Build_KG

May 13, 2021

[1]:

```
# NARENDRA SINGH BISHT
# AM.EN.P2ARI20043
# KG Final Assignment
```

[2]:

```
# Guidelines

# 1. Crawl the web sites for a particular topic;
# (hint: you can use a focussed crawling with the help vertical search engine.
# For e.g., https://thenextweb.com/news/
 ↪30-specialist-and-super-smart-search-engines
# 2. Extract the text  from the web pages and form into documents.
# 3. Build KG from documents, convert into Neo4J
# 4. Query the KG

# Submission should be in 2 files: .ipynb and its pdf
```

[3]:

```
# conda install spacy
```

[4]:

```
import spacy
from spacy.matcher import Matcher

from pathlib import Path

import pandas as pd
import re
import requests, bs4
import html2text

from py2neo import Graph

from googlesearch import search

import itertools

from tqdm import tqdm
# Instantly make your loops show a smart progress meter -
# just wrap any iterable with tqdm(iterable), and you're done!
```

[5]:

```python
# !python -m spacy download en_core_web_lg

nlp = spacy.load('en_core_web_lg')
```

[6]:

```python
query = 'India'
urls = []
for url in search(query,
                  tld="co.in",
                  num=25,
                  stop=25,
                  pause=2):
    urls.append(url)
```

[7]:

```python
print('Search Results:')
for idx, url in enumerate(urls):
    print(idx + 1, url)
```

```
Search Results:
1 https://indianexpress.com/article/india/coronavirus-india-live-updates-covid-second-wave-vaccination-7313012/
2 https://timesofindia.indiatimes.com/india/who-warns-variant-in-india-could-be-highly-contagious/articleshow/82594425.cms
3 https://www.moneycontrol.com/news/india/coronavirus-india-news-live-updates-lockdown-in-maharashtra-extended-until-may-31-to-curb-the-spread-of-covid-19-in-the-state-6886201.html
4 https://en.wikipedia.org/wiki/India
5 https://en.wikipedia.org/wiki/Politics_of_India
6 https://en.wikipedia.org/wiki/Outline_of_India
7 https://en.wikipedia.org/wiki/Names_for_India
8 https://en.wikipedia.org/wiki/South_India
9 https://www.india.gov.in/
10 https://www.incredibleindia.org/
11 https://www.britannica.com/place/India
12 https://www.india.com/
13 https://timesofindia.indiatimes.com/
14 https://mohfw.gov.in/
15 https://www.theguardian.com/world/india
16 http://www.airindia.in/
17 https://wikitravel.org/en/India
18 https://www.bbc.com/news/world/asia/india
19 https://www.usnews.com/news/best-countries/india
20 https://tourism.gov.in/
21 https://www.aljazeera.com/where/india/
22 https://sbi.co.in/
23 https://www.lonelyplanet.com/india
24 https://www.rbi.org.in/
25 https://www.digitalindia.gov.in/
```

