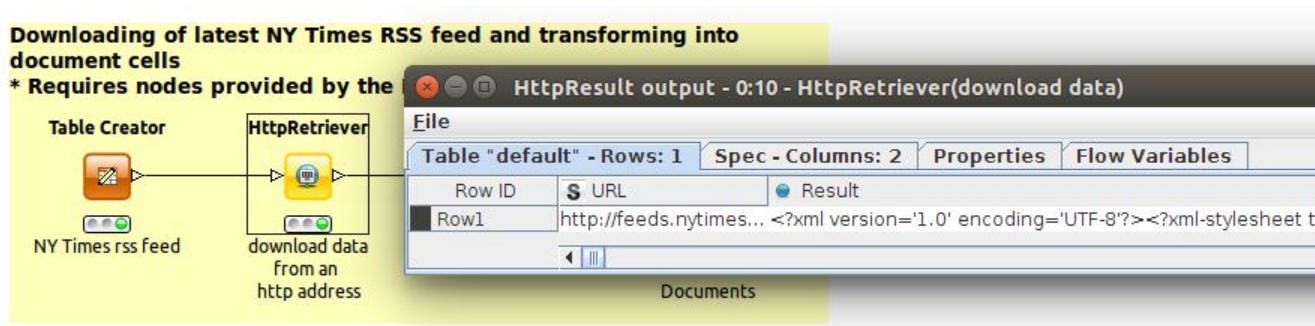
KNIME - Coleta via RSS

❖ Coleta de RSS

- Fluxo: NYTimesRSSFeed
- Prática em coleta e armazenamento
- Exercício de coleta de feeds específicos

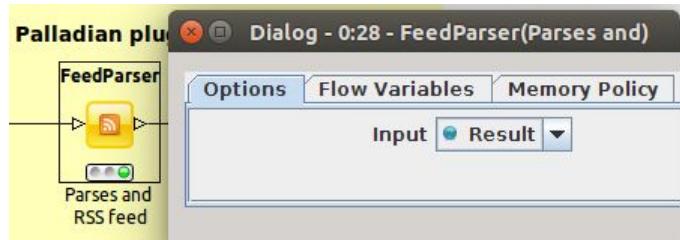
Leitura de Feeds RSS

- ❖ A tabela possui a URL a ser coletada e o node *HttpRetriever* faz o trabalho de coletar o feed (recupera o XML do conteúdo)



Leitura de Feeds RSS

- ❖ O FeedParser extrai o conteúdo em XML e transforma em tabela

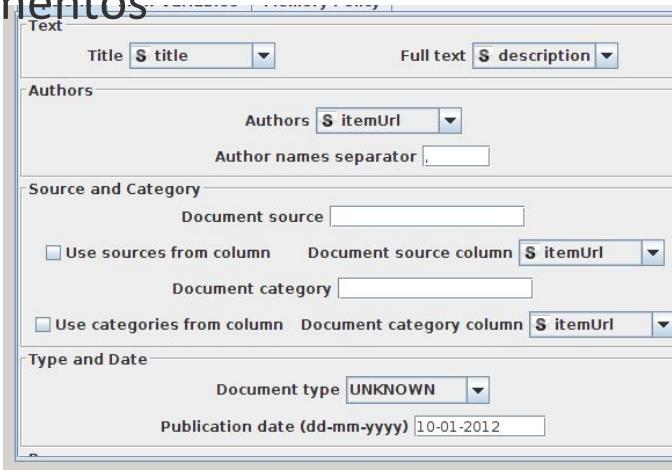


The screenshot shows a table titled 'Parsed feeds - 0:28 - FeedParser(Parses and)'. The table has a header row with columns: Row ID, feedUrl, title, description, published, and itemUrl. Below the header are 20 data rows. The data is partially visible, showing columns for feedUrl, title, description, published, and itemUrl. The 'published' column contains dates like '16.jul.2015 21:32:33.0...' and URLs like 'http://rss.nytimes.co...'. The 'itemUrl' column also contains URLs like 'http://rss.nytimes.co...'. The table has tabs for 'File', 'Table "default" - Rows: 20', 'Spec - Columns: 5', 'Properties', and 'Flow Variables'.

| Row ID | feedUrl | title | description | published | itemUrl |
|--------|--------------------|----------------------------|-------------------------|---------------------------|--------------------------|
| Row0 | http://feeds.ny... | World Briefing Euro... | Using materials pr... | 16.jul.2015 21:32:33.0... | http://rss.nytimes.co... |
| Row1 | http://feeds.ny... | 'Snapback' Is Easy ... | In the event that lr... | 16.jul.2015 21:29:06.0... | http://rss.nytimes.co... |
| Row2 | http://feeds.ny... | Putin Rejects U.N. Tr... | Vladimir V. Putin h... | 16.jul.2015 21:28:02.0... | http://rss.nytimes.co... |
| Row3 | http://feeds.ny... | Sea Warming Leads ... | Five nations have ... | 16.jul.2015 21:22:19.0... | http://rss.nytimes.co... |
| Row4 | http://feeds.ny... | Iranian Hard-Liners ... | Iranian analysts a... | 16.jul.2015 21:18:25.0... | http://rss.nytimes.co... |
| Row5 | http://feeds.ny... | World Briefing Afric... | A senior command... | 16.jul.2015 21:09:21.0... | http://rss.nytimes.co... |
| Row6 | http://feeds.ny... | World Briefing Afric... | Two bombings beli... | 16.jul.2015 21:06:03.0... | http://rss.nytimes.co... |
| Row7 | http://feeds.ny... | Brazil Adds to Tally o... | Federal prosecuto... | 16.jul.2015 20:51:38.0... | http://rss.nytimes.co... |
| Row8 | http://feeds.ny... | Event ISIS Affiliate Cl... | The group Sinai Pr... | 16.jul.2015 20:38:16.0... | http://rss.nytimes.co... |

Leitura de Feeds RSS

- ❖ Transformando os dados em documentos



Assim como na prática de similaridades, precisamos de **objetos do tipo “Document”**.

Utilizamos o node “**Strings to document**”

Qual coluna da tabela de dados representa o texto de um documento? qual coluna seria o autor? e o título?

| Table "default" - Rows: 20 | | | | | | | Spec - Columns: 6 | Properties | Flow Variables |
|----------------------------|---------------------------------|-------------------------------|---|--------------------------|--|---------------------------|-------------------|------------|----------------|
| Row ID | \$ feedUrl | \$ title | \$ description | \$ published | \$ itemUrl | Document | | | |
| Row0 | http://feeds.nytimes.com/nyt... | Memo From Afghanistan: ... | President Ashraf Ghani is seeking t... | 09.Mar.2015 09:27:02.... | http://rss.nytimes.com/c/34625/f/642565/s... | "Memo From Afghanistan | | | |
| Row1 | http://feeds.nytimes.com/nyt... | South Korea Split Over Ho... | The initial contrition, led by a parad... | 09.Mar.2015 09:17:21.... | http://rss.nytimes.com/c/34625/f/642565/s... | "South Korea Split Over T | | | |
| Row2 | http://feeds.nytimes.com/nyt... | Op-Ed Contributor: Embra... | Could the flow of young Italian emi... | 09.Mar.2015 08:57:57.... | http://rss.nytimes.com/c/34625/f/642565/s... | "Op-Ed Contributor: Emb | | | |
| Row3 | http://feeds.nytimes.com/nyt... | Contributing Op-Ed Writer:... | Alienated voters faced with some u... | 09.Mar.2015 08:24:29.... | http://rss.nytimes.com/c/34625/f/642565/s... | "Contributing Op-Ed Writ | | | |
| Row4 | http://feeds.nytimes.com/nyt... | Chechen Strongman Ties K... | Ramzan A. Kadyrov suggested that... | 09 Mar 2015 05:51:11 ... | http://rss.nytimes.com/c/34625/f/642565/s... | "Chechen Strongman Tie | | | |

Exercício

❖ Coletar outra fonte (**tarefa**)

- Encontre um feed rss diferente para teste
- Salvar dados em uma tabela ou db
- Pré-processar o texto e verificar frequência de termos

❖ Utilidade no projeto final

- Coletar e extrair conteúdo de interesse em blogs e sites de notícias
 - Escolha fontes relevantes para o assunto escolhido
- Transformar em documentos que podem em seguida ser processados para busca e outras técnicas

Coletores Web



Coletores Web

❖ Na ausência de feeds

- E quando não existem feeds disponíveis para o site pretendido?
- Necessário aplicar técnicas de WebScraping ou Crawling

Coletores Web

◆ Coletores

- Navegadores automáticos entre páginas web que visam armazenar uma cópia local das páginas encontradas
- Devem obedecer algumas restrições ao visitar sites
- O Protocolo de Exclusão de Robôs (Robot Exclusion Protocol) especifica algumas regras de acesso
- Principal regra é deixar um intervalo de tempo entre acessos a cada servidor.

