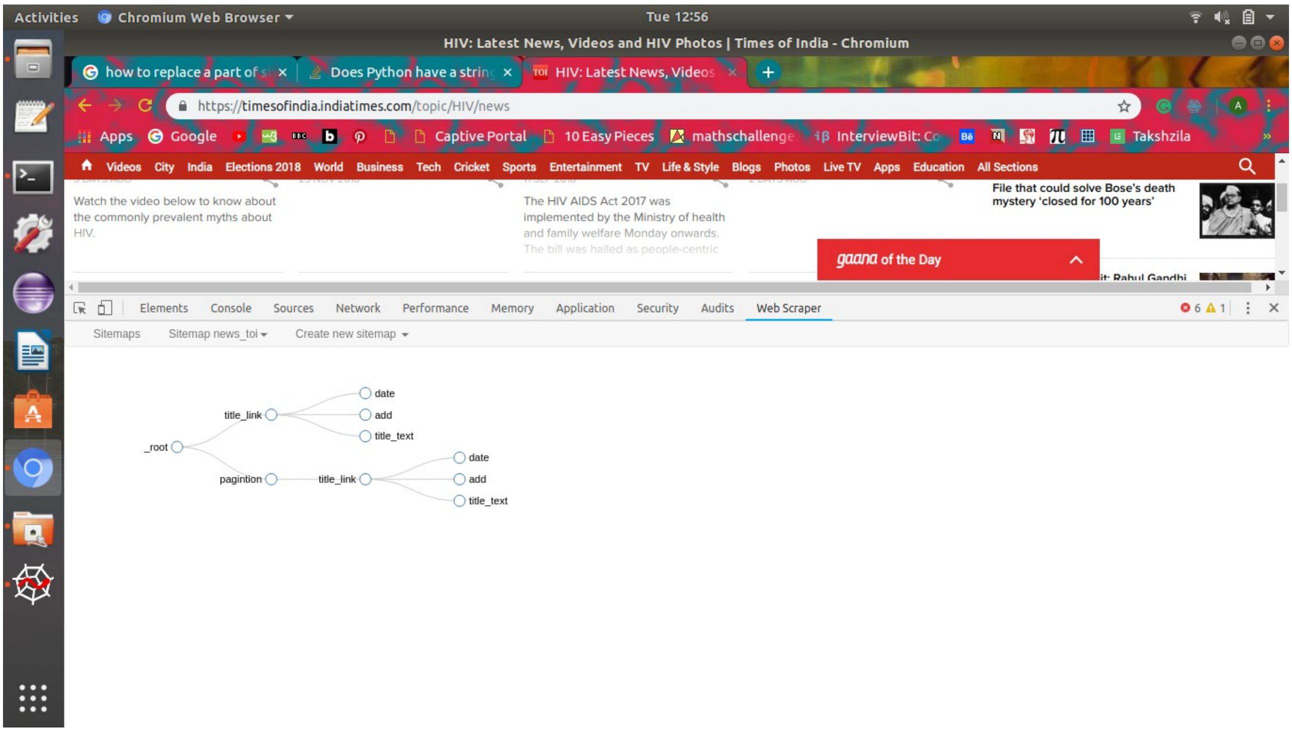
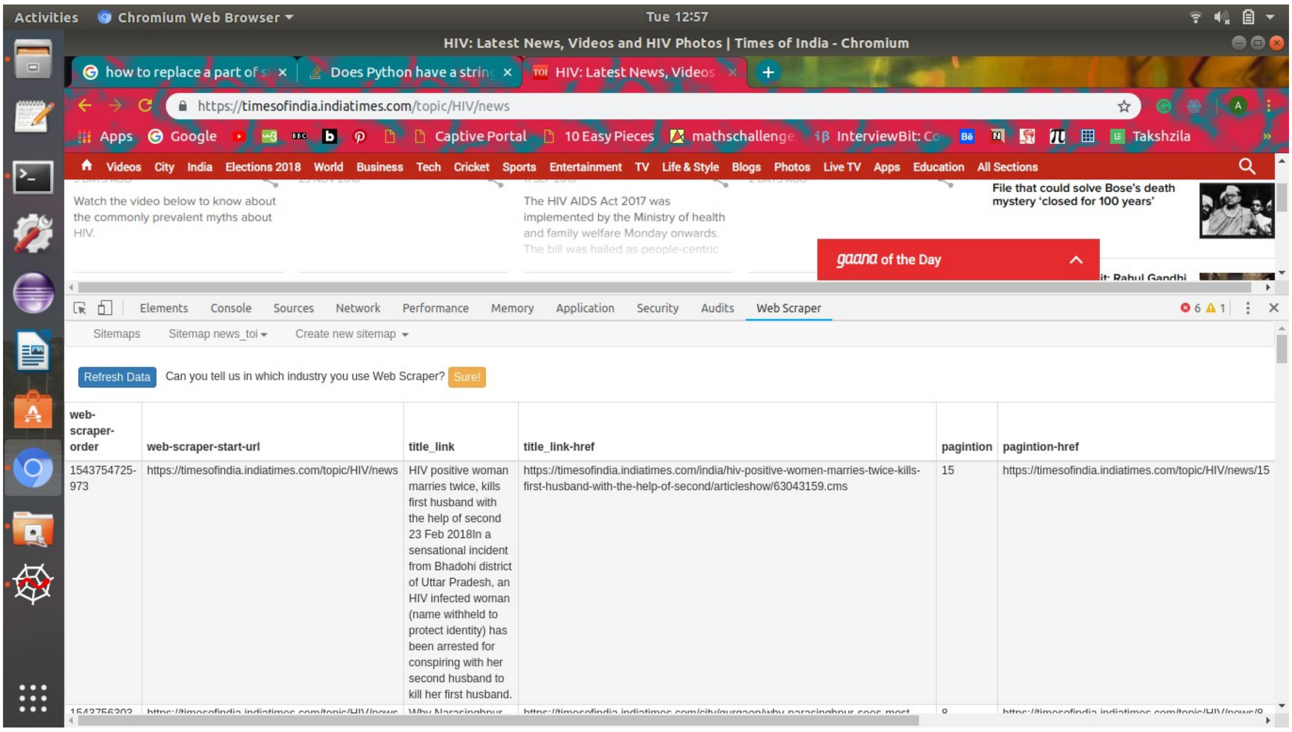
Assignment Report

