ASSIGNMENT - CYBERBOXER

Q1. How can we transform a text to numerical representation? (Write any three of them with one-line description)

SOLUTION:-

• BAG OF WORDS: BOW works using text vectorization, it takes the count of occurences of a particular token of the text and each token will have a feature column. eg; a good movie(1110), not a good movie(1111).

• Integer Encoding & One hot Encoding: applicable for data with ordinal and non ordinal relationship, that represents categorical text data into binary vectors.

• Skip-gram: this model takes each and every word in a large focus word and one-by-one the words that surrounds it within a defined window is then then feed to a neural network that after training predicts the probability for each word to actually appear in the window around the focus word.

Q2. List three(min) things to evaluate a classification model.

SOLUTION:-

• Accuracy represents the proportion of correctly classified observations for a classification model.

• Confusion matrix a 2x2 table showing four parameters: the number of true positives (TP), true negatives (TN), false negatives (FN) and false positives (FP).

• Precision, Recall and Specificity, the major performance metrics that describes a predictive classification model.

• ROC curve and Area Under the Curve (AUC) to summarize the overall performance of the classifier.

• For the evaluation of overall efficacy of a classification test: Balanced Accuracy and Diagnostic Odd Ratio (DOR). DOR is a term taken from medical domain. It checks the overall efficacy of a classification test.

1. Develop a code for the following problem:

a. Create a crawler for any news portal which can extract news events/information. b. Save all the extracted information into a csv/excel file with file name:news*portal*(date or page no). c. Parameters to extract are as follows, Article Content, Date Posted, Tags, Author etc.

SOLUTION: I have considered news portal : sciencenews.org for making the crawler. I am extracting News Title, Link of the news, Author Name, Date of posting and Comtent of the articles. I have extracted the information in "sciencenews_all_data.csv" file.

In [1]:

```
#.................using libraby BeautifulSoup...................for news crawle
r.............................#
from bs4 import BeautifulSoup
import requests, csv

source = requests.get('https://www.sciencenews.org/').text
 # data extracted from science news .com
soup = BeautifulSoup(source)

#.........................for extracting data into cs
v.............................................................#
csv_file = open('sciencenews_all_data.csv', 'w',encoding="utf-8")
csv_writer = csv.writer(csv_file)

csv_writer.writerow(['Topic Name', 'Link', 'Author name', 'Date of Posting', 'Summary'
])

# ...............................................For DATA CRAWLER
 ..............................................#
for item in soup.findAll('h2', {'class' :'node-title' }):
            for link in item('a'):
                #title = "TITLE: " + item.string + "."
                title = item.string
                #print("LINK:-")
                href ="https://www.sciencenews.org/" + link.get('href')
                print(title)
                print(href)
                source = requests.get(href)
                text = source.text
                soup = BeautifulSoup(text)
                for link in soup.findAll('span' , {'itemprop' : 'name' }):
                    author = link.string
                    for lin in soup.findAll('div', {'class' : 'view view-article-relate
d-content view-id-article_related_content view-display-id-attachment_1'}):
                        for t in lin('span' , {'class' : 'field-content'}):
                            #time = "DATE & TIME OF POSTING:" + t.string
                            time = t.string
                    print(author)
                    print(time + "\n")
                for link in soup.findAll('span', {'itemprop':'description'}):
                    for i in link('p'):
                        #a = "ABOUT :- \n" + i.text
                        a = i.text
                        #print(a)
            csv_writer.writerow([title, href, author, time, a])
csv_file.close()
print('Scraping Done!')
```

Exploding stars scattered traces of iron over Antarctic snow
https://www.sciencenews.org//article/exploding-stars-scattered-traces-iron-over-antarctic-snow
Emily Conover
6:00am, August 9, 2019

How these tiny insect larvae leap without legs
https://www.sciencenews.org//article/how-these-tiny-insect-larvae-leap-without-legs
Susan Milius
6:20pm, August 8, 2019

The worst wildfires can send smoke high enough to affect the ozone layer
https://www.sciencenews.org//article/worst-wildfires-can-send-smoke-high-enough-affect-ozone-layer
Megan Sever
2:00pm, August 8, 2019

How pieces of live human brain are helping scientists map nerve cells
https://www.sciencenews.org//article/experiment-live-human-brain-helps-scientists-map-nerve-cells
Laura Sanders
6:00am, August 7, 2019

50 years ago, Fermilab turned to bubbles
https://www.sciencenews.org//article/50-years-ago-fermilab-turned-bubbles
Bethany Brookshire
8:00am, August 8, 2019

With nowhere to hide from rising seas, Boston prepares for a wetter future
https://www.sciencenews.org//article/boston-adapting-rising-sea-level-coastal-flooding
Mary Caperton Morton
6:00am, August 6, 2019

One in 4 people live in places at high risk of running out of water
https://www.sciencenews.org//article/one-4-people-live-places-high-risk-running-out-water
Carolyn Wilke
6:00am, August 8, 2019

Why people with celiac disease suffer so soon after eating gluten
https://www.sciencenews.org//article/why-people-celiac-disease-suffer-so-soon-after-eating-gluten
Tina Hesman Saey
2:00pm, August 7, 2019

Giant, active galaxies from the early universe may have finally been found
https://www.sciencenews.org//article/giant-active-galaxies-early-universe-may-have-finally-been-found
Maria Temming
1:00pm, August 7, 2019

Racist words and acts, like the El Paso shooting, harm children’s health
https://www.sciencenews.org//article/racism-words-acts-el-paso-shooting-harm-children-health-longterm
Aimee Cunningham
3:37pm, August 6, 2019

A fungus makes a chemical that neutralizes the stench of skunk spray
https://www.sciencenews.org//article/fungus-makes-chemical-neutralizes-ste

nch-skunk-spray
Carolyn Wilke
10:00am, August 6, 2019

How the 5 riskiest U.S. cities for coastal flooding are preparing for risi
ng tides
https://www.sciencenews.org//article/top-five-us-coastal-cities-risk-flood
ing-rising-sea-levels
Mary Caperton Morton
6:00am, August 6, 2019

Ancient Maya warfare flared up surprisingly early
https://www.sciencenews.org//article/ancient-maya-warfare-flared-surprisin
gly-early
Bruce Bower
11:00am, August 5, 2019

A new map is the best view yet of how fast Antarctica is shedding ice
https://www.sciencenews.org//article/new-map-best-view-yet-how-fast-antarc
tica-shedding-ice
Maria Temming
8:00am, August 5, 2019

Decades of dumping acid suggest acid rain may make trees thirstier
https://www.sciencenews.org//article/decades-dumping-acid-suggest-acid-rai
n-may-make-trees-thirstier
Carolyn Wilke
6:00am, August 5, 2019

Satellites are transforming how archaeologists study the past
https://www.sciencenews.org//article/space-satellites-transforming-how-arc
haeologists-study-past
Erin Wayman
8:00am, August 4, 2019

The Arctic is burning and Greenland is melting, thanks to record heat
https://www.sciencenews.org//article/arctic-burning-greenland-melting-than
ks-record-heat
Carolyn Gramling
3:52pm, August 2, 2019

Hospitalizations highlight potential dangers of e-cigs to teens’ lungs
https://www.sciencenews.org//article/hospitalizations-highlight-potential-
dangers-e-cigs-teen-lungs
Aimee Cunningham
3:09pm, August 2, 2019

Stars may keep spinning fast, long into old age
https://www.sciencenews.org//article/stars-may-keep-spinning-fast-long-old
-age
Lisa Grossman
6:00am, August 2, 2019

Public trust that scientists work for the good of society is growing
https://www.sciencenews.org//article/public-trust-scientists-work-good-soc
iety-growing
Katy Daigle
10:45am, August 2, 2019

A new study challenges the idea that the placenta has a microbiome
https://www.sciencenews.org//article/new-study-challenges-idea-placenta-mi

```
crobiome-bacteria
Laura Sanders
1:00pm, July 31, 2019

A 3-D map of stars reveals the Milky Way’s warped shape
https://www.sciencenews.org//article/3-d-map-stars-reveals-milky-way-warpe
d-shape
Emily Conover
2:00pm, August 1, 2019

There’s more to pufferfish than that goofy spiked balloon
https://www.sciencenews.org//article/pufferfish-biology-mating-goofy-spike
d-balloon
Susan Milius
12:07pm, August 1, 2019

Monkeys can use basic logic to decipher the order of items in a list
https://www.sciencenews.org//article/monkeys-can-use-basic-logic-decipher-
order-items-list
Bruce Bower
2:03pm, July 31, 2019

Scraping Done!
```

1. Perform exploratory data analysis on the csv file you have created in previous question, retrieve named entities from articles as well.(EDA is a broad term.Example:- word count, topic modelling etc .Extra points for good analysis.)

SOLUTION:-

Statistical Features analysis from the data

1. head()
2. describe()
3. value_counts()
4. info()
5. isnull()
6. word_count
7. char_count()
8. word_density()

Exploratory Analysis

1. WordCloud Visualization
2. Semantic Analysis: Polarity & Subjectivity

In [9]:

```
#.....................................ED
A................................................#
import sys
import matplotlib.pyplot as plt
%matplotlib inline
import pandas as pd
df = pd.read_csv('sciencenews_all_data.csv')
# head() : it displays the first 5 rows, first 5 index values, of every column
df.head()
```

Out[9]:

	Topic Name	Link	Author name	Date of Posting	Summary
0	Exploding stars scattered traces of iron over ...	https://www.sciencenews.org//article/exploding...	Emily Conover	6:00am, August 9, 2019	“This is actually quite a profound thing,” say...
1	How these tiny insect larvae leap without legs	https://www.sciencenews.org//article/how-these...	Susan Milius	6:20pm, August 8, 2019	Poppinga and colleagues recently showed that C...
2	The worst wildfires can send smoke high enough...	https://www.sciencenews.org//article/worst-wil...	Megan Sever	2:00pm, August 8, 2019	Given that climate change is increasing fire f...
3	How pieces of live human brain are helping sci...	https://www.sciencenews.org//article/experimen...	Laura Sanders	6:00am, August 7, 2019	This article appears in the August 17, 2019 is...
4	50 years ago, Fermilab turned to bubbles	https://www.sciencenews.org//article/50-years-...	Bethany Brookshire	8:00am, August 8, 2019	NAL was renamed Fermilab in 1974 for physicist...

In [10]:

```
df.describe()
# describe() : view basic statistical details like count, frequency, etc.
```

Out[10]:

	Topic Name	Link	Author name	Date of Posting	Summa
count	24	24	24	24	2
unique	24	24	15	23	2
top	Satellites are transforming how archaeologists...	https://www.sciencenews.org//article/exploding...	Carolyn Wilke	6:00am, August 6, 2019	In 202 NAS and tl Indi Spa Resear Or
freq	1	1	3	2	

In [11]:

```
df['Author name'].value_counts()
# value_counts(): displays the number of times each specific value in a data frame is p
resent in descending order
# ANALYSIS: Certains auythors have contributed more than one news article
```

Out[11]:

```
Carolyn Wilke           3
Laura Sanders           2
Bruce Bower             2
Mary Caperton Morton    2
Susan Milius            2
Emily Conover           2
Maria Temming           2
Aimee Cunningham        2
Bethany Brookshire      1
Erin Wayman             1
Megan Sever             1
Carolyn Gramling        1
Lisa Grossman           1
Tina Hesman Saey        1
Katy Daigle             1
Name: Author name, dtype: int64
```

In [12]:

```
df['Link'].value_counts()
#ANALYSIS : Unique links
```

Out[12]:

```
https://www.sciencenews.org//article/exploding-stars-scattered-traces-iron
-over-antarctic-snow           1
https://www.sciencenews.org//article/giant-active-galaxies-early-universe-
may-have-finally-been-found    1
https://www.sciencenews.org//article/hospitalizations-highlight-potential-
dangers-e-cigs-teen-lungs      1
https://www.sciencenews.org//article/stars-may-keep-spinning-fast-long-old
-age                           1
https://www.sciencenews.org//article/decades-dumping-acid-suggest-acid-rai
n-may-make-trees-thirstier     1
https://www.sciencenews.org//article/pufferfish-biology-mating-goofy-spike
d-balloon                      1
https://www.sciencenews.org//article/why-people-celiac-disease-suffer-so-s
oon-after-eating-gluten        1
https://www.sciencenews.org//article/50-years-ago-fermilab-turned-bubbles
1
https://www.sciencenews.org//article/fungus-makes-chemical-neutralizes-ste
nch-skunk-spray                1
https://www.sciencenews.org//article/space-satellites-transforming-how-arc
haeologists-study-past         1
https://www.sciencenews.org//article/racism-words-acts-el-paso-shooting-ha
rm-children-health-longterm    1
https://www.sciencenews.org//article/one-4-people-live-places-high-risk-ru
nning-out-water                1
https://www.sciencenews.org//article/monkeys-can-use-basic-logic-decipher-
order-items-list               1
https://www.sciencenews.org//article/new-map-best-view-yet-how-fast-antarc
tica-shedding-ice              1
https://www.sciencenews.org//article/ancient-maya-warfare-flared-surprisin
gly-early                      1
https://www.sciencenews.org//article/arctic-burning-greenland-melting-than
ks-record-heat                 1
https://www.sciencenews.org//article/experiment-live-human-brain-helps-sci
entists-map-nerve-cells        1
https://www.sciencenews.org//article/3-d-map-stars-reveals-milky-way-warpe
d-shape                        1
https://www.sciencenews.org//article/worst-wildfires-can-send-smoke-high-e
nough-affect-ozone-layer       1
https://www.sciencenews.org//article/public-trust-scientists-work-good-soc
iety-growing                   1
https://www.sciencenews.org//article/top-five-us-coastal-cities-risk-flood
ing-rising-sea-levels          1
https://www.sciencenews.org//article/new-study-challenges-idea-placenta-mi
crobiome-bacteria              1
https://www.sciencenews.org//article/boston-adapting-rising-sea-level-coas
tal-flooding                   1
https://www.sciencenews.org//article/how-these-tiny-insect-larvae-leap-wit
hout-legs                      1
Name: Link, dtype: int64
```

In [13]:

```
df['Date of Posting'].value_counts()
# ANALYSIS: two articles were posted at exactly same time
```

Out[13]:

```
6:00am, August 6, 2019     2
10:00am, August 6, 2019    1
8:00am, August 8, 2019     1
6:00am, August 8, 2019     1
2:00pm, August 1, 2019     1
6:00am, August 9, 2019     1
3:09pm, August 2, 2019     1
3:37pm, August 6, 2019     1
6:00am, August 2, 2019     1
10:45am, August 2, 2019    1
12:07pm, August 1, 2019    1
3:52pm, August 2, 2019     1
6:00am, August 7, 2019     1
8:00am, August 4, 2019     1
8:00am, August 5, 2019     1
11:00am, August 5, 2019    1
2:00pm, August 8, 2019     1
1:00pm, July 31, 2019      1
2:00pm, August 7, 2019     1
6:20pm, August 8, 2019     1
1:00pm, August 7, 2019     1
6:00am, August 5, 2019     1
2:03pm, July 31, 2019      1
Name: Date of Posting, dtype: int64
```

In [14]:

```
df['Topic Name'].value_counts()
# ANALYSIS: unique article topics
```

Out[14]:

```
Satellites are transforming how archaeologists study the past
1
Racist words and acts, like the El Paso shooting, harm children’s health
1
Hospitalizations highlight potential dangers of e-cigs to teens’ lungs
1
A new map is the best view yet of how fast Antarctica is shedding ice
1
A new study challenges the idea that the placenta has a microbiome
1
Giant, active galaxies from the early universe may have finally been found
1
50 years ago, Fermilab turned to bubbles
1
Exploding stars scattered traces of iron over Antarctic snow
1
The Arctic is burning and Greenland is melting, thanks to record heat
1
There’s more to pufferfish than that goofy spiked balloon
1
The worst wildfires can send smoke high enough to affect the ozone layer
1
Monkeys can use basic logic to decipher the order of items in a list
1
With nowhere to hide from rising seas, Boston prepares for a wetter future
1
Why people with celiac disease suffer so soon after eating gluten
1
How pieces of live human brain are helping scientists map nerve cells
1
A fungus makes a chemical that neutralizes the stench of skunk spray
1
Stars may keep spinning fast, long into old age
1
Public trust that scientists work for the good of society is growing
1
One in 4 people live in places at high risk of running out of water
1
Decades of dumping acid suggest acid rain may make trees thirstier
1
How the 5 riskiest U.S. cities for coastal flooding are preparing for risi
ng tides    1
A 3-D map of stars reveals the Milky Way’s warped shape
1
Ancient Maya warfare flared up surprisingly early
1
How these tiny insect larvae leap without legs
1
Name: Topic Name, dtype: int64
```

