Trabalho de Banco de Dados

Matheus Augusto Marques

13 de junho de 2018

Resumo

O objetivo desse trabalho é o desenvolvimento de um sistema em 3 camadas que apresente relatórios de análise de dados obtidos de sites na internet a partir de ferramentas de extração de dados, no estilo web crawlers.

1 Descrição do Trabalho

O presente documento possui o objetivo de compor uma parte do trabalho que será realizado na disciplina de Banco de Dados, ele se destina na especificação do que será desenvolvido ao longo do ano. O trabalho possui o propósito de desenvolver no aluno as seguintes habilidades:

- técnicas e ferramentas para aquisição automatizada de dados publicados na internet;
- manipulação de dados semiestruturados obtidos de fontes variadas;
- desenvolvimento de aplicações web em 3 camadas utilizando bancos de dados e J2EE.

O primeiro passo é escolher um conjunto de fontes de dados (sites) que oferecem informações que podem ser extraídas a partir de ferramentas automatizadas e que possibilitem gerar relatórios interessantes. Por exemplo, extrair informações a respeito de produtos (preços, características, prazos, avaliações), hoteis (preços, acomodações, características, procura), clima (histórico ou dados de diferentes lugares), ou qualquer área que lhe interesse.

As informações devem ser extraídas usando ferrametas/scripts de terceiros e importadas em um sistema em Java baseado em banco de dados criado por você. O sistema deve permitir visualizar e editar os dados carregados e também apresentar via web relatórios gerados a partir dos dados, incluindo tabelas, gráficos, etc.

O desenvolvimento do trabalho será dividido em etapas a serem desenvolvidas durante os 4 bimestres do curso, conforme segue.

1º BIMESTRE Objetivo: Especificação do sistema e realização da extração a partir das fontes de dados escolhidas.

Data de entrega: 06/06/2018 (apresentação individual em sala e submissão via Moodle)

Detahamento:

Escolha da área de aplicação, e faça a especificação da aplicação a ser desenvolvida, incluindo: - descrição das fontes de dados (sites) e dos dados em si (informações a serem coletadas); - descrição dos relatórios e gráficos a serem desenvolvidos na aplicação.

Escolha uma ferramenta de extração automatizada e a configure para que extraia os dados desejados das fontes de dados e exporte em arquivos JSON: - você não precisa desenvolver a ferramenta de extração, há várias disponíveis (em python, perl, etc. ou até via web) que só precisam ser configuradas/ajustadas para extrair o que você precisa; - os dados devem ter informações que permitam gerar relatórios interessantes, então ter dados numéricos e/ou datas e/ou estatísticas e outros é importante; - pense nos relatórios e os descreva e textualmente e por meio de imagens que apresentem o protótipo das telas de relatório (pode fazer o desenho da tela até à mão mesmo e tirar foto); - as fontes de dados e os relatórios devem ter escopo ajustado para que não fique grande demais e para que permita explorar conceitos abordados durante a disciplina.

O que deve ser apresentado e submetido via Moodle: - Especificação da aplicação (arquivo textual - Word/Libre Office/etc.); - Descrição da configuração da ferramenta de extração: código de configuração e, se for o caso, modificações/customizações no código da ferramenta, e um roteiro de execução (um exemplo de sequência de passos para executar uma extração) (arquivo textual - txt/Word/Libre Office/etc.) - Apresentação de um JSON resultante de uma extração (arquivo JSON).

2º BIMESTRE Objetivo: Projeto e desenvolvimento das funcionalidades referentes à importação no sistema dos arquivos JSON resultantes do processo de extração e edição dos dados armazenados (CRUD).

Data de entrega: 31/07/2018 (apresentação individual em sala e submissão via Moodle)

3º BIMESTRE Objetivo: Projeto e desenvolvimento dos relatórios, incluindo operações de consulta e visualização dos relatórios em um browser.

Data de entrega: 03/10/2018 (apresentação individual em sala e submissão via Moodle)

4º BIMESTRE Objetivo: Realização de correções e ajustes e finalização do sistema.