[8]:

```
path = Path("Converted_Pages")
path.mkdir(exist_ok=True)
```

Extract the text from the web pages and store in text documents

[9]:

```
converted_pages = []
converted_page_count = 0

for url in tqdm(urls):

    print('\nRoot URL: ', url)

    linked_urls = []

    try:

        res = requests.get(url, timeout=2.5)
        responseSoup = bs4.BeautifulSoup(res.text,"lxml")
        count_a = len(responseSoup.find_all('a', href=True))
        limit = min(5, count_a)

        if limit > 0:

            for a in responseSoup.find_all('a', href=True):

                if (a['href'].find('www.') == -1) and (a['href'].find('https://') ==-1):

                    embedded_url = f"{url}{a['href']}"
                    linked_urls.append(embedded_url)

                else:
                    linked_urls.append(a['href'])

                if len(linked_urls) == limit:
                    break

        linked_urls.append(url)

        linked_urls = list(set(linked_urls))

        print('Related URLs crawled:')

        for idx, linked_url in enumerate(linked_urls):

            print(idx + 1, linked_url)

            try:
                r = requests.get(linked_url)
                html = r.text
                text = html2text.html2text(html)
                text = re.sub(r'[^a-zA-Z. ]', ' ', text)
                text = re.sub(r'\b\w{1,3}\b', '', text)
                text = re.sub('\s+',' ', text)

                #Create a file and save the text

                with open(f"Converted_Pages/Page_{converted_page_count}.
↪txt",'w', encoding = 'utf-8') as f:
                    f.write(text)

                converted_page_count += 1

            except Exception as ex:
                print(type(ex))
                print(linked_url)

    except Exception as ex:
        print(type(ex))
        print(url)
```

```
  0%|
| 0/25 [00:00<?, ?it/s]


Root URL:  https://indianexpress.com/article/india/coronavirus-india-live-
updates-covid-second-wave-vaccination-7313012/
Related URLs crawled:
1 https://tamil.indianexpress.com/
2 https://indianexpress.com/
3 https://indianexpress.com/article/india/coronavirus-india-live-updates-covid-
second-wave-vaccination-7313012/
4 https://malayalam.indianexpress.com/
5 https://bengali.indianexpress.com/
```

```
  4%|
| 1/25 [00:01<00:35,  1.48s/it]
```

```
6 https://www.jansatta.com/

Root URL:  https://timesofindia.indiatimes.com/india/who-warns-variant-in-india-
could-be-highly-contagious/articleshow/82594425.cms
Related URLs crawled:
1 https://timesofindia.indiatimes.com/us
2 https://timesofindia.indiatimes.com/india
3 https://timesofindia.indiatimes.com/india/who-warns-variant-in-india-could-be-
```

```
highly-contagious/articleshow/82594425.cms
4 https://timesofindia.indiatimes.com/

  8%|
| 2/25 [00:02<00:31,  1.36s/it]


Root URL:  https://www.moneycontrol.com/news/india/coronavirus-india-news-live-
updates-lockdown-in-maharashtra-extended-until-may-31-to-curb-the-spread-of-
covid-19-in-the-state-6886201.html
Related URLs crawled:
1 https://www.moneycontrol.com/news/india/coronavirus-india-news-live-updates-
lockdown-in-maharashtra-extended-until-may-31-to-curb-the-spread-of-covid-19-in-
the-state-6886201.html#HamburgerMenu_0
2 https://www.moneycontrol.com/
3 https://www.moneycontrol.com/news/india/coronavirus-india-news-live-updates-
lockdown-in-maharashtra-extended-until-may-31-to-curb-the-spread-of-covid-19-in-
the-state-6886201.html
4 https://www.moneycontrol.com/news/india/coronavirus-india-news-live-updates-
lockdown-in-maharashtra-extended-until-may-31-to-curb-the-spread-of-covid-19-in-
the-state-6886201.html#HamburgerMenu_17
5 https://www.moneycontrol.com/news/india/coronavirus-india-news-live-updates-
lockdown-in-maharashtra-extended-until-may-31-to-curb-the-spread-of-covid-19-in-
the-state-6886201.html#HamburgerMenu_3
6 https://www.moneycontrol.com/india/bestportfoliomanager/investment-tool

 12%|
| 3/25 [00:10<01:11,  3.23s/it]


Root URL:  https://en.wikipedia.org/wiki/India
