Web Semântica- Trabalho prático 2 Relatório Grupo 5

Bernardo Rodrigues - nº 88835 - Mestrado em Engenharia Informática

Diogo Andrade - nº 89265 - Mestrado em Engenharia Informática

Gonçalo Freixinho - nº 89251 - Mestrado em Engenharia Informática

Índice

Introdução ao tema	3
Dados, suas fontes e sua transformação	4
A definição da ontologia (RDFS e OWL)	5
Exemplo da definição das ontologias	5
Conjunto de inferências (SPIN)	6
Novas operações sobre os dados (SPARQL)	7
O uso e integração de dados da Wikidata e/ou DBpedia	9
A publicação da semânticos dos dados através de RDFa e/ou microfo	rmatos 11
Funcionalidades da Aplicação (UI)	12
Conclusões	16
Configuração para executar a aplicação	17
Criar repositório na graphDB	17
Importar os dados na graphDB	17
Executar aplicação	17
Executar ficheiro de criação de inferências	17

Introdução ao tema

Um sistema de informação é um conjunto organizado de elementos, podendo ser pessoas, dados, atividades ou recursos materiais. Estes elementos interagem entre si para processar informação e divulgá-la de forma adequada em função dos objetivos de uma determinada organização.

Assim, no âmbito deste tema e desta disciplina, foi proposto um problema que consiste em focarmo-nos no Sistema de Informação baseado em Web que foi desenvolvido no primeiro trabalho onde era possível realizar processos de exposição e gestão de informação presentes numa aplicação, acrescentando agora um conjunto de funcionalidades, tais como a criação, descrição e uso de ontologias, a criação de um conjunto de regras de inferência, a complementação do conjunto de dados do Sistema de Informação e a publicação da semântica dos dados do Sistema de Informação.

Para atingirmos todas estas funcionalidades foram adotadas várias tecnologias tais como: Python/Django para a programação da aplicação, dados em formato RDF (mais especificamente em NT), Triplestore GraphDB (utilizado como repositório de dados), SPARQL (para a realização das pesquisas e alterações dos dados na triplestore), uso de RDFS e OWL (para a realização das ontologias), SPIN (para representar o conjunto de inferências), SPARQLwrapper (para o acesso ao endpoint da dbpedia e wikidata) e o RDFa e /ou Micro-formatos (para lidar com a publicação da semântica).

Dados, suas fontes e sua transformação

Para o desenvolvimento do nosso sistema de informação, utilizou-se o mesmo dataset que tinha sido utilizado no trabalho 1, acerca de shows da plataforma netflix, link do dataset:

https://www.kaggle.com/shivamb/netflix-shows

Este dataset contém informação relativa aos vários filmes e séries aí contidos. A informação contida no dataset está no formato csv, deste modo é necessário proceder ao tratamento de dados para um formato rdf, ficheiros neste formato são caracterizados por cada entrada consistir num conjunto de 3 elementos, um sujeito, um predicado para o sujeito e o objeto.

Dados no formato CSV:

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
s1	TV Show	3%		João Miguel, Bianca	Brazil	August 14, 202	2020	TV-MA	4 Seasons	Internationa
s2	Movie	07:19	Jorge Michel Grau	Oscar Serrano, Azalia	Mexico	December 23,	2016	TV-MA	93 min	Dramas, Inte
s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Ch	Singapore	December 20,	2011	R	78 min	Horror Movi
s4	Movie	9	Shane Acker	Elijah Wood, John C.	United Sta	November 16,	2009	PG-13	80 min	Action & Adv
s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin S	United Sta	January 1, 202	2008	PG-13	123 min	Dramas
s6	TV Show	46	Serdar Akar	Yasemin Allen, Melis	Turkey	July 1, 2017	2016	TV-MA	1 Season	Internationa
s7	Movie	122	Yasir Al Yasiri	Amina Khalil, Ahmed	Egypt	June 1, 2020	2019	TV-MA	95 min	Horror Movi
s8	Movie	187	Kevin Reynolds	Samuel L. Jackson, Jo	United Sta	November 1, 2	1997	R	119 min	Dramas
s9	Movie	706	Shravan Kumar	Divya Dutta, Atul Ku	India	April 1, 2019	2019	TV-14	118 min	Horror Movie
s10	Movie	1920	Vikram Bhatt	Rajneesh Duggal, Ad	India	December 15,	2008	TV-MA	143 min	Horror Movie
s11	Movie	1922	Zak Hilditch	Thomas Jane, Molly	United Sta	October 20, 20	2017	TV-MA	103 min	Dramas, Thri
s12	TV Show	1983		Robert Więckiewic	Poland, Ui	November 30,	2018	TV-MA	1 Season	Crime TV Sho
s13	TV Show	1994	Diego Enrique Osc	rno	Mexico	May 17, 2019	2019	TV-MA	1 Season	Crime TV Sho
s14	Movie	2,215	Nottapon Boonpra	Artiwara Kongmalai	Thailand	March 1, 2019	2018	TV-MA	89 min	Documentar
s15	Movie	3022	John Suits	Omar Epps, Kate Wa	United Sta	March 19, 202	2019	R	91 min	Independent
s16	Movie	Oct-01	Kunle Afolayan	Sadiq Daba, David B	Nigeria	September 1, 2	2014	TV-14	149 min	Dramas, Inte

