Webpage Improvement for TV Series Data

Dokumentacja Projektu

Przygotowanego na przedmiot

Semantyczne Przetwarzanie Danych

Aleksandra Bułka

Aleksandra Byczyńska

28 stycznia 2020

1 Wstęp

W obecnych czasach, seriale telewizyjne i internetowe cieszą się taką samą popularnością jak i filmy. Popularny portal, *www.imdb.com* zawiera informacje o ponad 6.5 milionach filmów i seriali.

W ramach opisanego w tym dokumencie projektu, za cel postawiłyśmy sobie uczynienie serwisu *www.imdb.com* bardziej przyjaznym zwolennikom seriali. Wybrane podstrony serwisu poświęcone serialom zostały wzbogacone o mikrodane i dodatkowe, interesujące dla fanów informacje.

2 Założenia projektu

Celem projektu było wzbogacenie treści stron znanego portalu *https://www.imdb.com/*, poprzez dodanie informacji pochodzących z innych źródeł danych oraz dodanie mikrodanych. Wykonane zostały następujące kroki:

1. Pobrany został kod źródłowy stron i wyodrębnione zostały z niego pewne identyfikatory semantyczne.
2. Znalezione w poprzednim kroku identyfikatory semantyczne zostały podlinkowane do odpowiednich jednostek HTML i umieszczone z powrotem w kodzie strony jako mikrodane.
3. Dodatkowo, dla odnalezionych identyfikatorów, zebrane zostały dodatkowe informacje z niezależnych źródeł danych (DBPedii, DBTropes i WikiData). Informacje te zostały umieszczone w kodzie strony jako widoczne elementy HTML.

3 Zakres projektu

3.1 Dodane dane

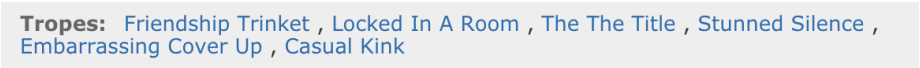
3.1.1 Kanał

Informacja o kanale, przez którego produkowany jest serial została uzyskana poprzez odpytanie bazy wiedzy DBPedia odpowiednim zapytaniem SPARQL. Jako identyfikator semantyczny posłużył w tym przypadku tytuł serialu. Element umieszczony w kodzie strony jest jednocześnie linkiem do odpowiedniej jednostki w DBPedii.

Rysunek 1: Dodana informacja o kanale, przez który dany serial jest produkowany. Tutaj: dla serialu "Przyjaciele", kanałem (network) jest NBC.

3.1.2 Motywy

Dodane również została informacja o motywach występujących w danym serialu. Została ona uzyskana poprzez odpytanie bazy wiedzy DBTropes. Odpowiednie zapytanie zostało napisane przy użyciu biblioteki rdflib. Jako identyfikator semantyczny posłużył znowu przypadku tytuł serialu. Wylistowane zostało kilka motywów, za każdym z nich kryje się link do odpowiedniej jednostki w DBTropes.



Rysunek 2: Dodana informacja o motywach obecnych w serialu. W widocznym przykładzie, obecne w serialu "Przyjaciele" motywy to np. Locked in a Room (zamknięci w pokoju).

3.1.3 Obsada

Najwięcej informacji dodawanych jest o obsadzie serialu. Dane o występujących w serialu aktorach i aktorkach uzyskane zostały poprzez odpytanie DBPedii, posługując się tytułem serialu jako identyfikatorem semantycznym. Odpytując bazę wiedzy WikiData uzyskano następnie datę urodzenia danego aktora/aktorki oraz URL do jego/jej konta na różnych serwisach społecznościowych, takich jak Instagram, Twitter lub Facebook. Umieszczone zostały one na stronie jako linki.

Actor	Date of birth	Handles
Jennifer Aniston	1969-02-11	RottenTomatoes Instagram Facebook
Courteney Cox	1964-06-15	Instagram Twitter
Lisa Kudrow	1963-07-30	Instagram Twitter Facebook
Matt LeBlanc	1967-07-25	RottenTomatoes Instagram Twitter Facebook
Matthew Perry	1969-08-19	Instagram Twitter
David Schwimmer	1966-11-02	Instagram Twitter
James Michael Tyler	1962-05-28	
Maggie Wheeler	1961-08-07	
Christina Pickles	1935-02-17	
Elliott Gould	1938-08-29	
Swati Anand		

Rysunek 3: Dodana informacja o aktorach i aktorkach grających w serialu. Widoczna na przykładzie jest obsada serialu "Przyjaciele", wraz z datami urodzenia i linkami do kont w serwisach społecznościowych

3.2 Dodane mikrodane

Do pobranego kodu źródłowego strony, dodane zostały dodatkowe mikrodane. Mikrodane to uzupełnienie HTML, schematy danych strukturalnych, które wzbogacają semantykę danych widocznych na stronie internetowej.