In [15]:

```
df['Summary'].value_counts()
# ANALYSIS : no repeated context present
```

Out[15]:

```
In 2021, NASA and the Indian Space Research Organization plan to launch a
satellite that will gather enough data to update this map every few months
– allowing scientists to better monitor how ice flow across Antarctica cha
nges as the climate changes.
1
That’s probably a valuable ability in the wild, she says, because many ani
mals need to monitor where group mates stand in the social pecking order.
“An ability to construct, retain, manipulate and reference ordered informa
tion may be an evolutionarily ancient, efficient [mental] mechanism for ke
eping track of relationships between individuals,” she says.
1
“Trust is important to legitimacy, credibility and effectiveness,” Boykoff
says. “Without trust, scientists would just be screaming into the wind.”
1
Buy Archaeology from Space from Amazon.com. Science News is a participant
in the Amazon Services LLC Associates Program. Please see our FAQ for more
details.
1
This article appears in the August 17, 2019 issue of Science News with the
headline, "A Menagerie of Neurons: Studies of living brain cells aim to de
termine what sets humans apart."
1
Unfortunately, the result might mean that astronomers can’t use stars’ spi
n speeds to guess ages anymore. “If that stops working in old stars, tha
t’s a bummer,” Curtis says.
1
The United States is considered to have relatively low risk; overall, it u
ses less than 20 percent of its available water. However, some western sta
tes including California, Arizona, New Mexico, Colorado and Nebraska typic
ally use 40 percent or more of current water supplies each year.
1
Given that climate change is increasing fire frequency and intensity in so
me places like the North American West (SN: 12/22/18, p. 18), we can proba
bly expect to see more of these fire clouds reaching the stratosphere, Fro
mm says. But, he cautions, “we are still on the learning curve when it com
es to understanding pyroCbs.”
1
With neighborhood-level projections for future sea level rise in hand, the
city of Boston has district-level projects completed for East Boston, Char
lestown and South Boston. A deployable flood wall is being installed along
the East Boston Greenway and a section of Main Street in Charlestown is be
ing elevated to protect the adjacent neighborhood. In several areas, inclu
ding around South Boston and the Seaport, concrete is being removed and re
placed by floodable parks and green space. Mayor Martin Walsh has pledged
10 percent of the city’s $3.49 billion capital budget in 2020 for such res
iliency projects.
1
3-D VISION  The Milky Way’s Cepheid stars are plotted in three dimensions,
revealing the galaxy’s warped shape. Unlike other stars, Cepheids vary in
brightness in a particular way that helps scientists make more precise est
imates of their distances from Earth. Brighter colors represent Cepheids c
loser to the warped plane of the galaxy, indicated by the grid. The star i
con indicates the sun.
1
Knowing that certain T cells, and cytokines in particular, cause celiac sy
mptoms may lead to therapies that could block the gluten-reacting T cells,
Anderson says. And doctors may be able to diagnose celiac disease by measu
ring IL-2 levels in the blood, sparing patients the need for tests in whic
h they’re repeatedly given gluten.
```

1
Meanwhile, increasingly frequent winter warm spells, insect outbreaks and wildfires have also caused many Arctic plants to lose their resistance to freezing, dry out and die, turning large parts of the Arctic brown (SN: 4/13/19, p. 16). That, in turn, increases the region's susceptibility to more wildfires: Normally, the icy peatlands are soggy enough to be fire-resistant, but they are thawing and drying out. Once set ablaze, the carbon-rich peat can burn for months, releasing large amounts of CO_2 back into the atmosphere and fueling the warming feedback loop (SN: 3/17/18, p. 20).
1
In the intimidating body part catalog, pufferfishes are perhaps best known for turning into spiky balls when outraged. These spines perk upright when puffers gulp water to balloon out their abdomens. Some of the same gene networks that put feathers on birds and hairs on mammals turn out to put the protective spines on puffers, Fraser and colleagues report July 25 in iScience. Those spines have evolved from the scales that covered distant fish ancestors. But between today's skinny spines, modern pufferfishes are totally naked. Try not to stare.
1
Trent: As pediatricians, we will be there to help families with these discussions and both the direct and indirect trauma that these events have caused. We will also continue our advocacy efforts to encourage our government leaders to adopt policies that broadly address gun violence and change the climate of racism impacting children, adolescents and families.
1
The investigation of the Wisconsin teens could provide some answers that will aid research. More details about the teens' e-cigarette use, such as the type of device, the e-liquid, the flavors, how much they vaped and so on, "would be very helpful in trying to understand what's going on and who else might be at risk," Crotty Alexander says.
1
This article appears in the August 17, 2019 issue of Science News with the headline, "Wicked High Tides: Boston is taking action to adapt to sea level rise."
1
"This is actually quite a profound thing," says astrophysicist Brian Fields of the University of Illinois at Urbana-Champaign, who was not involved with the research. "It's telling us about the recent history of our whole neighborhood in the galaxy and about the lives and deaths of massive stars."
1
Aagaard is convinced there are small amounts of bacteria in the placenta, but remains unsure about what biological role those microbes play, if any.
1
NAL was renamed Fermilab in 1974 for physicist Enrico Fermi. The lab's first accelerator produced protons in April 1969, and was shooting subatomic particles into a 76-centimeter bubble chamber filled with liquid hydrogen by 1972. Such chambers track bubble trails left by speeding particles. The lab began upgrading to a 4.5-meter chamber detector in 1973, which helped in the study of neutrinos and turned up evidence for bottom and top quarks. As accelerators modernized, bubble detectors were phased out, and Fermilab's chamber became an art installation. But SNOLAB's bubble chamber in Sudbury, Canada, still searches for weakly interacting massive particles, or WIMPS — a proposed type of dark matter. 1
Poppinga and colleagues recently showed that Chinese witch hazel trees build up forces in the mature fruit that suddenly shoot out a seed rotating a bit like a bullet from a rifle. Unlike gall midge launches, though, these tree latches break when they let go. The leap of a legless seed is fast and dramatic, but it's not repeatable.
1
Adding several common cosmetic ingredients also sped up pericosine's abili

```
ty to cut the skunk spray smell, the team found. That “was really thrillin
g,” Cichewicz says. “This now looks a lot more like a personal-care produc
t than it does an organic chemistry reaction.”
1
Prior to 800, Maya people may have considered it dishonorable to kill or w
ound others from a distance, Graham suspects. Classic Maya culture probabl
y discouraged killing large numbers of opponents in battle with any type o
f weapon, since no mass burials of war victims have been found, Inomata sa
ys.
1
Wang and colleagues now plan to take a larger census of ancient massive ga
laxies with ALMA. That work could give theorists more information about ho
w to tweak cosmological simulations to match early-universe observations.
1
Soils are typically slow to recover calcium they’ve lost, so the study may
also point to legacy effects of acid rain that we didn’t already know abou
t, says Charles Driscoll, a biogeochemist at Syracuse University in New Yo
rk who was not involved in the study.
1
Name: Summary, dtype: int64
```

In [16]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24 entries, 0 to 23
Data columns (total 5 columns):
Topic Name         24 non-null object
Link               24 non-null object
Author name        24 non-null object
Date of Posting    24 non-null object
Summary            24 non-null object
dtypes: object(5)
memory usage: 1.1+ KB
```

In [17]:

```
df.isnull()
# ANALYSIS: no missing values
```

Out[17]:

	Topic Name	Link	Author name	Date of Posting	Summary
0	False	False	False	False	False
1	False	False	False	False	False
2	False	False	False	False	False
3	False	False	False	False	False
4	False	False	False	False	False
5	False	False	False	False	False
6	False	False	False	False	False
7	False	False	False	False	False
8	False	False	False	False	False
9	False	False	False	False	False
10	False	False	False	False	False
11	False	False	False	False	False
12	False	False	False	False	False
13	False	False	False	False	False
14	False	False	False	False	False
15	False	False	False	False	False
16	False	False	False	False	False
17	False	False	False	False	False
18	False	False	False	False	False
19	False	False	False	False	False
20	False	False	False	False	False
21	False	False	False	False	False
22	False	False	False	False	False
23	False	False	False	False	False

In [163]:

```
##.................................statistical count of feature
s.......................##

df['word_count'] = df['Topic Name'].apply(lambda x : len(x.split()))
df['char_count'] = df['Topic Name'].apply(lambda x : len(x.replace(" ","")))
df['word_density'] = df['word_count'] / (df['char_count'] + 1)
print(df[['word_count', 'char_count', 'word_density']].head(24))
# calcuting word count, char count and word density for Topic Name.
# It can be further analysed by caluating average of word density or word count or char
count.
# maximum to limit of words/ character used for TOPIC NAME
```

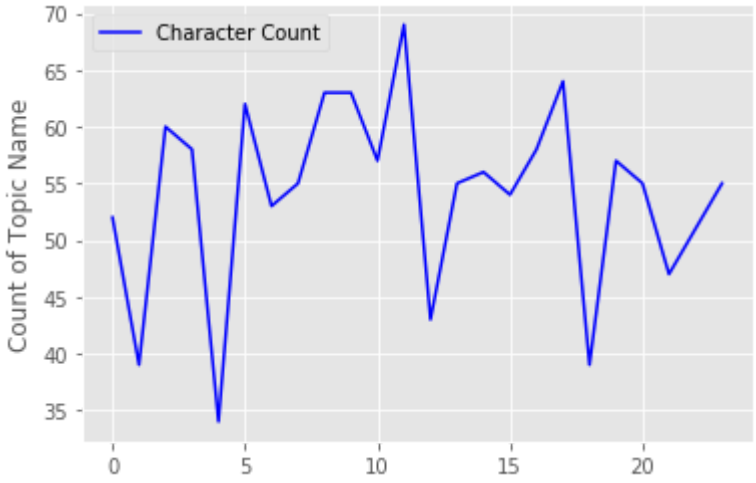
```
    word_count  char_count  word_density
0            9          52      0.169811
1            8          39      0.200000
2           13          60      0.213115
3           12          58      0.203390
4            7          34      0.200000
5           13          62      0.206349
6           15          53      0.277778
7           11          55      0.196429
8           12          63      0.187500
9           12          63      0.187500
10          12          57      0.206897
11          14          69      0.200000
12           7          43      0.159091
13          15          55      0.267857
14          11          56      0.192982
15           8          54      0.145455
16          12          58      0.203390
17           9          64      0.138462
18           9          39      0.225000
19          12          57      0.206897
20          12          55      0.214286
21          11          47      0.229167
22           9          51      0.173077
23          14          55      0.250000
```

In [164]:

```
%matplotlib inline
import matplotlib.pyplot as plt

plt.plot(df['char_count'] , 'b-', label='Character Count')
plt.legend(loc='upper left')
plt.ylabel('Count of Topic Name')

# charcter count min: less than 35 characters, max: more than 65 characters for TOPIC N
AME
```

Out[164]:

```
Text(0, 0.5, 'Count of Topic Name')
```

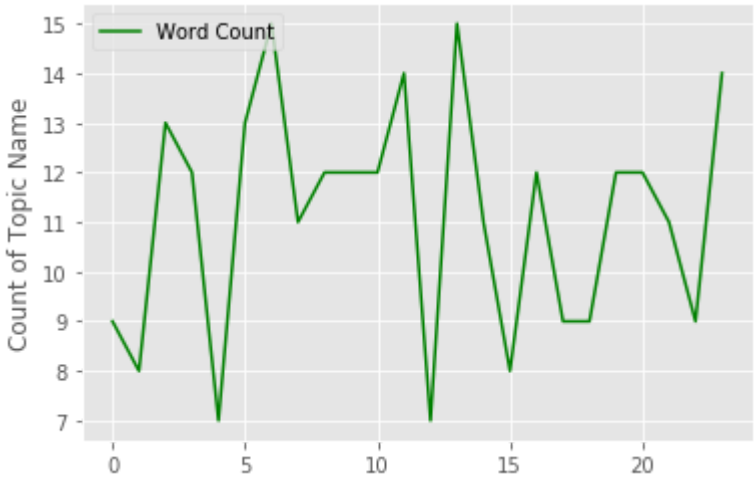


In [166]:

```
plt.plot(df['word_count'], 'g-', label='Word Count')
plt.legend(loc='upper left')
plt.ylabel('Count of Topic Name')
# word count min: 7 words, max: 15 words for TOPIC NAME
```

Out[166]:

```
Text(0, 0.5, 'Count of Topic Name')
```



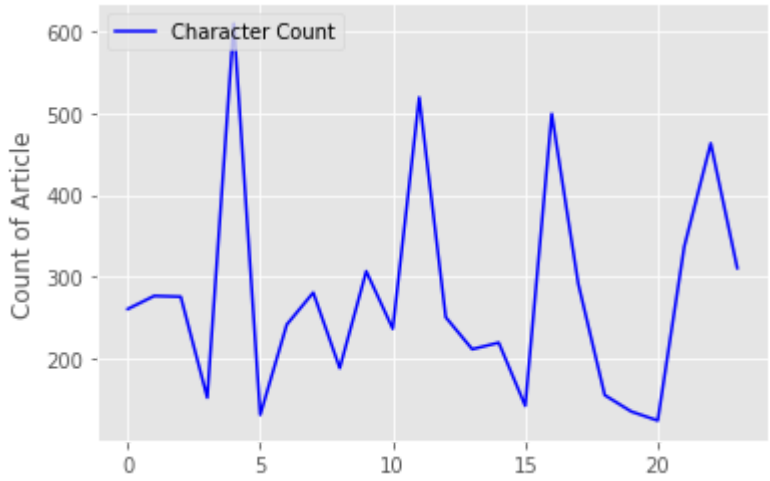
In [167]:

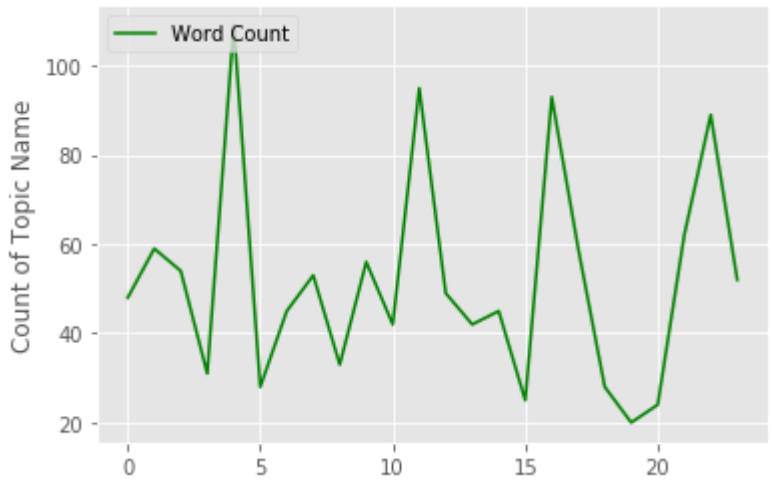
```
df['word_count'] = df['Summary'].apply(lambda x : len(x.split()))
df['char_count'] = df['Summary'].apply(lambda x : len(x.replace(" ","")))
df['word_density'] = df['word_count'] / (df['char_count'] + 1)

print(df[['word_count', 'char_count', 'word_density']].head(24))
```

```
    word_count  char_count  word_density
0           48         261      0.183206
1           59         277      0.212230
2           54         276      0.194946
3           31         153      0.201299
4          109         609      0.178689
5           28         132      0.210526
6           45         242      0.185185
7           53         281      0.187943
8           33         189      0.173684
9           56         307      0.181818
10          42         237      0.176471
11          95         519      0.182692
12          49         251      0.194444
13          42         212      0.197183
14          45         220      0.203620
15          25         143      0.173611
16          93         499      0.186000
17          59         292      0.201365
18          28         156      0.178344
19          20         136      0.145985
20          24         125      0.190476
21          62         337      0.183432
22          89         463      0.191810
23          52         311      0.166667
```

In [157]:

```
plt.plot(df['char_count'] , 'b-', label='Character Count')
plt.legend(loc='upper left')
plt.ylabel('Count of Article')
plt.show()
# Character count : min- less than 200 character and max - 600 characters for ARTICLE T
EXT
plt.plot(df['word_count'] , 'g-', label='Word Count')
plt.legend(loc='upper left')
plt.ylabel('Count of Article')
plt.show()
# Word count : min- 20 words and max - more than 100 words for ARTICLE TEXT
```





In [23]:

```
df['time'] = df['Date of Posting'].apply(lambda x : str(x)[:7])
df['month_date_year'] = df['Date of Posting'].apply(lambda x : str(x)[8:])
df[['time', 'month_date_year']].tail(23)
```

Out[23]:

	time	month_date_year
1	6:20pm,	August 8, 2019
2	2:00pm,	August 8, 2019
3	6:00am,	August 7, 2019
4	8:00am,	August 8, 2019
5	6:00am,	August 6, 2019
6	6:00am,	August 8, 2019
7	2:00pm,	August 7, 2019
8	1:00pm,	August 7, 2019
9	3:37pm,	August 6, 2019
10	10:00am	August 6, 2019
11	6:00am,	August 6, 2019
12	11:00am	August 5, 2019
13	8:00am,	August 5, 2019
14	6:00am,	August 5, 2019
15	8:00am,	August 4, 2019
16	3:52pm,	August 2, 2019
17	3:09pm,	August 2, 2019
18	6:00am,	August 2, 2019
19	10:45am	August 2, 2019
20	1:00pm,	July 31, 2019
21	2:00pm,	August 1, 2019
22	12:07pm	August 1, 2019
23	2:03pm,	July 31, 2019

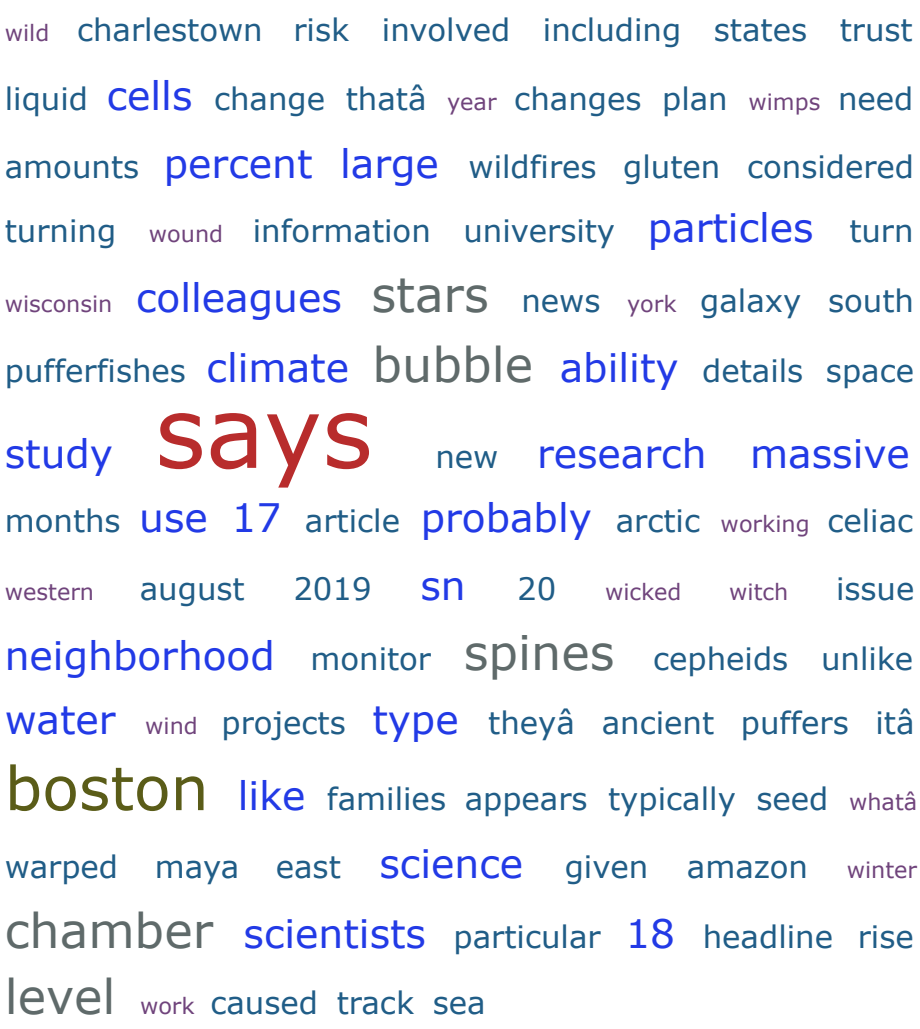
In [24]:

```
# program to generate WordCloud- Visualization
from word_cloud.word_cloud_generator import WordCloud
from IPython.core.display import HTML
from nltk.corpus import reuters
import nltk
import pandas as pd
```

In [25]:

```
ENGLISH_STOP_WORDS = frozenset([
    "a", "about", "above", "across", "after", "afterwards", "again", "against",
    "all", "almost", "alone", "along", "already", "also", "although", "always",
    "am", "among", "amongst", "amoungst", "amount", "an", "and", "another",
    "any", "anyhow", "anyone", "anything", "anyway", "anywhere", "are",
    "around", "as", "at", "back", "be", "became", "because", "become",
    "becomes", "becoming", "been", "before", "beforehand", "behind", "being",
    "below", "beside", "besides", "between", "beyond", "bill", "both",
    "bottom", "but", "by", "call", "can", "cannot", "cant", "co", "con",
    "could", "couldnt", "cry", "de", "describe", "detail", "do", "done",
    "down", "due", "during", "each", "eg", "eight", "either", "eleven", "else",
    "elsewhere", "empty", "enough", "etc", "even", "ever", "every", "everyone",
    "everything", "everywhere", "except", "few", "fifteen", "fifty", "fill",
    "find", "fire", "first", "five", "for", "former", "formerly", "forty",
    "found", "four", "from", "front", "full", "further", "get", "give", "go",
    "had", "has", "hasnt", "have", "he", "hence", "her", "here", "hereafter",
    "hereby", "herein", "hereupon", "hers", "herself", "him", "himself", "his",
    "how", "however", "hundred", "i", "ie", "if", "in", "inc", "indeed",
    "interest", "into", "is", "it", "its", "itself", "keep", "last", "latter",
    "latterly", "least", "less", "ltd", "made", "many", "may", "me",
    "meanwhile", "might", "mill", "mine", "more", "moreover", "most", "mostly",
    "move", "much", "must", "my", "myself", "name", "namely", "neither",
    "never", "nevertheless", "next", "nine", "no", "nobody", "none", "noone",
    "nor", "not", "nothing", "now", "nowhere", "of", "off", "often", "on",
    "once", "one", "only", "onto", "or", "other", "others", "otherwise", "our",
    "ours", "ourselves", "out", "over", "own", "part", "per", "perhaps",
    "please", "put", "rather", "re", "same", "see", "seem", "seemed",
    "seeming", "seems", "serious", "several", "she", "should", "show", "side",
    "since", "sincere", "six", "sixty", "so", "some", "somehow", "someone",
    "something", "sometime", "sometimes", "somewhere", "still", "such",
    "system", "take", "ten", "than", "that", "the", "their", "them",
    "themselves", "then", "thence", "there", "thereafter", "thereby",
    "therefore", "therein", "thereupon", "these", "they", "thick", "thin",
    "third", "this", "those", "though", "three", "through", "throughout",
    "thru", "thus", "to", "together", "too", "top", "toward", "towards",
    "twelve", "twenty", "two", "un", "under", "until", "up", "upon", "us",
    "very", "via", "was", "we", "well", "were", "what", "whatever", "when",
    "whence", "whenever", "where", "whereafter", "whereas", "whereby",
    "wherein", "whereupon", "wherever", "whether", "which", "while", "whither",
    "who", "whoever", "whole", "whom", "whose", "why", "will", "with",
    "within", "without", "would", "yet", "said","you", "your", "yours", "yourself",
    "yourselves"])

wc=WordCloud(use_tfidf=False,stopwords=ENGLISH_STOP_WORDS)
import pandas as pd

df = pd.read_csv(r"sciencenews_all_data.csv", encoding ="latin-1")
#don't randomize color, show only top 50
embed_code=wc.get_embed_code(text=df['Summary'],random_color=False,topn=100)
HTML(embed_code)

# highlights words used commanly in the articles like "says","boston","cells", "stars",
etc.
```

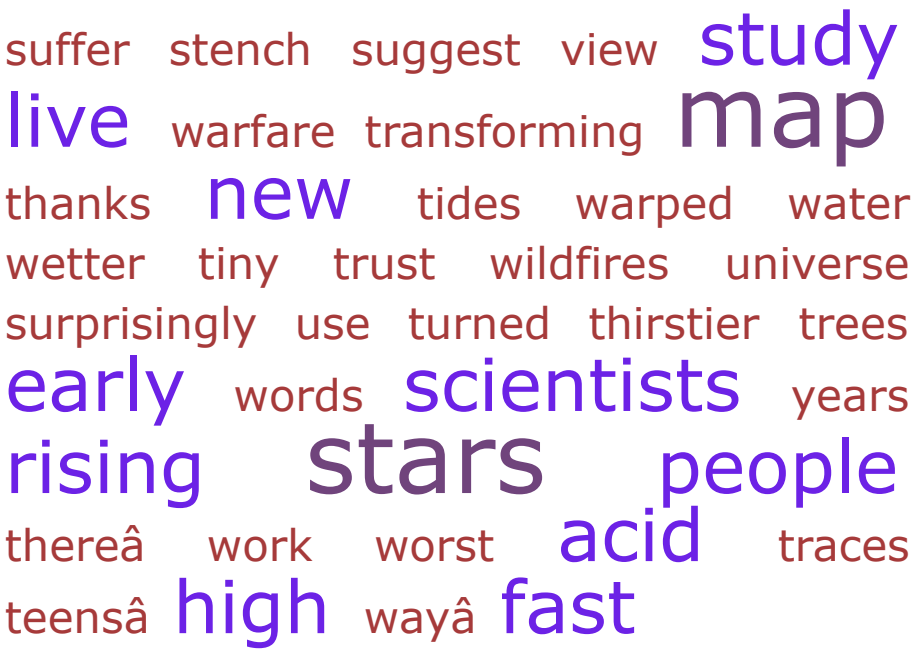
Out[25]:


wild charlestown risk involved including states trust
liquid cells change thatâ year changes plan wimps need
amounts percent large wildfires gluten considered
turning wound information university particles turn
wisconsin colleagues stars news york galaxy south
pufferfishes climate bubble ability details space
study says new research massive
months use 17 article probably arctic working celiac
western august 2019 sn 20 wicked witch issue
neighborhood monitor spines cepheids unlike
water wind projects type theyâ ancient puffers itâ
boston like families appears typically seed whatâ
warped maya east science given amazon winter
chamber scientists particular 18 headline rise
level work caused track sea

In [26]:

```
embed_code=wc.get_embed_code(text=df['Topic Name'],random_color=True,topn=40)
HTML(embed_code)
# common words used in Title articles : "live", "new", "early","study",etc.
```

Out[26]:



In [27]:

```
from gensim import matutils, models
import scipy.sparse
top_dict = {}
for c in df.columns:
    top = df[c].sort_values(ascending=False).head(30)
    top_dict[c]= list(zip(top.index, top.values))

top_dict
```

Out[27]:


```
{'Topic Name': [(5,
   'With nowhere to hide from rising seas, Boston prepares for a wetter fu
ture'),
  (7, 'Why people with celiac disease suffer so soon after eating glute
n'),
  (22, 'Thereâ\x80\x99s more to pufferfish than that goofy spiked balloo
n'),
  (2,
   'The worst wildfires can send smoke high enough to affect the ozone lay
er'),
  (16,
   'The Arctic is burning and Greenland is melting, thanks to record hea
t'),
  (18, 'Stars may keep spinning fast, long into old age'),
  (15, 'Satellites are transforming how archaeologists study the past'),
  (9,
   'Racist words and acts, like the El Paso shooting, harm childrenâ\x80\x
99s health'),
  (19,
   'Public trust that scientists work for the good of society is growing
'),
  (6, 'One in 4 people live in places at high risk of running out of wate
r'),
  (23, 'Monkeys can use basic logic to decipher the order of items in a li
st'),
  (1, 'How these tiny insect larvae leap without legs'),
  (11,
   'How the 5 riskiest U.S. cities for coastal flooding are preparing for
rising tides'),
  (3, 'How pieces of live human brain are helping scientists map nerve cel
ls'),
  (17,
   'Hospitalizations highlight potential dangers of e-cigs to teensâ\x80\x
99 lungs '),
  (8,
   'Giant, active galaxies from the early universe may have finally been f
ound'),
  (0, 'Exploding stars scattered traces of iron over Antarctic snow'),
  (14, 'Decades of dumping acid suggest acid rain may make trees thirstier
'),
  (12, 'Ancient Maya warfare flared up surprisingly early'),
  (20, 'A new study challenges the idea that the placenta has a microbiom
e'),
  (13,
   'A new map is the best view yet of how fast Antarctica is shedding ic
e'),
  (10,
   'A fungus makes a chemical that neutralizes the stench of skunk spray
'),
  (21, 'A 3-D map of stars reveals the Milky Wayâ\x80\x99s warped shape'),
  (4, '50 years ago, Fermilab turned to bubbles')],
 'Link': [(2,
   'https://www.sciencenews.org//article/worst-wildfires-can-send-smoke-hi
gh-enough-affect-ozone-layer'),
  (7,
   'https://www.sciencenews.org//article/why-people-celiac-disease-suffer-
so-soon-after-eating-gluten'),
  (11,
   'https://www.sciencenews.org//article/top-five-us-coastal-cities-risk-f
```

```
looding-rising-sea-levels'),
  (18,
   'https://www.sciencenews.org//article/stars-may-keep-spinning-fast-long
-old-age'),
  (15,
   'https://www.sciencenews.org//article/space-satellites-transforming-how
-archaeologists-study-past'),
  (9,
   'https://www.sciencenews.org//article/racism-words-acts-el-paso-shootin
g-harm-children-health-longterm'),
  (22,
   'https://www.sciencenews.org//article/pufferfish-biology-mating-goofy-s
piked-balloon'),
  (19,
   'https://www.sciencenews.org//article/public-trust-scientists-work-good
-society-growing'),
  (6,
   'https://www.sciencenews.org//article/one-4-people-live-places-high-ris
k-running-out-water'),
  (20,
   'https://www.sciencenews.org//article/new-study-challenges-idea-placent
a-microbiome-bacteria'),
  (13,
   'https://www.sciencenews.org//article/new-map-best-view-yet-how-fast-an
tarctica-shedding-ice'),
  (23,
   'https://www.sciencenews.org//article/monkeys-can-use-basic-logic-decip
her-order-items-list'),
  (1,
   'https://www.sciencenews.org//article/how-these-tiny-insect-larvae-leap
-without-legs'),
  (17,
   'https://www.sciencenews.org//article/hospitalizations-highlight-potent
ial-dangers-e-cigs-teen-lungs'),
  (8,
   'https://www.sciencenews.org//article/giant-active-galaxies-early-unive
rse-may-have-finally-been-found'),
  (10,
   'https://www.sciencenews.org//article/fungus-makes-chemical-neutralizes
-stench-skunk-spray'),
  (0,
   'https://www.sciencenews.org//article/exploding-stars-scattered-traces-
iron-over-antarctic-snow'),
  (3,
   'https://www.sciencenews.org//article/experiment-live-human-brain-helps
-scientists-map-nerve-cells'),
  (14,
   'https://www.sciencenews.org//article/decades-dumping-acid-suggest-acid
-rain-may-make-trees-thirstier'),
  (5,
   'https://www.sciencenews.org//article/boston-adapting-rising-sea-level-
coastal-flooding'),
  (16,
   'https://www.sciencenews.org//article/arctic-burning-greenland-melting-
thanks-record-heat'),
  (12,
   'https://www.sciencenews.org//article/ancient-maya-warfare-flared-surpr
isingly-early'),
  (4,
   'https://www.sciencenews.org//article/50-years-ago-fermilab-turned-bubb
les'),
```