1. Web Scraping using Chrome Extension

I have attached the screenshot of the sitemap selector graph for web scraping.



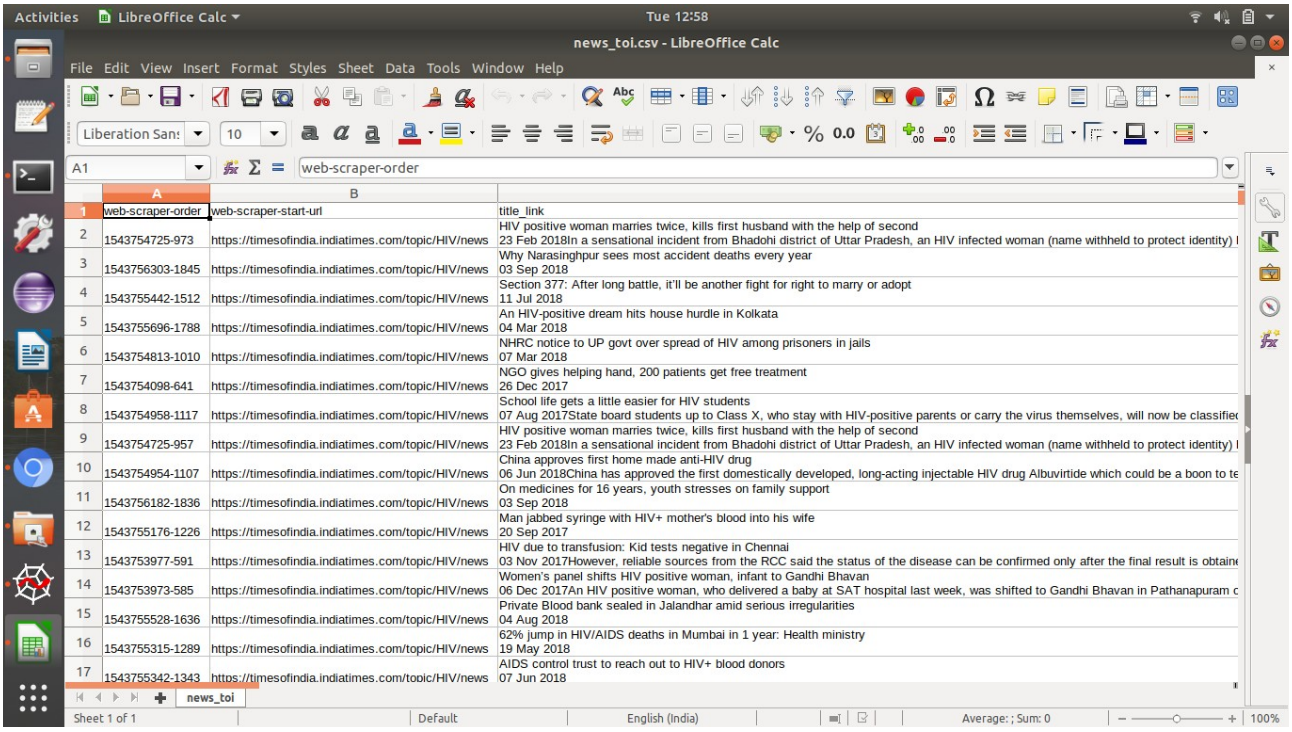
After this I have exported the file in CSV (Comma Separated Values) format for further analysis.