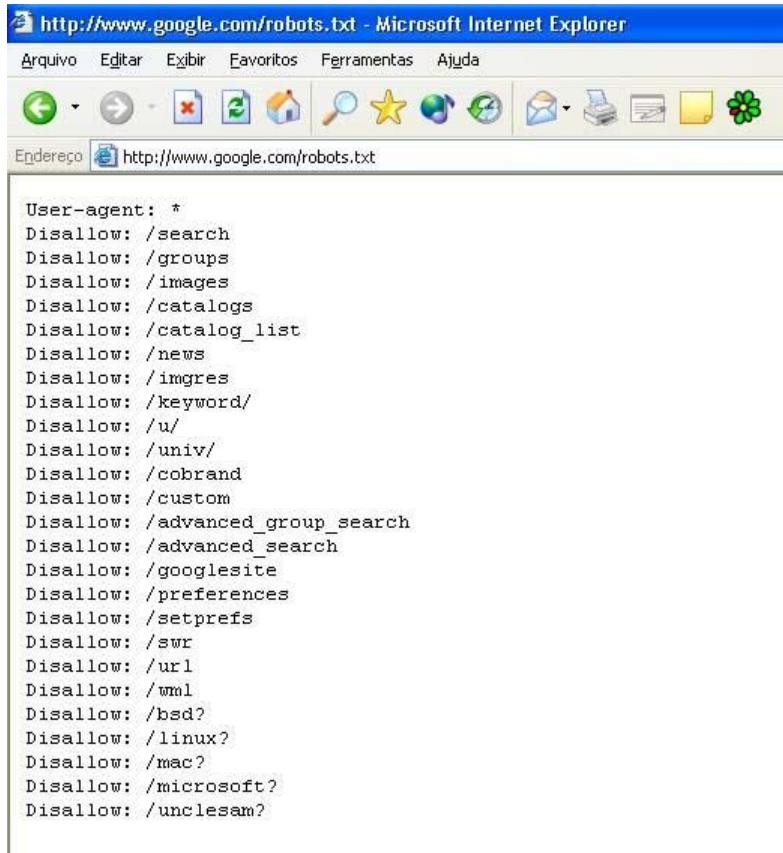
O Protocolo de Exclusão de Robôs

❖ Robots.txt

- Padrão definido em 30 de junho de 1994
- Define as permissões do coletor em um determinado site
- As diretrizes são descritas em um arquivo chamado “**robots.txt**”, localizado no servidor web coletado

O Protocolo de Exclusão de Robôs

➤ <http://NOME-SITE/robots.txt>

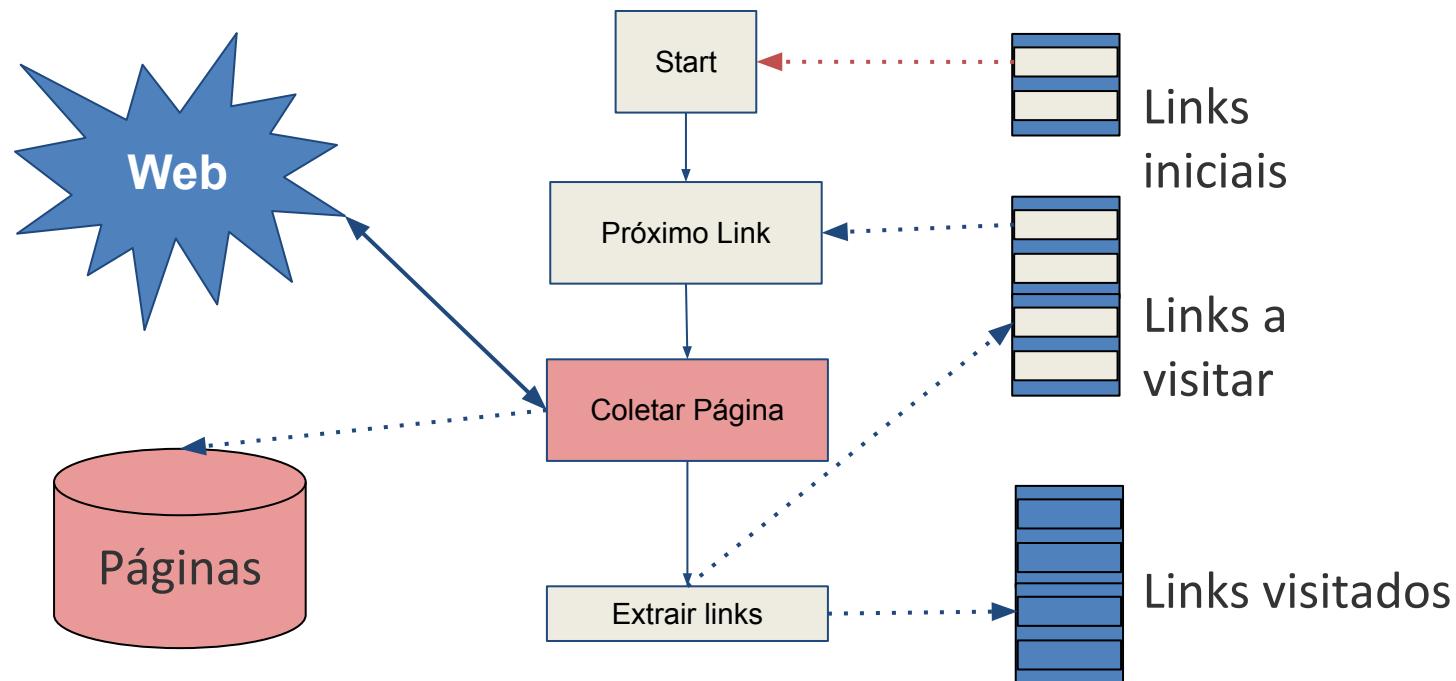


A screenshot of Microsoft Internet Explorer version 6.0 displaying the robots.txt file for the website www.google.com. The title bar reads "http://www.google.com/robots.txt - Microsoft Internet Explorer". The address bar shows the same URL. The page content area displays the following text:

```
User-agent: *
Disallow: /search
Disallow: /groups
Disallow: /images
Disallow: /catalogs
Disallow: /catalog_list
Disallow: /news
Disallow: /imgres
Disallow: /keyword/
Disallow: /u/
Disallow: /univ/
Disallow: /cobrand
Disallow: /custom
Disallow: /advanced_group_search
Disallow: /advanced_search
Disallow: /googlesite
Disallow: /preferences
Disallow: /setprefs
Disallow: /swr
Disallow: /url
Disallow: /wml
Disallow: /bsd?
Disallow: /linux?
Disallow: /mac?
Disallow: /microsoft?
Disallow: /unclesam?
```

Coletores Web

◆ Esquema gráfico do funcionamento de um coletor



Coletores de Conteúdo Específico

❖ Web Scraping

- Fetching
- Técnicas de Extração

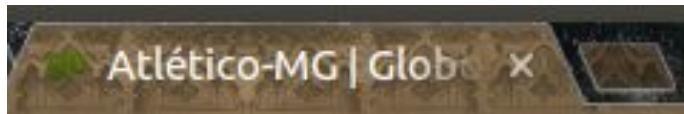
Técnicas de Extração

❖ Fetching

- Processo de realizar uma conexão ao servidor pretendido e requisitar o conteúdo para download
- Normalmente na forma de conteúdo HTML, ou com uso de APIs dados na forma XML, JSON entre outros.

Técnicas de Extração

O primeiro passo é entender a estrutura do documento HTML em que as informações estão.



```
<title> Atlético-mg | Globoesporte.com </title>
<li class="menu-item " id="menu-2-brasileirao-serie-a">
    <a href="http://globoesporte.globo.com/futebol/brasileirao-serie-a/">
        class="menu-item-link">
            <span class="menu-item-title"> brasileirão série a </span>
        </a>
    </li>
```

Técnicas de Extração - XPath

O **XPath** é uma sintaxe para navegar entre atributos e elementos em um documento html/xml.

```
<table>
    <tr>
        <td>Cell A</td>
        <td>Cell B</td>
    </tr>
</table>
```

Técnicas de Extração - XPath

O XPath

é uma sintaxe para navegar entre atributos e elementos em um documento html/xml.

Exemplos de notação XPath:

//td[1] : retorna o primeiro td

//td[position()=1] : retorna o primeiro td

//table[@class='tabelaX'] : retorna a table com a classe 'tabelaX'

//table/td[1] : retorna o primeiro td de uma <table>



KNIME - Fluxo de Coleta

KNIME - Coleta da Web

❖ Coleta

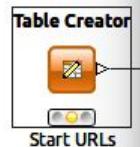
- Fluxo: WebDataExtraction-pratica
- Prática em coleta e extração de páginas web
- Exercícios extraindo dados específicos

Coleta da Web

❖ Seed - Lista de URLs

URLs podem ser lidas a partir de uma tabela.

