Extração de Categorias Shopee (n8n - Nós Nativos)

26/10/2025, 19:33:16

Introdução

Este documento detalha um workflow n8n projetado para extrair informações de categorias, subcategorias, 3º nível e IDs da plataforma Shopee, utilizando apenas nós nativos (`HTTP Request`, `Code`, `Set`, `Merge`). A solução contorna a necessidade de nós como `Cheerio` ou `HTML Extract` ao empregar expressões regulares (Regex) para parsear o conteúdo HTML das páginas.

1. Arquitetura do Workflow

O workflow segue uma sequência lógica para navegar pelas páginas de categorias da Shopee, extrair o HTML e, em seguida, parsear os dados desejados usando Regex.

```
Start !“ HTTP Request (Obter Total de Páginas) !“ Set (Extrair Total de Páginas) !“ HTTP Request (Obter HTML da Catego
cada página) !“ Set (Extrair Blocos de Categor
blocos individuais de categoria) !“ Set (Anali
os detalhes Categoria, Subcategoria, Nível3,
Categorias) -- (Coleta todos os dados extraídos em um único resultado)
```

2. Configuração Detalhada de Cada Nó

A seguir, a configuração de cada nó no workflow:

1. Start

Descrição: O ponto de partida do workflow.

2. HTTP Request (Obter Total de Páginas)

Descrição
: Faz a primeira requisição GET para a página de categorias da Shopee para obter o HTML e determinar o número total de páginas para iteração.

Configuração:

Authentication: `None`

Method: `GET`

URL: `https://seller.shopee.com.br/edu/category-guide?keyword=Decoracao`

Response Format: `String`

Headers:

`User-Agent`: `Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.4896.127 Safari/537.36` (Recomendado para simular um navegador real)

Options:

`Follow Redirects`: `true`

Saída Exemplo: O HTML completo da primeira página.

3. Code (Preparar Iteração de Páginas)

Descrição
: Extrai o número total de páginas do HTML da primeira requisição e gera uma lista de URLs de páginas a serem visitadas.

Configuração:

Code:

```
const html = $input.first().json.data; const totalPagesMatch =
html.match(/active">(\d+) \/ (\d+)</); const totalPages = totalPagesMatch ?
parseInt(totalPagesMatch[2], 10) : 1; const baseURL =
"https://seller.shopee.com.br/edu/category-guide?keyword=Decoracao&page="; const
mainCategory = "Decoração"; const outputItems = []; for (let i = 1; i <=
```

```
totalPages; i++) { outputItems.push({ page: i, baseURL: baseURL, mainCategory: mainCategory }); } return outputItems;
```

Saída Exemplo
: Uma lista de objetos, cada um representando uma página a ser processada:

```
[ { "page": 1, "baseURL": "...", "mainCategory": "Decoração" }, { "page": 2, "baseURL": "...", "mainCategory": "Decoração" }, // ... até a última página ]
```

4. HTTP Request (Obter HTML da Categoria)

Descrição
: Para cada item gerado pelo nó `Code` (representando uma página), este nó faz uma requisição para a URL da página específica e obtém seu HTML.

Configuração:

Authentication: `None`

Method: `GET`

URL: `{{ $json.baseURL }}{{ $json.page }}`

Response Format: `String`

Headers:

`User-Agent`: `Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.4896.127 Safari/537.36`

Options:

`Follow Redirects`: `true`

`Retry On Fail`: `true` (Com `Max Retries` de `3` e `Retry Interval` de `1000` ms)

Saída Exemplo
: O HTML da página atual, mantendo `page`, `baseURL` e `mainCategory` no item.

5. Set (Extrair Blocos de Categoria)

Descrição
: Extrai todos os blocos HTML que representam uma linha de categoria individual na página. Cada bloco será transformado em um item separado para processamento posterior.

Configuração:

Mode: `Split Out Items`

Add Field:

Name: `categoryBlock`

Value: `{{ $json.htmlContent.match(/(?s)<div class="shopee-tree-table-row">.?<\/div>/g) || [] }}`

Add Field:

Name: `mainCategory`

Value: `{{ $json.mainCategory }}`

Saída Exemplo
: Múltiplos itens, cada um contendo um bloco HTML de categoria e a categoria principal:

```
[ { "categoryBlock": "<div class=\"shopee-tree-table-row\">...</div>", "mainCategory": "Decoração" }, { "categoryBlock": "<div class=\"shopee-tree-table-row\">...</div>", "mainCategory": "Decoração" }, // ... para cada categoria encontrada na página ]
```

6. Set (Analisar Blocos de Categoria)

Descrição
: Para cada bloco de categoria, este nó utiliza expressões regulares para extrair a subcategoria, o 3º nível de categoria e o ID da categoria.

