Eksploracja grafu wiedzy

Marcin Szulc

Semantyczne Przetwarzanie Danych 2019/2020

1 Opis

Projekt zakłada stworzenie aplikacji umożliwiającej interaktywną eksplorację grafu gęsto połączonych danych. Wybraną domeną jest domena filmów i popkultury.

2 Zbiory danych

Koncepcja zakłada wykorzystanie następujących zbiorów:

1. **DBTropes** - `http://skipforward.opendfki.de/wiki/DBTropes`
   Zbiór zawiera informacje z serwisu `https://tvtropes.org/`. Serwis ten zbiera opisy oraz relacje między różnymi dziełami oraz twórcami popkultury ze szczególnym uwzględnieniem konwencji fabularnych. Dane w formacie *N-triples*
2. **Linked Movie Database (Linkedmdb)** - `https://data.world/linked-data/linkedmdb`
   Zbiór zawiera dane o filmach, aktorach oraz reżyserach. Dane w formacie *N-triples*

3 Narzędzia i biblioteki

Projekt składa z aplikacji prezentującej graf (aplikacja działająca w przeglądarce) oraz z bazy danych, przechowującej dane opisane w Sekcji 2.
Wykorzystane biblioteki oraz technologie:

1. **Vue.js** (JavaScript) - aplikacja prezentującą graf oraz zapewniająca iteraktywną eksplorację
   - **vis-network** - prezentacja grafu
   - **bootstrap-vue** - style CSS
   - **axios** - zapytania HTTP
   - **vue-notification** - dynamiczne notyfikacje
   - **n3** - parsowanie plików *.nt*
2. **Apache Jena Fuseki** - baza danych
3. **docker** + **docker-compose** - konteneryzacja całej aplikacji

4 Instalacja

Aplikacja pracuje w wyizolowanych kontenerach - aby ją uruchomić wymagane jest zainstalowane środowisko **docker** oraz **docker-compose**.
Poniższe kroki powinny być wykonane bedąc w głownym katalogu.

1. **Budowanie kontenerów**

   ```
   docker-compose build
   ```

2. **Uruchomienie aplikacji**

   ```
   docker-compose up -d
   ```

   Eksplorator dostępny jest pod adresem `localhost:8080`. Dodatkowo interfejs bazy danych dostępny jest pod adresem `localhost:3030`. Domyślne hasło do konta *admin* (*VyADgCvgP54l0vm*) znajduje się w pliku *db/fuseki/shiro.ini*. Możliwa jest jego zmiana (w pliku) przed uruchomieniem kontenerów.

3. **Wgranie danych**
   Dane powinny zostać sciągnięte z niżej wymienionych źródeł:
   `http://dbtropes.org/static/dbtropes.zip`
   `https://query.data.world/s/n5acuyvz4gmvjguqidnz6jjuhjewdn`

   Pliki *.nt* następnie powinny zostać umieszczone w folderze *input*.
   Następnie z pliku z bazą DBTropes należy usunąć linię numer 18012846, zawierającą błąd uniemożliwiający wgranie danych do bazy:

   ```
   docker-compose exec db sed -i '18012846d' \
       /staging/{nazwa pliku dbtropes}
   ```

   Uruchomienie komend

   ```
   docker-compose exec db ./load.sh dbtropes \
       /staging/{nazwa pliku dbtropes}

   docker-compose exec db ./load.sh linkedmdb \
       /staging/{nazwa pliku linkedmdb}
   ```

   spowoduje wgranie zbiorów danych do bazy.
   Możliwe jest także wgranie ww. zbiorów za pomocą przygotowanego skryptu. Z uwagi jednak na systemowe ograniczenia dot. pamięci kontenerów dockerowych na niektórych dystrybucjach (m.in. na Fedorze), skrypt może niewypakować jednego z archiwów.

   ```
   docker-compose exec db bash /fuseki/populate_db.sh
   ```

   By aplikacja działała poprawnie, wymagane jest dodanie nowych *datasetów* w bazie danych Apache Jena. W tym celu należy zalogować się w panelu administracyjnym (`localhost:3030`), następnie przejść do zakładki *datasets* i stworzyć dwa nowe obiekty, za każdym razem wybierając opcję *persistent*. Datasety powinny nazywać się odpowiednio *dbtropes* oraz *linkedmdb*. Aby dane były widoczne jako stworzone uprzednio *datasety* wymagany jest restart kontenerów:

```
docker-compose restart
```

Działająca aplikacja dostępna jest pod adresem `localhost:8080`.

5 Funkcjonalności

Aplikacja składa się z jednego widoku - grafu oraz przycisków sterujących.
Gdy zaznaczony wierzchołek posiada typ (`http://www.w3.org/1999/02/22-rdf-syntax-ns#type`), możliwe jest zaznaczenie *checkboxa*, znajdującego się z lewej strony ekranu, co spowoduje, że część operacji wykonywana będzie dla wszystkich wierzchołków tego typu.