Tabelas podem inclusive ser criadas a partir do node *Table Creator*

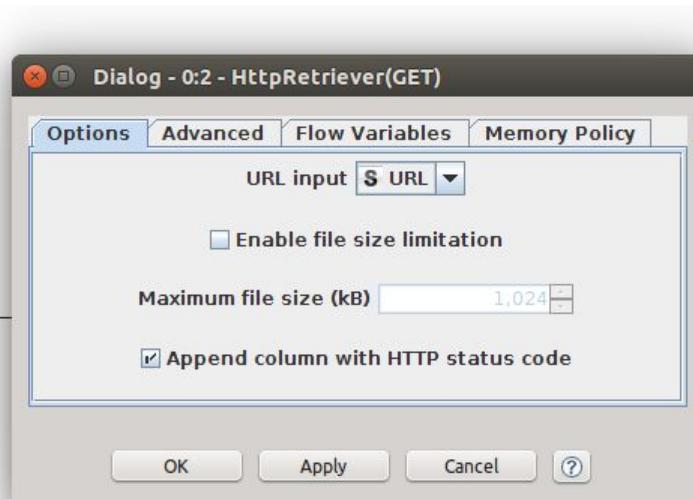
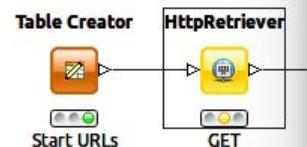


| Input line: | |
|-------------|---|
| | URL |
| Row0 | http://www.nytimes.com/pages/science/index.html |
| Row1 | |
| Row2 | |
| Row3 | |
| Row4 | |
| Row5 | |
| Row6 | |
| Row7 | |
| Row8 | |
| Row9 | |
| Row10 | |
| Row11 | |
| Row12 | |
| Row13 | |
| Row14 | |
| Row15 | |
| Row16 | |
| Row17 | |

Coleta da Web

❖ Fetch HTTP - Download de Páginas

O node ***HttpRetriever*** cuida automaticamente da requisição do conteúdo da página ao servidor.



| Table "default" - Rows: 1 | | Spec - Columns: 3 | Properties | Flow Variables |
|---------------------------|--|--|------------|----------------|
| Row ID | S URL | Result | HTTP ... | |
| Row0 | http://www.nytimes.com/pages/science/index.ht... | <!DOCTYPE html><!--[if (gt IE 9) (IE)]> <!--> <html lang="en" class... 200 | | |

Extração de Dados

❖ Parsing HTML -> XML

O *HtmlParser* extrai o conteúdo HTML e transforma em tabela



Parsed Documents - 0:3 - HtmlParser(Create XML)

| Row ID | HTTP... | XML Document |
|--------|---------|--|
| Row0 | 200 | <?xml version="1.0" encoding="UTF-8"?> <html class="no-js section-science tone-news a...><head><title>Science - The New York Times</title><meta content="IE=edge,chrome=1" http-e...</meta> |

Properties Flow Variables

Table "default" - Rows: 1 Spec - Columns: 4

Job Manager Selection Memory Policy

Options Flow Variables

Input Result ▾

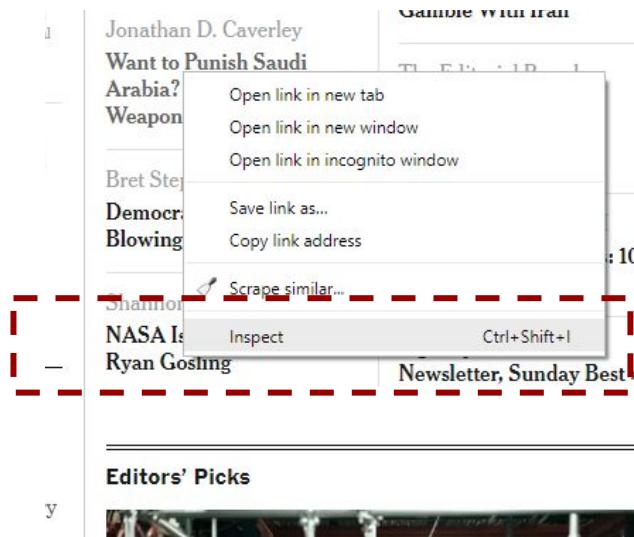
Make absolute URLs

Extraindo conteúdo de interesse com XPath

❖ Extraindo artigos

Primeiramente procuramos no html da página (Ctrl + U) onde está o conteúdo de interesse (Ctrl + F) e quais são as tags e classes associadas ao mesmo

Ou “botão direito-> Inspecionar”



Extraindo conteúdo de interesse com XPath

❖ Extraindo artigos - Onde está a notícia “Pluto’s....” ?

```
<li>
<article class="story theme-summary">
    <figure class="media photo" aria-label="media" role="group">
        <span class="visually-hidden">Photo</span>
        <a href="http://www.nytimes.com/interactive/2015/science/summer-of-science.html">
            
        </a>
        <figcaption class="caption" itemprop="caption description">
            <span class="credit" itemprop="copyrightHolder"><span class="visually-hidden">Credi
    ics Laboratory/Southwest Research Institute</span>
        </figcaption>
    </figure>
    <div class="story-body">
        <h3 class="kicker">Summer of Science </h3>
        <h2 class="story-heading">
            <a href="http://www.nytimes.com/interactive/2015/science/summer-of-science.html">Pluto
        </h2>
            <div class="thumb">
                <a href="http://www.nytimes.com/interactive/2015/science/summer-of-science.html" ar
                    
                </a>
            </div>
            <p class="summary">Just 2.5 million miles away, NASA's New Horizons
    egions near Pluto's equator.</p>
            <p class="byline">
                <span class="freshness"><time data-utc-timestamp="1434925274" datetime="1434925274"></t
                                <span class="divider"></span>
                <span class="author" itemprop="author">By KENNETH CHANG</sp
            </p>
        </div>
    </article>
</li>
```



NASA/JOHNS HOPKINS UNIVERSITY APPLIED PHYSICS
LABORATORY/SOUTHWEST RESEARCH INSTITUTE

SUMMER OF SCIENCE

Pluto's Dark Spots

Just 2.5 million miles away, NASA's New Horizons spacecraft got a sharper look at Missouri-size dark regions near Pluto's equator.

June 21, 2015 • By KENNETH CHANG

Climate Change Is Shrinking Bumblebees Range, Research Shows

Warming temperatures have driven bumblebee populations to retreat from the southern limits of their travels in North America and Europe, according to a new study.

2d ago · By NICHOLAS ST. FLEUR

Outside Psychologists S Torture Program, Rep

The scathing report, commissioned by the American Psychological Association, says the C.I.A. used prominent outside psychologists to quell internal objections.

1d ago · By JAMES RISEN

Extraindo conteúdo de interesse com XPath

- ❖ Nesta página os artigos estão em uma “div” e classe “story”
-

```
<div class="story-body">
    <h3 class="kicker">Summer of Science </h3>
    <h2 class="story-heading">
Pluto's Dark Spots
    <div class="thumb">
```

Extraindo conteúdo de interesse com XPath

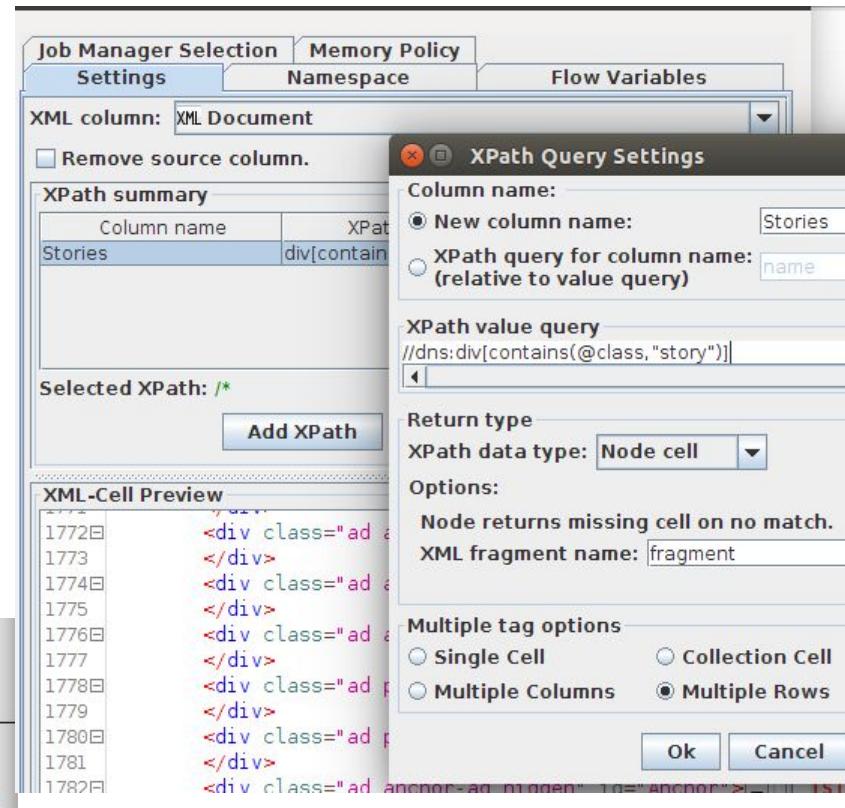
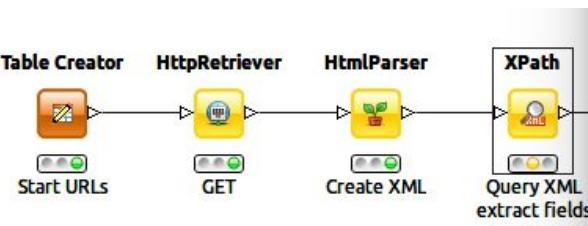
- ❖ Extraindo conteúdo de interesse com XPath.