```
   (21,
    'https://www.sciencenews.org//article/3-d-map-stars-reveals-milky-way-w
arped-shape')],
 'Author name': [(7, 'Tina Hesman Saey'),
  (1, 'Susan Milius'),
  (22, 'Susan Milius'),
  (2, 'Megan Sever'),
  (11, 'Mary Caperton Morton'),
  (5, 'Mary Caperton Morton'),
  (8, 'Maria Temming'),
  (13, 'Maria Temming'),
  (18, 'Lisa Grossman'),
  (20, 'Laura Sanders'),
  (3, 'Laura Sanders'),
  (19, 'Katy Daigle'),
  (15, 'Erin Wayman'),
  (0, 'Emily Conover'),
  (21, 'Emily Conover'),
  (10, 'Carolyn Wilke'),
  (6, 'Carolyn Wilke'),
  (14, 'Carolyn Wilke'),
  (16, 'Carolyn Gramling'),
  (12, 'Bruce Bower'),
  (23, 'Bruce Bower'),
  (4, 'Bethany Brookshire'),
  (9, 'Aimee Cunningham'),
  (17, 'Aimee Cunningham')],
 'Date of Posting': [(4, '8:00am, August 8, 2019'),
  (13, '8:00am, August 5, 2019'),
  (15, '8:00am, August 4, 2019'),
  (1, '6:20pm, August 8, 2019'),
  (0, '6:00am, August 9, 2019'),
  (6, '6:00am, August 8, 2019'),
  (3, '6:00am, August 7, 2019'),
  (5, '6:00am, August 6, 2019'),
  (11, '6:00am, August 6, 2019'),
  (14, '6:00am, August 5, 2019'),
  (18, '6:00am, August 2, 2019'),
  (16, '3:52pm, August 2, 2019'),
  (9, '3:37pm, August 6, 2019'),
  (17, '3:09pm, August 2, 2019'),
  (23, '2:03pm, July 31, 2019'),
  (2, '2:00pm, August 8, 2019'),
  (7, '2:00pm, August 7, 2019'),
  (21, '2:00pm, August 1, 2019'),
  (20, '1:00pm, July 31, 2019'),
  (8, '1:00pm, August 7, 2019'),
  (22, '12:07pm, August 1, 2019'),
  (12, '11:00am, August 5, 2019'),
  (19, '10:45am, August 2, 2019'),
  (10, '10:00am, August 6, 2019')],
 'Summary': [(19,
   'â\x80\x9cTrust is important to legitimacy, credibility and effectivene
ss,â\x80\x9d Boykoff says. â\x80\x9cWithout trust, scientists would just b
e screaming into the wind.â\x80\x9d'),
  (0,
   'â\x80\x9cThis is actually quite a profound thing,â\x80\x9d says astrop
hysicist Brian Fields of the University of Illinois at Urbana-Champaign, w
ho was not involved with the research. â\x80\x9cItâ\x80\x99s telling us ab
out the recent history of our whole neighborhood in the galaxy and about t
he lives and deaths of massive stars.â\x80\x9d'),
```

(11,
'With neighborhood-level projections for future sea level rise in hand, the city of Boston has district-level projects completed for East Boston, Charlestown and South Boston. A deployable flood wall is being installed a long the East Boston Greenway and a section of Main Street in Charlestown is being elevated to protect the adjacent neighborhood. In several areas, including around South Boston and the Seaport, concrete is being removed a nd replaced by floodable parks and green space. Mayor Martin Walsh has ple dged 10 percent of the cityâ\x80\x99s $3.49 billion capital budget in 2020 for such resiliency projects.'),
(8,
'Wang and colleagues now plan to take a larger census of ancient massiv e galaxies with ALMA. That work could give theorists more information abou t how to tweak cosmological simulations to match early-universe observatio ns.'),
(18,
'Unfortunately, the result might mean that astronomers canâ\x80\x99t us e starsâ\x80\x99 spin speeds to guess ages anymore. â\x80\x9cIf that stops working in old stars, thatâ\x80\x99s a bummer,â\x80\x9d Curtis says.'),
(9,
'Trent: As pediatricians, we will be there to help families with these discussions and both the direct and indirect trauma that these events have caused. We will also continue our advocacy efforts to encourage our govern ment leaders to adopt policies that broadly address gun violence and chang e the climate of racism impacting children, adolescents and families.'),
(5,
'This article appears in the August 17, 2019 issue of Science News with the headline, "Wicked High Tides: Boston is taking action to adapt to sea level rise."Â\xa0'),
(3,
'This article appears in the August 17, 2019 issue of Science News with the headline, "A Menagerie of Neurons:Â\xa0Studies of living brain cells a im to determine what sets humans apart."'),
(17,
'The investigation of the Wisconsin teens could provide some answers th at will aid research. More details about the teensâ\x80\x99 e-cigarette us e, such as the type of device, the e-liquid, the flavors, how much they va ped and so on, â\x80\x9cwould be very helpful in trying to understand what â\x80\x99s going on and who else might be at risk,â\x80\x9d Crotty Alexand er says.'),
(6,
'The United States is considered to have relatively low risk; overall, it uses less than 20 percent of its available water. However, some western states including California, Arizona, New Mexico, Colorado and Nebraska ty pically use 40 percent or more of current water supplies each year.'),
(23,
'Thatâ\x80\x99s probably a valuable ability in the wild, she says, beca use many animals need to monitor where group mates stand in the social pec king order. â\x80\x9cAn ability to construct, retain, manipulate and refer ence ordered information may be an evolutionarily ancient, efficient [ment al] mechanism for keeping track of relationships between individuals,â\x80 \x9d she says.'),
(14,
'Soils are typically slow to recover calcium theyâ\x80\x99ve lost, so t he study may also point to legacy effects of acid rain that we didnâ\x80\x 99t already know about, says Charles Driscoll, a biogeochemist at Syracuse University in New York who was not involved in the study.'),
(12,
'Prior to 800, Maya people may have considered it dishonorable to kill or wound others from a distance, Graham suspects. Classic Maya culture pro bably discouraged killing large numbers of opponents in battle with any ty

```
pe of weapon, since no mass burials of war victims have been found, Inomat
a says.'),
  (1,
   'Poppinga and colleagues recently showed that Chinese witch hazel trees
build up forces in the mature fruit that suddenly shoot out a seed rotatin
g a bit like a bullet from a rifle. Unlike gall midge launches, though, th
ese tree latches break when they let go. The leap of a legless seed is fas
t and dramatic, but itâ\x80\x99s not repeatable.'),
  (4,
   'NAL was renamed Fermilab in 1974 for physicist Enrico Fermi. The labâ
\x80\x99s first accelerator produced protons in April 1969, and was shooti
ng subatomic particles into a 76-centimeter bubble chamber filled with liq
uid hydrogen by 1972. Such chambers track bubble trails left by speeding p
articles. The lab began upgrading to a 4.5-meter chamber detector in 1973,
which helped in the study of neutrinos and turned up evidence for bottom a
nd top quarks. As accelerators modernized, bubble detectors were phased ou
t, and Fermilabâ\x80\x99s chamber became an art installation. But SNOLABâ
\x80\x99s bubble chamber in Sudbury, Canada, still searches for weakly int
eracting massive particles, or WIMPS â\x80\x94 a proposed type of dark mat
ter.Â\xa0'),
  (16,
   'Meanwhile, increasingly frequent winter warm spells, insect outbreaks
and wildfires have also caused many Arctic plants to lose their resistance
to freezing, dry out and die, turning large parts of the Arctic brown (SN:
4/13/19, p. 16). That, in turn, increases the regionâ\x80\x99s susceptibil
ity to more wildfires: Normally, the icy peatlands are soggy enough to be
fire-resistant, but they are thawing and drying out. Once set ablaze, the
carbon-rich peat can burn for months, releasing large amounts of COâ\x82\x
82 back into the atmosphere and fueling the warming feedback loop (SN: 3/1
7/18, p. 20).'),
  (7,
   'Knowing that certain T cells, and cytokines in particular, cause celia
c symptoms may lead to therapies that could block the gluten-reacting T ce
lls, Anderson says. And doctors may be able to diagnose celiac disease by
measuring IL-2 levels in the blood, sparing patients the need for tests in
which theyâ\x80\x99re repeatedly given gluten.'),
  (22,
   'In the intimidating body part catalog, pufferfishes are perhaps best k
nown for turning into spiky balls when outraged. These spines perk upright
when puffers gulp water to balloon out their abdomens. Some of the same ge
ne networks that put feathers on birds and hairs on mammals turn out to pu
t the protective spines on puffers, Fraser and colleagues report July 25 i
n iScience. Those spines have evolved from the scales that covered distant
fish ancestors. But between todayâ\x80\x99s skinny spines, modern pufferfi
shes are totally naked. Try not to stare.'),
  (13,
   'In 2021, NASA and the Indian Space Research Organization plan to launc
h a satellite that will gather enough data to update this map every few mo
nths â\x80\x94 allowing scientists to better monitor how ice flow across A
ntarctica changes as the climate changes.'),
  (2,
   'Given that climate change is increasing fire frequency and intensity i
n some places like the North American West (SN: 12/22/18, p. 18), we can p
robably expect to see more of these fire clouds reaching the stratosphere,
Fromm says. But, he cautions, â\x80\x9cwe are still on the learning curve
when it comes to understanding pyroCbs.â\x80\x9d'),
  (15,
   'Buy Archaeology from SpaceÂ\xa0from Amazon.com.Â\xa0Science NewsÂ\xa0i
s a participant in the Amazon Services LLC Associates Program. Please see
ourÂ\xa0FAQÂ\xa0for more details.'),
  (10,
```

'Adding several common cosmetic ingredients also sped up pericosineâ\x80\x99s ability to cut the skunk spray smell, the team found. That â\x80\x9cwas really thrilling,â\x80\x9d Cichewicz says. â\x80\x9cThis now looks a lot more like a personal-care product than it does an organic chemistry reaction.â\x80\x9d'),
(20,
'Aagaard is convinced there are small amounts of bacteria in the placenta, but remains unsure about what biological role those microbes play, if any.'),
(21,
'3-D VISIONÂ\xa0 The Milky Wayâ\x80\x99s Cepheid stars are plotted in three dimensions, revealing the galaxyâ\x80\x99s warped shape. Unlike other stars, Cepheids vary in brightness in a particular way that helps scientists make more precise estimates of their distances from Earth. Brighter colors represent Cepheids closer to the warped plane of the galaxy, indicated by the grid. The star icon indicates the sun. ')]}

In [178]:

```
data =df
from textblob import TextBlob # sentiment function of textblob returns two properties,
 polarity, and subjectivity
import sys
pol = lambda x: TextBlob(x).sentiment.polarity
sub = lambda x: TextBlob(x).sentiment.subjectivity

data['polarity'] = data['Summary'].apply(pol)
data['subjectivity'] = data['Summary'].apply(sub)
print(data)
```

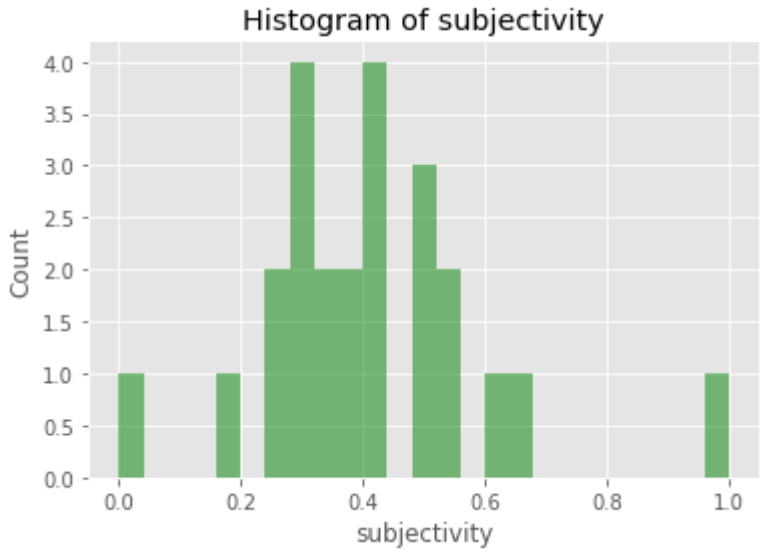
	Topic Name	Link	Author name
0	Exploding stars scattered traces of iron over ...	https://www.sciencenews.org//article/exploding...	Emily Conover
1	How these tiny insect larvae leap without legs	https://www.sciencenews.org//article/how-these...	Susan Milius
2	The worst wildfires can send smoke high enough...	https://www.sciencenews.org//article/worst-wil...	Megan Sever
3	How pieces of live human brain are helping sci...	https://www.sciencenews.org//article/experimen...	Laura Sanders
4	50 years ago, Fermilab turned to bubbles	https://www.sciencenews.org//article/50-years-...	Bethany Brookshire
5	With nowhere to hide from rising seas, Boston ...	https://www.sciencenews.org//article/boston-ad...	Mary Caperton Morton
6	One in 4 people live in places at high risk of...	https://www.sciencenews.org//article/one-4-peo...	Carolyn Wilke
7	Why people with celiac disease suffer so soon ...	https://www.sciencenews.org//article/why-peopl...	Tina Hesman Saey
8	Giant, active galaxies from the early universe...	https://www.sciencenews.org//article/giant-act...	Maria Temming
9	Racist words and acts, like the El Paso shooti...	https://www.sciencenews.org//article/racism-wo...	Aimee Cunningham
10	A fungus makes a chemical that neutralizes the...	https://www.sciencenews.org//article/fungus-ma...	Carolyn Wilke
11	How the 5 riskiest U.S. cities for coastal flo...	https://www.sciencenews.org//article/top-five-...	Mary Caperton Morton
12	Ancient Maya warfare flared up surprisingly early	https://www.sciencenews.org//article/ancient-m...	Bruce Bower
13	A new map is the best view yet of how fast Ant...	https://www.sciencenews.org//article/new-map-b...	Maria Temming
14	Decades of dumping acid suggest acid rain may ...	https://www.sciencenews.org//article/decades-d...	Carolyn Wilke
15	Satellites are transforming how archaeologists...	https://www.sciencenews.org//article/space-sat...	Erin Wayman
16	The Arctic is burning and Greenland is melting...	https://www.sciencenews.org//article/arctic-bu...	Carolyn Gramlin
17	Hospitalizations highlight potential dangers o...		
18	Stars may keep spinning fast, long into old age		
19	Public trust that scientists work for the good...		
20	A new study challenges the idea that the place...		
21	A 3-D map of stars reveals the Milky Wayâ□□s w...		
22	Thereâ□□s more to pufferfish than that goofy s...		
23	Monkeys can use basic logic to decipher the or...		