Can you tell us in which industry you use Web Scraper?

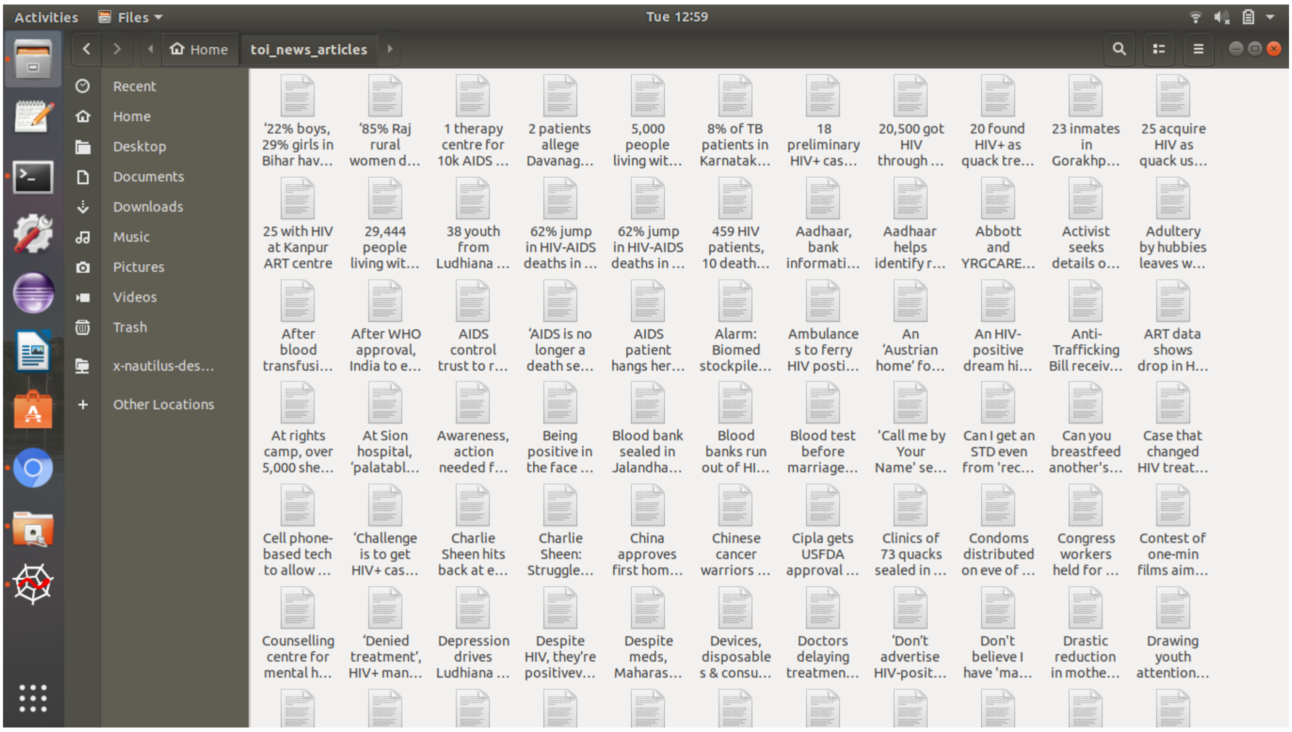
web-scraper-order	web-scraper-start-url	title_link	title_link-href	pagintion	pagintion-href
1543754725-973	https://timesofindia.indiatimes.com/topic/HIV/news	HIV positive woman marries twice, kills first husband with the help of second 23 Feb 2018In a sensational incident from Bhadohi district of Uttar Pradesh, an HIV infected woman (name withheld to protect identity) has been arrested for conspiring with her second husband to kill her first husband.	https://timesofindia.indiatimes.com/india/hiv-positive-women-marries-twice-kills-first-husband-with-the-help-of-second/articleshow/63043159.cms	15	https://timesofindia.indiatimes.com/topic/HIV/news/15
1543756303	https://timesofindia.indiatimes.com/topic/HIV/news	Why Narasinghpur	https://timesofindia.indiatimes.com/city/gurgaon/why-narasinghpur-sees-most	8	https://timesofindia.indiatimes.com/topic/HIV/news/8

The extracted CSV file is shown below.



