REPORT
IRE ASSIGNMENT – 2
RISHABH MALIK (2020201074)

SPARQL

SPARQL, pronounced 'sparkle', is the standard query language and protocol for Linked Open Data on the web or for RDF triplestores. SPARQL, short for "SPARQL Protocol and RDF Query Language", enables users to query information from databases or any data source that can be mapped to RDF. Just like SQL allows users to retrieve and modify data in a relational database, SPARQL provides the same functionality for NoSQL graph databases like Ontotext's GraphDB, wikidata.

SPARQL has four types of queries. It can be used to:

1. ASK whether there is at least one match of the query pattern in the RDF graph data;
2. SELECT all or some of those matches in a tabular form (including aggregation, sampling and pagination through OFFSET and LIMIT);
3. CONSTRUCT an RDF graph by substituting the variables in these matches in a set of triple templates;
4. DESCRIBE the matches found by constructing a relevant RDF graph.

WIKIDATA

Wikidata is a collaboratively edited multilingual knowledge graph hosted by the Wikimedia Foundation. It is a common source of open data that Wikimedia projects such as Wikipedia, and anyone else, can use under the CC0 public domain license. Wikidata is a wiki powered by the software MediaWiki, and is also powered by the set of knowledge graph MediaWiki extensions known as Wikibase.

APPROACH USED:

Link to sandbox : https://en.wikipedia.org/wiki/User:Rishabh_malik_007/sandbox
I have used some samples queries about world geography like population and capital of most and least populated countries and cities in the world and the common languages spoken in Asian countries.

- To get data in Hindi language I assigned "hi" parameter to language option in the query as shown below:

```
SERVICE wikibase:label { bd:serviceParam wikibase:language "hi" }
```

- This is a sample data of triplets that I received for **country – capital – country population query** from wikidata.

{'head': {'vars': ['**countryLabel', 'capitalLabel', 'population**']},

'results': {'bindings': [{**'capitalLabel**': {'type': 'literal',

'value': **'बीजिंग'**, 'xml:lang': 'hi'},

'countryLabel': {'type': 'literal', 'value': **'चीनी जनवादी गणराज्य'**, 'xml:lang': 'hi'},

'population': {'datatype': 'http://www.w3.org/2001/XMLSchema#decimal',

'type': 'literal', **'value**': **'1443497378**'}},

{**'capitalLabel**': {'type': 'literal', **'value**': **'नई दिल्ली'**, 'xml:lang': 'hi'},

'countryLabel': {'type': 'literal', **'value**': **'भारत'**, 'xml:lang': 'hi'},

'population': {'datatype': 'http://www.w3.org/2001/XMLSchema#decimal',

'type': 'literal', **'value**': **'1326093247**'}},

{**'capitalLabel**': {'type': 'literal', **'value**': **'वॉशिंगटन डी॰ सी॰'**, 'xml:lang': 'hi'},

'countryLabel': {'type': 'literal', **'value**': **'संयुक्त राज्य अमेरिका'**, 'xml:lang': 'hi'},

'population': {'datatype': 'http://www.w3.org/2001/XMLSchema#decimal',

'type': 'literal', **'value**': **'331449281**'}},

.....

.....

}

- Converted the received data of triplets in a tabular form as shown below :

COUNTRY	CAPITAL	COUNTRY POPULATION
चीनी जनवादी गणराज्य	बीजिंग	1443497378
भारत	नई दिल्ली	1326093247
संयुक्त राज्य अमेरिका	वॉशिंगटन डी॰ सी॰	331449281
इंडोनेशिया	जकार्ता	270625568
पाकिस्तान	इस्लामाबाद	216565318
ब्राज़ील	ब्रासीलिया	213317639
नाईजीरिया	अबुजा	190886311
बांग्लादेश	ढाका	165775000
रूस	मास्को	146804372
जापान	टोक्यो	126434565

- Explored rule based methods. What that means is some if-else conditions based on manually seeing the data. So, observing some patterns with triplets, and formulating some rules that could cover a lot of cases.
- As you can see, this is sort of table to text conversion
- To convert the above table to text I have done following :
  - **INPUT** : भारत नई दिल्ली 1326093247
  - **OUTPUT** : भारत की आबादी लगभग 1326093247 है और इसकी राजधानी नई दिल्ली है।
  - भारत + " की आबादी लगभग " + 1326093247 + " है और इसकी राजधानी " + नई दिल्ली " + "है।"
  - Similarly, sentences of other data is made in such a manner