Com o node **XPath** utilizamos a regra:

➤ `//dns:div[contains(@class,"story")]`
]

para extrair artigos na página (*Stories*)

`//dns:` é um
prefixo padrão



Extraindo conteúdo de interesse com XPath

Artigos extraídos com as regras XPath.

| Row ID | XML Stories |
|--------|--|
| Row0_1 | <?xml version="1.0" encoding="UTF-8"?> <div class="story-body" xmlns="http://www.w3.org/1999/xhtml"> <h2 class="story-heading"> </h2> <div class="thumb"> |
| Row0_2 | <?xml version="1.0" encoding="UTF-8"?> <div class="story-body" xmlns="http://www.w3.org/1999/xhtml"> <h3 class="kicker">Summer of Science </h3> <h2 class="story-heading"> Plut |
| Row0_3 | <?xml version="1.0" encoding="UTF-8"?> <div class="story-body" xmlns="http://www.w3.org/1999/xhtml"> <h2 class="story-heading"> </h2> <div class="thumb"> |
| Row0_4 | <?xml version="1.0" encoding="UTF-8"?> <div class="story-body" xmlns="http://www.w3.org/1999/xhtml"> <h2 class="story-heading"> |
| Row0_5 | <?xml version="1.0" encoding="UTF-8"?> <div class="story-tabs tabs" id="story-tabs" xmlns="http://www.w3.org/1999/xhtml"> <nav class="tab-navigation"> <ul class="tab-menu" role="tablist"> |
| Row0_6 | <?xml version="1.0" encoding="UTF-8"?> <div class="story-body" xmlns="http://www.w3.org/1999/xhtml"> <h2 class="story-heading" itemprop="headline"> |

Extraindo conteúdo de interesse com XPath

❖ Extraindo Headlines

E se quisermos os headlines das notícias?

Nesta página os headlines dos artigos estão dentro de tags “h2”

```
<h2 class="story-heading">  
  <a href="http://www.nytimes.com/interactive/2015/science/summer-of-science.html">Pluto's Dark Spots</a>  
</h2>
```

Extraindo conteúdo de interesse com XPath

- ❖ Extraindo Headlines - regra: `//dns:h2`

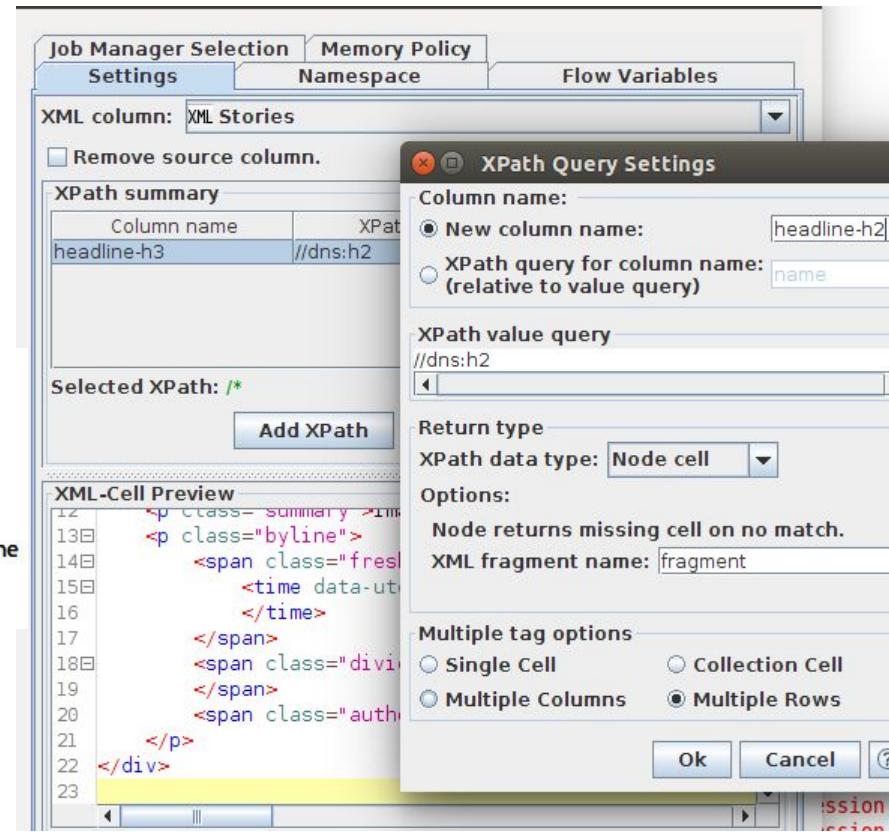
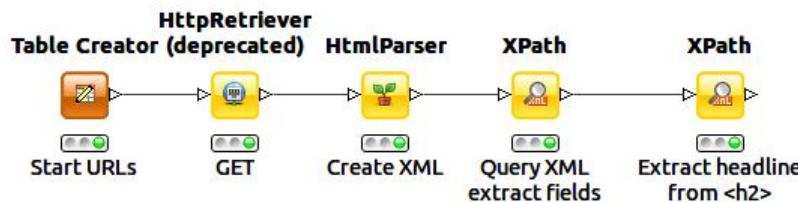
Ou seja, todos **h2** dentro da extração anterior (Story) possuem os headlines das notícias

| | |
|--------|---|
| | <pre><h2 class="story-heading"> Pluto's Dark Spots </h2></pre> |
| Row0_2 | <pre><?xml version="1.0" encoding="UTF-8"?> <div class="story-body" xmlns="http://www.w3.org/1999/xhtml"> <h3 class="kicker">Summer of Science </h3> <h2 class="story-heading"> Pluto's Dark Spots </h2></pre> |
| Row0_3 | <pre><?xml version="1.0" encoding="UTF-8"?> <div class="story-body" xmlns="http://www.w3.org/1999/xhtml"> <h2 class="story-heading"> </h2> <div class="thumb"></pre> |
| Row0_4 | <pre><?xml version="1.0" encoding="UTF-8"?> <div class="story-body" xmlns="http://www.w3.org/1999/xhtml"> <h2 class="story-heading"> </h2> <div class="thumb"></pre> |

Extraindo conteúdo de interesse com XPath

❖ Extraindo Headlines

regra: `//dns:h2`

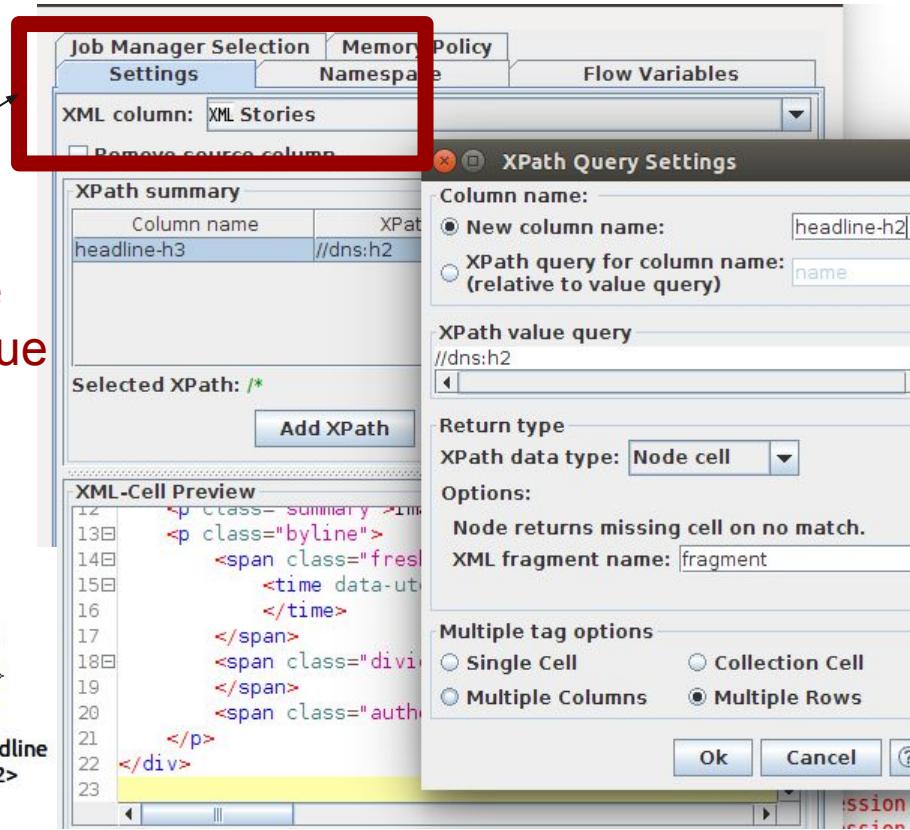
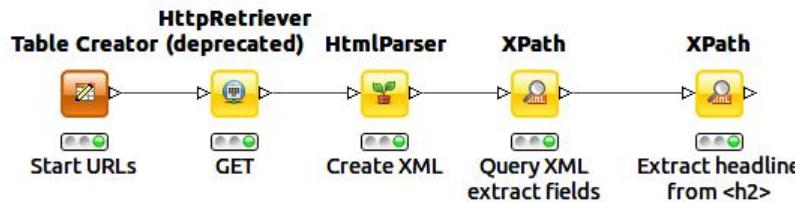


Lembre-se de especificar a coluna certa

❖ Extraindo Headlines

regra: `//dns:h2`

Não esqueça de fornecer a coluna que você criou na extração anterior, para que esta seja extraída em sequência.