	A	B	
1	web-scraper-order	web-scraper-start-url	title_link
2	1543754725-973	https://timesofindia.indiatimes.com/topic/HIV/news	HIV positive woman marries twice, kills first husband with the help of second 23 Feb 2018In a sensational incident from Bhadohi district of Uttar Pradesh, an HIV infected woman (name withheld to protect identity) I
3	1543756303-1845	https://timesofindia.indiatimes.com/topic/HIV/news	Why Narasinghpur sees most accident deaths every year 03 Sep 2018
4	1543755442-1512	https://timesofindia.indiatimes.com/topic/HIV/news	Section 377: After long battle, it'll be another fight for right to marry or adopt 11 Jul 2018
5	1543755696-1788	https://timesofindia.indiatimes.com/topic/HIV/news	An HIV-positive dream hits house hurdle in Kolkata 04 Mar 2018
6	1543754813-1010	https://timesofindia.indiatimes.com/topic/HIV/news	NHRC notice to UP govt over spread of HIV among prisoners in jails 07 Mar 2018
7	1543754098-641	https://timesofindia.indiatimes.com/topic/HIV/news	NGO gives helping hand, 200 patients get free treatment 26 Dec 2017
8	1543754958-1117	https://timesofindia.indiatimes.com/topic/HIV/news	School life gets a little easier for HIV students 07 Aug 2017State board students up to Class X, who stay with HIV-positive parents or carry the virus themselves, will now be classifiec
9	1543754725-957	https://timesofindia.indiatimes.com/topic/HIV/news	HIV positive woman marries twice, kills first husband with the help of second 23 Feb 2018In a sensational incident from Bhadohi district of Uttar Pradesh, an HIV infected woman (name withheld to protect identity) I
10	1543754954-1107	https://timesofindia.indiatimes.com/topic/HIV/news	China approves first home made anti-HIV drug 06 Jun 2018China has approved the first domestically developed, long-acting injectable HIV drug Albuvirtide which could be a boon to te
11	1543756182-1836	https://timesofindia.indiatimes.com/topic/HIV/news	On medicines for 16 years, youth stresses on family support 03 Sep 2018
12	1543755176-1226	https://timesofindia.indiatimes.com/topic/HIV/news	Man jabbed syringe with HIV+ mother's blood into his wife 20 Sep 2017
13	1543753977-591	https://timesofindia.indiatimes.com/topic/HIV/news	HIV due to transfusion: Kid tests negative in Chennai 03 Nov 2017However, reliable sources from the RCC said the status of the disease can be confirmed only after the final result is obtaine
14	1543753973-585	https://timesofindia.indiatimes.com/topic/HIV/news	Women's panel shifts HIV positive woman, infant to Gandhi Bhavan 06 Dec 2017An HIV positive woman, who delivered a baby at SAT hospital last week, was shifted to Gandhi Bhavan in Pathanapuram c
15	1543755528-1636	https://timesofindia.indiatimes.com/topic/HIV/news	Private Blood bank sealed in Jalandhar amid serious irregularities 04 Aug 2018
16	1543755315-1289	https://timesofindia.indiatimes.com/topic/HIV/news	62% jump in HIV/AIDS deaths in Mumbai in 1 year: Health ministry 19 May 2018
17	1543755342-1343	https://timesofindia.indiatimes.com/topic/HIV/news	AIDS control trust to reach out to HIV+ blood donors 07 Jun 2018

