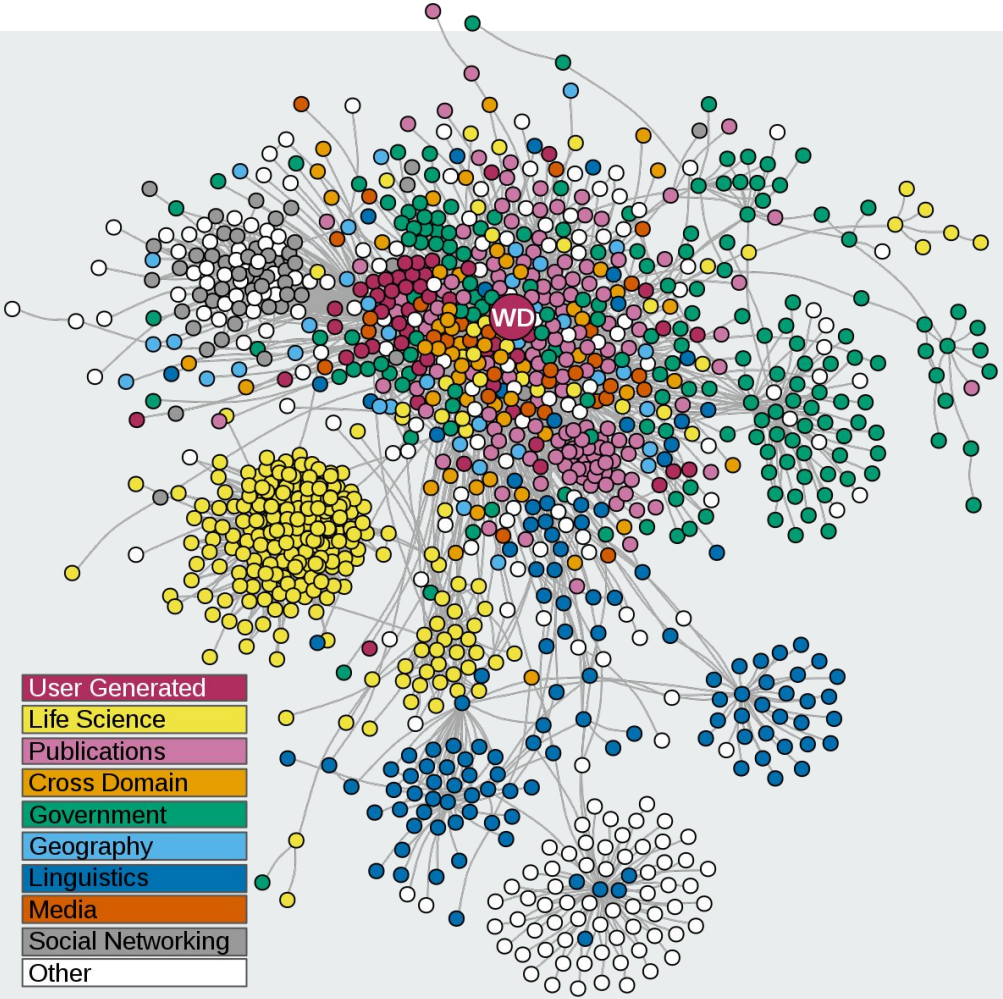
Linked Open Data [LOD]

Data Science (Additional Topics) Project Presentation
Topic : SEMI

Group 126 (Additional Topics):

Ajay Mathew Joseph (SEMI & TS) (s2435462)
Nishchit Mahajan (SEMI & IENLP) (s2466104)

Linked Data


WD
User Generated
Life Science
Publications
Cross Domain
Government
Geography
Linguistics
Media
Social Networking
Other

OBJECTIVE

- Retrieving non-trivial data related to Movies from DBPedia
- To demonstrate the added value of semantic web technology by running remote SPARQL queries to query Linked Open Data on the web and retrieving data in a way not possible with Google search and thus enriching the original dataset.