```
g
17  https://www.sciencenews.org//article/hospitali...      Aimee Cunningha
m
18  https://www.sciencenews.org//article/stars-may...         Lisa Grossma
n
19  https://www.sciencenews.org//article/public-tr...           Katy Daigl
e
20  https://www.sciencenews.org//article/new-study...         Laura Sander
s
21  https://www.sciencenews.org//article/3-d-map-s...         Emily Conove
r
22  https://www.sciencenews.org//article/pufferfis...          Susan Miliu
s
23  https://www.sciencenews.org//article/monkeys-c...           Bruce Bowe
r

            Date of Posting  \
0    6:00am, August 9, 2019
1    6:20pm, August 8, 2019
2    2:00pm, August 8, 2019
3    6:00am, August 7, 2019
4    8:00am, August 8, 2019
5    6:00am, August 6, 2019
6    6:00am, August 8, 2019
7    2:00pm, August 7, 2019
8    1:00pm, August 7, 2019
9    3:37pm, August 6, 2019
10  10:00am, August 6, 2019
11   6:00am, August 6, 2019
12  11:00am, August 5, 2019
13   8:00am, August 5, 2019
14   6:00am, August 5, 2019
15   8:00am, August 4, 2019
16   3:52pm, August 2, 2019
17   3:09pm, August 2, 2019
18   6:00am, August 2, 2019
19  10:45am, August 2, 2019
20    1:00pm, July 31, 2019
21   2:00pm, August 1, 2019
22  12:07pm, August 1, 2019
23    2:03pm, July 31, 2019

                                                Summary  polarity  subjectiv
ity  \
0   â□□This is actually quite a profound thing,â□□...  0.056667      0.55
0000
1   Poppinga and colleagues recently showed that C... -0.022222      0.341
667
2   Given that climate change is increasing fire f...  0.250000      0.250
000
3   This article appears in the August 17, 2019 is...  0.000000      0.000
000
4   NAL was renamed Fermilab in 1974 for physicist...  0.078125      0.532
292
5   This article appears in the August 17, 2019 is...  0.130000      0.320
000
6   The United States is considered to have relati...  0.078114      0.291
246
7   Knowing that certain T cells, and cytokines in...  0.293651      0.509
921
8   Wang and colleagues now plan to take a larger ...  0.166667      0.666
```

```
667
9   Trent: As pediatricians, we will be there to h...  0.081250      0.356
250
10  Adding several common cosmetic ingredients als...  0.100000      0.300
000
11  With neighborhood-level projections for future...  0.019444      0.259
722
12  Prior to 800, Maya people may have considered ...  0.126984      0.198
413
13  In 2021, NASA and the Indian Space Research Or...  0.100000      0.366
667
14  Soils are typically slow to recover calcium th... -0.081818      0.427
273
15  Buy Archaeology from SpaceÂ from Amazon.com.Â ...  0.500000      0.500
000
16  Meanwhile, increasingly frequent winter warm s...  0.191991      0.418
831
17  The investigation of the Wisconsin teens could...  0.225000      0.375
000
18  Unfortunately, the result might mean that astr... -0.237500      0.629
167
19  â□□Trust is important to legitimacy, credibili...  0.400000       1.000
000
20  Aagaard is convinced there are small amounts o... -0.250000      0.400
000
21  3-D VISIONÂ  The Milky Wayâ□□s Cepheid stars a...  0.235417       0.502
083
22  In the intimidating body part catalog, pufferf...  0.033333      0.412
500
23  Thatâ□□s probably a valuable ability in the wi...  0.133333       0.291
667

    word_count  char_count  word_density
0            2          12      0.153846
1            2          11      0.166667
2            2          10      0.181818
3            2          12      0.153846
4            2          17      0.111111
5            3          18      0.157895
6            2          12      0.153846
7            3          14      0.200000
8            2          12      0.153846
9            2          15      0.125000
10           2          12      0.153846
11           3          18      0.157895
12           2          10      0.181818
13           2          12      0.153846
14           2          12      0.153846
15           2          10      0.181818
16           2          15      0.125000
17           2          15      0.125000
18           2          12      0.153846
19           2          10      0.181818
20           2          12      0.153846
21           2          12      0.153846
22           2          11      0.166667
23           2          10      0.181818
```

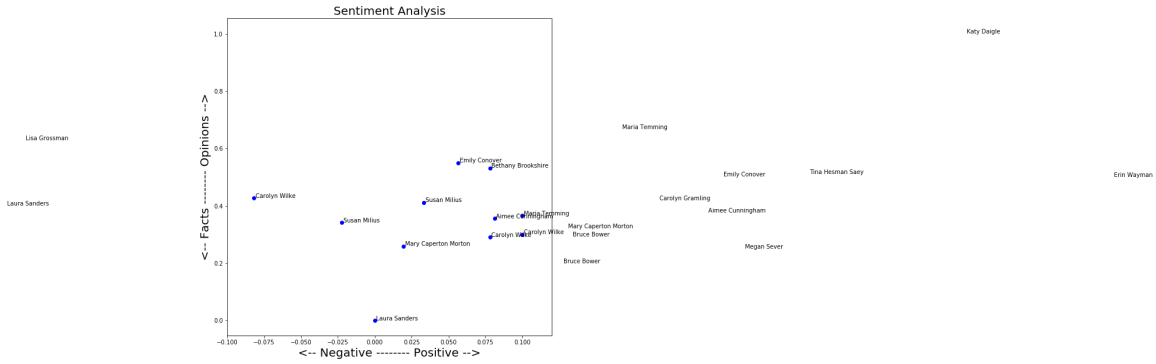
In [187]:

```
num_bins = 25
n, bins, patches = plt.hist(data['polarity'], num_bins, facecolor='red', alpha=0.5)
plt.xlabel('Polarity')
plt.ylabel('Count')
plt.title('Histogram of polarity')
plt.show()

n, bins, patches = plt.hist(data['subjectivity'], num_bins, facecolor='green', alpha=0.
5)
plt.xlabel('subjectivity')
plt.ylabel('Count')
plt.title('Histogram of subjectivity')
plt.show()

# Polarity is float which lies in the range of [-1,1] where 1 means positive statement
 and -1 means a negative statement.
# by looking at graph and data . the one with high polarity is a POSITIVE ARTICLE.

# Subjectivity refers to judgement / personal opinion:
# if it is 0.8, then the statement is positive and 0.75 subjectivity is a public opinio
n and not a factual information.
```



In [29]:

```
#.................................SENTIMENT NALAYSIS.............................#
import matplotlib.pyplot as plt

plt.rcParams['figure.figsize'] = [10, 10]

for index, comedian in enumerate(data.index):
    x = data.polarity.loc[comedian]
    y = data.subjectivity.loc[comedian]
    plt.scatter(x, y, color='blue')
    plt.text(x+.001, y+.001, data['Author name'][index], fontsize=10)
    plt.xlim(-.10, .12)

plt.title('Sentiment Analysis', fontsize=20)
plt.xlabel('<-- Negative -------- Positive -->', fontsize=20)
plt.ylabel('<-- Facts -------- Opinions -->', fontsize=20)
# representing positive and negative facts and analysis in graph based on POlarity and
 Subjectivity
plt.show()
```



In [30]:

```
#.........................................NAMED ENTITY RECOGNITION (NER)
 ................................................#
#NER using nltk and spact library is done to udentity certain things like to identify t
he names of things,
#.............such as persons, organizations, or locations from the csv extraxted.
import nltk
nltk.download('averaged_perceptron_tagger')
from nltk.tokenize import word_tokenize
from nltk.tag import pos_tag
```

```
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\AKSHIT\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

In [31]:

```
ex = df['Summary'] # taking articles data only from the entire csv for NER
new = str(ex)
def preprocess(sent):
    sent = nltk.word_tokenize(sent)
    sent = nltk.pos_tag(sent)
    return sent
sent = preprocess(new)
sent
#word tokenization
#part of speech tagging
```

Out[31]:

```
[('0', 'CD'),
 ('â\x80\x9cThis', 'NN'),
 ('is', 'VBZ'),
 ('actually', 'RB'),
 ('quite', 'RB'),
 ('a', 'DT'),
 ('profound', 'JJ'),
 ('thing', 'NN'),
 (',', ','),
 ('â\x80\x9d', 'VBP'),
 ('...', ':'),
 ('1', 'CD'),
 ('Poppinga', 'NNP'),
 ('and', 'CC'),
 ('colleagues', 'NNS'),
 ('recently', 'RB'),
 ('showed', 'VBD'),
 ('that', 'IN'),
 ('C', 'NNP'),
 ('...', ':'),
 ('2', 'CD'),
 ('Given', 'NNP'),
 ('that', 'WDT'),
 ('climate', 'VBP'),
 ('change', 'NN'),
 ('is', 'VBZ'),
 ('increasing', 'VBG'),
 ('fire', 'NN'),
 ('f', 'NN'),
 ('...', ':'),
 ('3', 'CD'),
 ('This', 'DT'),
 ('article', 'NN'),
 ('appears', 'VBZ'),
 ('in', 'IN'),
 ('the', 'DT'),
 ('August', 'NNP'),
 ('17', 'CD'),
 (',', ','),
 ('2019', 'CD'),
 ('is', 'VBZ'),
 ('...', ':'),
 ('4', 'CD'),
 ('NAL', 'NNP'),
 ('was', 'VBD'),
 ('renamed', 'VBN'),
 ('Fermilab', 'NNP'),
 ('in', 'IN'),
 ('1974', 'CD'),
 ('for', 'IN'),
 ('physicist', 'NN'),
 ('...', ':'),
 ('5', 'CD'),
 ('This', 'DT'),
 ('article', 'NN'),
 ('appears', 'VBZ'),
 ('in', 'IN'),
 ('the', 'DT'),
 ('August', 'NNP'),
```

```
    ('17', 'CD'),
    (',', ','),
    ('2019', 'CD'),
    ('is', 'VBZ'),
    ('...', ':'),
    ('6', 'CD'),
    ('The', 'DT'),
    ('United', 'NNP'),
    ('States', 'NNPS'),
    ('is', 'VBZ'),
    ('considered', 'VBN'),
    ('to', 'TO'),
    ('have', 'VB'),
    ('relati', 'NNS'),
    ('...', ':'),
    ('7', 'CD'),
    ('Knowing', 'VBG'),
    ('that', 'DT'),
    ('certain', 'JJ'),
    ('T', 'NNP'),
    ('cells', 'NNS'),
    (',', ','),
    ('and', 'CC'),
    ('cytokines', 'NNS'),
    ('in', 'IN'),
    ('...', ':'),
    ('8', 'CD'),
    ('Wang', 'NNP'),
    ('and', 'CC'),
    ('colleagues', 'NNS'),
    ('now', 'RB'),
    ('plan', 'VBP'),
    ('to', 'TO'),
    ('take', 'VB'),
    ('a', 'DT'),
    ('larger', 'JJR'),
    ('...', ':'),
    ('9', 'CD'),
    ('Trent', 'NN'),
    (':', ':'),
    ('As', 'IN'),
    ('pediatricians', 'NNS'),
    (',', ','),
    ('we', 'PRP'),
    ('will', 'MD'),
    ('be', 'VB'),
    ('there', 'RB'),
    ('to', 'TO'),
    ('h', 'VB'),
    ('...', ':'),
    ('10', 'CD'),
    ('Adding', 'NNP'),
    ('several', 'JJ'),
    ('common', 'JJ'),
    ('cosmetic', 'JJ'),
    ('ingredients', 'NNS'),
    ('als', 'NNS'),
    ('...', ':'),
    ('11', 'CD'),
    ('With', 'IN'),
    ('neighborhood-level', 'JJ'),
```

```
    ('projections', 'NNS'),
    ('for', 'IN'),
    ('future', 'JJ'),
    ('...', ':'),
    ('12', 'CD'),
    ('Prior', 'NNP'),
    ('to', 'TO'),
    ('800', 'CD'),
    (',', ','),
    ('Maya', 'NNP'),
    ('people', 'NNS'),
    ('may', 'MD'),
    ('have', 'VB'),
    ('considered', 'VBN'),
    ('...', ':'),
    ('13', 'CD'),
    ('In', 'IN'),
    ('2021', 'CD'),
    (',', ','),
    ('NASA', 'NNP'),
    ('and', 'CC'),
    ('the', 'DT'),
    ('Indian', 'JJ'),
    ('Space', 'NNP'),
    ('Research', 'NNP'),
    ('Or', 'NNP'),
    ('...', ':'),
    ('14', 'CD'),
    ('Soils', 'NNS'),
    ('are', 'VBP'),
    ('typically', 'RB'),
    ('slow', 'JJ'),
    ('to', 'TO'),
    ('recover', 'VB'),
    ('calcium', 'NN'),
    ('th', 'NN'),
    ('...', ':'),
    ('15', 'CD'),
    ('Buy', 'NNP'),
    ('Archaeology', 'NNP'),
    ('from', 'IN'),
    ('SpaceÂ', 'NNP'),
    ('from', 'IN'),
    ('Amazon.com.Â', 'NNP'),
    ('...', ':'),
    ('16', 'CD'),
    ('Meanwhile', 'RB'),
    (',', ','),
    ('increasingly', 'RB'),
    ('frequent', 'JJ'),
    ('winter', 'NN'),
    ('warm', 'NN'),
    ('s', 'NN'),
    ('...', ':'),
    ('17', 'CD'),
    ('The', 'DT'),
    ('investigation', 'NN'),
    ('of', 'IN'),
    ('the', 'DT'),
    ('Wisconsin', 'NNP'),
    ('teens', 'NNS'),
```

```
    ('could', 'MD'),
    ('...', ':'),
    ('18', 'CD'),
    ('Unfortunately', 'RB'),
    (',', ','),
    ('the', 'DT'),
    ('result', 'NN'),
    ('might', 'MD'),
    ('mean', 'VB'),
    ('that', 'IN'),
    ('astr', 'NN'),
    ('...', ':'),
    ('19', 'CD'),
    ('â\x80\x9cTrust', 'NN'),
    ('is', 'VBZ'),
    ('important', 'JJ'),
    ('to', 'TO'),
    ('legitimacy', 'NN'),
    (',', ','),
    ('credibili', 'NN'),
    ('...', ':'),
    ('20', 'CD'),
    ('Aagaard', 'NNP'),
    ('is', 'VBZ'),
    ('convinced', 'VBN'),
    ('there', 'EX'),
    ('are', 'VBP'),
    ('small', 'JJ'),
    ('amounts', 'NNS'),
    ('o', 'VBP'),
    ('...', ':'),
    ('21', 'CD'),
    ('3-D', 'JJ'),
    ('VISIONÂ', 'NNP'),
    ('The', 'DT'),
    ('Milky', 'NNP'),
    ('Wayâ\x80\x99s', 'NNP'),
    ('Cepheid', 'NNP'),
    ('stars', 'VBZ'),
    ('a', 'DT'),
    ('...', ':'),
    ('22', 'CD'),
    ('In', 'IN'),
    ('the', 'DT'),
    ('intimidating', 'NN'),
    ('body', 'NN'),
    ('part', 'NN'),
    ('catalog', 'NN'),
    (',', ','),
    ('pufferf', 'NN'),
    ('...', ':'),
    ('23', 'CD'),
    ('Thatâ\x80\x99s', 'NNP'),
    ('probably', 'RB'),
    ('a', 'DT'),
    ('valuable', 'JJ'),
    ('ability', 'NN'),
    ('in', 'IN'),
    ('the', 'DT'),
    ('wi', 'NN'),
    ('...', ':'),
```

```
    ('Name', 'NN'),
    (':', ':'),
    ('Summary', 'NNP'),
    (',', ','),
    ('dtype', 'NN'),
    (':', ':'),
    ('object', 'NN')]
```