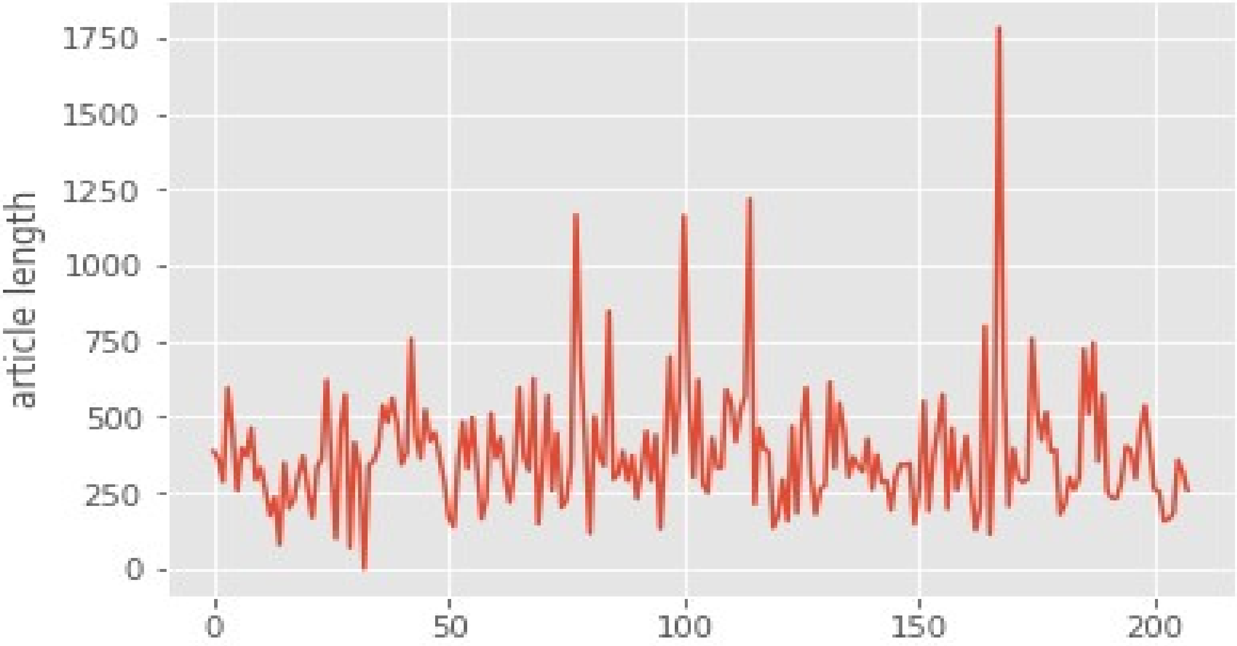
After this step I extracted the articles using BeautifulSoup Web scraping module in Python.