Conversão de dados para NT:

```
<http://shows.org/type/tv_show> <http://shows.org/pred/type> "TV Show" .
<http://shows.org/shows/s1> <http://shows.org/pred/type> <http://shows.org/type/tv_show> .
<http://shows.org/shows/s1> <http://shows.org/pred/title> "3%" .
<http://shows.org/people/joão_miguel> <http://shows.org/pred/name> "João Miguel" .
<http://shows.org/shows/s1> <http://shows.org/pred/cast> <http://shows.org/people/joão_miguel> .
<http://shows.org/people/bianca_comparato> <http://shows.org/pred/name> "Bianca Comparato" .
<http://shows.org/shows/s1> <http://shows.org/pred/cast> <http://shows.org/people/bianca_comparato> .
<http://shows.org/shows/s1> <http://shows.org/pred/cast> <http://shows.org/people/bianca_comparato> .
<http://shows.org/shows/s1> <http://shows.org/pred/name> "Michel Gomes" .
<http://shows.org/shows/s1> <http://shows.org/pred/name> "Rodolfo Valente" .
<http://shows.org/people/rodolfo_valente> <http://shows.org/pred/name> "Rodolfo Valente" .
<http://shows.org/people/vaneza_oliveira> <http://shows.org/pred/name> "Vaneza Oliveira" .
<http://shows.org/people/vaneza_oliveira> <http://shows.org/people/vaneza_oliveira> .
```

A definição da ontologia (RDFS e OWL)

Foi criado um ficheiro (no formato N3) de ontologias para classificar automaticamente os dados. Deste modo podemos facilitar a criação das queries e otimizar as pesquisas.

A Ontologia tem as seguintes classes: People, Type, Country, ListedIn, Duration, ReleaseYear e Show. Existem as seguintes subclasses: Director, Actor. Apresenta as seguintes propriedades: type, cast, director, country, release_year, duration, listed_in, date_added, description, title. Propriedade inferidas: dualrole, popularity.

Por exemplo, neste caso em específico, é criada a classe Pessoas e tanto os Actores como os Diretores agora estão classificados sob a classe Pessoas.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix show: <http://show.org/show/> .
@prefix pred: <http://shows.org/pred/> .
pred:name rdfs:subPropertyOf foaf:name .
foaf:Person rdf:type rdfs:Class .
show:People rdf:type rdfs:Class ;
            rdfs:subClassOf foaf:Person ;
            rdfs:domain pred:name .
show:Type rdf:type rdfs:Class .
show:Country rdf:type rdfs:Class .
show:ListedIn rdf:type rdfs:Class .
show:Duration rdf:type rdfs:Class .
show:ReleaseYear rdf:type rdfs:Class .
show:Show rdf:type rdfs:Class
```

Exemplo da definição das ontologias

Conjunto de inferências (SPIN)

Recorrendo ao uso de inferências podemos descobrir novas relações entre as entidades ontológicas e entre as suas propriedades. Estas descobertas são baseadas nos dados já existentes e na informação adicional descrita pelas ontologias.

Com o uso de inferências conseguimos então criar novos triplos, para construir e adicionar esses novos triplos à base de dados criamos um script python (inferencias.py).

Nos nossos dados vimos que havia entidades que atuavam no mesmo filme em que eram diretores, deste modo usamos inferências para atribuir a propriedade dualrole para saber quais as pessoas que eram simultaneamente ator e diretor de um filme.

Outra inferência que criamos foi com a intenção de descobrir quais os atores/diretores mais populares, tendo isto em conta calculamos em quantos filmes eles estavam envolvidos e através desse número de filmes medimos o nível de popularidade de uma celebridade, por exemplo um ator que participou em 20 ou mais filmes é considerado extremamente popular.

Inferência para criar a propriedade dualrole

Novas operações sobre os dados (SPARQL)

Para tirar proveito das ontologias criadas, desenvolvemos novas queries SPARQL.

Por exemplo, como definimos Atores e Diretores como instâncias da classe Pessoas, criamos uma query onde pesquisamos todas por todas pessoas, retornando assim tanto os diretores e atores sem necessidade de estar a realizar queries com unions.