Configuração:

Mode: `Merge`

Add Field:

Name: `Categoria`

Value: `{{ $json.mainCategory }}`

Add Field:

Name: `Subcategoria`

Value: `{{ $json.categoryBlock.match(/(?s)(.
?)<\/span>/)?.[1] || null }}`

Add Field:

Name: `Nivel3`

Value: `{{ $json.categoryBlock.match(/(?s)(.
?)<\/span>/)?.[1] || null }}`

Add Field:

Name: `ID`

Value: `{{ $json.categoryBlock.match(/ID: (\d+)/)?.[1] || null }}`

Saída Exemplo: Um objeto com os dados estruturados para cada categoria:

```
{ "Categoria": "Decoração", "Subcategoria": "Móveis de Decoração", "Nivel3":
"Estantes e Prateleiras", "ID": "12345" }
```

7. Merge (Combinar Categorias)

Descrição
: Combina todos os itens de categoria processados de todas as páginas em um único array no final do workflow.

Configuração:

Mode: `Append`

Input 1: Conectado à saída do nó `Set (Analisar Blocos de Categoria)`.

Saída Exemplo: Um único item contendo um array com todas as categorias extraídas.

```
[ { "Categoria": "Decoração", "Subcategoria": "Móveis de Decoração", "Nivel3":
"Estantes e Prateleiras", "ID": "12345" }, { "Categoria": "Decoração",
```

```
"Subcategoria": "Móveis de Decoração", "Nivel3": "Mesas de Centro", "ID":
"67890" }, // ... todas as 164 categorias ]
```

3. Expressões Regex Utilizadas

As expressões regulares são cruciais para a extração de dados sem Cheerio:

Para `totalPages` (Nó: Code):

`active">(\d+) \/ (\d+)<`

Explicação
: Captura os números do texto como "1 / 11". O `[2]` no código JavaScript seleciona o segundo grupo de captura, que é o número total de páginas.

Para `categoryBlock` (Nó: Set - Extrair Blocos de Categoria):

`(?s)<div class="shopee-tree-table-row">.?<\/div>` com a flag `g` (global) e `s` (dotall).

Explicação
: Encontra todas as ocorrências de divs com a classe `shopee-tree-table-row` e todo o seu conteúdo interno até o fechamento da div. O `(?s)` torna o `.` (ponto) capaz de corresponder a quebras de linha.

Para `Subcategoria` (Nó: Set - Analisar Blocos de Categoria):

`(?s)<span class="first-level">(.?)<\/span>`

Explicação
: Captura o texto dentro da primeira tag `span` com a classe `first-level` dentro de cada `categoryBlock`.

Para `Nivel3` (Nó: Set - Analisar Blocos de Categoria):

`(?s)<span class="second-level">(.?)<\/span>`

Explicação
: Captura o texto dentro da primeira tag `span` com a classe `second-level` dentro de cada `categoryBlock`. Retornará `null` se não for encontrado.

Para `ID` (Nó: Set - Analisar Blocos de Categoria):

`ID: (\d+)`

Explicação: Captura a sequência de dígitos após "ID: " dentro de cada `categoryBlock`.

4. Tratamento de Erros

HTTP Request (Obter HTML da Categoria):

`Retry On Fail`: Ativado com 3 tentativas e 1 segundo de intervalo para lidar com falhas de rede temporárias ou picos de carga do servidor.

Regex com Nullish Coalescing (`|| null`):

Nas expressões Regex dentro do nó `Set (Analisar Blocos de Categoria)`, o uso de `?.[1] || null` garante que, se um padrão não for encontrado (por exemplo, um 3º nível de categoria não existir), o campo correspondente será definido como `null` em vez de causar um erro no workflow.

5. Dicas de Otimização

Delay entre Requisições
: Para evitar ser bloqueado pelo servidor da Shopee ou sobrecarregá-lo, considere adicionar um nó `Wait` (com um atraso de 1 a 3 segundos) logo após o nó `HTTP Request (Obter HTML da Categoria)`.

User-Agent
: Sempre utilize um `User-Agent` realista nos seus `HTTP Request` para simular um navegador e evitar ser detectado como um bot.

Monitoramento
: Monitore a execução do workflow. Se notar falhas frequentes ou dados incompletos, pode ser necessário ajustar o atraso, o User-Agent ou as expressões Regex.

Limite de Dados
: Para conjuntos de dados muito grandes, considere salvar os resultados intermediários em um banco de dados ou armazenamento de arquivos, especialmente se o n8n estiver sendo executado em um ambiente com recursos limitados de memória.