Extraindo conteúdo de interesse com XPath

- ❖ Extraindo Headlines - regra: `//dns:h2`

Portanto conseguimos extrair os headlines dos artigos (Stories)

```
<h2 class="story-heading">  
  <a href="http://www.nytimes.com/interactive/2015/science/summer-of-science.html">Pluto's Dark Spots</a>  
</h2>
```

=

| | |
|----------|---|
| Row0_2_1 | <?xml version="1.0" encoding="UTF-8"?> <h2 class="story-heading" xmlns="http://www.w3.org/1999/xhtml"> Pluto's Dark Spots </h2> |
|----------|---|

Extraindo conteúdo de interesse com XPath

❖ Texto dos Headlines (Tarefa)

```
<h2 class="story-heading">  
  <a href="http://www.nytimes.com/interactive/2015/science/summer-of-science.html">Pluto's Dark Spots</a>  
  2>
```

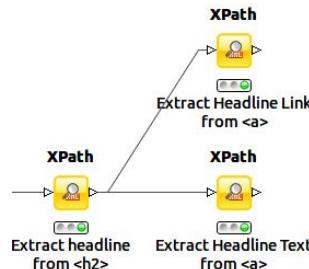
- Uma vez que temos o conteúdo dos headlines (entre tags h2) extraídos, como podemos extrair o texto do título?

Dica: Após detectar como extrair o local do título você poderá utilizar a opção  para recuperar apenas o texto da resposta, sem adição de tags html.

Extraindo conteúdo de interesse com XPath

❖ Link dos headlines

- O último passo nos trouxe um texto sem outros dados.
- Para extrairmos os links precisamos buscar dados um passo antes, ou seja criar um **fluxo paralelo** à extração do texto e **não em sequencial**.



| Table "default" - Rows: 89 | |
|----------------------------|--|
| S | Headline-Text |
| | Japan's New Satellite Captures an Image of Earth Every 10 Minut... |
| | Pluto's Dark Spots |
| | Climate Change Is Shrinking Where Bumblebees Range, Research... |

Extraindo conteúdo de interesse com XPath

❖ Link dos headlines

The screenshot shows a software interface for extracting data from XML. On the left, there's a sidebar with an 'XML column' dropdown set to 'XML headline-h2'. Below it is a checkbox 'Remove source column.' and an 'XPath summary' section. Under 'XPath summary', the 'Column name' is 'Headline-Link'. The 'Selected XPath' field contains the expression `/*`. The 'XML-Cell Preview' area shows the first few lines of an XML document:

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <h2 class="story">
3   <a href="http://www.nytimes.com/interactive/2015/07/10/science/An-Image-of-Earth-Every-Ten-Minutes.html">
4     An Image of Earth Every Ten Minutes
5   </a>
6 </h2>
```

A modal dialog titled 'XPath Query Settings' is open in the center. It has several sections:

- Column name:** Radio buttons for 'New column name:' (selected) and 'XPath query for column name:' (relative to value query). The 'New column name:' field is set to 'Headline-Link'.
- XPath value query:** The expression `//dns:a/@href` is entered.
- Return type:** A dropdown menu shows 'String cell'.
- Options:** A checkbox 'Return missing cell on empty string.' is unchecked.
- Multiple tag options:** Radio buttons for 'Single Cell', 'Collection Cell', 'Multiple Columns' (selected), and 'Multiple Rows' (selected).

On the right side of the interface, there are four horizontal lines, each containing a URL. The URLs are:

- <http://www.nytimes.com/interactive/2015/07/10/science/An-Image-of-Earth-Every-Ten-Minutes.html>
- <http://www.nytimes.com/interactive/2015/science/summer-of-science.html>
- <http://www.nytimes.com/2015/07/10/science/bumblebees-global-warming-shrinking-habitats.html>
- <http://www.nytimes.com/2015/07/11/us/psychologists-shielded-us-torture-program-report-finds.html>

At the bottom of the dialog are 'Ok', 'Cancel', and a question mark icon.

Extraindo conteúdo de interesse com XPath

❖ Autores (**Tarefa**)

Extrair o nome dos autores nos artigos

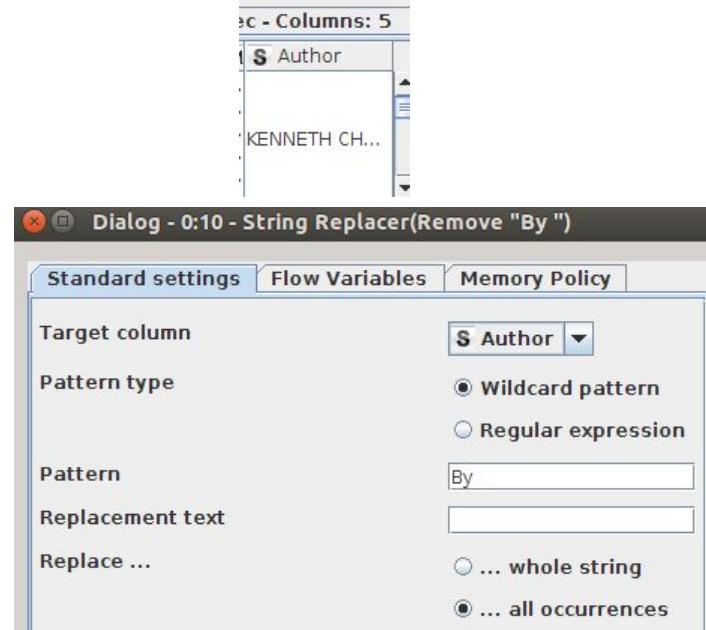
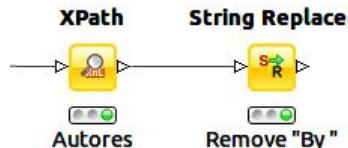
```
<p class="byline">
    <span class="freshness"><time data-utc-timestamp="1434925274" datetime="1434925274"></time></span>
        <span class="divider"></span>
        <span class="author" itemprop="author">By KENNETH CHANG</span>
    ..</p>
```

Extraindo conteúdo de interesse com XPath

- ❖ Substituindo strings - removendo “by”

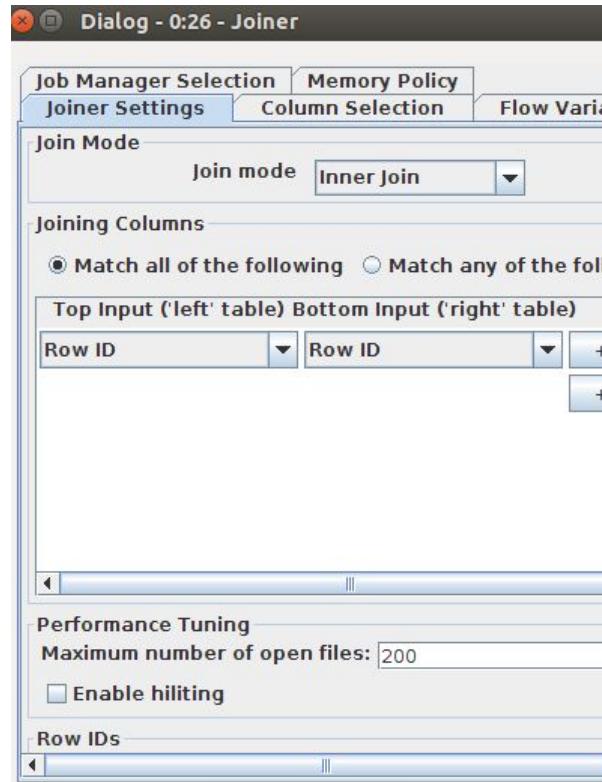
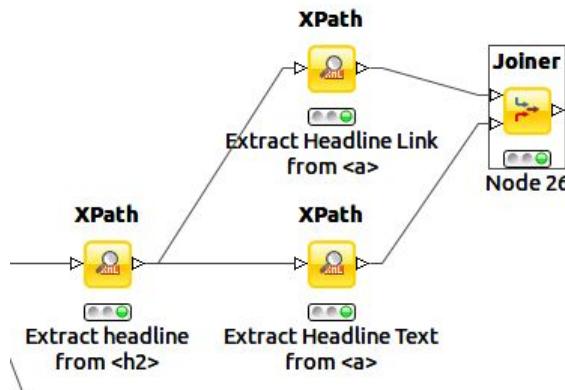
```
<p class="byline">
  <span class="freshness"><time data-utc-timestamp="1434925274" datetime="1434925274"></time></span>
    <span class="divider"></span>
    <span class="author" itemprop="author">By KENNETH CHANG</span>
</p>
```

O node *String Replacer* substitui valores de strings.
Nesse exemplo estamos substituindo as ocorrências de “By” por “” (string vazia)



Extraindo conteúdo de interesse com XPath

❖ Join - Unificando dados



Extraindo conteúdo de interesse com XPath

❖ Detectando falhas com o node Rule Engine

Cria coluna Filter

➤ **MISSING \$headline-h2\$ => “yes”**

Atribui valor “yes” caso a coluna Headline não esteja preenchida.

➤ **TRUE => “no”**

Existem dados em Headline portanto cria campo com valor “no”

The screenshot shows the KNIME Rule Engine node configuration window. The top bar includes tabs for 'Job Manager Selection', 'Memory Policy', 'Rule Editor' (selected), and 'Flow Variables'. The 'Rule Editor' tab has sections for 'Column List', 'Flow Variable List', 'Category', 'Function', 'Description', and 'Expression'. The 'Column List' contains columns like ROWID, ROWINDEX, ROWCOUNT, URL, Result, and XML Document. The 'Flow Variable List' contains 'knime.workspace'. The 'Category' dropdown is set to 'All'. The 'Function' dropdown lists various XPath functions: ? < ?, ? <= ?, ? = ?, ? > ?, ? >= ?, ? AND ?, ? IN ?, ? LIKE ?, ? MATCHES ?, ? OR ?, and ? XOR ?. The 'Description' column is empty. The 'Expression' section contains the following code:

```
? 1 // enter ordered set of rules,  
? 2 // $double column name$ > 5.0 =  
? 3 // $string column name$ LIKE "*"  
? 4 // TRUE => "default outcome"  
S 5 MISSING $headline-h2$ => "yes"  
S 6 TRUE => "no"
```

At the bottom, there is an 'Append Column' button with a 'Filter' input field.

Extraindo conteúdo de interesse com XPath

❖ Filtrando Falhas

Seleciona apenas linhas completas (com valor “no” na coluna Filter)

The screenshot shows a software interface for filtering data. At the top, there are tabs for "Filter Criteria", "Flow Variables", and "Memory Policy". The "Filter Criteria" tab is active, showing the following configuration:

- Set filter parameter:** Column value matching
- select the column to test:** S Filter
- filter based on collection elements
- matching criteria:**
 - use pattern matching
 - pattern: no
 - contains wild cards
 - case sensitive match
 - regular expression
 - use range checking
 - lower bound: []
 - upper bound: []

To the right of the main window, there is a vertical toolbar with a "Row Filter" icon. Below the toolbar, the text "Filter missing headlines and summaries" is displayed.

The title bar of the application window says "Filtered - 0:15 - Row Filter(Filter missing)".

The bottom part of the interface shows a table with the following data:

| Row ID | S Headline | S Href | S Author | S Filter |
|--------|--|------------------|-----------------|----------|
| Row0_1 | Japan's New Satellite Captures an Image of Earth Every 10 Minut... | http://www.ny... | DEREK WATK... | ...no |
| Row0_2 | Pluto's Dark Spots | http://www.ny... | KENNETH CH... | ...no |
| Row0_3 | Climate Change Is Shrinking Where Bumblebees Range, Research Fi... | http://www.ny... | NICHOLAS ST.... | ...no |

Exercício

- ❖ **Parte I** - Extraia os dados de resumo (summary) das notícias
 - Dica: Os dados com resumo já foram extraídos das notícias (no parsing da classe *story*)
 - ❖ Adicione uma coluna com o resumo na tabela com todos os dados
 - ❖ Adicione uma regra para eliminar dados incompletos de resumo
- ❖ **Parte II** - Extraia os dados de um site diferente (notícias, blogs...)
- ❖ Utilidade no projeto final
 - Coletar e extrair conteúdo de interesse
 - Escolha um site/blog com base na relevância com seu assunto



Anexo: Extensão Scraper (Chrome)

Extensão Scraper (Chrome)

❖ Scraper

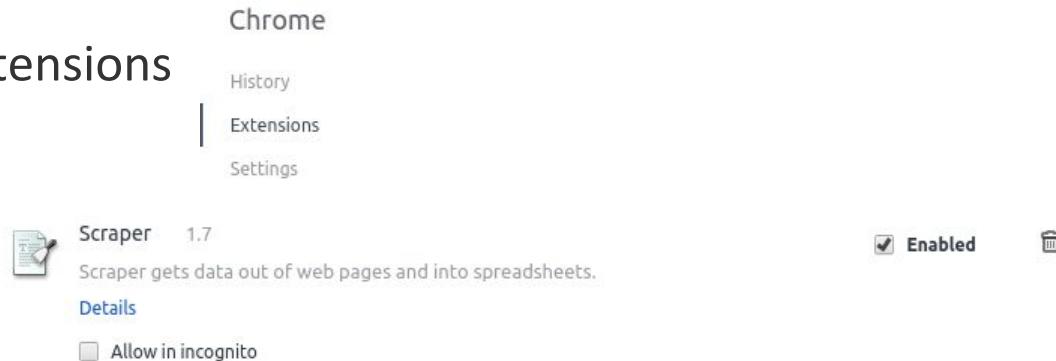
- Ferramenta que simplifica a extração de regras (Xpaths) para Web Scraping

❖ Instalação no google chrome ([Link](#))

- Menu -> Settings -> Extensions

Chrome
History
Extensions
Settings

- procure por “*Scraper*”
e instale



Extensão Scraper (Chrome)

❖ Exemplo no site do NYTimes

The screenshot shows the NYTimes Science homepage with the URL www.nytimes.com/section/science?action=click&pgtype=Homepage®ion=TopBar&module=HPMiniNav&contentCollection=Science&WT.nav=page. The page features a grid of five news articles:

- Breast Cancer Treatment and D.C.I.S.: Answers to Questions About New Findings
- Tracking a Rare Nautilus in Papua New Guinea
- Red Pandas Are Adorable and in Trouble
- The Butterfly, the Ant and the Oregano
- Coaxing Children With Selective Mutism to Find Their Voices

Below the grid, there are two main sections of news articles:

- Latest**:
 - Giant Panda Gives Birth to Two Cubs at the National Zoo
 - California Dam Lets Water Shared by Farms Flow to Salmon
 - Vaccinations Bring Hope, Bracelets Deliver Reminders
- Search**:
 - August 22, 2015: Giant Panda Gives Birth to Two Cubs at the National Zoo (by LIAM STACK) - Includes a photo of a panda cub.
 - August 22, 2015: California Dam Lets Water Shared by Farms Flow to Salmon (by THE ASSOCIATED PRESS) - Includes a photo of a dam.
 - August 22, 2015: Vaccinations Bring Hope, Bracelets Deliver Reminders (by DONALD G. MCNEIL, JR.) - Includes a photo of a woman looking down.

To the right, there is a sidebar for the **Crossword** section, which includes the text "The best puzzle in the world just got better." and a "TRY IT FREE" button. The sidebar also has links to follow the Times on social media: [NYTimesScience](#) on Facebook and [@NYTScience](#) on Twitter.

Extensão Scraper (Chrome)

❖ Exemplo no site do NYTimes

Encontre um elemento na página (ex: Head da notícia e clique com o botão direito).

Depois clique em:



Screenshot of the NYTimes website showing several news articles. The first article, "Giant Panda Gives Birth to Two Cubs at the National Zoo", is highlighted with a red box and has its title and author information (By LIAM STACK) displayed below it. To the right of the article is a small thumbnail image of a giant panda.

The other articles listed are:

- California Dam Lets Water Shared by Farms Flow to Salmon**
By THE ASSOCIATED PRESS
Releases from the Lewiston Dam are intended to protect fish in the Klamath River that sustain Indian tribes, but they could revive a fight over competing claims for irrigation.
- Vaccinations Bring Hope, Bracelets Deliver Reminders**
By DONALD G. MCNEIL, Jr.
A simple, cheap silicon bracelet could help mothers in developing countries get their babies vaccinated on schedule.
- Decades of Data Fail to Resolve Debate on Treating Tiny Breast Lesions**
By GINA KOLATA
The study's authors said the data indicates that treatment has not

Extensão Scraper (Chrome)

❖ Exemplo no site do NYTimes

Em selector *Xpath* você encontra a regra para itens similares ao selecionado anteriormente.

Assim basta copiar a regra e colar no node *Xpaths* do knime!

Além disso é possível visualizar os dados que essa regra está gerando.

The screenshot shows the 'Scraper - Science - The New York Times' window. On the left, there's a configuration panel with a 'Selector' dropdown set to 'XPath' containing the path '/div[2]/section[1]/div[1]/ol[1]/li'. Below it are 'Columns' (Link, @href) and 'Filters' (Exclude empty results). On the right, a table lists seven news items with columns for Link, URL, and a small preview icon. The entire table is highlighted with a red border. Above the table, the article 'Giant Panda Gives Birth to Two Cubs at the National Zoo' is displayed with its author, date (August 22, 2015), and a snippet of the text.

| Link | URL |
|--|---|
| 1 Giant Panda Gives Birth to Two Cubs at the National Zoo | http://www.nytimes.com/2015/08/23/us/giant-panda-gives-birth-at-national-zoo-in-washington.html?rref=collection%2Fsectioncollection%2Fscience |
| 2 California Dam Lets Water Shared by Farms Flow to Salmon | http://www.nytimes.com/2015/08/23/us/california-dam-lets-water-shared-by-farms-flow-to-salmon.html?rref=collection%2Fsectioncollection%2Fscience |
| 3 Vaccinations Bring Hope, Bracelets Deliver Reminders | http://www.nytimes.com/2015/08/25/health/vaccinations-bring-hope-bracelets-deliver-reminders.html?rref=collection%2Fsectioncollection%2Fscience |
| 4 Decades of Data Fail to Resolve Debate on Treating Tiny Breast Lesions | http://www.nytimes.com/2015/08/22/science/study-fuels-debate-over-treating-breast-lesion-called-stage-0-cancer.html?rref=collection%2Fsectioncollection%2Fscience |
| 5 Jacob Bekenstein, Physicist Who Revolutionized Theory of Black Holes, Dies at 68 | http://www.nytimes.com/2015/08/22/science/space/jacob-bekenstein-physicist-who-revolutionized-theory-of-black-holes-dies-at-68.html?rref=collection%2Fsectioncollection%2Fscience |
| 6 Breast Cancer Treatment and D.C.I.S.: Readers React | http://www.nytimes.com/2015/08/21/health/breast-cancer-treatment-and-dcis-readers-react.html?rref=collection%2Fsectioncollection%2Fscience |
| 7 A Racial Gap in Attitudes Toward Hospice Care | http://www.nytimes.com/2015/08/25/health/a-racial-gap-in-attitudes-toward-hospice-care.html?rref=collection%2Fsectioncollection%2Fscience |

Extensão Scraper (Chrome)

❖ Modificando uma regra na ferramenta

A regra identificada para o autor retornou apenas um resultado. Isso porque foi gerado “[1]” para as tags, ou seja, apenas o primeiro registro.
(Leia mais sobre [notação XPath](#))



Extensão Scraper (Chrome)

❖ Modificando uma regra na ferramenta

The screenshot shows the Chrome DevTools Elements tab. In the Selector section, the XPath is set to `//div[2]/section[1]/div[1]/ol[1]/li[1]/article/div/p`. A red box highlights the last `[1]` in the XPath. To the right, the results pane displays two items:

- 1 By LIAM STACK
- 2 Mei Xiang, a giant panda at the National Zoo in Washington, gave birth to healthy twin cubs on Saturday, three days after the zoo's staff discovered she was pregnant.

Retirando o último “[1]” antes de article

The screenshot shows the Chrome DevTools Elements tab after modifying the XPath. The Selector section now contains the XPath `//div[2]/section[1]/div[1]/ol[1]/li/article/div/p`, with the last `[1]` removed. A red box highlights the removed `[1]`. To the right, the results pane displays six items:

- 1 By LIAM STACK
- 2 Mei Xiang, a giant panda at the National Zoo in Washington, gave birth to healthy twin cubs on Saturday, three days after the zoo's staff discovered she was pregnant.
- 3 By THE ASSOCIATED PRESS
- 4 Releases from the Lewiston Dam are intended to protect fish in the Klamath River that sustain Indian tribes, but they could revive a fight over competing claims for irrigation.
- 5 By DONALD G. McNEIL Jr.
- 6 A simple, cheap silicon bracelet could help mothers in developing countries get their babies vaccinated on schedule.

Extensão Scraper (Chrome)

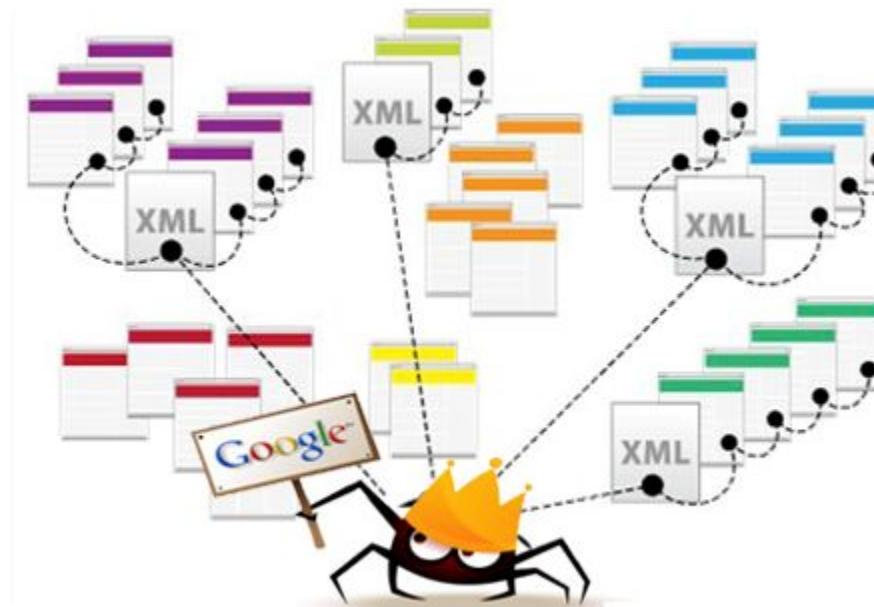
❖ Modificando uma regra na ferramenta

Especificando a regra em p para restringir a classe “byline” que possui apenas o nome do autor.

