IRE Assignment 4

Wikidata

Wikidata is a free and open knowledge base that can be read and edited by both humans and machines.It is hosted by Wikimedia Foundation and provides support to many other sites and services beyond just Wikimedia projects.It is a Knowledge Graph that is collaboratively developed and edited to collect structed-data for Wikipedia.

Features:

- **Collaborative -** Data is entered and maintained by Wikidata editors and automated bots.
- **Multilingual -** Editing, consuming, browsing, and reusing the data is fully multilingual. Data entered in any language is immediately available in all other languages.
- **Structured data** - Having Structured organization of data it provides easy reuse of data by wikimedia projects, third parties and enables computers to process and understand it.

Structure of data:

The Wikidata repository consists mainly of items, each one having a label, a description and any number of aliases. Items are uniquely identified by a Q followed by a number, such as 'Sachin Tendulkar' (Q9488).

Statements describe detailed characteristics of an Item and consist of a property and a value. Properties in Wikidata have a P followed by a number, such as with 'instance of'(P31).

Each statement can be thought of a triplet (item -> property = value)

Eg: 1) Sachin Tendulkar -> instance of = human

(Q9488) (P31) (Q5)

2) Charminar -> location = Hyderabad

(Q1771696) (P276) (Q1361)

The value is the value of property and it can also be an item with Q_identifier as in the above two examples.It can also be just a value as shown below.

Jawaharlal Nehru -> date of birth = 14 November 1889

(Q1047) (P569)

Linked Data:
Wikidata also supports linked data as the structured data can be interlinked.
Wikidata uses unique identifiers or URIs for all its items.Wikidata uses a unique data model,but its content can also be exported in RDF(Resource Description Framework) which is a standard format for linked data.

We can relate Wikidata statements with linked data triplets as:

Item -> property = value
(Subject) (predicate) (object)

But Wikidata statementing RDF.s using also have elements beyond the triplet like references, so cannot fully represent Wikidata using RDF.

Ways to access Wikidata:
Wikidata has several built-in tools, external tools, or programming interfaces to access its data.
Popular among them are:
-> Wikidata Query Service (SPARQL)
-> Reasonator
-> Wikidata API

SPARQL

SPARQL (SPARQL Protocol and RDF Query Language) is an RDF query language for dealing with linked data.
It semantic query language and a standard to retrieve and manipulate data stored in RDF format.

- SPARQL allows users to write queries against what can loosely be called "key-value" data or, more specifically, data in RDF. So, database is a set of "subject-predicate-object" triples.
- Unlike SQL, RDF can have multiple entries per predicate.
- SPARQL provides a full set of analytic query operations such as JOIN, SORT, AGGREGATE for data whose schema is intrinsically part of the data rather than requiring a separate schema definition.However, schema information (the ontology) is often provided externally, to allow joining of different datasets unambiguously.
- SPARQL contains capabilities for querying required and optional graph patterns along with their conjunctions and disjunctions.

Query forms
SPARQL language specifies four different query variations for different purposes:

➔ SELECT query
Used to extract values from a SPARQL endpoint, the results are returned in a table format.

➔ CONSTRUCT query
Used to extract information from the SPARQL endpoint and transform the results into valid RDF.

- ➔ ASK query
  Used to provide a simple True/False result for a query on a SPARQL endpoint.
- ➔ DESCRIBE query
  Used to extract an RDF graph from the SPARQL endpoint.

Each of these query forms takes a WHERE block to restrict the query, although, in the case of the DESCRIBE query, the WHERE is optional.

Syntax

A SPARQL query comprises, in order:

I. Prefix declarations, for abbreviating URIs
II. Result clause, identifying what information to return from the query
III. The query pattern, specifying what to query for in the underlying dataset
IV. Query modifiers, slicing, ordering, and otherwise rearranging query results

Syntax Structure described above:

```
# prefix declarations
PREFIX foo: <http://example.com/resources/>
...
# result clause
SELECT ...
# query pattern
WHERE {
    ...
}
# query modifiers
ORDER BY …
```

Each query pattern is a triple pattern.Triple patterns are just like triples, except that any of the parts of a triple can be replaced with a variable.

Example

```
PREFIX vcard:      <http://www.w3.org/2001/vcard-rdf/3.0#>
SELECT ?givenName
WHERE
 {
   ?y vcard:Given  ?givenName .
 }
```

Here vcard is a URI which is defined using PREFIX and used.
SPARQL variables start with a ? and can match any node (resource or literal) in the RDF dataset.Each query involves a triple pattern, each triple ends in a '.' .
The SELECT result clause returns a table of variables and values that satisfy the query.

SPARQL for Wikidata

As Wikidata supports RDF , we can use the standardized SPARQL for querying Wikidata which is called Wikidata Query Service(WDQS)

Structure of SPARQL statements is similar but we have predefined prefixes in WDQS .The following are some prefixes:

Items are prefixed with wd: and properties with wdt: . Similarly, we p: for property , ps: for statement and many others.

Example

Query to find all the States in India in Wikidata.

```
SELECT ?state ?stateLabel WHERE {
  ?state wdt:P31 wd:Q13390680.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "te","en". }
}
```

The statement " ?state wdt:P31 wd:Q13390680. " collects all items which have the property 'instance of'(P31) with object 'state of india' (Q13390680) into the variable state.

The code snippet with 'SERVICE' is responsible for retrieving labels for the collected items into an additional variable with Label postfix in the specified language(here stateLabel).We can change the languge in 'wikibase:language' and obtain labels in required language.We can also order the languages so that if any label is not present in a language ,it will be shown in the language given next and so on.Here I have choosen languages telugu followed by english.

LSJBOT

Lsjbot is an automated Wikipedia article-creating program, or Internet bot, developed by Sverker Johansson for the Swedish Wikipedia. The bot primarily focuses on articles about living organisms and geographical entities (such as rivers, dams and mountains).

According to its description page on the Swedish Wikipedia, Lsjbot was active in the Swedish Wikipedia and in the Cebuano and Waray Wikipedias, and has created most Wikipedia articles in those languages.It is said that it can write about 10,000 articles per day.

Wikipedia Page Generated

Example 1
International Institute of Information Technology, Hyderabad

International Institute of Information Technology, Hyderabad ఒక ఇన్స్టిట్యూట్ ఆఫ్ టెక్నాలజీ.ఇది హైదరాబాద్,భారత దేశం లో ఉంది. దీనికి డా.రాజ్ రెడ్డి అధ్యక్షుడు. ఇది 1998-01-01 లో ప్రారంభం అయింది.

Example2
Jawaharlal Nehru Technological University, Hyderabad

Jawaharlal Nehru Technological University, Hyderabad ఒక విశ్వవిద్యాలయం.ఇది కూకట్ పల్లి,భారత దేశం లో ఉంది. ఇది 1965-01-01 లో ప్రారంభం అయింది.

Method:
I have used wikidata api using php request to obtain Q id number of the item.
Using the Q id I formed the query .I have formed the query with fields related to colleges.
Using the json requests I got the response for the query.I composed the text for the page using the properties obtained in response and using the words in telugu to form sentences.
For now, I have hardcoded the words of telugu needed.But as an extension I can search if the word is present in wikidata otherwise use hardcoded one.But some words hardcoded can be used repetatively and can make some sentences for which we know the format.

Command to run:
python3 main.py