Data de entrega: 30/11/2018 (apresentação individual em sala e submissão via Moodle)

2 Especificação do Trabalho

Essa seção se destina a cumprir a primeira etapa do trabalho. Ou seja, aqui estará contido a área selecionada para reunir os dados e também qual foi a ferramenta escolhida para buscar os dados nos sites selecionados.

2.1 Área de aplicação

Para se ajustar no escopo do tempo e do que foi proposto no trabalho, foi escolhido para ser o tema de geração de informações o produto relógio masculino. Existem diversas outras temáticas, mas essa foi escolhidas apenas pelo gosto pessoal do autor e que de certo ponto parece ser interessante de se pesquisar.

2.1.1 Sites Selecionados

Os sites selecionados para a busca dos dados são os que seguem:

- Submarino
- Americanas
- ShopTime
- Amazon

2.1.2 Dados escolhidos para serem extraídos

Os dados extraidos dos sites escolhidos serão relacionados com o produto relógio masculino. Os dados buscados serão:

- link do produto
- Nome
- Preço
- Parcelas
- Quantidade de Avaliações
- Nota da Avaliação

2.1.3 Formato da extração dos dados

A coleta dos itens se dará da seguinte forma em cada um dos sites selecionados.

- Coleta dos 40 itens mais vendidos
- Coleta dos 40 itens mais avaliados
- Coleta dos 40 itens com maior preço

O critério de seleção dos itens, será no formato que os sites buscados apresetarem os resultados. Ou seja, quando a busca for feita no site pelo crawler, o mesmo selecionará os x primeiros itens do resultado da busca.

A quantidade de itens selecionada não é resultado de nenhum cálculo estatístico. O valor escolhido é inicial e poderá ser modificado a medida que o projeto se desenvolva.

Para cada um dois itens buscados será extraido os atributos definidos na seção 2.1.2.

2.1.4 Relatórios e Gráficos

Os relatórios e gráficos serão desenvolvidos com base nos itens buscados da seção 2.1.3. Em geral, os relatórios e dados estarão associados com a listagem a seguir.

- Verificar quantos itens são iguais em cada conjunto selecionado
- Média de preços em cada conjunto selecionado
- Média das parcelas em cada conjunto selecionado
- Verificar quais as marcas mais vendidas
- Faixa de preços em cada conjunto selecionado
- Faixa das parcelas em cada conjunto selecionado
- Cor de relógios mais presete nos produtos
- Quantos produtos são aprova de água
- País de origem das marcas mais vendidas
- Comparar os valores dos produtos brasileiros com o valor dos produtos estadunidenses

Em geral os relatórios possuirão textos e gráficos para simbolizar os resultados. Exemplo,

3 ferramenta de extração

A ferramenta de extração escolhida foi um módulo de nodejs que se chama crawler. Ele pode ser encontrado no link "https://www.npmjs.com/package/crawler". O módulo é escrito em javascript e utiliza comandos em jquery para selecionar os dados da página. Sua utilização é simples e rápida.

3.1 Exemplo de uso

Vejamos os passos para fazermos uso da ferramenta. Para utiliza-lá precisamos instalar o npm e nodejs no computador - isso é simples e o leitor pode fazer sozinho. Depois de instalado precisamos baixar e importar o módulo Crawler para nosso projeto - isso também é simples.

Com o ambiente configurado, precisamos criar um arquivo com um nome qualquer para utilizarmos o módulo. Faça esse passo. Depois, importe no arquivo criado o módulo do Crawler. Agora podemos utilizá-lo.

```
1  var Crawler = require("crawler");
2  const fs = require("fs");
```