Technologies Used

- Python
- Pandas
- DBPedia Spotlight API
- Rdflib
- SPARQLWrapper

Methodology

1. Retrieve the movie dataset
2. Link each item with its corresponding DBPedia links using DBPedia Spotlight API
3. Construct an RDF database from the dataset using rdflib
4. Run five complex queries using SPARQLWrapper

Construction and linking of RDF dataset

- Obtained from Kaggle: link
- Data about 45,466 movies
- Subset of 10,000 movies taken
- Using DBPedia Spotlight API, link each movie title with it's DBPedia link
- Using rdflib, construct an RDF database with the movies' name, release date and DBPedia link

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
    xmlns:movie="https://www.imdb.com/"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
>
  <rdf:Description rdf:about="#@tt0061581">
    <movie:title>Divorce American Style</movie:title>
    <movie:releaseDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1967-06-21</movie:releaseDate>
    <owl:sameAs rdf:resource="http://dbpedia.org/resource/Divorce_American_Style"/>
  </rdf:Description>
  <rdf:Description rdf:about="#@tt0110913">
    <movie:title>Pumpkinhead II: Blood Wings</movie:title>
    <movie:releaseDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1994-03-16</movie:releaseDate>
    <owl:sameAs rdf:resource="http://dbpedia.org/resource/Jack_Pumpkinhead"/>
  </rdf:Description>
```

(Results) Example SPARQL queries

1. Get other movies related to a particular movie from it's Wikipedia page.
   a. Using dbo:wikiPageWikiLink (link from a wiki page to another wiki page)
   b.

```
    MENU
    1. Get related movies
    2. Get average budget of other movies of a particular movie's director
    3. Get other movies where the coactors of a particular movie acted together
    4. Get youngest main crew member of a movie
    5. Get the longest movie of a movie's top actor

Enter your choice: 1

Enter the movie name:The Machinist

Other movies related to The Machinist are:

Kirill Lavrov
The Brothers Karamazov
A Beautiful Mind
Insomnia
Batman Begins
```

2. Find the average budget of all movies of a particular movie's director.

Input: A movie's title

Output: The average budget of all movies of the input movie's director

```
    MENU
    1. Get related movies
    2. Get average budget of all movies of a particular movie's director
    3. Get other movies where the coactors of a particular movie acted together
    4. Get youngest main crew member of a movie
    5. Get the longest movie of a movie's top actor

Enter your choice: 2

Enter the movie name:Before Sunrise
The average budget of Richard_Linklater's movies is: Dollars 9943937.5
```

3. Find other movies where a particular movie's actors acted together

Input: A movie's title

Output: All movies where the main two actors acted together

```
    MENU
    1. Get related movies
    2. Get average budget of all movies of a particular movie's director
    3. Get other movies where the coactors of a particular movie acted together
    4. Get youngest main crew member of a movie
    5. Get the longest movie of a movie's top actor

Enter your choice: 3

Enter the movie name:The Matrix

The top 2 actors are:
Keanu Reeves
Hugo Weaving

The movies where they acted together are:
The Matrix Reloaded
The Matrix Revolutions
The Matrix
```

4. Find the youngest main crew member of a particular movie

Input: A movie's title

Output: The youngest main crew member and his/her Date of Birth

```
    MENU
    1. Get related movies
    2. Get average budget of all movies of a particular movie's director
    3. Get other movies where the coactors of a particular movie acted together
    4. Get youngest main crew member of a movie
    5. Get the longest movie of a movie's top actor

Enter your choice: 4

Enter the movie name:The Matrix
The youngest main crew member is Keanu Reeves born on 1964-09-02
```

5. Find the longest movie of a particular movie's main actor

Input: A movie's title

Output: The longest movie of the movie's lead actor

```
    MENU
    1. Get related movies
    2. Get average budget of all movies of a particular movie's director
    3. Get other movies where the coactors of a particular movie acted together
    4. Get youngest main crew member of a movie
    5. Get the longest movie of a movie's top actor

Enter your choice: 5

Enter the movie name:Jumanji
The longest movie of Robin Williams is Hamlet running 4.033333333333333 hours
```

Conclusions

- The added value of semantic web technology was demonstrated by linking a dataset with Linked Open Data
- Non-trivial queries which cannot be answered by general Google search could be answered with Linked Open Data
- Linked Open Data could be used in developing recommender systems (**get_related_movies()**)

Limitations

- Details can only be queried with a movie's name (and not actor's /director's name)
- Some movies with ambiguous names were linked incorrectly by the Spotlight API (eg: JFK)
- Some DBPedia entries have missing, redundant and inconsistent keys

Source Code

SPARQL LOD MOVIES

THANK YOU!