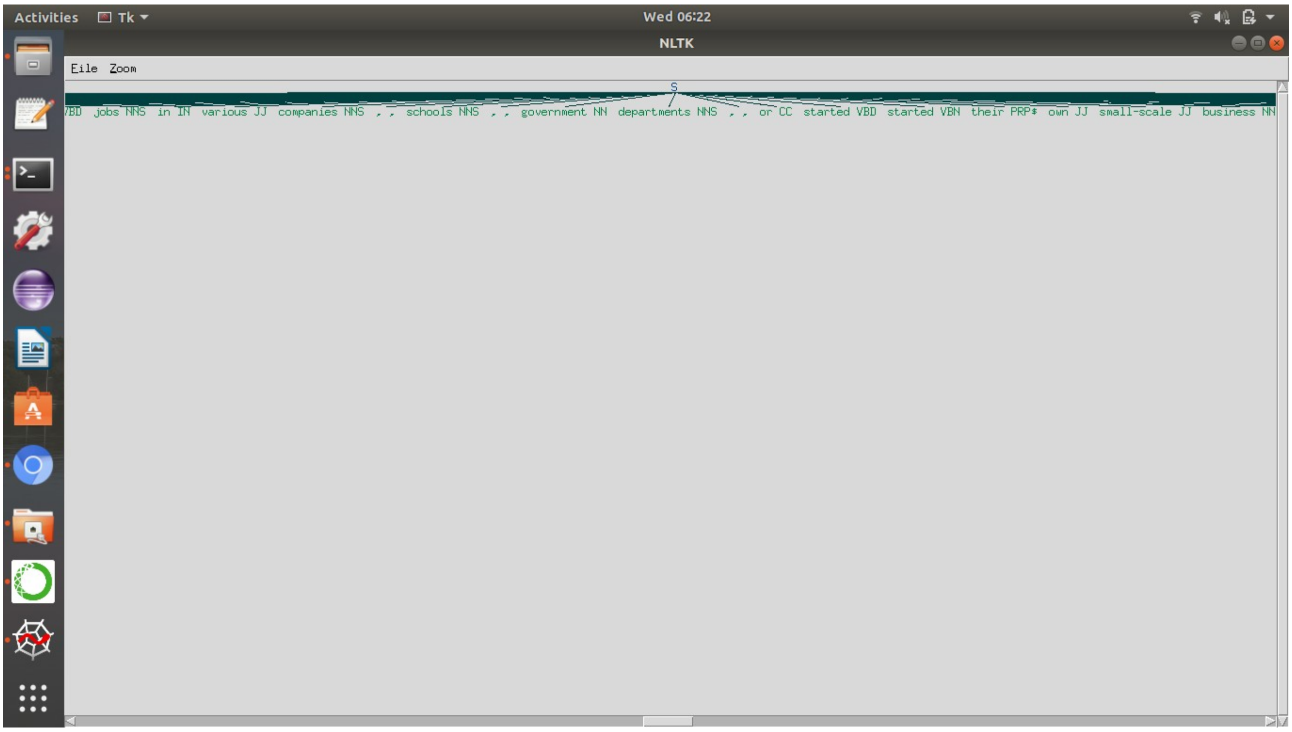
'22% boys, 29% girls in Bihar hav…
'85% Raj rural women d…
1 therapy centre for 10k AIDS …
2 patients allege Davanag…
5,000 people living wit…
8% of TB patients in Karnatak…
18 preliminary HIV+ cas…
20,500 got HIV through …
20 found HIV+ as quack tre…
23 inmates in Gorakhp…
25 acquire HIV as quack us…
25 with HIV at Kanpur ART centre
29,444 people living wit…
38 youth from Ludhiana …
62% jump in HIV-AIDS deaths in …
62% jump in HIV-AIDS deaths in …
459 HIV patients, 10 death…
Aadhaar, bank informati…
Aadhaar helps identify r…
Abbott and YRGCARE…
Activist seeks details o…
Adultery by hubbies leaves w…
After blood transfusi…
After WHO approval, India to e…
AIDS control trust to r…
'AIDS is no longer a death se…
AIDS patient hangs her…
Alarm: Biomed stockpile…
Ambulances to ferry HIV posti…
An 'Austrian home' fo…
An HIV-positive dream hi…
Anti-Trafficking Bill receiv…
ART data shows drop in H…
At rights camp, over 5,000 she…
At Sion hospital, 'palatabl…
Awareness, action needed f…
Being positive in the face …
Blood bank sealed in Jalandha…
Blood banks run out of HI…
Blood test before marriage…
'Call me by Your Name' se…
Can I get an STD even from 'rec…
Can you breastfeed another's…
Case that changed HIV treat…
Cell phone-based tech to allow …
'Challenge is to get HIV+ cas…
Charlie Sheen hits back at e…
Charlie Sheen: Struggle…
China approves first hom…
Chinese cancer warriors …
Cipla gets USFDA approval …
Clinics of 73 quacks sealed in …
Condoms distributed on eve of …
Congress workers held for …
Contest of one-min films aim…
Counselling centre for mental h…
'Denied treatment', HIV+ man…
Depression drives Ludhiana …
Despite HIV, they're positivev…
Despite meds, Maharas…
Devices, disposables & consu…
Doctors delaying treatmen…
'Don't advertise HIV-posit…
Don't believe I have 'ma…
Drastic reduction in mothe…
Drawing youth attention…

2. Analysing data

The average word lenght of articles is 383 words with a maximum of 1782 words. The following graph shows an analysis on word length of articles.