W tym projekcie jako mikrodane dodały zostane dane o aktorach i aktorkach występujących w serialu, a dokładniej - daty ich urodzenia. Dane te zostały usyskane w podobny sposów jak w poprzednim kroku - po odpytaniu bazy wiedzy WikiData. Następnie dodano je do JSONa, w którym znajdowały się już inne mikrodane tej osoby: jej "@type" czyli "Person", jej url i imię.

Rysunek 4: Mikrodane aktorki Jennifer Anniston przed dodaniem jej daty urodzenia

```
"actor": [
    {
        "@type": "Person",
        "url": "/name/nm0000098/",
        "name": "Jennifer Aniston",
        "birthDate": "1969-02-11T00:00:00Z"
    },
```


Rysunek 5: Mikrodane aktorki Jennifer Anniston po dodaniu jej daty urodzenia

4 Źródła danych

Jako niezależne źródła danych, posłużyły nam następujące bazy wiedzy:

4.1 DBPedia

DBPedia to jeden z najbardziej znanych części projektu Semantic Web. Jest projekt mający na celu usystematyzowanie i powiązanie ze sobą danych z Wikipedii, a następnie udostępnienie tych zorganizowanych informacji w Internecie. DBpedia umożliwia tworzenie zaawansowanych kwerend relacji i właściwości do zasobów Wikipedii, w tym linków do innych zbiorów danych.

Z DBPedii można uzyskać informacje wysyłając zapytanie na udostępnioną przez projekt końcówkę *http://dbpedia.org/sparql*.

4.2 WikiData

WikiData, a po polsku angielskim Wikidane jest to projekt internetowy mający na celu stworzenie wolnej, otwartej, wielojęzycznej bazy różnorodnych danych. Głównym zastosowaniem tej bazy danych jest używanie jej w projektach Wikimedia Foundation, czyli przede wszystkim w Wikipedii.

Podobnie jak w przypadku DBPedii, do WikiData można wysłać zapytanie SPARQLowe na udostępniony endpoint *https://query.wikidata.org/sparql*.

4.3 DBTropes

DBTropes to wersja TVTropes w postaci Linked Data. Natomiast samo TV Tropes (Television Tropes and Idioms) to wiki, która opisuje rozmaite motywy i konwencje oraz opisuje pod ich kątem dzieła różnych twórców. Z początku na stronie opisywano (zgodnie z nazwą) motywy telewizyjne i filmowe, ale z czasem strona objęła także literaturę, komiks, gry komputerowe, a nawet reklamy i zabawki. Jest znana z podchodzenia do tematów w lekki i zabawny sposób.

W przeciwieństwie do powyższych źródeł danych, żeby przeszukiwać DBTropes, należy pobrać zrzut tej bazy i lokalnie wysyłać zapytania. Dodatkowo mankamentem jest fakt, iż ostani zrzut tej bazy pochodzi z 2016 roku.

5 Użyte technologie

Program napisany został w języku programowania Python 3.8. Zapytania użyte do przeszukiwania baz wiedzy napisane zostały w SPARQL.