In [32]:

```
# OUTPUT: a list of tuples containing the individual words and their associated part-of
-speech.
# that indicate how sentences should be chunked.
# noun phrase: NP
# an optional determiner: DT
# adjectives: JJ
# a noun: N ,etc different codes have different meaning
pattern = 'NP: {<DT>?<JJ>*<NN>}'
cp = nltk.RegexpParser(pattern)
cs = cp.parse(sent)
print(cs)
# chunking
```

```
(S
  0/CD
  (NP â□□This/NN)
  is/VBZ
  actually/RB
  quite/RB
  (NP a/DT profound/JJ thing/NN)
  ,/,
  â□□/VBP
  .../:
  1/CD
  Poppinga/NNP
  and/CC
  colleagues/NNS
  recently/RB
  showed/VBD
  that/IN
  C/NNP
  .../:
  2/CD
  Given/NNP
  that/WDT
  climate/VBP
  (NP change/NN)
  is/VBZ
  increasing/VBG
  (NP fire/NN)
  (NP f/NN)
  .../:
  3/CD
  (NP This/DT article/NN)
  appears/VBZ
  in/IN
  the/DT
  August/NNP
  17/CD
  ,/,
  2019/CD
  is/VBZ
  .../:
  4/CD
  NAL/NNP
  was/VBD
  renamed/VBN
  Fermilab/NNP
  in/IN
  1974/CD
  for/IN
  (NP physicist/NN)
  .../:
  5/CD
  (NP This/DT article/NN)
  appears/VBZ
  in/IN
  the/DT
  August/NNP
  17/CD
  ,/,
  2019/CD
  is/VBZ
  .../:
```

```
6/CD
The/DT
United/NNP
States/NNPS
is/VBZ
considered/VBN
to/TO
have/VB
relati/NNS
.../:
7/CD
Knowing/VBG
that/DT
certain/JJ
T/NNP
cells/NNS
,/,
and/CC
cytokines/NNS
in/IN
.../:
8/CD
Wang/NNP
and/CC
colleagues/NNS
now/RB
plan/VBP
to/TO
take/VB
a/DT
larger/JJR
.../:
9/CD
(NP Trent/NN)
:/:
As/IN
pediatricians/NNS
,/,
we/PRP
will/MD
be/VB
there/RB
to/TO
h/VB
.../:
10/CD
Adding/NNP
several/JJ
common/JJ
cosmetic/JJ
ingredients/NNS
als/NNS
.../:
11/CD
With/IN
neighborhood-level/JJ
projections/NNS
for/IN
future/JJ
.../:
12/CD
```

```
Prior/NNP
to/TO
800/CD
,/,
Maya/NNP
people/NNS
may/MD
have/VB
considered/VBN
.../:
13/CD
In/IN
2021/CD
,/,
NASA/NNP
and/CC
the/DT
Indian/JJ
Space/NNP
Research/NNP
Or/NNP
.../:
14/CD
Soils/NNS
are/VBP
typically/RB
slow/JJ
to/TO
recover/VB
(NP calcium/NN)
(NP th/NN)
.../:
15/CD
Buy/NNP
Archaeology/NNP
from/IN
SpaceÂ/NNP
from/IN
Amazon.com.Â/NNP
.../:
16/CD
Meanwhile/RB
,/,
increasingly/RB
(NP frequent/JJ winter/NN)
(NP warm/NN)
(NP s/NN)
.../:
17/CD
(NP The/DT investigation/NN)
of/IN
the/DT
Wisconsin/NNP
teens/NNS
could/MD
.../:
18/CD
Unfortunately/RB
,/,
(NP the/DT result/NN)
might/MD
```

```
mean/VB
that/IN
(NP astr/NN)
.../:
19/CD
(NP â□□Trust/NN)
is/VBZ
important/JJ
to/TO
(NP legitimacy/NN)
,/,
(NP credibili/NN)
.../:
20/CD
Aagaard/NNP
is/VBZ
convinced/VBN
there/EX
are/VBP
small/JJ
amounts/NNS
o/VBP
.../:
21/CD
3-D/JJ
VISIONÂ/NNP
The/DT
Milky/NNP
Wayâ□□s/NNP
Cepheid/NNP
stars/VBZ
a/DT
.../:
22/CD
In/IN
(NP the/DT intimidating/NN)
(NP body/NN)
(NP part/NN)
(NP catalog/NN)
,/,
(NP pufferf/NN)
.../:
23/CD
Thatâ□□s/NNP
probably/RB
(NP a/DT valuable/JJ ability/NN)
in/IN
(NP the/DT wi/NN)
.../:
(NP Name/NN)
:/:
Summary/NNP
,/,
(NP dtype/NN)
:/:
(NP object/NN))
```

In [33]:

```
from nltk.chunk import conlltags2tree, tree2conlltags
from pprint import pprint
iob_tagged = tree2conlltags(cs)
pprint(iob_tagged) # IOB TAGS FOR CHUNKS
```

```
[('0', 'CD', 'O'),
 ('â\x80\x9cThis', 'NN', 'B-NP'),
 ('is', 'VBZ', 'O'),
 ('actually', 'RB', 'O'),
 ('quite', 'RB', 'O'),
 ('a', 'DT', 'B-NP'),
 ('profound', 'JJ', 'I-NP'),
 ('thing', 'NN', 'I-NP'),
 (',', ',', 'O'),
 ('â\x80\x9d', 'VBP', 'O'),
 ('...', ':', 'O'),
 ('1', 'CD', 'O'),
 ('Poppinga', 'NNP', 'O'),
 ('and', 'CC', 'O'),
 ('colleagues', 'NNS', 'O'),
 ('recently', 'RB', 'O'),
 ('showed', 'VBD', 'O'),
 ('that', 'IN', 'O'),
 ('C', 'NNP', 'O'),
 ('...', ':', 'O'),
 ('2', 'CD', 'O'),
 ('Given', 'NNP', 'O'),
 ('that', 'WDT', 'O'),
 ('climate', 'VBP', 'O'),
 ('change', 'NN', 'B-NP'),
 ('is', 'VBZ', 'O'),
 ('increasing', 'VBG', 'O'),
 ('fire', 'NN', 'B-NP'),
 ('f', 'NN', 'B-NP'),
 ('...', ':', 'O'),
 ('3', 'CD', 'O'),
 ('This', 'DT', 'B-NP'),
 ('article', 'NN', 'I-NP'),
 ('appears', 'VBZ', 'O'),
 ('in', 'IN', 'O'),
 ('the', 'DT', 'O'),
 ('August', 'NNP', 'O'),
 ('17', 'CD', 'O'),
 (',', ',', 'O'),
 ('2019', 'CD', 'O'),
 ('is', 'VBZ', 'O'),
 ('...', ':', 'O'),
 ('4', 'CD', 'O'),
 ('NAL', 'NNP', 'O'),
 ('was', 'VBD', 'O'),
 ('renamed', 'VBN', 'O'),
 ('Fermilab', 'NNP', 'O'),
 ('in', 'IN', 'O'),
 ('1974', 'CD', 'O'),
 ('for', 'IN', 'O'),
 ('physicist', 'NN', 'B-NP'),
 ('...', ':', 'O'),
 ('5', 'CD', 'O'),
 ('This', 'DT', 'B-NP'),
 ('article', 'NN', 'I-NP'),
 ('appears', 'VBZ', 'O'),
 ('in', 'IN', 'O'),
 ('the', 'DT', 'O'),
 ('August', 'NNP', 'O'),
 ('17', 'CD', 'O'),
 (',', ',', 'O'),
```

```
    ('2019', 'CD', 'O'),
    ('is', 'VBZ', 'O'),
    ('...', ':', 'O'),
    ('6', 'CD', 'O'),
    ('The', 'DT', 'O'),
    ('United', 'NNP', 'O'),
    ('States', 'NNPS', 'O'),
    ('is', 'VBZ', 'O'),
    ('considered', 'VBN', 'O'),
    ('to', 'TO', 'O'),
    ('have', 'VB', 'O'),
    ('relati', 'NNS', 'O'),
    ('...', ':', 'O'),
    ('7', 'CD', 'O'),
    ('Knowing', 'VBG', 'O'),
    ('that', 'DT', 'O'),
    ('certain', 'JJ', 'O'),
    ('T', 'NNP', 'O'),
    ('cells', 'NNS', 'O'),
    (',', ',', 'O'),
    ('and', 'CC', 'O'),
    ('cytokines', 'NNS', 'O'),
    ('in', 'IN', 'O'),
    ('...', ':', 'O'),
    ('8', 'CD', 'O'),
    ('Wang', 'NNP', 'O'),
    ('and', 'CC', 'O'),
    ('colleagues', 'NNS', 'O'),
    ('now', 'RB', 'O'),
    ('plan', 'VBP', 'O'),
    ('to', 'TO', 'O'),
    ('take', 'VB', 'O'),
    ('a', 'DT', 'O'),
    ('larger', 'JJR', 'O'),
    ('...', ':', 'O'),
    ('9', 'CD', 'O'),
    ('Trent', 'NN', 'B-NP'),
    (':', ':', 'O'),
    ('As', 'IN', 'O'),
    ('pediatricians', 'NNS', 'O'),
    (',', ',', 'O'),
    ('we', 'PRP', 'O'),
    ('will', 'MD', 'O'),
    ('be', 'VB', 'O'),
    ('there', 'RB', 'O'),
    ('to', 'TO', 'O'),
    ('h', 'VB', 'O'),
    ('...', ':', 'O'),
    ('10', 'CD', 'O'),
    ('Adding', 'NNP', 'O'),
    ('several', 'JJ', 'O'),
    ('common', 'JJ', 'O'),
    ('cosmetic', 'JJ', 'O'),
    ('ingredients', 'NNS', 'O'),
    ('als', 'NNS', 'O'),
    ('...', ':', 'O'),
    ('11', 'CD', 'O'),
    ('With', 'IN', 'O'),
    ('neighborhood-level', 'JJ', 'O'),
    ('projections', 'NNS', 'O'),
    ('for', 'IN', 'O'),
```

```
    ('future', 'JJ', 'O'),
    ('...', ':', 'O'),
    ('12', 'CD', 'O'),
    ('Prior', 'NNP', 'O'),
    ('to', 'TO', 'O'),
    ('800', 'CD', 'O'),
    (',', ',', 'O'),
    ('Maya', 'NNP', 'O'),
    ('people', 'NNS', 'O'),
    ('may', 'MD', 'O'),
    ('have', 'VB', 'O'),
    ('considered', 'VBN', 'O'),
    ('...', ':', 'O'),
    ('13', 'CD', 'O'),
    ('In', 'IN', 'O'),
    ('2021', 'CD', 'O'),
    (',', ',', 'O'),
    ('NASA', 'NNP', 'O'),
    ('and', 'CC', 'O'),
    ('the', 'DT', 'O'),
    ('Indian', 'JJ', 'O'),
    ('Space', 'NNP', 'O'),
    ('Research', 'NNP', 'O'),
    ('Or', 'NNP', 'O'),
    ('...', ':', 'O'),
    ('14', 'CD', 'O'),
    ('Soils', 'NNS', 'O'),
    ('are', 'VBP', 'O'),
    ('typically', 'RB', 'O'),
    ('slow', 'JJ', 'O'),
    ('to', 'TO', 'O'),
    ('recover', 'VB', 'O'),
    ('calcium', 'NN', 'B-NP'),
    ('th', 'NN', 'B-NP'),
    ('...', ':', 'O'),
    ('15', 'CD', 'O'),
    ('Buy', 'NNP', 'O'),
    ('Archaeology', 'NNP', 'O'),
    ('from', 'IN', 'O'),
    ('SpaceÂ', 'NNP', 'O'),
    ('from', 'IN', 'O'),
    ('Amazon.com.Â', 'NNP', 'O'),
    ('...', ':', 'O'),
    ('16', 'CD', 'O'),
    ('Meanwhile', 'RB', 'O'),
    (',', ',', 'O'),
    ('increasingly', 'RB', 'O'),
    ('frequent', 'JJ', 'B-NP'),
    ('winter', 'NN', 'I-NP'),
    ('warm', 'NN', 'B-NP'),
    ('s', 'NN', 'B-NP'),
    ('...', ':', 'O'),
    ('17', 'CD', 'O'),
    ('The', 'DT', 'B-NP'),
    ('investigation', 'NN', 'I-NP'),
    ('of', 'IN', 'O'),
    ('the', 'DT', 'O'),
    ('Wisconsin', 'NNP', 'O'),
    ('teens', 'NNS', 'O'),
    ('could', 'MD', 'O'),
    ('...', ':', 'O'),
```

```
    ('18', 'CD', 'O'),
    ('Unfortunately', 'RB', 'O'),
    (',', ',', 'O'),
    ('the', 'DT', 'B-NP'),
    ('result', 'NN', 'I-NP'),
    ('might', 'MD', 'O'),
    ('mean', 'VB', 'O'),
    ('that', 'IN', 'O'),
    ('astr', 'NN', 'B-NP'),
    ('...', ':', 'O'),
    ('19', 'CD', 'O'),
    ('â\x80\x9cTrust', 'NN', 'B-NP'),
    ('is', 'VBZ', 'O'),
    ('important', 'JJ', 'O'),
    ('to', 'TO', 'O'),
    ('legitimacy', 'NN', 'B-NP'),
    (',', ',', 'O'),
    ('credibili', 'NN', 'B-NP'),
    ('...', ':', 'O'),
    ('20', 'CD', 'O'),
    ('Aagaard', 'NNP', 'O'),
    ('is', 'VBZ', 'O'),
    ('convinced', 'VBN', 'O'),
    ('there', 'EX', 'O'),
    ('are', 'VBP', 'O'),
    ('small', 'JJ', 'O'),
    ('amounts', 'NNS', 'O'),
    ('o', 'VBP', 'O'),
    ('...', ':', 'O'),
    ('21', 'CD', 'O'),
    ('3-D', 'JJ', 'O'),
    ('VISIONÂ', 'NNP', 'O'),
    ('The', 'DT', 'O'),
    ('Milky', 'NNP', 'O'),
    ('Wayâ\x80\x99s', 'NNP', 'O'),
    ('Cepheid', 'NNP', 'O'),
    ('stars', 'VBZ', 'O'),
    ('a', 'DT', 'O'),
    ('...', ':', 'O'),
    ('22', 'CD', 'O'),
    ('In', 'IN', 'O'),
    ('the', 'DT', 'B-NP'),
    ('intimidating', 'NN', 'I-NP'),
    ('body', 'NN', 'B-NP'),
    ('part', 'NN', 'B-NP'),
    ('catalog', 'NN', 'B-NP'),
    (',', ',', 'O'),
    ('pufferf', 'NN', 'B-NP'),
    ('...', ':', 'O'),
    ('23', 'CD', 'O'),
    ('Thatâ\x80\x99s', 'NNP', 'O'),
    ('probably', 'RB', 'O'),
    ('a', 'DT', 'B-NP'),
    ('valuable', 'JJ', 'I-NP'),
    ('ability', 'NN', 'I-NP'),
    ('in', 'IN', 'O'),
    ('the', 'DT', 'B-NP'),
    ('wi', 'NN', 'I-NP'),
    ('...', ':', 'O'),
    ('Name', 'NN', 'B-NP'),
    (':', ':', 'O'),
```

```
    ('Summary', 'NNP', 'O'),
    (',', ',', 'O'),
    ('dtype', 'NN', 'B-NP'),
    (':', ':', 'O'),
    ('object', 'NN', 'B-NP')]
```

In [34]:

```
from nltk import ne_chunk
nltk.download('maxent_ne_chunker')
nltk.download('words')
ne_tree = ne_chunk(pos_tag(word_tokenize(new))) # CATEGORY LABELS
print(ne_tree)
```

```
[nltk_data] Downloading package maxent_ne_chunker to
[nltk_data]     C:\Users\AKSHIT\AppData\Roaming\nltk_data...
[nltk_data]   Package maxent_ne_chunker is already up-to-date!
[nltk_data] Downloading package words to
[nltk_data]     C:\Users\AKSHIT\AppData\Roaming\nltk_data...
[nltk_data]   Package words is already up-to-date!
```

```
(S
  0/CD
  â□□This/NN
  is/VBZ
  actually/RB
  quite/RB
  a/DT
  profound/JJ
  thing/NN
  ,/,
  â□□/VBP
  .../:
  1/CD
  Poppinga/NNP
  and/CC
  colleagues/NNS
  recently/RB
  showed/VBD
  that/IN
  C/NNP
  .../:
  2/CD
  Given/NNP
  that/WDT
  climate/VBP
  change/NN
  is/VBZ
  increasing/VBG
  fire/NN
  f/NN
  .../:
  3/CD
  This/DT
  article/NN
  appears/VBZ
  in/IN
  the/DT
  August/NNP
  17/CD
  ,/,
  2019/CD
  is/VBZ
  .../:
  4/CD
  (ORGANIZATION NAL/NNP)
  was/VBD
  renamed/VBN
  (PERSON Fermilab/NNP)
  in/IN
  1974/CD
  for/IN
  physicist/NN
  .../:
  5/CD
  This/DT
  article/NN
  appears/VBZ
  in/IN
  the/DT
  August/NNP
  17/CD
```

```
,/,
2019/CD
is/VBZ
.../:
6/CD
The/DT
(GPE United/NNP States/NNPS)
is/VBZ
considered/VBN
to/TO
have/VB
relati/NNS
.../:
7/CD
Knowing/VBG
that/DT
certain/JJ
T/NNP
cells/NNS
,/,
and/CC
cytokines/NNS
in/IN
.../:
8/CD
(PERSON Wang/NNP)
and/CC
colleagues/NNS
now/RB
plan/VBP
to/TO
take/VB
a/DT
larger/JJR
.../:
9/CD
Trent/NN
:/:
As/IN
pediatricians/NNS
,/,
we/PRP
will/MD
be/VB
there/RB
to/TO
h/VB
.../:
10/CD
Adding/NNP
several/JJ
common/JJ
cosmetic/JJ
ingredients/NNS
als/NNS
.../:
11/CD
With/IN
neighborhood-level/JJ
projections/NNS
for/IN
```

```
future/JJ
.../:
12/CD
Prior/NNP
to/TO
800/CD
,/,
Maya/NNP
people/NNS
may/MD
have/VB
considered/VBN
.../:
13/CD
In/IN
2021/CD
,/,
(ORGANIZATION NASA/NNP)
and/CC
the/DT
(GPE Indian/JJ)
(ORGANIZATION Space/NNP Research/NNP Or/NNP)
.../:
14/CD
Soils/NNS
are/VBP
typically/RB
slow/JJ
to/TO
recover/VB
calcium/NN
th/NN
.../:
15/CD
Buy/NNP
Archaeology/NNP
from/IN
(ORGANIZATION SpaceÂ/NNP)
from/IN
Amazon.com.Â/NNP
.../:
16/CD
Meanwhile/RB
,/,
increasingly/RB
frequent/JJ
winter/NN
warm/NN
s/NN
.../:
17/CD
The/DT
investigation/NN
of/IN
the/DT
(ORGANIZATION Wisconsin/NNP)
teens/NNS
could/MD
.../:
18/CD
Unfortunately/RB
```

