IRE Assignment - 4
Generating wikipedia page from wikidata

Name: Avani Gupta
Roll number: 2019121004

How to create Wikipedia pages from Wiki Data?

- *Study Wiki Data*
- *Study SPARQL for Wiki Data*
- *Study about LSJBot (Swedish Wikipedia creation)*
- *Look into Hindi (or your mother tongue) Wiki Data*
- *Create one Wikipedia page automatically in your mother tongue*
- *Submit a report + submit the Wiki page*

Wikidata is a free, collaborative, multilingual, secondary database, collecting structured data to provide support for Wikipedia, Wikimedia Commons, the other wikis of the Wikimedia movement, and to anyone in the world.It is independent of language.

SPARQL is an RDF query language—that is, a semantic query language for databases—able to retrieve and manipulate data stored in Resource Description Framework (RDF)format. In other words it s a language to formulate questions (queries) for knowledge databases.

WDQS Wikidata Query Service is a tool to provide insight into Wikidata's content. It integrates Wikidata and SPARQL. User enters a SPARQL query which runs against Wikidata's dataset and result is displayed.

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).

Lsjbot was active in the Swedish Wikipedia and in the Cebunoa Waray Wikipedias, and has created most Wikipedia articles in those languages(between 80% and 99% of the total.

Source code of LSJ bot is available at https://sv.wikipedia.org/wiki/Wikipedia:Projekt_DotNetWikiBot_Framework/Lsjbot/Makespecies

It is written in C#. As one can see it doesnot use any ML algorithm but simply extracts data and converts it to text based on know grammatical rules.

Wikidata is language independent. 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 Q

After having studied Wikidata, SPARQL for wikidata, LSJ bot creating a Wikipedia page I came up with strategy to use spaqrl query to extract the propertiy value pairs from wikidata then convert them to Natural Language sentence.

The script for my code is as follows:

In [12]:
```
import requests
url = 'https://query.wikidata.org/sparql'
query = """SELECT ?itemLabel ?wdLabel ?propLabel ?wd_Label ?pq_Label {
  VALUES (?item) {(wd:Q229058)}

  ?item ?p ?statement .
  ?statement ?ps ?prop .

  ?wd wikibase:claim ?p.
  ?wd wikibase:statementProperty ?ps.

  OPTIONAL {
  ?statement ?pq ?pq_ .
  ?wd_ wikibase:qualifier ?pq .
  }

  SERVICE wikibase:label { bd:serviceParam wikibase:language "hi,en" }
} ORDER BY ?wd ?statement ?prop
"""

r = requests.get(url, params = {'format':'json', 'query': query})
data = r.json()
print (data['results']['bindings'])
```

I used SPARQL query to fetch the properties and values of item(here ISRO wd: Q229058.

Q229058 is id of ISRO on Wikidata)

In [33]:
```
largestr = " "
str = " "
pre = " "
v = ""
for li in data['results']['bindings']:
    if pre == li['wdLabel']['value']:
        if li['propLabel']['value'] != v:
            str += ", " + li['propLabel']['value']
    else:
        if(li['wdLabel']['value'][-3:-1]=='है'):
            str += " है|"
        if(len(str)>2):
            largestr += str +"| "
        s = li['wdLabel']['value'].split()
        if(s[0][-1]=='ा' or s[0][-1]=='ी'):
            str = li['itemLabel']['value'] + ' की ' + li['wdLabel']['value'] + " " + li['propLabel']['value']
        else:
            str = li['itemLabel']['value'] + ' का ' + li['wdLabel']['value'] + " " + li['propLabel']['value']
    v = li['propLabel']['value']
    pr = li['wdLabel']['value']
    f = open("page.txt",'w')
    f.write(largestr)
```

Then I processed the data obtained and stored it in form of Natural Language sentences. I used simple grammar rules of hindi like suffixes 'आ' and 'ई' generally occur in word which is feminine. More Natural Grammer rules can be applied to refine the page created. Also, hindi instances of some properties were not available, hence enlish values were fetched, which can be converted to equivalent hindi words by using simple English to Hindi Translator.

The approach I used is general and generates wikipedia page on any Item in wikidata.

The page generated is attached as follows:

(page.txt file)