Listing 1: Importação dos módulos

```
1  var jsonObject = [];
2  var fileName = "americanas-exemplo.json";
3
4  var c1 = new Crawler({
5    maxConnections: 2,
6    callback: function(error, res, done) {
7      if (error) {
8        console.log(error);
9      } else {
10       var $ = res.$;
11       $(".card-product-url")
12         .slice(0, 5)
13         .each(function() {
14           var details = $(this).children(".card-product-details");
15           if (details.children(".card-product-info-unavailable").length == 0) {
16             jsonObject.push({
17               nome: details.children(".card-product-name").text(),
18               preco: details
19                 .children(".card-product-offers")
20                 .children(".card-product-prices")
21                 .children(".card-product-price")
22                 .children(".value")
23                 .text(),
24               parcelas: details
25                 .children(".card-product-offers")
26                 .children(".card-product-prices")
27                 .children(".placeholder")
28                 .children(".card-product-installments")
29                 .text()
30             });
31           }
32         });
33         done();
34     }
35
36     var json = JSON.stringify(jsonObject);
37     fs.createWriteStream(fileName).write(json);
38
39   }
40 });
41
42 c1.queue({
43   uri:
44     "https://www.americanas.com.br/busca/relogio-masculino?ordenacao=rating"
45 });
```

Listing 2: Configuração do crawler. O código busca dados de relógios no site da Americanas.

O código 1 mostra como podemos importar o módulo Crawler. Já o código 2 mostra como configurar e usar o módulo. Em geral, devemos (Linha 4) instânciar um objeto do tipo Crawler e dentro dele criarmos a configuração da requisição a uma determinada página. Na linha 42 especificamos a url que queremos buscar os dados e nela ocorre a requisição no site. Da linha 10 a 40 do código 2 especificamos que tipo de ação queremos realizar depois da requisição ter ocorrido com sucesso. Todos os comando estão em formato de jquery, pois a ferramenta as utiliza. Nas linhas 36 e 37 exportamos os dados para json.

Resumidamente, devemos seguir os passos a seguir para utilizar o crawler.

- Instalar o npm e nodejs
- Baixar o módulo Crawler
- Criar um arquivo e importar o módulo Crawler
- Configurar o Crawler instânciando um objeto do mesmo
- Especificar a url que será analisada
- Exportar os dados para json

Para mais informações acesse o link da ferramenta.

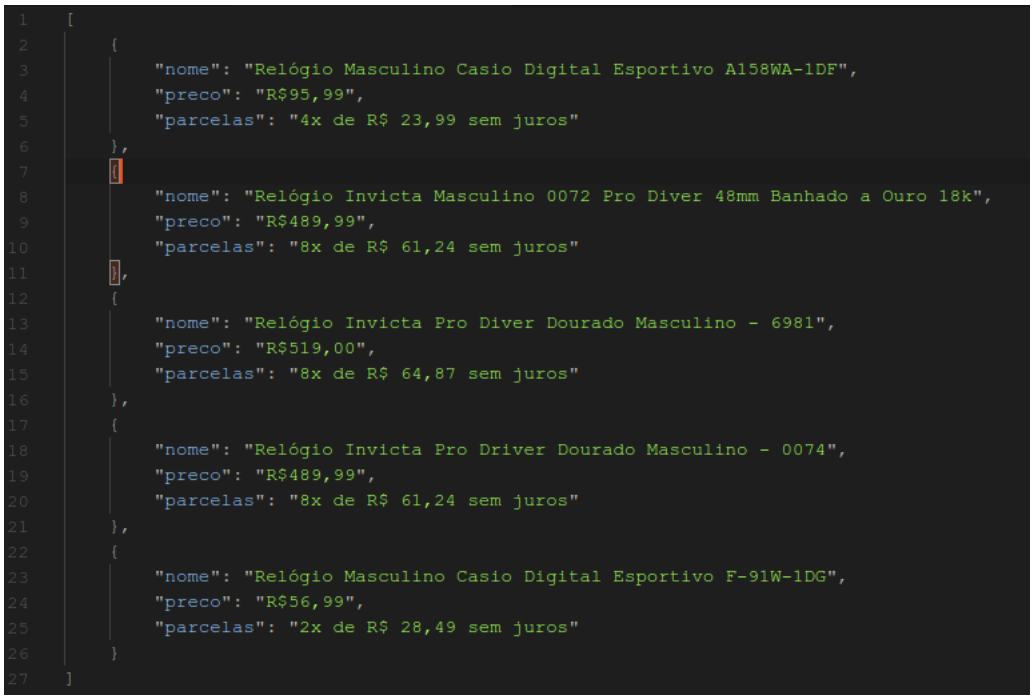

Figura 1: Arquivo json de saida