```
<p class="byline">
    <span class="freshness"><time data-utc-timestamp="1434925274" datetime="1434925274"></time></span>
        <span class="divider"></span>
        <span class="author" itemprop="author">By KENNETH CHANG</span>
</p>
```

| Text |
|---------------------------|
| 1 By LIAM STACK |
| 2 By THE ASSOCIATED PRESS |
| 3 By DONALD G. McNEIL Jr. |
| 4 By GINA KOLATA |
| 5 By DENNIS OVERBYE |
| 6 By LELA MOORE |
| 7 By SARAH VARNEY |
| 8 By PAM BELLUCK |

Coletores para Máquinas de Busca



Coletores para Máquinas de Busca

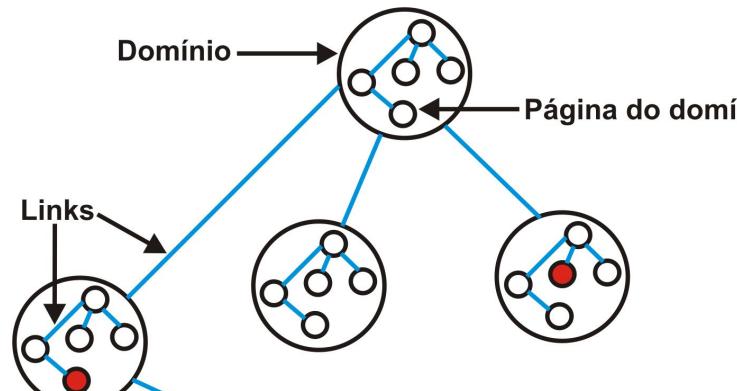
❖ WebCrawling

- Coletores
- Desafios de Projeto

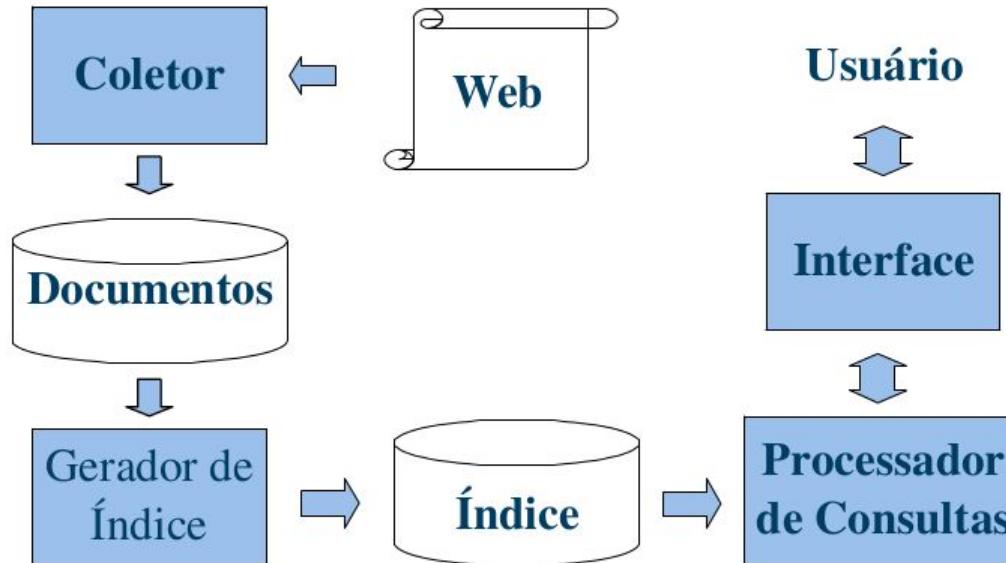
Coletores para Máquinas de Busca

❖ WebCrawling

- Processo de se navegar entre páginas www usando estrutura de hyperlinks.
- A estrutura de hyperlinks da web pode ser modelada como um grafo



Arquitetura da Máquina de Busca



Desafios de Projeto

❖ Desafios ao se projetar um coletor

- Coleta contínua: definir a periodicidade de atualização das páginas (*Freshness*) versus encontrar novas url's
- Usar o máximo de largura de banda sem sobrecarregar os sites visitados (*politeness*)
- Identificar páginas redundantes (*Mirror*)
- Coletar boas páginas.

Problemas práticos

◆ Sobrecarga do DNS

- Coletores geram número muito alto de requisições de DNS
- Normalmente os servidores de DNS viram gargalos para o coletor
- Solução é manter um cache com DNSs previamente resolvidos

Problemas práticos

❖ **Erros de acesso repetido (falsos ataques)**

- O coletor pode criar um falso ataque a um servidor web por problemas como
 - Uso de diferentes nomes para um servidor
 - Vários servidores em um mesmo local

Problemas práticos

◆ Links infinitos

- Problemas na extração de links podem gerar erros que levam a links infinitos que são validados pelo servidor Web

www.aa.bb.com/musica

www.aa.bb.com/musica/musica

Problemas práticos

◆ Páginas dinâmicas

- Alguns sites podem gerar número infinitos de páginas válidas
- Exemplo, um site que fornece um HTML com o dia da semana de qualquer data, onde a data entra na URL

Problemas práticos

◆ Normalização de URLs

- URLs devem ser normalizadas para evitar repetições

<http://www.exemplo.br/home>

<http://www.exemplo.br/home/>

<http://www.exemplo.br/home/index.html>

Problemas práticos

❖ Diferenças de velocidade

- Servidores mais lentos podem prejudicar processo de coleta por travar robôs coletores

Coletores Open Source

◆ SCRAPY (<http://scrapy.org/>)

- Scrapy é um framework Python para Web Scraping
- Formatos de exportação, como JSON, linhas JSON, XML e CSV. Scrapy foi construído para extrair informações específicas de sites, não necessariamente ficar para um dump completo do HTML
- Ele não tem a funcionalidade para executar em um ambiente distribuído uma vez que seu caso de uso primário são coletas focadas.

Coletores Open Source

◆ SCRAPY (<http://scrapy.org/>)

Prós:

- Fácil de configurar e usar, se você é familiar ao Python
- Boa documentação

Contras:

- Não há suporte para execução em um ambiente distribuído
- Não há suporte para coletas contínuas
- Exportação de grandes quantidades de dados é difícil

Coletores Open Source

◆ HERITRIX

(<https://webarchive.jira.com/wiki/display/Heritrix/Heritrix>)

- Heritrix é desenvolvido, mantido e utilizado pelo The Internet Archive. Vem sendo mantido desde o seu lançamento em 2004 e é utilizado em produção por vários outros sites
- O formato de saída são arquivos WARC, um formato eficiente para escrever vários recursos (como HTML) e seus metadados em arquivos.
- Possui interface web que pode ser usada para monitoramento e configuração da coleta.

Coletores Open Source

◆ HERITRIX (<https://webarchive.jira.com/wiki/display/Heritrix/Heritrix>)

Prós:

- Boa documentação e fácil de instalar
- Plataforma madura e estável. Ele tem sido usado na produção **archive.org** por mais de uma década
- Bom desempenho e suporta coleta distribuída.

Contras:

- Não é escalável dinamicamente. Você deve decidir previamente o número de servidores e esquema de particionamento inicial