भारतीय अंतरिक्ष अनुसंधान संगठन

भारतीय अंतरिक्ष अनुसंधान संगठन का संस्थापक विक्रम अंबालाल साराभाई है|

भारतीय अंतरिक्ष अनुसंधान संगठन का स्वामी भारत सरकार है|
भारतीय अंतरिक्ष अनुसंधान संगठन का प्रशासनिक इकाई में है कर्नाटक है|
भारतीय अंतरिक्ष अनुसंधान संगठन का लोगो चित्र http://commons.wikimedia.org/wiki/Special:FilePath/Indian%20Space%20Research%20Organisation%20Logo.svg है|
भारतीय अंतरिक्ष अनुसंधान संगठन का पिछला है Indian National Committee for Space Research है|
भारतीय अंतरिक्ष अनुसंधान संगठन का मुख्यालय का स्थान बंगलौर है|
भारतीय अंतरिक्ष अनुसंधान संगठन का पुरस्कार प्राप्त Space Pioneer Awards, इंदिरा गांधी शांति पुरस्कार, गाँधी शांति पुरस्कार, Space Pioneer Awards है|
भारतीय अंतरिक्ष अनुसंधान संगठन का देश भारत है|
भारतीय अंतरिक्ष अनुसंधान संगठन का चित्र http://commons.wikimedia.org/wiki/Special:FilePath/ISRO.JPG है|
भारतीय अंतरिक्ष अनुसंधान संगठन का ISNI (ISO 27729) 0000 0004 0500 9274 है|
भारतीय अंतरिक्ष अनुसंधान संगठन का VIAF अभिज्ञापक 133195349 है|
भारतीय अंतरिक्ष अनुसंधान संगठन का GND अभिज्ञापक 2078014-X है|
भारतीय अंतरिक्ष अनुसंधान संगठन का LCCN अभिज्ञापक n78092155 है|
भारतीय अंतरिक्ष अनुसंधान संगठन का उदहारण है space agency है|
भारतीय अंतरिक्ष अनुसंधान संगठन का product or material produced अंतरिक्ष यान है|
भारतीय अंतरिक्ष अनुसंधान संगठन का कर्मचारी 16072 है|
भारतीय अंतरिक्ष अनुसंधान संगठन का Gran Enciclopèdia Catalana ID 0246479 है|
भारतीय अंतरिक्ष अनुसंधान संगठन का Encyclopædia Britannica Online ID topic/Indian-Space-Research-Organisation है|
भारतीय अंतरिक्ष अनुसंधान संगठन का topic's main template Template:ISRO facilities, साँचा:भारतीय अंतरिक्ष कार्यक्रम है|
भारतीय अंतरिक्ष अनुसंधान संगठन का आधिकारिक नाम Indian Space Research Organisation है|
भारतीय अंतरिक्ष अनुसंधान संगठन का legal form निगम है|
भारतीय अंतरिक्ष अनुसंधान संगठन का short name ISRO, ఇస్రో, इसरो है|
भारतीय अंतरिक्ष अनुसंधान संगठन का owner of भारतीय गहन अंतरिक्ष नेटवर्क, मुख्य नियंत्रण सुविधा, Indian Space Research Organisation Telemetry, Tracking and Command Network, थुम्बा इक्वेटोरियल रॉकेट लॉन्चिंग स्टेशन, राष्ट्रीय दूरसंवेदी केंद्र है|
भारतीय अंतरिक्ष अनुसंधान संगठन का ट्विटर सदस्य नाम isro है|
भारतीय अंतरिक्ष अनुसंधान संगठन का Instagram username indianspacetime है|
भारतीय अंतरिक्ष अनुसंधान संगठन का Facebook ID isro है|
भारतीय अंतरिक्ष अनुसंधान संगठन का total revenue 256.58 है|
भारतीय अंतरिक्ष अनुसंधान संगठन का GRID ID grid.418654.a है|
भारतीय अंतरिक्ष अनुसंधान संगठन का Crossref funder ID 501100001413 है|
भारतीय अंतरिक्ष अनुसंधान संगठन का NE.se ID isro है|
भारतीय अंतरिक्ष अनुसंधान संगठन का Quora topic ID Indian-Space-Research-Organisation-ISRO है|
भारतीय अंतरिक्ष अनुसंधान संगठन का Ringgold ID 29123 है|
भारतीय अंतरिक्ष अनुसंधान संगठन का subsidiary ISRO Inertial Systems Unit, सतीश धवन अंतरिक्ष केंद्र, एंट्रिक्स, इसरो उपग्रह केंद्र, अंतरिक्ष अनुप्रयोग केंद्र, विक्रम साराभाई अंतरिक्ष केंद्र, Laboratory for Electro-Optics Systems, भारतीय सुदूर संवेदन संस्थान, ISRO Propulsion Complex, द्रव नोदन प्रणाली केंद्र है|
भारतीय अंतरिक्ष अनुसंधान संगठन का का भाग प्रधानमंत्री कार्यालय (भारत) है|
भारतीय अंतरिक्ष अनुसंधान संगठन का कॉमन्स श्रेणी Indian Space Research Organisation है|
भारतीय अंतरिक्ष अनुसंधान संगठन का industry space industry है|
भारतीय अंतरिक्ष अनुसंधान संगठन का का सदस्य International Space University, Inter-Agency Space Debris Coordination Committee, International Astronautical Federation, Committee on Space Research, International Cospas-Sarsat Programme है|
भारतीय अंतरिक्ष अनुसंधान संगठन का chairperson कैलासवटिवु शिवन है|
भारतीय अंतरिक्ष अनुसंधान संगठन का निर्माण की तिथि 1969-08-15T00:00:00Z है|
भारतीय अंतरिक्ष अनुसंधान संगठन का स्थान का समन्वय Point(77.698055555 12.965555555) है|
भारतीय अंतरिक्ष अनुसंधान संगठन का Microsoft Academic ID 1289461252 है|
भारतीय अंतरिक्ष अनुसंधान संगठन का फ्रीबेस पहचानकर्ता /m/03zwkz है|
भारतीय अंतरिक्ष अनुसंधान संगठन का HAL structure ID 301842 है|
भारतीय अंतरिक्ष अनुसंधान संगठन का ROR ID 00cwrns71 है|
भारतीय अंतरिक्ष अनुसंधान संगठन का मूल कंपनी अंतरिक्ष विभाग है|
भारतीय अंतरिक्ष अनुसंधान संगठन का आधिकारिक वेबसाइट http://www.isro.gov.in/ है|