Related URLs crawled:
1 https://en.wikipedia.org/wiki/India/wiki/Wikipedia:Featured_articles
2 https://en.wikipedia.org/wiki/India#mw-head
3 https://en.wikipedia.org/wiki/India#searchInput
4 https://en.wikipedia.org/wiki/India/wiki/Wikipedia:Protection_policy#extended
5 https://en.wikipedia.org/wiki/India
6 https://en.wikipedia.org/wiki/India/wiki/India_(disambiguation)

 16%|
| 4/25 [00:16<01:25,  4.07s/it]


Root URL:  https://en.wikipedia.org/wiki/Politics_of_India
Related URLs crawled:
1 https://en.wikipedia.org/wiki/Politics_of_India/wiki/Wikipedia:Citing_sources
2 https://en.wikipedia.org/wiki/Politics_of_India
3 https://en.wikipedia.org/wiki/Politics_of_India/wiki/Wikipedia:Citing_sources#
Inline_citations
4 https://en.wikipedia.org/wiki/Politics_of_India#searchInput
5 https://en.wikipedia.org/wiki/Politics_of_India/wiki/Wikipedia:WikiProject_Fac
```

```
t_and_Reference_Check
6 https://en.wikipedia.org/wiki/Politics_of_India#mw-head

 20%|
| 5/25 [00:19<01:14,  3.74s/it]


Root URL:  https://en.wikipedia.org/wiki/Outline_of_India
Related URLs crawled:
1 https://en.wikipedia.org/wiki/Outline_of_India#searchInput
2 https://en.wikipedia.org/wiki/Outline_of_India/wiki/File:Flag_of_India.svg
3 https://en.wikipedia.org/wiki/Outline_of_India/wiki/Flag_of_India
4 https://en.wikipedia.org/wiki/Outline_of_India/wiki/File:Emblem_of_India.svg
5 https://en.wikipedia.org/wiki/Outline_of_India#mw-head
6 https://en.wikipedia.org/wiki/Outline_of_India

 24%|
| 6/25 [00:22<01:10,  3.74s/it]


Root URL:  https://en.wikipedia.org/wiki/Names_for_India
Related URLs crawled:
1 https://en.wikipedia.org/wiki/Names_for_India/wiki/Names_of_India_in_its_offic
ial_languages
2 https://en.wikipedia.org/wiki/Names_for_India/wiki/Bharata_(disambiguation)
3 https://en.wikipedia.org/wiki/Names_for_India#searchInput
4 https://en.wikipedia.org/wiki/Names_for_India/wiki/India
5 https://en.wikipedia.org/wiki/Names_for_India#mw-head
6 https://en.wikipedia.org/wiki/Names_for_India

 28%|
| 7/25 [00:25<01:02,  3.48s/it]


Root URL:  https://en.wikipedia.org/wiki/South_India
Related URLs crawled:
1 https://en.wikipedia.org/wiki/South_India
2 https://en.wikipedia.org/wiki/South_India#mw-head
3 https://en.wikipedia.org/wiki/South_India/wiki/India
4 https://en.wikipedia.org/wiki/South_India#searchInput
5 https://en.wikipedia.org/wiki/South_India/wiki/File:India_South_India_Locator_
Map.svg
6 https://en.wikipedia.org/wiki/South_India/wiki/Wikipedia:Good_articles

 32%|
| 8/25 [00:29<01:03,  3.71s/it]


Root URL:  https://www.india.gov.in/
Related URLs crawled:
1 https://twitter.com/indiagovin
2 https://www.india.gov.in/
```

```
3 https://www.facebook.com/NationalPortalIndia
4 https://www.india.gov.in/user/register
5 https://www.india.gov.in/#main-content
6 https://www.india.gov.in/user/login

 36%|
| 9/25 [00:42<01:43,  6.48s/it]


Root URL:  https://www.incredibleindia.org/

 40%|
| 10/25 [00:45<01:17,  5.17s/it]

Related URLs crawled:
1 https://www.incredibleindia.org/

Root URL:  https://www.britannica.com/place/India
Related URLs crawled:
1 https://www.britannica.com/place/India/
2 https://www.britannica.com/place/India/on-this-day
3 https://www.britannica.com/place/India/games
4 https://premium.britannica.com/premium-
membership/?utm_source=house&utm_medium=mendel&utm_campaign=premium-pres-day
5 https://www.britannica.com/place/India
6 https://www.britannica.com/place/India/quiz/browse

 44%|
| 11/25 [00:48<01:07,  4.79s/it]


Root URL:  https://www.india.com/
Related URLs crawled:
1 https://www.india.com/entertainment/
2 https://www.india.com
3 https://www.india.com/
4 https://www.india.com/hindi-news/

 48%|
| 12/25 [00:51<00:52,  4.02s/it]

5 https://www.india.com/news/india/

Root URL:  https://timesofindia.indiatimes.com/
Related URLs crawled:
1 https://timesofindia.indiatimes.com/us
2 https://timesofindia.indiatimes.com/briefs
3 https://timesofindia.indiatimes.com/

 52%|
| 13/25 [00:52<00:37,  3.13s/it]
```

```
Root URL:  https://mohfw.gov.in/
Related URLs crawled:
1 https://mohfw.gov.in/#
2 https://mohfw.gov.in/#latest-update
3 https://mohfw.gov.in/
4 https://mohfw.gov.in/index.html
5 https://mohfw.gov.in/#site-advisories

 56%|
| 14/25 [00:53<00:29,  2.72s/it]


Root URL:  https://www.theguardian.com/world/india
Related URLs crawled:
1 https://support.theguardian.com/contribute?INTCMP=header_support_contribute&ac
quisitionData=%7B%22source%22:%22GUARDIAN_WEB%22,%22componentType%22:%22ACQUISIT
IONS_HEADER%22,%22componentId%22:%22header_support_contribute%22%7D
2 https://www.theguardian.com/world/india
3 https://www.theguardian.com/international
4 https://support.theguardian.com/subscribe?INTCMP=header_support_subscribe&acqu
isitionData=%7B%22source%22:%22GUARDIAN_WEB%22,%22componentType%22:%22ACQUISITIO
NS_HEADER%22,%22componentId%22:%22header_support_subscribe%22%7D

 60%|
| 15/25 [00:58<00:32,  3.23s/it]

5 https://www.theguardian.com/world/india#maincontent

Root URL:  http://www.airindia.in/

 64%|
| 16/25 [01:00<00:27,  3.03s/it]

<class 'requests.exceptions.ReadTimeout'>
http://www.airindia.in/

Root URL:  https://wikitravel.org/en/India

 68%|
| 17/25 [01:03<00:23,  2.92s/it]

<class 'requests.exceptions.ReadTimeout'>
https://wikitravel.org/en/India

Root URL:  https://www.bbc.com/news/world/asia/india
Related URLs crawled:
1 https://account.bbc.com/account
2 https://www.bbc.co.uk/accessibility/
3 https://www.bbc.com/news/world/asia/india#skip-to-content
4 https://www.bbc.co.uk
```

```
5 https://www.bbc.com/news/world/asia/india
6 https://www.bbc.com/news/world/asia/india#

 72%|
| 18/25 [01:10<00:28,  4.01s/it]


Root URL:  https://www.usnews.com/news/best-countries/india

76%|
| 19/25 [01:12<00:21,  3.58s/it]

<class 'requests.exceptions.ReadTimeout'>
https://www.usnews.com/news/best-countries/india

Root URL:  https://tourism.gov.in/
Related URLs crawled:
1 https://tourism.gov.in/#skipCont
2 https://india.gov.in/hi
<class 'requests.exceptions.SSLError'>
https://india.gov.in/hi
3 https://tourism.gov.in/#
4 https://india.gov.in/
<class 'requests.exceptions.SSLError'>
https://india.gov.in/
5 https://tourism.gov.in/

 80%|
           | 20/25 [01:14<00:14,  2.89s/it]


Root URL:  https://www.aljazeera.com/where/india/
Related URLs crawled:
1 https://www.aljazeera.com/where/india//middle-east/
2 https://www.aljazeera.com/where/india//news/
3 https://www.aljazeera.com/where/india//asia/
4 https://www.aljazeera.com/where/india//africa/
5 https://www.aljazeera.com/where/india//live

 84%|
          | 21/25 [01:20<00:15,  3.93s/it]

6 https://www.aljazeera.com/where/india/

Root URL:  https://sbi.co.in/
Related URLs crawled:
1 https://sbi.co.in/
2 https://sbi.co.in/javascript:history.back();

 88%|
         | 22/25 [01:21<00:08,  2.94s/it]
```

```
Root URL:  https://www.lonelyplanet.com/india
Related URLs crawled:
1 https://www.lonelyplanet.com/india/
2 https://www.lonelyplanet.com/india/places
3 https://www.lonelyplanet.com/india/search
4 https://www.lonelyplanet.com/india#footer
5 https://www.lonelyplanet.com/search
6 https://www.lonelyplanet.com/india
```

```
 92%|
                | 23/25 [01:24<00:06,  3.01s/it]
```

```
Root URL:  https://www.rbi.org.in/
Related URLs crawled:
1 https://www.rbi.org.in/
2 https://www.rbi.org.in/javascript:__doPostBack('UsrFontCntr$LinkBtnFontIncreas
e','')
3 https://www.rbi.org.in/javascript:__doPostBack('UsrFontCntr$LinkBtnAccessibilt
y','')
4 https://www.rbi.org.in/#mainsection
5 https://www.rbi.org.in/hindi/Home.aspx
6 https://www.rbi.org.in/javascript:__doPostBack('UsrFontCntr$LinkBtnFontDecreas
e','')
```

```
 96%|
              | 24/25 [01:34<00:05,  5.24s/it]
```

```
Root URL:  https://www.digitalindia.gov.in/
Related URLs crawled:
1 https://www.facebook.com/OfficialDigitalIndia
2 https://in.linkedin.com/company/digital-india
3 https://www.youtube.com/user/MyNeGP
4 https://twitter.com/_DigitalIndia
5 https://www.instagram.com/officialdigitalindia/
6 https://www.digitalindia.gov.in/
```

```
100%|
            | 25/25 [01:39<00:00,  4.00s/it]
```