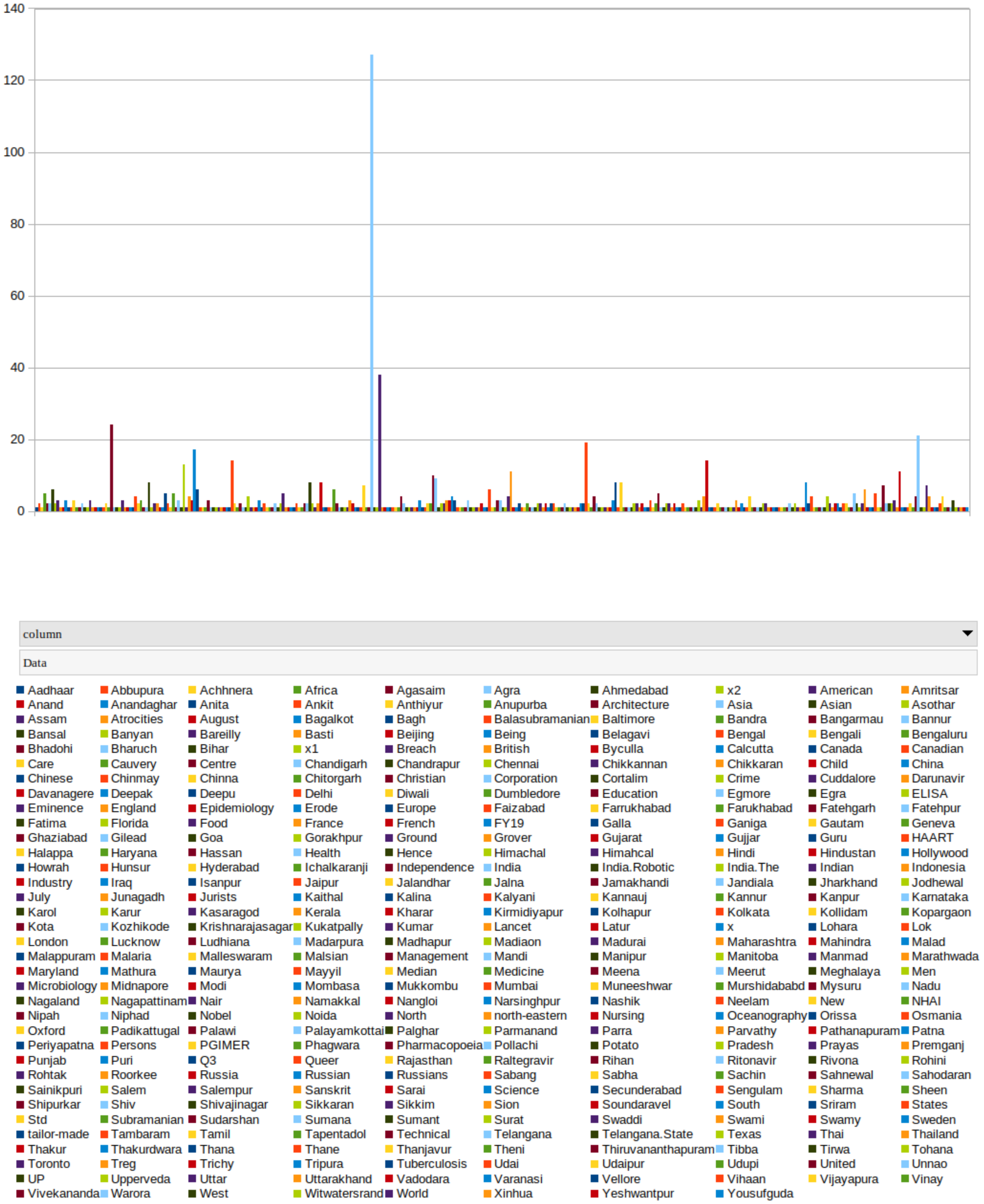


The NLTK tree for an article is shown below.



```
                                                  S
VBD  jobs NNS  in IN  various JJ  companies NNS  , ,  schools NNS  , ,  government NN  departments NNS  , ,  or CC  started VBD  started VBN  their PRP$  own JJ  small-scale JJ  business NN
```

Frequency distribution for city names mentioned in articles with their legends.



These named entities are extracted using the IO_tag and parts of speech(POS) tag in NLTK.

3. Modules Used:

1. BeautifulSoup4 Module(bs4)
2. Pandas
3. Requests
4. JSON Module
5. Natural Language Toolkit (NLTK)
6. os Module
7. Matplotlib

