IRE Assignment 2

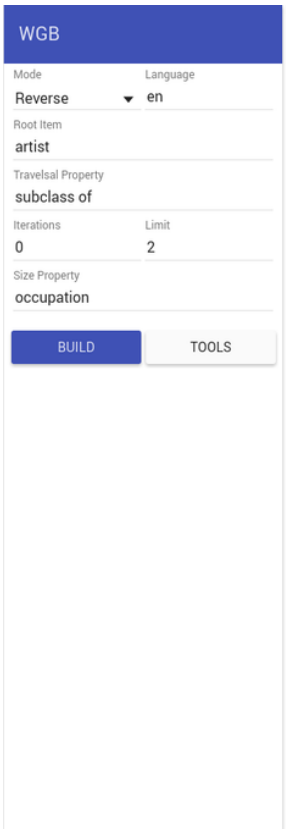
Snehal Kumar

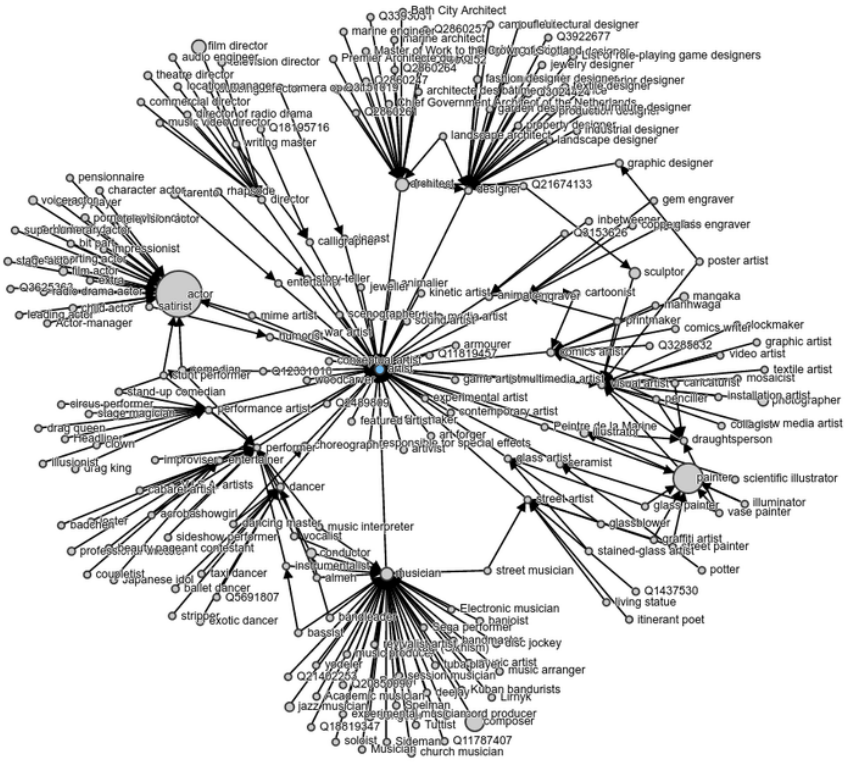
2019101003

Wikidata

Wikidata is an open and free knowledge base available for machines and human alike. It is essentially a database which contains around 50 million data items for all the Wiki projects including Wikipedia. It is automatically created when a Wikipedia page is published and stores data in a structured format by using a knowledge graph.

Each item in Wikidata is represented by a unique id: QID which covers the topic of the item. Each statement is stored as a key-pair of property and associated entity. A link between a property and entity can be stored in this manner. The property of the statement is the main data value or the category. To build the knowledge graph, Wikidata uses the properties and forms connections between the items and the values. This makes it easier to store and query the topics.