```
PREFIX rdfs: <a href="http://www.w3.org/2000/01/rdf-schema#">http://www.w3.org/1999/02/22-rdf-syntax-ns#></a>

prefix foaf: <a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/>>

prefix show: <a href="http://show.org/show/">http://show.org/show/">http://show.org/show/">prefix pred: <a href="http://shows.org/pred/">http://shows.org/pred/</a>

SELECT DISTINCT ?celname ?popularity (GROUP_CONCAT(?title;SEPARATOR=";") AS ?titles)

WHERE {

    ?cel a show:People .
    ?cel pred:name ?celname .
    ?cel pred:popularity ?popularity .
    ?cel pred:dualrole ?show .
    ?show pred:title ?title .
} GROUP bY ?celname ?popularity OFFSET """ + str(page * 30) + """ LIMIT 30
```

Exemplo de uma das novas queries recorrendo às ontologias

Como por exemplo com as ontologias definidas, adicionamos a página que lista todas as celebridades. Como tanto os atores como os diretores estão definidos como instâncias da classe Person, a query para listar as celebridades em vez de precisar de listar tanto atores como diretores e realizar a união dos dois, agora com as ontologias apenas precisamos de listar as entidades Person (visto que esta inclui atores e diretores).

Das pesquisas realizadas no primeiro trabalho, para listar o tipo de show, countries e categorias de um show, utilizamos as ontologias criadas de forma a aprimorar estas mesmas pesquisas. Pois enquanto que na versão inicial tínhamos de percorrer a lista de todos os filmes para ir buscar essas listas, agora basta apenas iterar usando a classe definida nas ontologias.

```
PREFIX pred: <http://shows.org/pred/>
SELECT DISTINCT ?typename
WHERE {
     ?show pred:type ?type .
     ?type pred:type ?typename .
}
```

Versão inicial, onde se percorria a lista de todos os filmes para ir buscar a lista de tipos dos shows

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

prefix foaf: <http://xmlns.com/foaf/0.1/>

prefix show: <http://show.org/show/>
prefix pred: <http://shows.org/pred/>

SELECT ?typename
WHERE {
    ?type a show:Type.
    ?type pred:type ?typename .
}
```

Versão atual onde recorremos à ontologia para pesquisar pelo tipo

O uso e integração de dados da Wikidata e/ou DBpedia

Para adicionar informação adicional aos nossos dados, decidimos complementar os nossos dados com dados presentes na Wikidata e na DBpedia. Para adquirir estes dados realizamos queries à Wikidata e à DBPedia pesquisando pelos nomes das entidades, por exemplo para ir buscar dados sobre um determinado ator pesquisamos pelo nome desse ator (o que caso existam pessoas/shows com nomes idênticos pode levar a retornar a entidade errada, porém algumas medidas foram tomadas para prevenir a situação como por exemplo certificarmos que a pessoa pesquisada pertence à classe actor e director nas wikis.)