Identify the entity pairs and corresponding relations

[10]:

```python
string = []
for p in path.iterdir():
    if str(p).find(".ipynb_checkpoints") == -1:

        with open(p, "r", encoding="utf8") as f:
            full_text = f.read()
            for l in re.split(r"\.", full_text):
                if l != ".":
                    string.append(l)

sentences = pd.DataFrame(string)
```

[11]:

```
def get_entities(sentence):
    doc = nlp(sentence)
    entities = []
    for ent in doc.ents:
        entities.append(ent.text)
    entities = list(dict.fromkeys(entities) )
    if len(entities) != 2:
        return None
    return [entities[0], entities[1]]
```

[12]:

```
def get_relation(sentence):

    doc = nlp(sentence)

    # Matcher class object
    matcher = Matcher(nlp.vocab)

    #define the pattern
    pattern = [{'DEP':'ROOT'},
               {'DEP':'prep','OP':"?"},
               {'DEP':'agent','OP':"?"},
               {'POS':'ADJ','OP':"?"}]

    matcher.add("matching_1", [pattern])

    matches = matcher(doc)
    k = len(matches) - 1

    span = doc[matches[k][1]:matches[k][2]]

    return(span.text)
```

[13]:

```
entity_pairs = []
relations = []

for sentence in tqdm(sentences[0]):
    entity_pair = get_entities(sentence)
    if entity_pair:
        entity_pair.append(get_relation(sentence))
        entity_pairs.append(entity_pair)
```

```
100%|
         | 98836/98836 [10:25<00:00, 157.98it/s]
```

[14]:

```
entity_pairs = list(entity_pairs for entity_pairs, _ in itertools.
 ↪groupby(entity_pairs))
```

[15]:

```
# extract subject
source = [i[0] for i in entity_pairs]
```

[16]:

```
# extract object
target = [i[1] for i in entity_pairs]
```

[17]:

```
# extract relations
relations = [i[2] for i in entity_pairs]
```

[18]:

```
kg_df = pd.DataFrame({'source':source, 'target':target, 'edge':relations})
```

[19]:

```
kg_df.sample(20)
```

[19]:

```
                                   source  \
165   Coronavirus India News Live Updates
6017                                india
2685                                India
1251                                India
3503                            Karnataka
2568                  Talbot India Before
5819                                AIIMS
3841                          Names India
3105                  Demographics Sikkim
359     Coronavirus India News Highlights
1252                                 July
4730                                India
2290                               Indian
5205                        Ruth Gilligan
514   Coronavirus India News Live Updates
4513                        CreateAccount
4924                            Instagram
5595                                India
2380                                 ISBN
4876                            Instagram

                                       target           edge
165                                     India        reports
6017                                   Punjab   timesofindia
2685                                     Page          index
1251                                     NCRB         deaths
3503                                Jharkhand      Jharkhand
2568                                   Europe   India Before
```

```
5819                                    Delhi         series
3841                                    India      languages
3105  Demographics Sikkim Demographics Sikkim         Sikkim
359                               Uttarakhand        decides
1252                                 December      retrieved
4730     Maurice Merlin Mahatma Gandhi Buddha         Buddha
2290                                    India          grown
5205                           Ondaatje Prize           wins
514                                     India           send
4513                              South India          index
4924                         lGODdhAQABAPAAAP       download
5595                                   Videos  warns variant
2380  Special BookSources Special BookSources           wiki
4876                         lGODdhAQABAPAAAP       download
```