- **Output text on of above table will look like this -**

दुनिया में कुछ सबसे बड़ी आबादी वाले देश हैं चीनी जनवादी गणराज्य, भारत, संयुक्त राज्य अमेरिका, इंडोनेशिया, पाकिस्तान, ब्राज़ील, नाईजीरिया, बांग्लादेश, रूस और जापान। चीनी जनवादी गणराज्य की आबादी लगभग 1443497378 है और इसकी राजधानी बीजिंग है। भारत की आबादी लगभग 1326093247 है और इसकी राजधानी नई दिल्ली है। संयुक्त राज्य अमेरिका की आबादी लगभग 331449281 है और इसकी राजधानी वॉशिंगटन डी॰ सी॰ है। इंडोनेशिया की आबादी लगभग 270625568 है और इसकी राजधानी जकार्ता है। किस्तान की आबादी लगभग 216565318 है और इसकी राजधानी इस्लामाबाद है। ब्राज़ील की आबादी लगभग 213317639 है और इसकी राजधानी ब्रासीलिया है। नाईजीरिया की आबादी लगभग 190886311 है और इसकी राजधानी अबुजा है। बांग्लादेश की आबादी लगभग 165775000 है और इसकी राजधानी ढाका है। रूस की आबादी लगभग 146804372 है और इसकी राजधानी मास्को है। जापान की आबादी लगभग 126434565 है और इसकी राजधानी टोक्यो है।

❖ Similarly we have other data also and we have applied similar rule based sentence forming for them.

List of 10 MOST populated CITIES

CITY	POPULATION
शंघाई	23390000
बीजिंग	21710000
ढाका	16800000
मुम्बई	15414288
कराची	14910352
लागोस	14862000
इस्तांबुल	14657434
टोक्यो	14049146
तिआंजिन	13245000

- To convert the above table to text I have done following :
  - **INPUT** : शंघाई 23390000
  - **OUTPUT** : शंघाई की आबादी लगभग 23390000 है।
  - शंघाई + " की आबादी लगभग " + 23390000 + "है।"

दुनिया के कुछ सबसे बड़े आबादी वाले शहर हैं शंघाई, बीजिंग, ढाका, मुम्बई, कराची, लागोस, इस्तांबुल, टोक्यो और तिआंजिन। शंघाई की आबादी लगभग 23390000 है। बीजिंग की आबादी लगभग 21710000 है। ढाका की आबादी लगभग 16800000 है। मुम्बई की आबादी लगभग 15414288 है। कराची की आबादी लगभग 14910352 है। लागोस की आबादी लगभग 14862000 है। इस्तांबुल की आबादी लगभग 14657434 है। टोक्यो की आबादी लगभग 14049146 है। तिआंजिन की आबादी लगभग 13245000 है।

List of 10 LEAST populated countries and their capital(s)

COUNTRY	CAPITAL	COUNTRY POPULATION
तुवालू	फुनाफुति	11192
नौरु	यारेन जिला	13650
पलाउ	ङुरूलमुड	21729
सान मारिनो	सैन मारिनो नगर	33400
लिक्टेन्स्टाइन	वादुज़	37922
मोनैको	Q55115	38695
मार्शल द्वीपसमूह	माजुरो	53127
सन्त किट्स और नेविस	बासेटेयर	55345
डोमिनिका	रोसीयू	73925
अण्डोरा	अण्डोरा ला वेला	78151

दुनिया में कुछ सबसे कम आबादी वाले देश हैं तुवालू, नौरु, पलाउ, सान मारिनो, लिक्टेन्स्टाइन, मोनैको, मार्शल द्वीपसमूह, सन्त किट्स और नेविस, डोमिनिका और अण्डोरा। तुवालू की आबादी लगभग 11192 है और इसकी राजधानी फुनाफुति है। नौरु की आबादी लगभग 13650 है और इसकी राजधानी यारेन जिला है। पलाउ की आबादी लगभग 21729 है और इसकी राजधानी ङुरूलमुड है। सान मारिनो की आबादी लगभग 33400 है और इसकी राजधानी सैन मारिनो नगर है। लिक्टेन्स्टाइन की आबादी लगभग 37922 है और इसकी राजधानी वादुज़ है। मोनैको की आबादी लगभग 38695 है और इसकी राजधानी Q55115 है। मार्शल द्वीपसमूह की आबादी लगभग 53127 है और इसकी राजधानी माजुरो है। सन्त किट्स और नेविस की आबादी लगभग 55345 है और इसकी राजधानी बासेटेयर है। डोमिनिका की आबादी लगभग 73925 है और इसकी राजधानी रोसीयू है। अण्डोरा की आबादी लगभग 78151 है और इसकी राजधानी अण्डोरा ला वेला है।