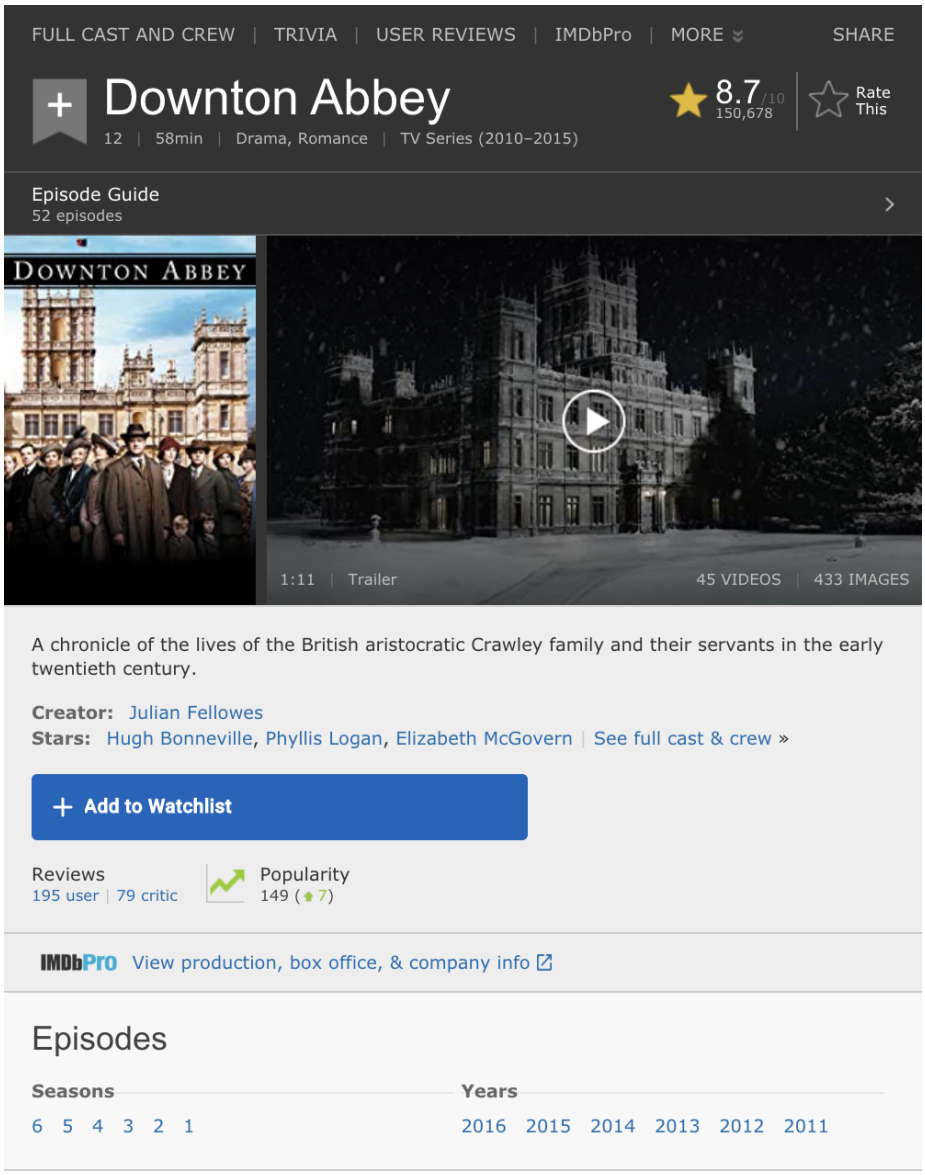
Dodatkowo, wykorzystane zostały biblioteki: *Beautiful Soup* do pobierania kodu źródłowego podstron, *requests* do wysyłania zapytań do zewnętrznego API.

Do pracy z DBTropes, została wykorzystana biblioteka *rdflib*. Był to optymalny sposób pracy z trójkową bazą danych, której kopie posiada się lokalnie.
Do wysyłania zapytań do zewnętrznego API (jak w przypadku WikiData i DBTropes) przydała się biblioteka *SPARQLWrapper*