,/,
the/DT
result/NN
might/MD
mean/VB
that/IN
astr/NN
.../:
19/CD
â□□Trust/NN
is/VBZ
important/JJ
to/TO
legitimacy/NN
,/,
credibili/NN
.../:
20/CD
(PERSON Aagaard/NNP)
is/VBZ
convinced/VBN
there/EX
are/VBP
small/JJ
amounts/NNS
o/VBP
.../:
21/CD
3-D/JJ
VISIONÂ/NNP
The/DT
(PERSON Milky/NNP)
Wayâ□□s/NNP
Cepheid/NNP
stars/VBZ
a/DT
.../:
22/CD
In/IN
the/DT
intimidating/NN
body/NN
part/NN
catalog/NN
,/,
pufferf/NN
.../:
23/CD
Thatâ□□s/NNP
probably/RB
a/DT
valuable/JJ
ability/NN
in/IN
the/DT
wi/NN
.../:
Name/NN
:/:
Summary/NNP
,/,

```
dtype/NN
:/:
object/NN)
```

In [35]:

```
# NER BY SPACY
import spacy
from spacy import displacy
from collections import Counter
import en_core_web_sm
nlp = en_core_web_sm.load()
```

In [46]:

```
doc = nlp(new)
pprint([(X, X) for X in doc.ents])
```

```
[(0, 0),
 (1, 1),
 (Poppinga, Poppinga),
 (2, 2),
 (3, 3),
 (the August 17, 2019, the August 17, 2019),
 (4, 4),
 (NAL, NAL),
 (Fermilab, Fermilab),
 (1974, 1974),
 (5, 5),
 (the August 17, 2019, the August 17, 2019),
 (6, 6),
 (The United States, The United States),
 (7, 7),
 (Knowing, Knowing),
 (8, 8),
 (Wang, Wang),
 (9, 9),
 (10, 10),
 (11, 11),
 (12, 12),
 (800, 800),
 (Maya, Maya),
 (13, 13),
 (2021, 2021),
 (NASA, NASA),
 (the Indian Space Research Or, the Indian Space Research Or),
 (14, 14),
 (Soils, Soils),
 (15, 15),
 (Amazon.com, Amazon.com),
 (16, 16),
 (17, 17),
 (Wisconsin, Wisconsin),
 (18, 18),
 (19, 19),
 (20, 20),
 (21, 21),
 (The Milky, The Milky),
 (22, 22),
 (23, 23)]
```

In [47]:

```
pprint([(X, X.ent_iob_, X.ent_type_) for X in doc])
#BILUO TAGGING "B" :the token begins an entity,
#"I": it is inside an entity, "O" :it is outside an entity, and "" : no entity tag is s
et.
```

```
[(0, 'B', 'CARDINAL'),
 (    , 'O', ''),
 (â□□This, 'O', ''),
 (is, 'O', ''),
 (actually, 'O', ''),
 (quite, 'O', ''),
 (a, 'O', ''),
 (profound, 'O', ''),
 (thing, 'O', ''),
 (,, 'O', ''),
 (â□□, 'O', ''),
 (..., 'O', ''),
 (
, 'O', ''),
 (1, 'B', 'CARDINAL'),
 (    , 'O', ''),
 (Poppinga, 'B', 'ORG'),
 (and, 'O', ''),
 (colleagues, 'O', ''),
 (recently, 'O', ''),
 (showed, 'O', ''),
 (that, 'O', ''),
 (C, 'O', ''),
 (..., 'O', ''),
 (
, 'O', ''),
 (2, 'B', 'CARDINAL'),
 (    , 'O', ''),
 (Given, 'O', ''),
 (that, 'O', ''),
 (climate, 'O', ''),
 (change, 'O', ''),
 (is, 'O', ''),
 (increasing, 'O', ''),
 (fire, 'O', ''),
 (f, 'O', ''),
 (..., 'O', ''),
 (
, 'O', ''),
 (3, 'B', 'CARDINAL'),
 (    , 'O', ''),
 (This, 'O', ''),
 (article, 'O', ''),
 (appears, 'O', ''),
 (in, 'O', ''),
 (the, 'B', 'EVENT'),
 (August, 'I', 'EVENT'),
 (17, 'I', 'EVENT'),
 (,, 'I', 'EVENT'),
 (2019, 'I', 'EVENT'),
 (is, 'O', ''),
 (..., 'O', ''),
 (
, 'O', ''),
 (4, 'B', 'CARDINAL'),
 (    , 'O', ''),
 (NAL, 'B', 'ORG'),
 (was, 'O', ''),
 (renamed, 'O', ''),
 (Fermilab, 'B', 'PERSON'),
 (in, 'O', ''),
```

```
    (1974, 'B', 'DATE'),
    (for, 'O', ''),
    (physicist, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (5, 'B', 'CARDINAL'),
    (    , 'O', ''),
    (This, 'O', ''),
    (article, 'O', ''),
    (appears, 'O', ''),
    (in, 'O', ''),
    (the, 'B', 'EVENT'),
    (August, 'I', 'EVENT'),
    (17, 'I', 'EVENT'),
    (,, 'I', 'EVENT'),
    (2019, 'I', 'EVENT'),
    (is, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (6, 'B', 'CARDINAL'),
    (    , 'O', ''),
    (The, 'B', 'GPE'),
    (United, 'I', 'GPE'),
    (States, 'I', 'GPE'),
    (is, 'O', ''),
    (considered, 'O', ''),
    (to, 'O', ''),
    (have, 'O', ''),
    (relati, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (7, 'B', 'CARDINAL'),
    (    , 'O', ''),
    (Knowing, 'B', 'GPE'),
    (that, 'O', ''),
    (certain, 'O', ''),
    (T, 'O', ''),
    (cells, 'O', ''),
    (,, 'O', ''),
    (and, 'O', ''),
    (cytokines, 'O', ''),
    (in, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (8, 'B', 'CARDINAL'),
    (    , 'O', ''),
    (Wang, 'B', 'ORG'),
    (and, 'O', ''),
    (colleagues, 'O', ''),
    (now, 'O', ''),
    (plan, 'O', ''),
    (to, 'O', ''),
    (take, 'O', ''),
    (a, 'O', ''),
    (larger, 'O', ''),
    (..., 'O', ''),
    (
```

```
   , 'O', ''),
    (9, 'B', 'CARDINAL'),
    (    , 'O', ''),
    (Trent, 'O', ''),
    (:, 'O', ''),
    (As, 'O', ''),
    (pediatricians, 'O', ''),
    (,, 'O', ''),
    (we, 'O', ''),
    (will, 'O', ''),
    (be, 'O', ''),
    (there, 'O', ''),
    (to, 'O', ''),
    (h, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (10, 'B', 'CARDINAL'),
    (   , 'O', ''),
    (Adding, 'O', ''),
    (several, 'O', ''),
    (common, 'O', ''),
    (cosmetic, 'O', ''),
    (ingredients, 'O', ''),
    (als, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (11, 'B', 'CARDINAL'),
    (   , 'O', ''),
    (With, 'O', ''),
    (neighborhood, 'O', ''),
    (-, 'O', ''),
    (level, 'O', ''),
    (projections, 'O', ''),
    (for, 'O', ''),
    (future, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (12, 'B', 'CARDINAL'),
    (   , 'O', ''),
    (Prior, 'O', ''),
    (to, 'O', ''),
    (800, 'B', 'CARDINAL'),
    (,, 'O', ''),
    (Maya, 'B', 'NORP'),
    (people, 'O', ''),
    (may, 'O', ''),
    (have, 'O', ''),
    (considered, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (13, 'B', 'CARDINAL'),
    (   , 'O', ''),
    (In, 'O', ''),
    (2021, 'B', 'CARDINAL'),
    (,, 'O', ''),
    (NASA, 'B', 'ORG'),
    (and, 'O', ''),
```

```
    (the, 'B', 'ORG'),
    (Indian, 'I', 'ORG'),
    (Space, 'I', 'ORG'),
    (Research, 'I', 'ORG'),
    (Or, 'I', 'ORG'),
    (..., 'O', ''),
    (
   , 'O', ''),
    (14, 'B', 'CARDINAL'),
    (   , 'O', ''),
    (Soils, 'B', 'PERSON'),
    (are, 'O', ''),
    (typically, 'O', ''),
    (slow, 'O', ''),
    (to, 'O', ''),
    (recover, 'O', ''),
    (calcium, 'O', ''),
    (th, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (15, 'B', 'CARDINAL'),
    (   , 'O', ''),
    (Buy, 'O', ''),
    (Archaeology, 'O', ''),
    (from, 'O', ''),
    (SpaceÂ, 'O', ''),
    ( , 'O', ''),
    (from, 'O', ''),
    (Amazon.com, 'B', 'PRODUCT'),
    (., 'O', ''),
    (Â, 'O', ''),
    ( , 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (16, 'B', 'CARDINAL'),
    (   , 'O', ''),
    (Meanwhile, 'O', ''),
    (,, 'O', ''),
    (increasingly, 'O', ''),
    (frequent, 'O', ''),
    (winter, 'O', ''),
    (warm, 'O', ''),
    (s, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (17, 'B', 'CARDINAL'),
    (   , 'O', ''),
    (The, 'O', ''),
    (investigation, 'O', ''),
    (of, 'O', ''),
    (the, 'O', ''),
    (Wisconsin, 'B', 'GPE'),
    (teens, 'O', ''),
    (could, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (18, 'B', 'CARDINAL'),
```

```
    (   , 'O', ''),
    (Unfortunately, 'O', ''),
    (,, 'O', ''),
    (the, 'O', ''),
    (result, 'O', ''),
    (might, 'O', ''),
    (mean, 'O', ''),
    (that, 'O', ''),
    (astr, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (19, 'B', 'CARDINAL'),
    (   , 'O', ''),
    (â□□Trust, 'O', ''),
    (is, 'O', ''),
    (important, 'O', ''),
    (to, 'O', ''),
    (legitimacy, 'O', ''),
    (,, 'O', ''),
    (credibili, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (20, 'B', 'CARDINAL'),
    (   , 'O', ''),
    (Aagaard, 'O', ''),
    (is, 'O', ''),
    (convinced, 'O', ''),
    (there, 'O', ''),
    (are, 'O', ''),
    (small, 'O', ''),
    (amounts, 'O', ''),
    (o, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (21, 'B', 'CARDINAL'),
    (   , 'O', ''),
    (3-D, 'O', ''),
    (VISIONÂ, 'O', ''),
    (  , 'O', ''),
    (The, 'B', 'ORG'),
    (Milky, 'I', 'ORG'),
    (Wayâ□□s, 'O', ''),
    (Cepheid, 'O', ''),
    (stars, 'O', ''),
    (a, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (22, 'B', 'CARDINAL'),
    (   , 'O', ''),
    (In, 'O', ''),
    (the, 'O', ''),
    (intimidating, 'O', ''),
    (body, 'O', ''),
    (part, 'O', ''),
    (catalog, 'O', ''),
    (,, 'O', ''),
    (pufferf, 'O', ''),
```

```
    (..., 'O', ''),
    (
   , 'O', ''),
    (23, 'B', 'CARDINAL'),
    (   , 'O', ''),
    (Thatâ□□s, 'O', ''),
    (probably, 'O', ''),
    (a, 'O', ''),
    (valuable, 'O', ''),
    (ability, 'O', ''),
    (in, 'O', ''),
    (the, 'O', ''),
    (wi, 'O', ''),
    (..., 'O', ''),
    (
   , 'O', ''),
    (Name, 'O', ''),
    (:, 'O', ''),
    (Summary, 'O', ''),
    (,, 'O', ''),
    (dtype, 'O', ''),
    (:, 'O', ''),
    (object, 'O', '')]
```

In [59]:

```
labels = [x for x in doc]
Counter(labels) # EXTRACTING UNIQUE ENTITY LABELS FROM THE DATASET
```

Out[59]:

```
Counter({0: 1,
             : 1,
         â□□This: 1,
         is: 1,
         actually: 1,
         quite: 1,
         a: 1,
         profound: 1,
         thing: 1,
         ,: 1,
         â□□: 1,
         ...: 1,
         : 1,
         1: 1,
             : 1,
         Poppinga: 1,
         and: 1,
         colleagues: 1,
         recently: 1,
         showed: 1,
         that: 1,
         C: 1,
         ...: 1,
         : 1,
         2: 1,
             : 1,
         Given: 1,
         that: 1,
         climate: 1,
         change: 1,
         is: 1,
         increasing: 1,
         fire: 1,
         f: 1,
         ...: 1,
         : 1,
         3: 1,
             : 1,
         This: 1,
         article: 1,
         appears: 1,
         in: 1,
         the: 1,
         August: 1,
         17: 1,
         ,: 1,
         2019: 1,
         is: 1,
         ...: 1,
         : 1,
         4: 1,
             : 1,
         NAL: 1,
         was: 1,
         renamed: 1,
         Fermilab: 1,
         in: 1,
         1974: 1,
         for: 1,
```

```
physicist: 1,
...: 1,
: 1,
5: 1,
    : 1,
This: 1,
article: 1,
appears: 1,
in: 1,
the: 1,
August: 1,
17: 1,
,: 1,
2019: 1,
is: 1,
...: 1,
: 1,
6: 1,
    : 1,
The: 1,
United: 1,
States: 1,
is: 1,
considered: 1,
to: 1,
have: 1,
relati: 1,
...: 1,
: 1,
7: 1,
    : 1,
Knowing: 1,
that: 1,
certain: 1,
T: 1,
cells: 1,
,: 1,
and: 1,
cytokines: 1,
in: 1,
...: 1,
: 1,
8: 1,
    : 1,
Wang: 1,
and: 1,
colleagues: 1,
now: 1,
plan: 1,
to: 1,
take: 1,
a: 1,
larger: 1,
...: 1,
: 1,
9: 1,
    : 1,
Trent: 1,
:: 1,
As: 1,
pediatricians: 1,
```

```
,: 1,
we: 1,
will: 1,
be: 1,
there: 1,
to: 1,
h: 1,
...: 1,
: 1,
10: 1,
   : 1,
Adding: 1,
several: 1,
common: 1,
cosmetic: 1,
ingredients: 1,
als: 1,
...: 1,
: 1,
11: 1,
   : 1,
With: 1,
neighborhood: 1,
-: 1,
level: 1,
projections: 1,
for: 1,
future: 1,
...: 1,
: 1,
12: 1,
   : 1,
Prior: 1,
to: 1,
800: 1,
,: 1,
Maya: 1,
people: 1,
may: 1,
have: 1,
considered: 1,
...: 1,
: 1,
13: 1,
   : 1,
In: 1,
2021: 1,
,: 1,
NASA: 1,
and: 1,
the: 1,
Indian: 1,
Space: 1,
Research: 1,
Or: 1,
...: 1,
: 1,
14: 1,
   : 1,
Soils: 1,
are: 1,
```