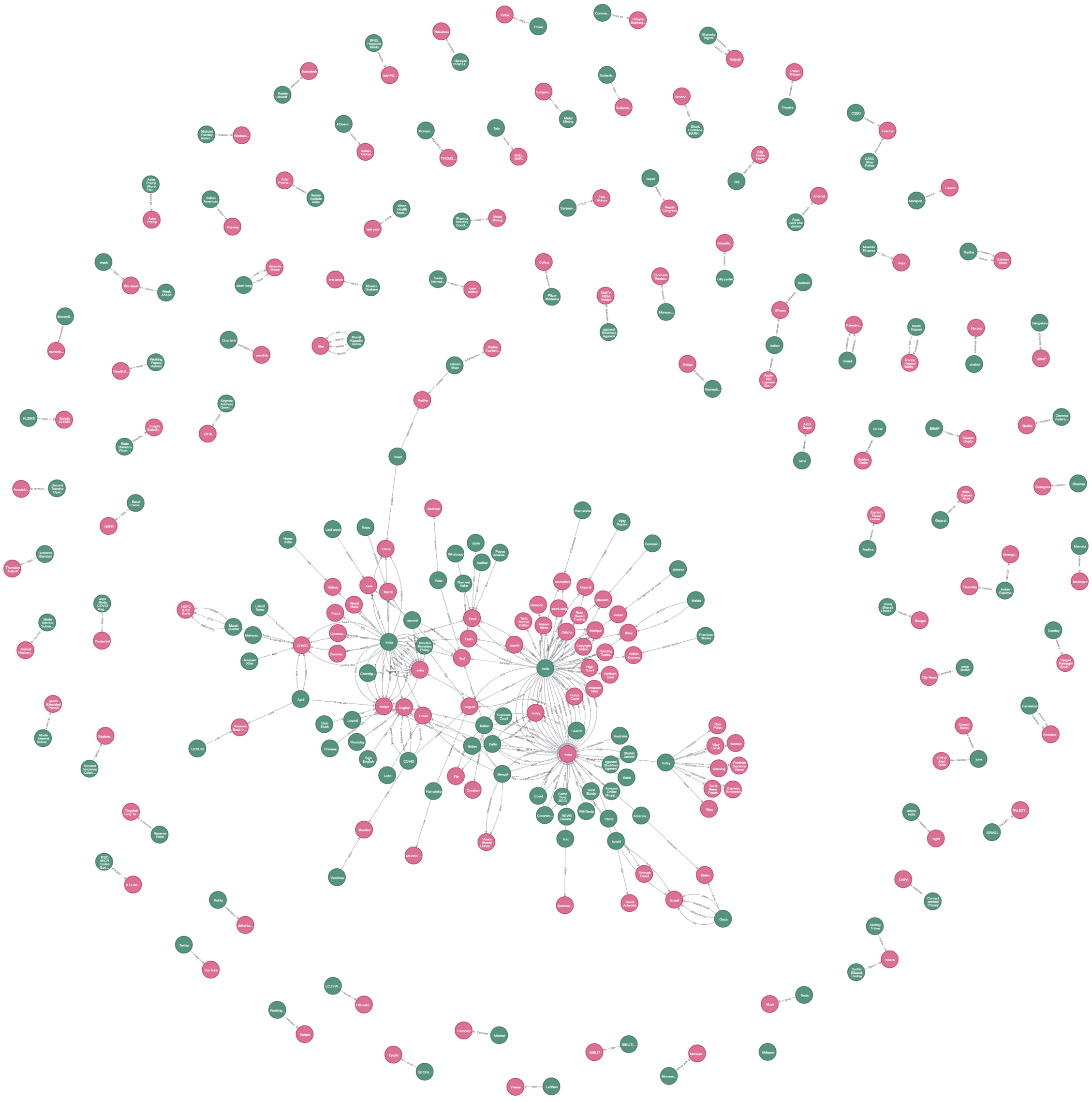
Create the Knowledge Graph in Neo4j

```
[20]: graph = Graph("bolt://localhost:7687", user="neo4j", password="password")
      transaction = graph.begin()
```

```
[21]: for index, row in kg_df.iterrows():

          transaction.evaluate(f'''
          MERGE (a:Subject {{name:$Subject}})
          MERGE (b:Object {{name:$Object}})
          MERGE (a)-[r:Relation {{name:$Relation}}]->(b)
          ''', parameters = {'Subject': row['source'],
                             'Object': row['target'],
                             'Relation': row['edge']})
      transaction.commit()
```

```
[21]: <py2neo.database.work.TransactionSummary at 0x22fe6fe3c40>
```