Com os dados das wikis providenciamos ao utilizador informação extra que complementa a informação da base de dados como por exemplo uma pequena descrição de um "show" ou lista de prêmios que ganhou.

```
PREFIX rdf: <a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#">http://dbpedia.org/ontology/>
PREFIX rdfs: <a href="http://www.w3.org/2000/01/rdf-schema#">http://www.w3.org/2000/01/rdf-schema#</a>
PREFIX foaf: <a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/>

SELECT * WHERE {
    # search can either be a film or a tvShow so we search for both
Optional {?item a ont:Film. ?item ?label """ + name + """@en. ?item rdfs:label ?itemLabel. FILTER(LANGMATCHES(LANG(?itemLabel), 'en')).}
Optional {?item a ont:TelevisionShow. ?item ?label """ + n + """@en. ?item rdfs:label ?itemLabel. FILTER(LANGMATCHES(LANG(?itemLabel), 'en')).}
#FILTER regex (str(?itemLabel), "name"). # exact name
Optional {?item ont:abstract ?abstract. FILTER(LANGMATCHES(LANG(?abstract), 'en')).}
```

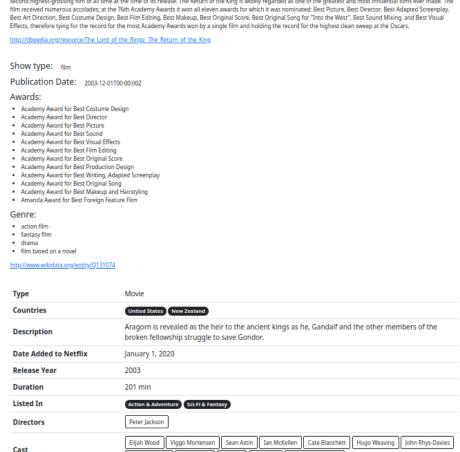
Query à DBpedia para a descrição de um "show"

As queries às wikis são realizadas através de queries SPARQL aos endpoints respetivos das wikis no exemplo anterior, vemos a query para pedir a descrição de um filme/série.



Abstract

The Lord of the Rings: The Return of the King is a 2003 epic fantasy adventure film directed by Peter Jackson, based on the third volume of J. R. R. Tolkien's The Lord of the Rings. The film is the third and final instalment in The Lord of the Rings trilogy and was produced by Barrie M. Osborne, Jackson and Fran Walsh, and written by Walsh, Philippa Boyens and Jackson. The film features an ensemble cast including Elijah Wood, Ian McKellen, Liv Tyker, Vigop Mortensen, Sean Astin, Cate Blanchett, John Rhys-Davies, Bernard Hill, Billy Boyd, Dorminic Monaghan, Orlando Bloom, Hugo Weaving, Miranda Otto, David Wenham, Karl Urban, John Noble, Andy Serkis, Ian Holm, and Sean Bean. It was preceded by The Fellowship of the Ring (2001) and The Two Towers, Frow Cowers, Frodo, Sam and Gollum are making their final way toward Mount Doom in Mordor in order to destroy the One Ring, unaware of Gollum's true intentions, while Gandalf, Aragorn, Legolas, Gimli and the rest are joining forces together against Sauron and his legions in Minas Tirith. The Return of the King was financed and distributed by American studio New Line Cinema, but filmed and edited entirely in Jackson's native New Zealand, concurrently with the other two parts of the trillogy. It premiered on 1 December 2003 at the Embassy Theatrie in Wellington and was theatrically released on 17 December 2003 in the United States, and on 18 December 2003 in New Zealand. The film was highly acclaimed by critics and fans alike, who considered it to be a landmark in filmmaking and the fantasy film genre. It grossed \$1.142 billion worldwide, making it the highest-grossing film of 2003 and the second highest-grossing film of all time at the time of its release. The Return of the King is widely regarded as one of the greatest and most influential films ever made. The film received numerous accolades; at the 76th Academy Awards it won all eleven awards for which it was nominated: Best Picture, Best Didnets, Best Adapted Screenplay, Best Art Direction, Best Costroin, Best Costroi



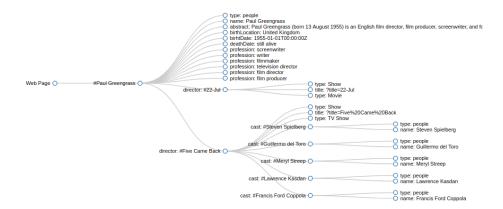
Exemplo dos dados das wikis numa página do site

Bernard Hill | Liv Tyler | Billy Boyd | Dominic Monaghan

No exemplo temos uma página relativa a um filme, com a descrição adquirida da DBpedia, assim como o link da página relativa ao filme. Alguns atributos extras como os prêmios ganhos pelo filme adquiridos pela Wikidata e finalmente os dados presentes na base de dados.