Names of Asian countries and common languages used there


```
{'अफ़्गानिस्तान': ['अरबी भाषा', 'उज़्बेक भाषा', 'तुर्कमेन भाषा', 'पश्तो भाषा'],
 'इंडोनेशिया': ['इंडोनेशियाई भाषा', 'जावा भाषा'],
 'इज़राइल': ['अरबी भाषा', 'इब्रानी भाषा'],
 'इराक': ['अरबी भाषा', 'कुर्दी भाषा'],
 'ईरान': ['फ़ारसी भाषा'],
 'उज़्बेकिस्तान': ['उज़्बेक भाषा'],
 'ओमान': ['अरबी भाषा'],
 'कज़ाख़िस्तान': ['कज़ाख़ भाषा', 'रूसी भाषा'],
 'कम्बोडिया': ['खमेर भाषा'],
 'क़तर': ['अरबी भाषा'],
 'किर्गिज़स्तान': ['किर्गिज़ भाषा', 'रूसी भाषा'],
 'चीनी गणराज्य': ['होक्का भाषा'],
 'जापान': ['जापानी भाषा'],
 'जॉर्डन': ['अरबी भाषा'],
 'ताजिकिस्तान': ['ताजिकी भाषा', 'रूसी भाषा'],
 'तुर्कमेनिस्तान': ['तुर्कमेन भाषा'],
 'तुर्की': ['तुर्कीयाई भाषा'],
 'थाईलैण्ड': ['थाई भाषा'],
 'दक्षिण कोरिया': ['कोरियाई भाषा'],
```


- The above image is a small potion of entire data received.

जापान में लोग जापानी भाषा बोलते हैं । साइप्रस में लोग आधुनिक, तुर्कीयाई और यूनानी बोलते हैं । कज़ाख़िस्तान में लोग कज़ाख़ और रूसी बोलते हैं । इंडोनेशिया में लोग इंडोनेशियाई और जावा बोलते हैं । तुर्की में लोग तुर्कीयाई भाषा बोलते हैं । मिस्र में लोग अरबी भाषा बोलते हैं । पहलवी वंश में लोग फ़ारसी भाषा बोलते हैं । नागोरो-कराबाख गणराज्य में लोग आर्मीनियाई और रूसी बोलते हैं । उज़्बेकिस्तान में लोग उज़्बेक भाषा बोलते हैं । सिंगापुर में लोग अंग्रेज़ी, तमिल और मलय बोलते हैं । बहरीन में लोग अरबी भाषा बोलते हैं । कम्बोडिया में लोग खमेर भाषा बोलते हैं । पूर्वी तिमोर में लोग तेतुम और पुर्तगाली बोलते हैं । भारत में लोग अंग्रेज़ी और हिन्दी बोलते हैं । मंगोलिया में लोग मंगोल भाषा और साहित्य बोलते हैं । ईरान में लोग फ़ारसी भाषा बोलते हैं । इराक में लोग अरबी और कुर्दी बोलते हैं । इज़राइल में लोग अरबी और इब्रानी बोलते हैं । यमन में लोग अरबी भाषा बोलते हैं । जॉर्डन में लोग अरबी भाषा बोलते हैं । किर्गिज़स्तान में लोग किर्गिज़ और रूसी बोलते हैं । लाओस में लोग लाओ भाषा बोलते हैं । लेबनान में लोग अरबी भाषा बोलते हैं । मालदीव में लोग महल् बोलते हैं । मलेशिया में लोग मलय भाषा बोलते हैं । म्यान्मार में लोग बर्मी भाषा बोलते हैं । नेपाल में लोग नेपाली भाषा बोलते हैं । ओमान में लोग अरबी भाषा बोलते हैं । पाकिस्तान में लोग अंग्रेज़ी, अरबी और उर्दू बोलते हैं । क़तर में लोग अरबी भाषा बोलते हैं । सउदी अरब में लोग अरबी भाषा बोलते हैं । श्रीलंका में लोग तमिल और सिंहली बोलते हैं । सीरिया में लोग अरबी भाषा बोलते हैं । ताजिकिस्तान में लोग ताजिकी और रूसी बोलते हैं । चीनी गणराज्य में लोग होक्का भाषा बोलते हैं । थाईलैण्ड में लोग थाई भाषा बोलते हैं । तुर्कमेनिस्तान में लोग तुर्कमेन भाषा बोलते हैं । संयुक्त अरब अमीरात में लोग अरबी भाषा बोलते हैं । वियतनाम में लोग वियतनामी भाषा बोलते हैं । दक्षिण कोरिया में लोग कोरियाई भाषा बोलते हैं । अफ़्गानिस्तान में लोग अरबी, उज़्बेक, तुर्कमेन और पश्तो बोलते हैं । बांग्लादेश में लोग बांग्ला भाषा बोलते हैं । भूटान में लोग जोंगखा बोलते हैं । ब्रुनेई में लोग अंग्रेज़ी और मलय बोलते हैं । फ़िलीपीन्स में लोग अंग्रेज़ी भाषा बोलते हैं ।