Query the Knowledge Graph

```
[22]: # Number of Nodes
      graph.run("""MATCH (n) RETURN COUNT(n)""").data()
```

```
[22]: [{'COUNT(n)': 2460}]
```

```
[29]: # Number of Relationships
      graph.run("""MATCH ()-->() RETURN COUNT(*)""").data()
```

```
[29]: [{'COUNT(*)': 2188}]
```

```
[31]: # List node labels
      graph.run("""CALL db.labels()""").data()
```

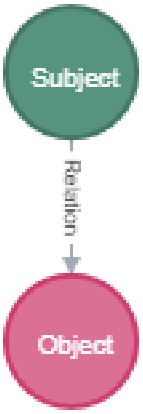
```
[31]: [{'label': 'Subject'}, {'label': 'Object'}]
```

```
[32]: # List relationship types
      graph.run("""CALL db.relationshipTypes()""").data()
```

```
[32]: [{'relationshipType': 'Relation'}]
```

```
[34]: # What is related, and how
      graph.run("""CALL db.schema.visualization()""").data()
```

```
[34]: [{'nodes': [Node('Object', constraints=[], indexes=[], name='Object'),
         Node('Subject', constraints=[], indexes=[], name='Subject')],
        'relationships': [Relation(Node(), Node())]}]
```



```
[26]: graph.run("""MATCH (n)-[r]->(m) RETURN n, r, m limit 10""").data()
```

```
[26]: [{'n': Node('Subject', name='India'),
        'r': Relation(Node('Subject', name='India'), Node('Object', name='English'),
      name='wiki'),
        'm': Node('Object', name='English')},
       {'n': Node('Subject', name='COVID Vaccine General'),
        'r': Relation(Node('Subject', name='COVID Vaccine General'), Node('Object',
      name='English'), name='https'),
        'm': Node('Object', name='English')},
       {'n': Node('Subject', name='Indians'),
        'r': Relation(Node('Subject', name='Indians'), Node('Object', name='English'),
      name='Indians'),
        'm': Node('Object', name='English')},
       {'n': Node('Subject', name='COVID'),
        'r': Relation(Node('Subject', name='COVID'), Node('Object', name='English'),
      name='distancing'),
        'm': Node('Object', name='English')},
       {'n': Node('Subject', name='COVID'),
        'r': Relation(Node('Subject', name='COVID'), Node('Object', name='English'),
      name='stigma'),
        'm': Node('Object', name='English')},
       {'n': Node('Subject', name='century'),
        'r': Relation(Node('Subject', name='century'), Node('Object', name='English'),
      name='referred'),
        'm': Node('Object', name='English')},
       {'n': Node('Subject', name='Santali'),
        'r': Relation(Node('Subject', name='Santali'), Node('Object', name='English'),
      name='simple'),
        'm': Node('Object', name='English')},
       {'n': Node('Subject', name='AIIMS'),
        'r': Relation(Node('Subject', name='AIIMS'), Node('Object', name='English'),
      name='https'),
        'm': Node('Object', name='English')},
       {'n': Node('Subject', name='Stevenson Angus Waite Maurice'),
        'r': Relation(Node('Subject', name='Stevenson Angus Waite Maurice'),
      Node('Object', name='English'), name='books'),
        'm': Node('Object', name='English')},
       {'n': Node('Subject', name='South India South India'),
        'r': Relation(Node('Subject', name='South India South India'), Node('Object',
      name='English'), name='wiki'),
        'm': Node('Object', name='English')}]
```

```
[ ]:
```