A publicação da semânticos dos dados através de RDFa e/ou microformatos

De modo a adicionar dados semânticos às páginas web, recorremos ao uso de Microformatos e RDFa. Deste modo toda a informação que o user consegue observar quando pesquisa por um ator/diretor/show está também embutida no código HTML.



Exemplo de os dados semânticos de um director

Podemos ver no exemplo os dados relativos a uma entidade, estes dados incluem tantos os presentes na base de dados como os adquiridos através da Wikidata/DBpedia.

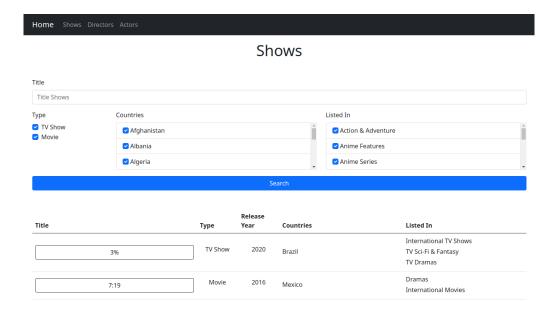
No exemplo vemos dados relativos ao diretor "Paul Greengrass" estes incluem uma descrição da pessoa, o local de nascimento, entre outros... (adquiridos por queries às wikis) e dados como os filmes que realizou e respectivos elencos (conseguidos através de queries SPARQL à base de dados).

Exemplo dos dados semânticos de um actor (nome e filmes em que participou)

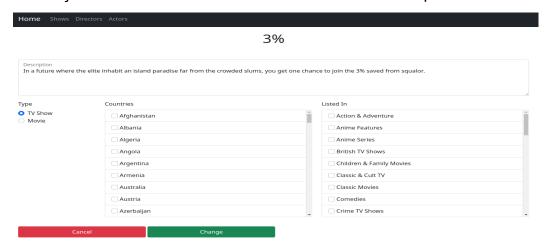
Funcionalidades da Aplicação (UI)

A app foi desenhada de modo a permitir aos utilizadores:

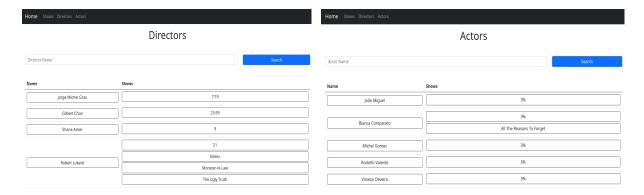
Observar e pesquisar todos os filmes/séries existentes, filtrar e procurar por séries/filmes dependendo da categoria, procurar por um filme específico e obter informação detalhada sobre este, tais como data de criação, o seu tipo e categoria.



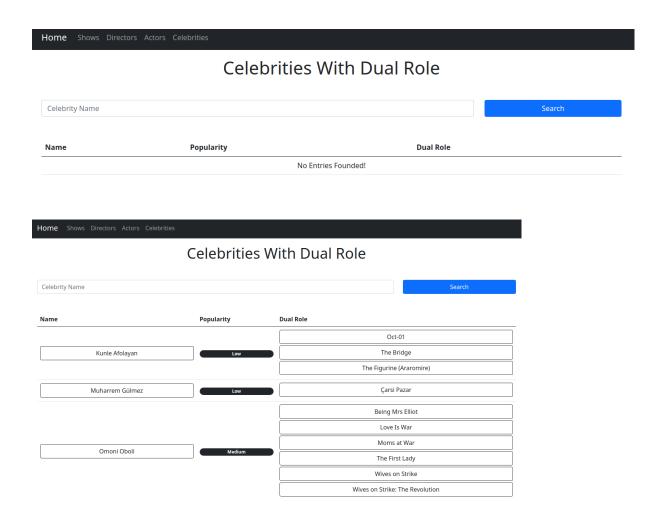
Editar informações relativamente aos filmes/séries existentes na plataforma.



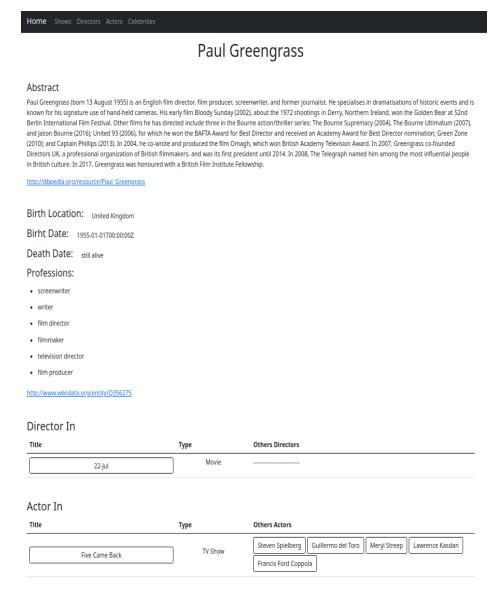
Pesquisar pela lista de atores ou diretores e obter informação sobre quais os filmes onde estes participaram, quer seja a atuar quer seja a dirigir o filme/série.