SPARQL

SPARQL is a query language and protocol for Resource Description Framework (RDF) databases and Linked Open Data. Similar to how SQL queries relational database by accessing tables, SPARQL queries RDF databases by accessing a web of Linked Data. It consists of four types of queries:

1. ASK - if there is at least one match of the query
2. SELECT - selects all matches in a tabular form
3. CONSTRUCT - constructs an RDF graph by substituting the variables with a set of triplets

4. DESCRIBE - describes the matches found by constructing an RDF graph.

Example of query: (Selecting the top 100 cities by population size)

```
SELECTDISTINCT ?cityLabel ?population ?gps
WHERE{
  ?citywdt:P31/wdt:P279*wd:Q515 .
  ?citywdt:P1082 ?population .
  ?citywdt:P625 ?gps .
SERVICEwikibase:label {
bd:serviceParamwikibase:language "en" .
  }
}
ORDER BYDESC(?population)LIMIT 100
```

Approach

I created a wikipedia page on my state: `Haryana` in my mother tongue: `Hindi` . Sandbox link:

User:Snehalku/sandbox - Wikipedia

हरियाणा भारत के राज्य का उदहारण है। हरियाणा का देश भारत है। हरियाणा का महाद्वीप एशिया है। हरियाणा हरियाणा की राजधानी चण्डीगढ़ है। हरियाणा की आधिकारिक भाषा हिन्दी है। हरियाणा के निर्माण की तिथि 1966-11-01T00:00:00Z हैं। हरियाणा का कार्यकारी मंडल हरियाणा विधानसभा है। हरियाणा का कार्यकारी निकाय हरियाणा सरकार है।

W https://en.wikipedia.org/wiki/User:Snehalku/sandbox

To get the appropriate information about the state, first we create the required SPARQL query to get all the statements having subject as 'Haryana':

```
SELECT ?objectLabel ?dataLabel { VALUES (?state) {(wd:Q1174)}
?state ?p ?statement .
?statement ?ps ?data .
?object wikibase:statementProperty ?ps.
SERVICE wikibase:label { bd:serviceParam wikibase:language "hi" }
}
```

Where wd:Q1174 denotes the state of Haryana and language is set to be Hindi.

Using this output, we generate the Resource Description Framework (RDF) triplets and store it to generate the natural language later. RDF triplets require subject, predicate and object:

```
{'subject': 'हरियाणा', 'predicate': 'सीमा लगती है', 'object': 'दिल्ली'},
 {'subject': 'हरियाणा', 'predicate': 'सीमा लगती है', 'object': 'उत्तराखण्ड'},
 {'subject': 'हरियाणा', 'predicate': 'सीमा लगती है', 'object': 'राजस्थान'},
 {'subject': 'हरियाणा', 'predicate': 'सीमा लगती है', 'object': 'पंजाब'},
 {'subject': 'हरियाणा', 'predicate': 'सीमा लगती है', 'object': 'उत्तर प्रदेश'}
```

Once we have the RDF triplets, we use a rule-based approach to generate the natural language. The Hindi language uses a subject-object-verb structuring in its sentences, different to English. Also there are different forms of verbs depending on gender and persons of authority for which separate functions were made checking the presence of particular verbs in the sentence. These changes have been dealt on case to case basis to generate language in a suitable form to be used as text in the Wikipedia article.

```
def female(predicate):
    return True if any(noun in predicate for noun in ["प्रशासनिक इकाई", "आईडी", "वेबसाइट", "राजधानी", "जनसंख्या","भाषा", "निकाय", "छवि"]) else False

def respect(predicate):
    return True if any(noun in predicate for noun in ["तिथि","राज्याध्यक्ष","निर्माण","स्थान","विषय", "शासनाध्यक्ष"]) else False

for data in RDFTriplets:
    predi = data['predicate']
```

```
    obj = data['object']
    sentence = ""
    if "उदहारण" in data["predicate"]:
        sentence = " ".join([SUBJECT, obj, GENDER["MALE"], data["predicate"], TENSE["PRESENT"]])
    .
    .
    elif female(data["predicate"]):
        sentence = " ".join([SUBJECT, GENDER["FEMALE"], data["predicate"], obj, TENSE["PRESENT"]])
    elif respect(data["predicate"]):
        sentence = " ".join([SUBJECT, GENDER["THEY"], data["predicate"], obj, TENSE["RESPECT"]])
    .
    .
    elif "सीमा लगती है" in data["predicate"]:
        sentence = " ".join([SUBJECT, GENDER["FEMALE"], obj, "से", data["predicate"]])
    else:
        sentence = " ".join([SUBJECT, obj, data["predicate"]])
    sentences.append(sentence)
```

