TP_RPCW

Trabalho Prático de RPCW

Relatório do Trabalho Prático

Introdução

Este trabalho prático envolve a criação de um ficheiro JSON com informações sobre festas tradicionais em Portugal, extraídas de um site específico, e a utilização desta informação para criar uma ontologia. O processo incluiu web scraping, processamento de texto para lidar com diferentes formatos de entrada e a transformação das datas em formatos padronizados. Além disso, o projeto inclui a construção de uma aplicação web para exibir essas festas utilizando a ontologia carregada em um servidor GraphDB.

Coleta de Dados

Fonte de Dados

As informações foram coletadas do site TerraMater.

Estrutura das Informações no Site

As informações no site estão organizadas da seguinte forma:

- Por regiões (uma página para cada província).
- Dentro de cada página, as festas são organizadas por trimestres e uma festa por linha.
- Cada linha de festa pode ter diferentes formatos.

Exemplos de Formatos das Linhas:

- freguesia , concelho , data - nome : descricao
- freguesia , concelho , data - nome . descricao
- freguesia , concelho , data : nome . descricao
- concelho , data - nome , +descricao
- concelho , data : nome , +descricao

Esses diferentes formatos representam um desafio significativo para a criação de um script capaz de extrair as informações de maneira consistente.

Desenvolvimento do Script de Web Scraping

Tecnologias Utilizadas

- **Bibliotecas:** BeautifulSoup para web scraping, Pandas para manipulação de dados, Dateutil para manipulação de datas, FuzzyWuzzy para correspondência aproximada de strings.
- **Python:** Linguagem de programação utilizada para o desenvolvimento do script.

Estrutura do Script

O script realiza as seguintes tarefas principais:

1. **Configuração do User-Agent:** Para evitar bloqueios durante o scraping.
2. **Leitura e Processamento do Ficheiro de Metadados:** Utilização de um ficheiro `FreguesiasPortugalMetadata.xlsx` para obter informações sobre regiões, distritos, concelhos e freguesias obtido do site dados.gov.
3. **Definição de Funções para Extração de Informações:** Funções específicas para lidar com diferentes formatos de linhas de festas e para converter datas em formatos padronizados.
4. **Função Principal de Web Scraping:** Coleta de dados de cada URL das páginas de regiões e armazena as informações extraídas em estruturas de dados apropriadas.

Casos de Linhas e Desafios

Cada linha de festa no site pode seguir um dos vários formatos identificados. Os exemplos apresentados representam alguns dos casos encontrados:

- freg , conc , data - nome : desc
- conc , data - nome , +desc
- conc , data - nome

Esses formatos exigem que o script de scraping seja flexível e robusto para capturar as informações corretamente. A função de parsing analisa a linha e tenta identificar os componentes (freguesia, concelho, data, nome, descrição) com base em padrões de texto.

Tratamento de Datas

As datas no site vêm em diferentes formatos, como "14 de novembro" ou "domingos de pascoa". Para padronizar as datas, foram criadas funções específicas para converter expressões como "primeiro domingo de julho" em um formato `dd/mm/yyyy`.

Estrutura do Ficheiro JSON

O ficheiro JSON final possui a seguinte estrutura:

```
{
  "regioes": [
    {
      "regiao": "beira_litoral",
      "distritos": [
        {
          "distrito": "Aveiro",
          "concelhos": [
            {
              "concelho": "Águeda",
              "freguesias": ["Aguada de Cima", "Fermentelos"]
            }
          ]
        }
      ]
    }
```

```
    ],
    "festas": [
      {
        "festa_id": 1,
        "Nome": "Cortejo dos Reis",
        "Descrição": "com pequeno auto teatral e leilão das ofertas",
        "Data Início": "06-01-2024",
        "Data Fim": "06-01-2024",
        "Região": "beira_litoral",
        "Distrito": "Aveiro",
        "Concelho": "Sever do Vouga",
        "Freguesia": "Talhadas"
      }
    ]
  }
```


Criação da Ontologia

Definição das Classes, Data Properties e Object Properties

Para a criação da ontologia, foram definidas várias classes, data properties e object properties no Protégé. A seguir, apresentamos um resumo das definições:

Classes:

- Regiao
- Distrito
- Concelho
- Freguesia
- Festa

Object Properties:

- regiaoTemFesta
- distritoTemFesta
- concelhoTemFesta
- freguesiaTemFesta
- ocorreRegiao
- ocorreDistrito
- ocorreConcelho
- ocorreFreguesia
- pertenceRegiao
- pertenceDistrito
- pertenceConcelho
- temDistrito
- temConcelho
- temFreguesia

Data Properties:

- nome
- descricao
- dataInicio
- dataFim

Inserção dos Indivíduos

Foi realizada a inserção dos indivíduos no arquivo TTL a partir do JSON criado. O script `ontologia.py` lê o ficheiro JSON e gera o arquivo TTL com os indivíduos.