4. Python Scripts

For Data Analysis

```
import requests
import pandas as pd
import bs4
import json

url_list=[]

field=["title_link-href"]

data=pd.read_csv(r"/home/arushi/
news_toi.csv",skipinitialspace=True,usecols=field,index_col=None)

list1=data.iloc[:,0].tolist()

#print(dic[:30])


title=[]
d_time=[]
content=[]

print(len(list1))
print("\n\n")



for i in range(len(list1)):
    res=requests.get(list1[i])

    print(i)
    print("------****res.text length****----\n")
    print(len(res.text))
    print("\n")
    soup=bs4.BeautifulSoup(res.text)
    ele_text=soup.select("div .Normal")
    ele_date=soup.select("time")
    ele_title=soup.select("arttitle")
    #print(ele_date)
    #print(ele_title)
   # print(type(ele_text))
    print(ele_date[0].attrs)
    print(ele_title[0].getText())
    print(ele_text[0].getText())
    print("--------new article------\n\n")
    #append lists
    title.append(ele_title[0].getText())
    d_time.append(ele_date[0].attrs)
    content.append(ele_text[0].getText())
    x=title[i-1152]
    if "HIV/AIDS" in x:
        y=x.replace("HIV/AIDS","HIV-AIDS")
    toi2=open(r"/home/arushi/toi_news_articles/%s"%y,"w")
    toi2.write(json.dumps(d_time[i])+"\n")
    toi2.write(content[i])



print(len(title))
print(title)
print("\n")

print(len(d_time))
print(d_time)
print("\n")

print(len(content))
print(content)
print("\n")

new_file=[title,d_time,content]
zip(*new_file)
print(zip(*new_file))

for t,d,c in zip(*new_file):
    print(t)
    print(d)
    print(c)


x=title[0]
print(content[0])
toi2=open(r"/home/arushi/toi_news_articles/%s"%x,"w")
toi2.write(json.dumps(d_time[0])+"\n")
toi2.write(content[0])
```

```
#str1=data[1:1]
#print(str1+"hello")




res=requests.get(r"https://timesofindia.indiatimes.com/india/hiv-positive-women-marries-twice-kills-first-husband-with-the-help-of-second/articleshow/63043159.cms")

print(type(res))

soup=bs4.BeautifulSoup(res.text)
print(type(soup))


ele_text=soup.select("div .Normal")
ele_date=soup.select("time")
print(ele_date)
ele_title=soup.select("arttitle")
print(ele_title)

print(type(ele_text))

print(ele_text[0].getText())


import nltk
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk import ne_chunk, pos_tag
import os
import matplotlib.pyplot as plt

plt.style.use('ggplot')
from nltk import conlltags2tree, tree2conlltags


#from spacy.en import English



#new fuction definition
def entities(text):
    return ne_chunk(
            pos_tag(
                    word_tokenize(text)))

#nltk.download('words')
```

```
#nltk.download('punkt')
#nltk.download('maxent_ne_chunker')
#nltk.download('averaged_perceptron_tagger')

path=r"/home/arushi/toi_news_articles"

len_art=[]
city_name=[]

for filename in os.listdir(path):
    print(filename)
    toi2=open(r"/home/arushi/toi_news_articles/"+filename,"r")
    data=toi2.read().replace('\n', '')
    #len_art.append(len(data.split()))
    words=word_tokenize(data)

    #print(nltk.pos_tag(words))

    tree=entities(data)
    iob_tags = tree2conlltags(tree)
    #print(iob_tags)

    for tup in iob_tags:
        if(tup[2]=="B-GPE" or tup[2]=="O_GPE" or tup[2]=="I-GPE"):
            city_name.append(tup[0])

    #print(tree)
    #tree.draw()



print(city_name)

import pandas as pd
df = pd.DataFrame(city_name, columns=["colummn"])
df.to_csv('city_list.csv', index=False)


city_set=set(city_name)
word_tag_fd=nltk.FreqDist(words)


plt.hist(city_name)

fig_size = plt.rcParams["figure.figsize"]
fig_size[0] = 12
fig_size[1] = 9
plt.rcParams["figure.figsize"] = fig_size
plt.xticks(city_set)
plt.show()
```

```
'''
print(len_art)
plt.plot(len_art)
plt.ylabel('article length')
plt.show()

print(max(len_art))
print(min(len_art))
print(sum(len_art)/len(len_art))
'''
```

Neural Network Model

```
import pandas as pd
import numpy as np
import pickle
from keras.preprocessing.text import Tokenizer
from keras.models import Sequential
from keras.layers import Activation, Dense, Dropout
from sklearn.preprocessing import LabelBinarizer
import sklearn.datasets as skds
from pathlib import Path

# For reproducibility
np.random.seed(1237)

# Source file directory
path_train = "/home/arushi/toi_news_articles/'85% Raj rural women don’t know about HIV-AIDS'"

files_train = skds.load_files(path_train,load_content=False)

label_index = files_train.target
label_names = files_train.target_names
labelled_files = files_train.filenames

data_tags = ["filename","category","news"]
data_list = []

# Read and add data from file to a list
i=0
for f in labelled_fmodel = Sequential()
model.add(Dense(512, input_shape=(vocab_size,)))
model.add(Activation('relu'))
model.add(Dropout(0.3))
model.add(Dense(512))
model.add(Activation('relu'))
model.add(Dropout(0.3))
model.add(Dense(num_labels))
model.add(Activation('softmax'))
model.summary()
model.compile(loss='categorical_crossentropy',
              optimizer='adam',
              metrics=['accuracy'])

history = model.fit(x_train, y_train,
                    batch_size=batch_size,
                    epochs=30,
                    verbose=1,
                    validation_split=0.1)iles:
    data_list.append((f,label_names[label_index[i]],Path(f).read_text()))
    i += 1

# We have training data available as dictionary filename, category, data
data = pd.DataFrame.from_records(data_list, columns=data_tags)

encoder = LabelBinarizer()
encoder.fit(train_tags)
y_train = encoder.transform(train_tags)
y_test = encoder.transform(test_tags)
```