There were a total of 9 cases used. Conditional statements for generating the language was used due to the small size of the output generated. This also made it easier to deal with different cases separately in a complex language like Hindi. Some of the sentences formed:

```
हरियाणा का महाद्वीप एशिया है।
हरियाणा भारत के राज्य का उदहारण है।
हरियाणा के राज्याध्यक्ष कप्तान सिंह सोलंकी हैं।
हरियाणा की राजधानी चण्डीगढ़ है।
हरियाणा की आधिकारिक भाषा हिन्दी है।
हरियाणा की चण्डीगढ़ से सीमा लगती है।
```

Once the sentences were formed, they were placed and formatted in the Wikipedia sandbox to publish the final article.

हरियाणा भारत के राज्य का उदहारण है। हरियाणा का देश भारत है। हरियाणा का महाद्वीप एशिया है। हरियाणा हरियाणा की राजधानी चण्डीगढ़ है। हरियाणा की आधिकारिक भाषा हिन्दी है। हरियाणा के निर्माण की तिथि 1966-11-01T00:00:00Z हैं। हरियाणा का कार्यकारी मंडल हरियाणा विधानसभा है। हरियाणा का कार्यकारी निकाय हरियाणा सरकार है।

हरियाणा भारत की प्रशासनिक इकाई में है। हरियाणा की जनसंख्या 27761063 है। हरियाणा के स्थान का समन्वय Point(76.324586 29.193657) हैं। हरियाणा का क्षेत्र 44212 है। हरियाणा का कॉमन्स श्रेणी Haryana है।

Contents [hide]
1 सरकारी अधिकारी
2 प्रशासनिक इकाई
3 हरियाणा के पड़ोसी
4 विविध जानकारी

सरकारी अधिकारी [edit source]

हरियाणा के राज्याध्यक्ष कप्तान सिंह सोलंकी हैं। हरियाणा के शासनाध्यक्ष मनोहर लाल खट्टर हैं।

प्रशासनिक इकाई [edit source]

गुरुग्राम संभाग हरियाणा की प्रशासनिक इकाई में है। जोशीक विभाग हरियाणा की प्रशासनिक इकाई में है। प्रशासनिक इकाई संख्या Q48732504 हरियाणा की प्रशासनिक इकाई में है। प्रशासनिक इकाई संख्या Q763616 हरियाणा की प्रशासनिक इकाई में है। गुरुग्राम संभाग हरियाणा की प्रशासनिक इकाई में है। प्रशासनिक इकाई संख्या Q48734314 हरियाणा की प्रशासनिक इकाई में है। प्रशासनिक इकाई संख्या Q3712130 हरियाणा की प्रशासनिक इकाई में है।

हरियाणा के पड़ोसी [edit source]

हरियाणा की चण्डीगढ़ से सीमा लगती है। हरियाणा की हिमाचल प्रदेश से सीमा लगती है। हरियाणा की दिल्ली से सीमा लगती है। हरियाणा की उत्तराखण्ड से सीमा लगती है। हरियाणा की राजस्थान से सीमा लगती है। हरियाणा की पंजाब से सीमा लगती है। हरियाणा की उत्तर प्रदेश से सीमा लगती है।

हरियाणा

राज्य

हरियाणा का चित्र.

हरियाणा की मानचित्र छवि

निर्माण की तिथि	1966-11-01
स्थान का समन्वय	(76.324586 29.193657)