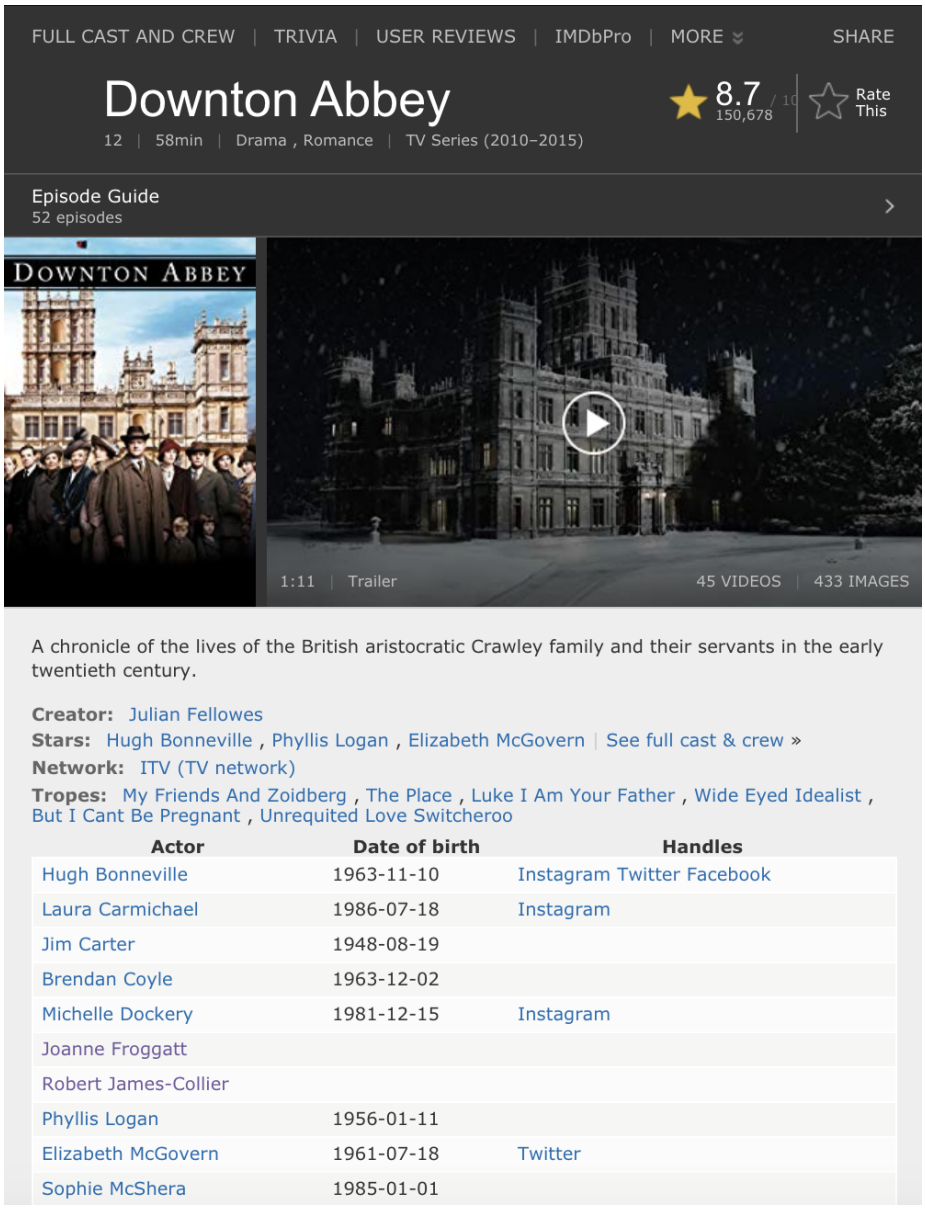
6 Eksperymenty

Zostały wykonane eksperymenty, polegające na wykonywaniu ulepszeń - czyli dodawaniu informacji o obsadzie, kanale i motywach w serialu oraz dodawanie mikrodanych dla wybranych podstron portalu *https://www.imdb.com/*.

Poniżej przykład jak wyglądało to dla jednej z ulepszonych podstron
www.imdb.com/title/tt1606375/ czyli stronie serialu "Downton Abbey".



Rysunek 6: Strona serialu "Downton Abbey" przed wzbogaceniem jej o dodatkowe informacje.



Rysunek 7: Strona serialu "Downton Abbey" wraz z dodatkowymi informacjami.

7 Wnioski

Projekt zakończył się sukcesem. Udało się przetworzyć wiele podstron portalu *https://www.imdb.com/* i wzbogacić je o dane seriali. Podstrony udało się również uzupełnić dodatkowymi mikrodanymi.

Niewątpliwie jednym z ważniejszych lekcji wyniesionych z realizacji tegoż projektu jest poszerzenie wiedzy o źródłach danych w postaci Linked Data. Bardziej niszowym z nich (takim jak DBTropes) brakuje aktualnych danych oraz ich zrzuty nie są robione regularnie. Dużo lepiej wypada w porównaniu duża baza wiedzy DBPedia, lecz jej także wiele brakuje do kompletności.