5.1 Wyświetlanie grafu

Graf wyświetlany jest na środkowej części ekranu. Graf można przesuwać, trzymając lewy przycisk myszki oraz przybliżać/oddalać kręcąc pokrętłem myszy. Wierzchołki, które są "puste" w środku reprezentują literały.

5.2 Aktywna baza

Aplikacja obsługuje dane pochodzące z dwóch źródeł: Linkedmdb oraz DBTropes. W jednym momencie wyświetlane mogą być jedynie dane pochodzące z jednego źródła. Przełącznik wskazujący na aktualnie aktywne źródło znajduje się po prawej stronie ekranu.

5.3 Ilość pobieranych danych

Aby aplikacja była responsywna, ilość ściąganych z bazy danych rekordów jest ograniczona. Domyślna wartość wynosi 20 - można ją zmienic w prawym górnym rogu ekranu.

5.4 Dodawanie nowych wierzchołków

Poczatkowo graf jest pusty, aby dodać nowe wierzchołki (można to zrobić także w innym, dowolnym momencie), należy wpisać interesującą nas frazę w pole tekstowe z prawej stronie ekranu (uwaga: pole jest wrażliwe na wielkie/małe litery), np. "Hobbit". Po wciśnięciu przycisku "Fetch" na ekranie powinny pojawić się wszystkie obiekty, które posiadają "Hobbit" w nazwie (`http://www.w3.org/2000/01/rdf-schema#label`).

5.5 Usuwanie wierzchołków

Po kliknięciu na dowolny wierzchołek, może on zostać usunięty poprzez kliknięcie przycisku "Delete", znajdującego się z lewej strony ekranu. Operację tę można także wykonać dla wszystkich wierzchołków, które posiadają taki sam typ jak zaznaczony wierzchołek.
Obydwie operacje usuwają wierzchołki jedynie z wizualizacji, dane dalej pozostają obecne w bazie danych.

5.6 Wyświetlanie właściwości

Po zaznaczeniu dowolnego wierzchołka, z lewej strony ekranu wyświetlana jest lista właściwości (properties), za pomocą których jest on połączony z innymi obiektami. Kliknięcie na konkretną właściwość powoduje dołączenie do grafu wszystkich obiektów, z którymi połączony jest zaznaczony wierzchołek za pomocą wybranej właściwości. Operację tę, można także wykonać dla wszystkich wierzchołków, które posiadają taki sam typ jak zaznaczony wierzchołek.

5.7 Eksportowanie i importowanie grafu

Aktualnie wyświetlany graf może zostać wyeksportowany do pliku w formacie *.nt*.
Uwaga: gdy graf zawiera wierzchołki, które nie są połączone żadną krawędzią, w wynikowym pliku reprezentowane są one jako trójki postaci:

```
{wierzcholek} http://www.w3.org/1999/02/22-rdf-syntax-ns#type _:x .
```

Analogicznie, dowolny plik *.nt* może zostać zaimportowany do aplikacji z dysku.

5.8 Znajdywanie najkrótszej ścieżki

Aplikacja pozwala na znalezienie najkrótszej ścieżki od jednego wierzchołka do drugiego. W tym celu należy zaznaczyć wierzchołek na grafie oraz kliknąc jeden z przycisków "Select as first node in path", analogicznie dla drugiego wierzchołka. Po wybraniu wierzchołków, kliknięcie przycisku "Find shortest path" spowoduje próbę znalezienia najkrótszej ścieżkę. Ścieżka poszukiwana jest wśród danych w bazie danych - nie na aktualnie wyświetlanym grafie. Gdy znaleziona ścieżka posiada wierzchołki nieobecne na aktualnym grafie, wierzchołki te zostaną do niego dodane. Aplikacja poszukuje jedynie ścieżek o maksymalnej długości 10. Gdy najkrótszych ścieżek jest więcej niż jedna, zwrócona zostaje jedna oraz wyświetlony stosowny komunikat.
Uwaga: zapytania ignorują ścieżki zawierające właściwość *http://www.w3.org/1999/02/22-rdf-syntax-nstype*. Gdy właściwośc ta jest uwzględniana, zwracane są mało interesujące ścieżki, np. przechodzące przez wierzchołek "typu" film (każde dwa filmy łączy wierzchołek mowiący o tym, że należa one do typu film).

6 Opis wejść i wyjść

Wejścia:

- załadowanie nowej bazy danych: pliki w formacie *N-triples*
- załadowanie nowego grafu do wyświetlenia w aplikacji: plik w formacie *N-triples*

Wyjścia:

- zapisanie aktualnie wyświetlanego na ekranie grafu: plik w formacie *N-triples*