- Exporta arquivos ARC / WARC. Adicionar suporte formatos personalizados exige alterar o código

Coletores Open Source

◆ APACHE NUTCH (<http://nutch.apache.org/>)

- O Nutch faz uso do ecossistema Hadoop MapReduce para processamento. Necessário ter um cluster Hadoop instalado e configurado
- Vem com integração para sistemas de indexação, como o ElasticSearch (via plugins)
- Possui um sistema de plugins flexível, permitindo extender funcionalidades personalizadas. (Exige codificação)

Coletores Open Source

◆ APACHE NUTCH (<http://nutch.apache.org/>)

Prós:

- Dinamicamente escalável (e tolerante a falhas) através Hadoop
- Sistema de plugins flexível
- Versão estável

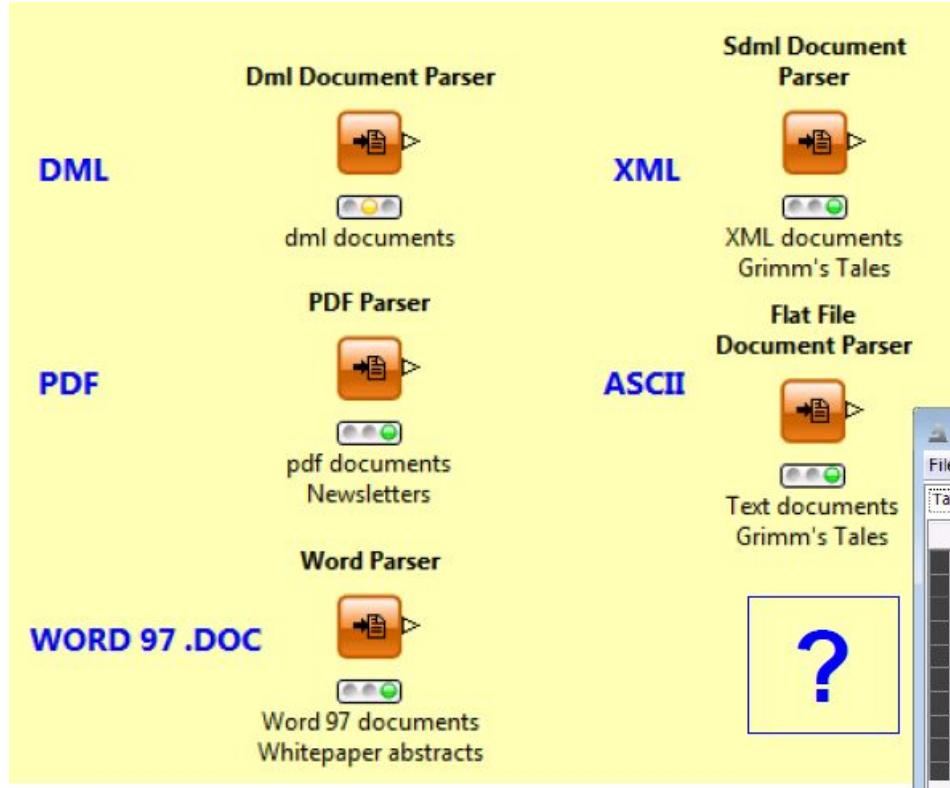
Contras:

- Documentação falta de bons exemplos
- Herda desvantagens de Hadoop (leituras em disco, configuração difícil)
- Não há suporte embutido para coleta contínua
- Exportação limitada a Solr** / ElasticSearch



KNIME - Outras Fontes de Documentos

Leitura em Diretórios



The output is
a list of
Documents

| Documents output table - 222 - Flat File Document Parser(Text do...) | |
|--|--|
| File | |
| Row ID | Document |
| Row 1 | "A shoemaker, by no fault of his own, had become so poor that at last he |
| Row 2 | "Allerlerauh." |
| Row 3 | "A Riddling Tale." |
| Row 4 | "Bearskin." |
| Row 5 | "Brides On Their Trial." |
| Row 6 | "Brother and Sister." |
| Row 7 | "Brother Lustig." |
| Row 8 | "Cat and Mouse in Partnership." |
| Row 9 | "Cinderella." |
| Row 10 | "Clever Esie." |

Leitura em Tabelas

- ❖ Lendo dados com o node *Table Reader*

The screenshot shows the KNIME workflow interface. On the left, there is a node icon for "Table Reader" with a small preview window below it. To the right, a larger window titled "Read table - 0:2 - Table Reader(read data)" displays a preview of the data. The window has tabs for "File", "Spec - Columns: 7", "Properties", and "Flow Variables". The "Spec - Columns: 7" tab is selected. The data preview shows 8 rows of a table with columns: Row ID, user name, review title, review text, stars, Reviews, Restaurant, and Category. The data includes reviews from users like travellerBruss..., coverdriven, Lula12783, Deise_Boy08, Tanguyatea, OlgaNottingham, Jack D, and Rick M, with reviews ranging from 2 to 29 stars and 3 to 29 reviews.

| Row ID | user name | review title | review text | stars | Reviews | Restaurant | Category |
|---------------|--------------------|--------------------------------|------------------------|-------|---------|-----------------|----------|
| Row0_1_Row... | 1travellerBruss... | Great food, interesting ser... | this restaurant is ... | 5 | 29 | Saigon and m... | Asian |
| Row0_2_Row... | coverdriven | Excellent Lunch Destination | Very much enjoy... | 4 | 2 | Saigon and m... | Asian |
| Row0_3_Row... | Lula12783 | Hidden treasure near KaDa... | We found this littl... | 5 | 5 | Saigon and m... | Asian |
| Row0_4_Row... | Deise_Boy08 | Excellent Food Very Reaso... | Food was top cla... | 5 | 12 | Saigon and m... | Asian |
| Row0_5_Row... | Tanguyatea | Good food, great prices! | I went there bec... | 4 | 6 | Saigon and m... | Asian |
| Row0_6_Row... | OlgaNottingham | Nice food at a reasonable ... | I am no expert o... | 4 | 12 | Saigon and m... | Asian |
| Row0_7_Row... | Jack D | Good food and entertainin... | From reading oth... | 4 | 3 | Saigon and m... | Asian |
| Row0_8_Row... | Rick M | Very good | A very tasty Viet... | 4 | 6 | Saigon and m... | Asian |

A disciplina - RI

❖ Plano de Ensino

- **Unidade 01:** Conceitos de inteligência competitiva e coletiva, crowdsourcing e redes sociais. Recuperação da informação e Máquinas de busca. Desafios da Mineração na web e nas redes. Exemplos de Projetos da disciplina.
- **Unidade 02:** Algoritmos e soluções para problemas de busca e extração de informação da WWW. Ferramenta e prática de processamento textual e recuperação de informação.
- **Unidade 03:** Tipos de coleta, arquitetura e componentes de coletores Web. Ferramenta e prática de coleta de dados na Web.

Esse assunto era o que você imaginava?



