Trabalho de Banco de Dados

Eduardo Henrique Giroto

14 de junho de 2018

1 Descrição do Trabalho

O trabalho consiste em desenvolver um sistema em 3 camadas que apresente relatórios de análise de dados utilizando ferramentas de extração(web crawlers) para obter os respectivos dados.

2 Especificação do Trabalho

Essa seção destina-se ao Primeiro Bimestre, tendo como objetivo a especificação do sistema e a realização da extração a partir das fontes de dados escolhidas.

2.1 Área de Aplicação

A área escolhida foi a Copa do Mundo, tendo como principal motivo da escolha estarmos em um ano de Copa, gosto pessoal e por ter vários dados interessantes a se extrair.

2.2 Sites Escolhidos

- Fifa.

2.3 Tipos de dados extraídos

- Seleções Campeãs do Mundo e seu respectivo número de títulos.
- Pontuação Geral das Seleções.
- Número de participações em Copas de cada seleção.
- Número de partidas de cada Seleção, tendo o número de vitórias, empates e derrotas.
- Número de cartões de cada seleção(total, amarelos e vermelhos).

2.4 Relatórios e Gráficos

Os relatórios e gráficos serão gerados com base nos tipos dados extraídos, sendo associados aos seguintes tópicos:

- Aproveitamento total de cada seleção na história das Copas.

- Média de cartões por jogo de cada seleção.
- Porcentagem de participação em Copas.
- Média de gols por jogo de cada seleção
- Saldo de gols de cada Seleção.

3 Ferramenta de Extração

A ferramenta de extração usada no trabalho foi o Scrapy, um framework de Web Scraping em python. O website da ferramenta é *scrapy.org*.

3.1 Instalação da Ferramenta

Primeiramente, baixe o Scrapy, segue o comando:

pip install scrapy

3.2 Criando Projeto

Para criar um projeto, basta utilizar o comando:

scrapy startproject nomeprojeto

O ”nomeprojeto”seria o nome que você deseja para o seu projeto.

3.3 Criando um Spider

Spiders são classes que vão percorrer um determinado website a procura dos dados que você definiu. Segue o comando para a criação de um spider:

scrapy genspider nome dominio

”nome”é o nome dado ao seu spider e ”dominio”é o nome do domínio do seu website a ser explorado.

3.4 Exemplo de Spider

Segue um exemplo de um spider que vai crawlear uma tabela do nosso site alvo:

```python
# -*- coding: utf-8 -*-
import scrapy


class PontuacaoSpider(scrapy.Spider):
    name = 'pontuacao'
    allowed_domains = ['fifa.com']
    start_urls = ['http://www.fifa.com/fifa-tournaments/statistics-and-records/worldcup/teams/index.html']

    def parse(self, response):
        for tabela2 in response.xpath('//table[@class="table tbl-alltimeranking"]/tbody/tr'):
            yield{
                'rank': tabela2.xpath('td[@class="tbl-rank"]/span/text()').extract_first(),
                'selecao': tabela2.xpath('td[@class="tbl-teamname teamname-link"]/a/div/div[@class="t-n"]/span/text()').extract_first(),
                'pontos': tabela2.xpath('td[@class="tbl-points"]/span/text()').extract_first(),
                'jogos': tabela2.xpath('td[@class="tbl-matches-num"]/span/text()').extract_first(),
                'vitorias': tabela2.xpath('td[@class="tbl-win"]/span/text()').extract_first(),
                'empates': tabela2.xpath('td[@class="tbl-draw"]/span/text()').extract_first(),
                'derrotas': tabela2.xpath('td[@class="tbl-lost"]/span/text()').extract_first()
            }
```

O spider acima possui a variável allowed domais que possui o domínio do website desejado, a

variável start urls possui a página específica do domínio que irá ser crawleado. O spider irá retornar o título da página que está querendo extrair, dentro da função parse.

3.5 Executando o Spider

Para executar o spider dentro de um projeto, basta fazer o comando a seguir:

scrapy crawl nome

”nome”é o nome do seu Spider, também podemos ter o parâmetro -o”no comando para colocar a saída no formato desejado, exemplo:

scrapy crawl nome -o saida.json

No nosso exemplo, a saída será no formato json, segue a saída do spider acima:

```
[
{"rank": "1", "selecao": "Brazil", "pontos": "227", "jogos": "104", "vitorias": "70", "empates": "17", "derrotas": "17"},
{"rank": "2", "selecao": "Germany", "pontos": "218", "jogos": "106", "vitorias": "66", "empates": "20", "derrotas": "20"},
{"rank": "3", "selecao": "Italy", "pontos": "156", "jogos": "83", "vitorias": "45", "empates": "21", "derrotas": "17"},
{"rank": "4", "selecao": "Argentina", "pontos": "140", "jogos": "77", "vitorias": "42", "empates": "14", "derrotas": "21"},
{"rank": "5", "selecao": "Spain", "pontos": "99", "jogos": "59", "vitorias": "29", "empates": "12", "derrotas": "18"},
{"rank": "6", "selecao": "England", "pontos": "98", "jogos": "62", "vitorias": "26", "empates": "20", "derrotas": "16"},
{"rank": "7", "selecao": "France", "pontos": "96", "jogos": "59", "vitorias": "28", "empates": "12", "derrotas": "19"},
{"rank": "8", "selecao": "Netherlands", "pontos": "93", "jogos": "50", "vitorias": "27", "empates": "12", "derrotas": "11"},
{"rank": "9", "selecao": "Uruguay", "pontos": "72", "jogos": "51", "vitorias": "20", "empates": "12", "derrotas": "19"},
{"rank": "10", "selecao": "Sweden", "pontos": "61", "jogos": "46", "vitorias": "16", "empates": "13", "derrotas": "17"},
{"rank": "11", "selecao": "Russia", "pontos": "59", "jogos": "40", "vitorias": "17", "empates": "8", "derrotas": "15"},
{"rank": "12", "selecao": "Serbia", "pontos": "59", "jogos": "43", "vitorias": "17", "empates": "8", "derrotas": "18"},
{"rank": "13", "selecao": "Mexico", "pontos": "56", "jogos": "53", "vitorias": "14", "empates": "14", "derrotas": "25"},
{"rank": "14", "selecao": "Belgium", "pontos": "51", "jogos": "41", "vitorias": "14", "empates": "9", "derrotas": "18"},
{"rank": "15", "selecao": "Poland", "pontos": "50", "jogos": "31", "vitorias": "15", "empates": "5", "derrotas": "11"},
{"rank": "16", "selecao": "Hungary", "pontos": "48", "jogos": "32", "vitorias": "15", "empates": "3", "derrotas": "14"},
{"rank": "17", "selecao": "Portugal", "pontos": "43", "jogos": "26", "vitorias": "13", "empates": "4", "derrotas": "9"},
{"rank": "18", "selecao": "Czech Republic", "pontos": "41", "jogos": "33", "vitorias": "12", "empates": "5", "derrotas": "16"},
{"rank": "19", "selecao": "Austria", "pontos": "40", "jogos": "29", "vitorias": "12", "empates": "4", "derrotas": "13"},
{"rank": "20", "selecao": "Chile", "pontos": "40", "jogos": "33", "vitorias": "11", "empates": "7", "derrotas": "15"},
{"rank": "21", "selecao": "Switzerland", "pontos": "39", "jogos": "33", "vitorias": "11", "empates": "6", "derrotas": "16"},
{"rank": "22", "selecao": "Paraguay", "pontos": "31", "jogos": "27", "vitorias": "7", "empates": "10", "derrotas": "10"},
```