6. JSON Completo do Workflow

{ "nodes": [{ "parameters": {}, "name": "Start", "type": "n8n-nodes-base.start", "typeVersion": 1, "position": [240, 300] }, { "parameters": { "url": "https://seller.shopee.com.br/edu/category-guide?keyword=Decoracao", "options": { "responseFormat": "string" }, "headerParameters": [{ "name": "User-Agent", "value": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.4896.127 Safari/537.36" }] }, "name": "HTTP Request (Obter Total de Páginas)", "type": "n8n-nodes-base.httpRequest", "typeVersion": 3, "position": [440, 300] }, { "parameters": { "functionCode": "const html = $input.first().json.data;\nconst totalPagesMatch = html.match(/active\">(\\d+)\\/ (\\d+)</);\nconst totalPages = totalPagesMatch ? parseInt(totalPagesMatch[2], 10) : 1;\nconst baseURL = \"https://seller.shopee.com.br/edu/category-guide?keyword=Decoracao&page=\";\nconst mainCategory = \"Decoração\";\n\nconst outputItems = [];\nfor (let i = 1; i <= totalPages; i++) {\n outputItems.push({\n page: i,\n baseURL: baseURL,\n mainCategory: mainCategory\n });\n}\nreturn outputItems;" }, "name": "Code (Preparar Iteração de Páginas)", "type": "n8n-nodes-base.code", "typeVersion": 1, "position": [680, 300] }, { "parameters": { "url": "={{ $json.baseURL }}{{ $json.page }}", "options": { "responseFormat": "string", "retryOnFail": true, "retryInterval": 1000 }, "headerParameters": [{ "name": "User-Agent", "value": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/100.0.4896.127 Safari/537.36" }] }, "name": "HTTP Request (Obter HTML da Categoria)", "type": "n8n-nodes-base.httpRequest", "typeVersion": 3, "position": [900, 300] }, { "parameters": { "mode": "splitOut", "options": {}, "values": { "string": [{ "name": "categoryBlock", "value": "{{ $json.htmlContent.match(/(?s)<div class=\"shopee-tree-table-row\">.*?<\\/div>/g) || [] }}" }, { "name": "mainCategory", "value": "={{ $json.mainCategory }}" }] } }, "name": "Set (Extrair Blocos de Categoria)", "type": "n8n-nodes-base.set", "typeVersion": 1, "position": [1120, 300] }, { "parameters": { "values": { "string": [{ "name": "Categoria", "value": "={{ $json.mainCategory }}" }, { "name": "Subcategoria", "value": "={{ $json.categoryBlock.match(/(?s)<span class=\"first-level\">(.*?)<\\/span>/)?.[1] || null }}" }, { "name": "Nivel3", "value": "={{ $json.categoryBlock.match(/(?s)<span class=\"second-level\">(.*?)<\\/span>/)?.[1] || null }}" }, { "name": "ID", "value": "={{ $json.categoryBlock.match(/ID: (\\d+)/)?.[1] || null }}" }] }, "options": {} }, "name": "Set (Analisar Blocos de Categoria)", "type":

"n8n-nodes-base.set", "typeVersion": 1, "position": [1340, 300] }, { "parameters": { "mode": "append" }, "name": "Merge (Combinar Categorias)", "type": "n8n-nodes-base.merge", "typeVersion": 1, "position": [1560, 300] }], "connections": { "Start": [["HTTP Request (Obter Total de Páginas)"]], "HTTP Request (Obter Total de Páginas)": [["Code (Preparar Iteração de Páginas)"]], "Code (Preparar Iteração de Páginas)": [["HTTP Request (Obter HTML da Categoria)"]], "HTTP Request (Obter HTML da Categoria)": [["Set (Extrair Blocos de Categoria)"]], "Set (Extrair Blocos de Categoria)": [["Set (Analisar Blocos de Categoria)"]], "Set (Analisar Blocos de Categoria)": [["Merge (Combinar Categorias)"]] }, "pinData": {} }

7. Instruções Passo a Passo

Copie o JSON: Copie todo o conteúdo JSON fornecido na seção anterior.

Abra o n8n: Abra sua instância do n8n (seja local ou na nuvem).

Crie um Novo Workflow
: No painel de controle do n8n, clique em "New" para criar um novo workflow vazio.

Importar do JSON
: Clique no ícone de "Options" (geralmente três pontos ou um menu hambúrguer) no canto superior direito do seu workflow. Selecione "Import from JSON" (ou "Import").

Cole o JSON: Cole o JSON copiado na caixa de texto que aparece e clique em "Import".

Salve o Workflow
: Dê um nome ao seu workflow (ex: "Extrair Categorias Shopee") e salve-o.

Execute o Workflow
: Clique no botão "Execute Workflow" (no canto superior direito) para iniciar a execução.

Verifique os Resultados
: Após a execução, o nó `Merge (Combinar Categorias)` conterá todos os dados extraídos em um único item, na forma de um array de objetos. Você pode inspecionar a saída clicando neste nó e verificando a aba "Output".

Com este setup, você poderá extrair as 164 categorias da Shopee conforme solicitado, utilizando apenas os nós nativos e o poder das expressões regulares no n8n.