Estrutura do Projeto

```
TP_RPCW
📦festas
├ 📂node_modules
├ 📂public
├ 📂src
│ ├ 📂assets
│ │ ├ 📂mapa
│ │ ├ 📜festas.json
│ │ ├ 📜fotos_juntas.png
│ │ ├ 📜fundo.jpg
│ │ └ 📜romaria.png
│ ├ 📂components
│ │ ├ 📜Descricao.js
│ │ ├ 📜Festa.js
│ │ ├ 📜Festas.js
│ │ ├ 📜Footer.js
│ │ ├ 📜Header.js
│ │ ├ 📜Mapa.css
│ │ └ 📜Mapa.js
│ ├ 📂pages
│ │ ├ 📜Criar.js
│ │ └ 📜Home.js
│ ├ 📜App.css
│ ├ 📜App.js
│ ├ 📜App.test.js
│ ├ 📜index.css
│ ├ 📜index.js
│ ├ 📜logo.svg
│ ├ 📜ontologia_teste.ttl
│ ├ 📜reportWebVitals.js
│ └ 📜setupTests.js
├ 📜.gitignore
├ 📜README.md
├ 📜package-lock.json
├ 📜package.json
├ 📜server.mjs
└ 📜tailwind.config.js
📦scrap
 ├ 📂dados
```

```
├ 📜.DS_Store
├ 📜README.md
├ 📜datas.json
├ 📜distritos_concelhos_freguesias.json
├ 📜festas.json
├ 📜ontologia.py
├ 📜outros.txt
├ 📜scraper.py
├ 📜scraper_fernando.py
├ 📜script_regioes.py
├ 📜sem_distritos.txt
├ 📜testes.py
└ 📜urls.txt
```

Estrutura do Projeto e Execução

Pasta `scrap`

A pasta `scrap` contém scripts e dados relacionados à criação do JSON e da ontologia. Esta parte do projeto envolve:

- **Coleta de Dados:** Scripts para realizar web scraping e extrair informações sobre as festas.
- **Processamento de Dados:** Arquivos JSON intermediários e finais, além de scripts para manipulação e limpeza de dados.
- **Criação da Ontologia:** Scripts como `ontologia.py` que geram o arquivo TTL a partir do JSON criado.

Pasta `festas`

A pasta `festas` contém o código da aplicação web que exibe as festas. Esta parte do projeto envolve:

- **Frontend:** Componentes React que constroem a interface de usuário.
- **Backend:** Um servidor para servir a aplicação, se necessário.
- **Configurações:** Arquivos de configuração para dependências, estilo e outros aspectos da aplicação.

Como Executar o Projeto

Passo 1: Configurar a Ontologia no GraphDB

1. **Importar a Ontologia:** A ontologia deve ser carregada no GraphDB com o nome `FestasRomarias`.
2. **Inserir Dados:** Insira o ficheiro JSON `festas.json` na ontologia.

Passo 2: Instalar Dependências

1. Navegue até a pasta `festas`.
2. Execute o comando:

```
npm install
```

Passo 3: Executar a Aplicação

Ainda na pasta festas, execute o comando:

```
npm start
```

O comando acima abrirá uma janela no navegador no endereço http://localhost:3000, onde a aplicação estará rodando.

Dependências

Certifique-se de que possui as seguintes ferramentas instaladas:

- **Node.js e npm:** Para gerenciar dependências e scripts de build.
- **GraphDB:** Para hospedar a ontologia e os dados das festas.

Descrição do server.mjs

O ficheiro server.mjs serve para fazer queries à base de dados (ontologia) hospedada no GraphDB, que está disponível no localhost:7200. O resultado das queries pode ser acessado no localhost:5000. A aplicação frontend faz fetch para essa porta para obter os dados necessários.

Consultar e Queries SPARQL

As queries SPARQL utilizadas foram testadas no GraphDB para garantir que os dados retornados estivessem corretos. Aqui estão alguns exemplos de queries utilizadas:

`/festas`

Descrição: Obter todas as festas.

`/distritos`

Descrição: Obter todos os distritos.

`/concelhos?distrito='Porto'`

Descrição: Obter todos os concelhos de um determinado distrito.

`/freguesias?concelho='Gaia'`

Descrição: Obter todas as freguesias de um determinado concelho.

`/festas?distrito='Porto'`

Descrição: Obter todas as festas de um determinado distrito.

`/criar_festa`

Descrição: Adicionar uma festa à ontologia.

Conclusão

Este trabalho prático permitiu a aplicação de várias técnicas de web scraping, processamento de dados e manipulação de ontologias para criar uma aplicação web completa que exibe informações sobre festas tradicionais em Portugal. A utilização de ferramentas como BeautifulSoup, Pandas, e GraphDB foi essencial para o sucesso do projeto