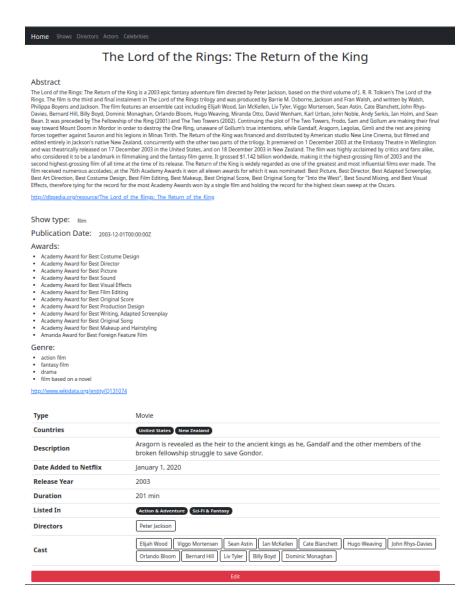
Pesquisar pelas celebridades que são simultaneamente diretor e ator num filme, assim como o quanto popular elas são, dependendo esta popularidade do número de filmes nos quais participaram. Inicialmente vazio mas mostra resultados após executados os comandos das inferências definidas no ficheiro python.



Visualizar informação detalhada acerca de uma pessoa, detalhando os filmes nos quais participou e os filmes em que foi diretor, juntamente com informação (extra) extraída da Wikidata e DBpedia.



Visualizar informações detalhadas de cada filme/série (como por exemplo, o seu tipo, descrição, data que foi adicionado à plataforma, a sua duração, entre outras). Assim como o diretor e atores que participaram neste.



Conclusões

Foi criada uma aplicação onde os utilizadores têm acesso à informação relativa aos filmes e séries assim como a possibilidade de os poderem gerir.

Conseguimos assim, criar um Sistema de Informação baseado em Web que recorre ao uso de inferências (nomeadamente o SPIN) e ontologias para otimizar as queries e encontrar novas relações entre os dados. Apesar de ser possível adicionar novas informações aos nosso dados recorrendo ao uso da Wikidata e DBpedia, devido a como as queries são realizadas e como a informação da wiki está estruturada, podem ocorrer situações em que um ator seja trocado por outro com um nome idêntico ou devido ao fatos de as pesquisas serem somente em inglês, alguns shows podem ser impossíveis de encontrar devido ao seu título se encontrar numa linguagem diferente.

Por último, o uso de Micro-Formatos e RDFa permitiu embutir os dados no código HTML da página, enriquecendo assim a nossa aplicação.

Em suma, aprendemos novas competências e técnicas relacionadas com a área de web semântica que nos permitiram desenvolver com sucesso um sistema de informação baseado na web, usando Python/Django para a programação da aplicação, dados no formato RDF (NT) guardados numa base de dados Triplestore (GraphDB), queries na linguagem SPARQL, uso de RDFS e OWL (para a realização das ontologias), SPIN (para representar o conjunto de inferências), SPARQLwrapper (para o acesso ao endpoint da dbpedia, wikidata e à base de dados) e o RDFa e Micro-formatos (para lidar com a publicação da semântica).

Configuração para executar a aplicação

Criar repositório na graphDB

- Setup > Repositories > Create new repository > GraphDB Free
 - o Repository ID: shows
 - o Ruleset: RDFS

Importar os dados na graphDB

- Import > RDF > Import RDF files
 - data/shows.nt
 - data/showsont.n3
 - Import

Executar aplicação

- \$ pip3 install -r requirements.txt
- \$ python3 app/manage.py runserver

Executar ficheiro de criação de inferências

- \$ pip3 install -r requirements.txt
- \$ python3 inferencias/inferencias.py

URL Repository

http://127.0.0.1:7200/repositories/shows