typically: 1,
slow: 1,
to: 1,
recover: 1,
calcium: 1,
th: 1,
...: 1,
: 1,
15: 1,
   : 1,
Buy: 1,
Archaeology: 1,
from: 1,
SpaceÂ: 1,
 : 1,
from: 1,
Amazon.com: 1,
.: 1,
Â: 1,
 : 1,
...: 1,
: 1,
16: 1,
   : 1,
Meanwhile: 1,
,: 1,
increasingly: 1,
frequent: 1,
winter: 1,
warm: 1,
s: 1,
...: 1,
: 1,
17: 1,
   : 1,
The: 1,
investigation: 1,
of: 1,
the: 1,
Wisconsin: 1,
teens: 1,
could: 1,
...: 1,
: 1,
18: 1,
   : 1,
Unfortunately: 1,
,: 1,
the: 1,
result: 1,
might: 1,
mean: 1,
that: 1,
astr: 1,
...: 1,
: 1,
19: 1,
   : 1,
â□□Trust: 1,
is: 1,
important: 1,

```
to: 1,
legitimacy: 1,
,: 1,
credibili: 1,
...: 1,
: 1,
20: 1,
   : 1,
Aagaard: 1,
is: 1,
convinced: 1,
there: 1,
are: 1,
small: 1,
amounts: 1,
o: 1,
...: 1,
: 1,
21: 1,
   : 1,
3-D: 1,
VISIONÂ: 1,
  : 1,
The: 1,
Milky: 1,
Wayâ□□s: 1,
Cepheid: 1,
stars: 1,
a: 1,
...: 1,
: 1,
22: 1,
   : 1,
In: 1,
the: 1,
intimidating: 1,
body: 1,
part: 1,
catalog: 1,
,: 1,
pufferf: 1,
...: 1,
: 1,
23: 1,
   : 1,
Thatâ□□s: 1,
probably: 1,
a: 1,
valuable: 1,
ability: 1,
in: 1,
the: 1,
wi: 1,
...: 1,
: 1,
Name: 1,
:: 1,
Summary: 1,
,: 1,
dtype: 1,
```

```
          :: 1,
          object: 1})
```

In [119]:

```
items = [x for x in doc]
Counter(items).most_common(20) # 20 most frequent tokens in article data
```

Out[119]:

```
[(0, 1),
 (    , 1),
 (â□□This, 1),
 (is, 1),
 (actually, 1),
 (quite, 1),
 (a, 1),
 (profound, 1),
 (thing, 1),
 (,, 1),
 (â□□, 1),
 (..., 1),
 (, 1),
 (1, 1),
 (    , 1),
 (Poppinga, 1),
 (and, 1),
 (colleagues, 1),
 (recently, 1),
 (showed, 1)]
```

In [116]:

```
displacy.render(nlp(str(new)), jupyter=True, style='ent') # generates raw markup  for
 "named entity visualization"
```

0 **CARDINAL** â□□This is actually quite a profound thing,â□□...

1 **CARDINAL** Poppinga **ORG** and colleagues recently showed that C...

2 **CARDINAL** Given that climate change is increasing fire f...

3 **CARDINAL** This article appears in the August 17, 2019 **EVENT** is...

4 **CARDINAL** NAL **ORG** was renamed Fermilab **PERSON** in 1974 **DATE** for physicist...

5 **CARDINAL** This article appears in the August 17, 2019 **EVENT** is...

6 **CARDINAL** The United States **GPE** is considered to have relati...

7 **CARDINAL** Knowing **GPE** that certain T cells, and cytokines in...

8 **CARDINAL** Wang **ORG** and colleagues now plan to take a larger ...

9 **CARDINAL** Trent: As pediatricians, we will be there to h...

10 **CARDINAL** Adding several common cosmetic ingredients als...

11 **CARDINAL** With neighborhood-level projections for future...

12 **CARDINAL** Prior to 800 **CARDINAL** , Maya **NORP** people may have considered ...

13 **CARDINAL** In 2021 **CARDINAL** , NASA **ORG** and the Indian Space Research Or **ORG** ...

14 **CARDINAL** Soils **PERSON** are typically slow to recover calcium th...

15 **CARDINAL** Buy Archaeology from SpaceÂ from Amazon.com **PRODUCT** .Â ...

16 **CARDINAL** Meanwhile, increasingly frequent winter warm s...

17 **CARDINAL** The investigation of the Wisconsin **GPE** teens could...

18 **CARDINAL** Unfortunately, the result might mean that astr...

19 **CARDINAL** â□□Trust is important to legitimacy, credibili...

20 **CARDINAL** Aagaard is convinced there are small amounts o...

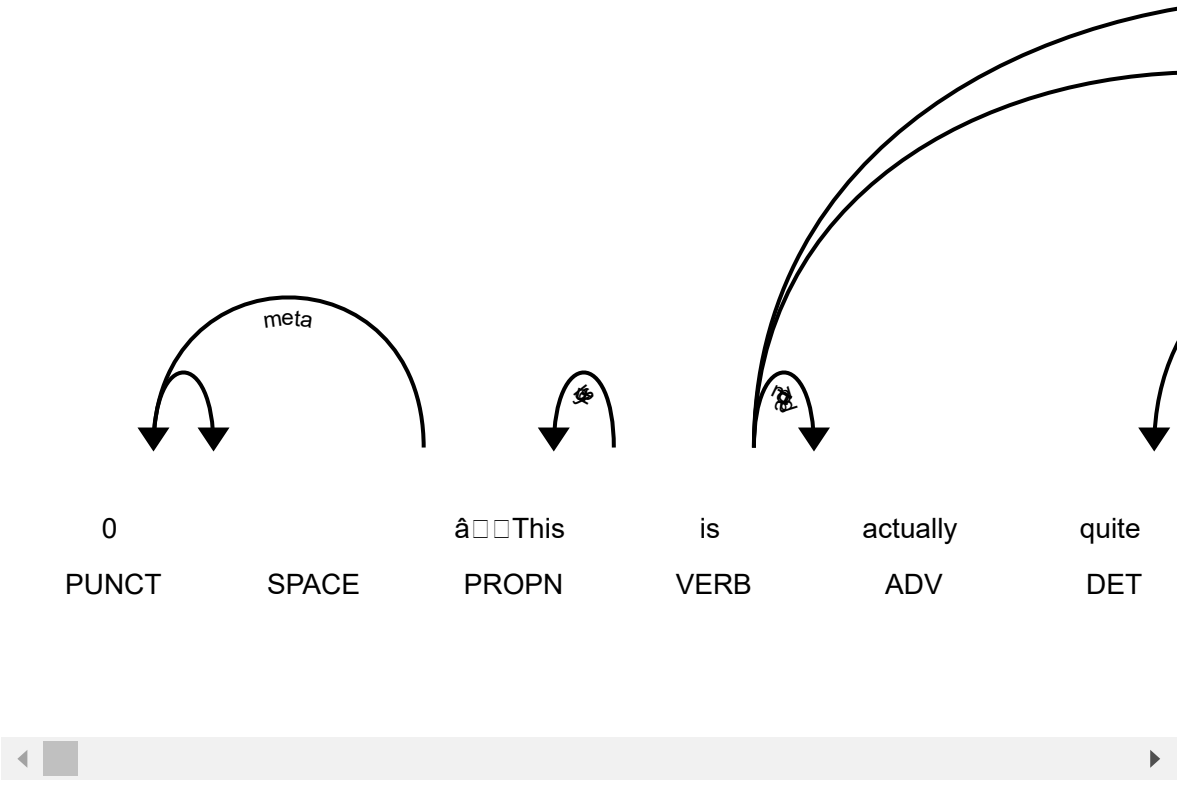
21 **CARDINAL** 3-D VISIONÂ The Milky **ORG** Wayâ□□s Cepheid stars a...

22 **CARDINAL** In the intimidating body part catalog, pufferf...

23 **CARDINAL** Thatâ□□s probably a valuable ability in the wi... Name: Summary, dtype: object

In [121]:

```
displacy.render(nlp(str(new)), style='dep', jupyter = True, options = {'distance': 100
})
# another type of display.render style: dep/ent
```



In [83]:

```
[(x.orth_,x.pos_, x.lemma_) for x in [y
                                      for y
                                      in nlp(str(doc))
                                      if not y.is_stop and y.pos_ != 'PUNCT']]
# extracting the part-of-speech and then lemmatizing this sentences
```

Out[83]:

```
[('    ', 'SPACE', '    '),
 ('â\x80\x9cThis', 'PROPN', 'â\x80\x9cThis'),
 ('actually', 'ADV', 'actually'),
 ('profound', 'ADJ', 'profound'),
 ('thing', 'NOUN', 'thing'),
 ('â\x80\x9d', 'NOUN', 'â\x80\x9d'),
 ('\n', 'SPACE', '\n'),
 ('1', 'NUM', '1'),
 ('    ', 'SPACE', '    '),
 ('Poppinga', 'PROPN', 'Poppinga'),
 ('colleagues', 'NOUN', 'colleague'),
 ('recently', 'ADV', 'recently'),
 ('showed', 'VERB', 'show'),
 ('C', 'NOUN', 'c'),
 ('\n', 'SPACE', '\n'),
 ('2', 'NUM', '2'),
 ('    ', 'SPACE', '    '),
 ('Given', 'VERB', 'give'),
 ('climate', 'NOUN', 'climate'),
 ('change', 'NOUN', 'change'),
 ('increasing', 'VERB', 'increase'),
 ('fire', 'NOUN', 'fire'),
 ('f', 'X', 'f'),
 ('\n', 'SPACE', '\n'),
 ('3', 'NUM', '3'),
 ('    ', 'SPACE', '    '),
 ('article', 'NOUN', 'article'),
 ('appears', 'VERB', 'appear'),
 ('August', 'PROPN', 'August'),
 ('17', 'NUM', '17'),
 ('2019', 'NUM', '2019'),
 ('\n', 'SPACE', '\n'),
 ('4', 'NUM', '4'),
 ('    ', 'SPACE', '    '),
 ('NAL', 'PROPN', 'NAL'),
 ('renamed', 'VERB', 'rename'),
 ('Fermilab', 'PROPN', 'Fermilab'),
 ('1974', 'NUM', '1974'),
 ('physicist', 'NOUN', 'physicist'),
 ('\n', 'SPACE', '\n'),
 ('5', 'NUM', '5'),
 ('    ', 'SPACE', '    '),
 ('article', 'NOUN', 'article'),
 ('appears', 'VERB', 'appear'),
 ('August', 'PROPN', 'August'),
 ('17', 'NUM', '17'),
 ('2019', 'NUM', '2019'),
 ('\n', 'SPACE', '\n'),
 ('6', 'NUM', '6'),
 ('    ', 'SPACE', '    '),
 ('United', 'PROPN', 'United'),
 ('States', 'PROPN', 'States'),
 ('considered', 'VERB', 'consider'),
 ('relati', 'NOUN', 'relati'),
 ('\n', 'SPACE', '\n'),
 ('7', 'NUM', '7'),
 ('    ', 'SPACE', '    '),
 ('Knowing', 'VERB', 'know'),
 ('certain', 'ADJ', 'certain'),
```

```
    ('T', 'PROPN', 'T'),
    ('cells', 'NOUN', 'cell'),
    ('cytokines', 'VERB', 'cytokine'),
    ('\n', 'SPACE', '\n'),
    ('8', 'NUM', '8'),
    ('    ', 'SPACE', '    '),
    ('Wang', 'PROPN', 'Wang'),
    ('colleagues', 'NOUN', 'colleague'),
    ('plan', 'VERB', 'plan'),
    ('larger', 'ADJ', 'large'),
    ('\n', 'SPACE', '\n'),
    ('9', 'NUM', '9'),
    ('    ', 'SPACE', '    '),
    ('Trent', 'PROPN', 'Trent'),
    ('pediatricians', 'NOUN', 'pediatrician'),
    ('h', 'NOUN', 'h'),
    ('\n', 'SPACE', '\n'),
    ('10', 'NUM', '10'),
    ('   ', 'SPACE', '   '),
    ('Adding', 'VERB', 'add'),
    ('common', 'ADJ', 'common'),
    ('cosmetic', 'ADJ', 'cosmetic'),
    ('ingredients', 'NOUN', 'ingredient'),
    ('als', 'NOUN', 'al'),
    ('\n', 'SPACE', '\n'),
    ('11', 'NUM', '11'),
    ('   ', 'SPACE', '   '),
    ('neighborhood', 'NOUN', 'neighborhood'),
    ('level', 'NOUN', 'level'),
    ('projections', 'NOUN', 'projection'),
    ('future', 'NOUN', 'future'),
    ('\n', 'SPACE', '\n'),
    ('12', 'NUM', '12'),
    ('   ', 'SPACE', '   '),
    ('Prior', 'ADV', 'prior'),
    ('800', 'NUM', '800'),
    ('Maya', 'PROPN', 'Maya'),
    ('people', 'NOUN', 'people'),
    ('considered', 'VERB', 'consider'),
    ('\n', 'SPACE', '\n'),
    ('13', 'NUM', '13'),
    ('   ', 'SPACE', '   '),
    ('2021', 'NUM', '2021'),
    ('NASA', 'PROPN', 'NASA'),
    ('Indian', 'PROPN', 'Indian'),
    ('Space', 'PROPN', 'Space'),
    ('Research', 'PROPN', 'Research'),
    ('\n', 'SPACE', '\n'),
    ('14', 'NUM', '14'),
    ('   ', 'SPACE', '   '),
    ('Soils', 'PROPN', 'Soils'),
    ('typically', 'ADV', 'typically'),
    ('slow', 'ADJ', 'slow'),
    ('recover', 'VERB', 'recover'),
    ('calcium', 'NOUN', 'calcium'),
    ('th', 'X', 'th'),
    ('\n', 'SPACE', '\n'),
    ('15', 'NUM', '15'),
    ('   ', 'SPACE', '   '),
    ('Buy', 'PROPN', 'Buy'),
    ('Archaeology', 'PROPN', 'Archaeology'),
```

```
    ('SpaceÂ', 'PROPN', 'SpaceÂ'),
    ('\xa0', 'SPACE', '  '),
    ('Amazon.com', 'PROPN', 'Amazon.com'),
    ('Â', 'PROPN', 'Â'),
    ('\xa0', 'SPACE', '  '),
    ('\n', 'SPACE', '\n'),
    ('16', 'NUM', '16'),
    ('   ', 'SPACE', '   '),
    ('increasingly', 'ADV', 'increasingly'),
    ('frequent', 'ADJ', 'frequent'),
    ('winter', 'NOUN', 'winter'),
    ('warm', 'ADJ', 'warm'),
    ('s', 'NOUN', 's'),
    ('\n', 'SPACE', '\n'),
    ('17', 'NUM', '17'),
    ('   ', 'SPACE', '   '),
    ('investigation', 'NOUN', 'investigation'),
    ('Wisconsin', 'PROPN', 'Wisconsin'),
    ('teens', 'NOUN', 'teen'),
    ('\n', 'SPACE', '\n'),
    ('18', 'NUM', '18'),
    ('   ', 'SPACE', '   '),
    ('Unfortunately', 'ADV', 'unfortunately'),
    ('result', 'NOUN', 'result'),
    ('mean', 'VERB', 'mean'),
    ('astr', 'ADV', 'astr'),
    ('\n', 'SPACE', '\n'),
    ('19', 'NUM', '19'),
    ('   ', 'SPACE', '   '),
    ('â\x80\x9cTrust', 'X', 'â\x80\x9ctrust'),
    ('important', 'ADJ', 'important'),
    ('legitimacy', 'VERB', 'legitimacy'),
    ('credibili', 'NOUN', 'credibili'),
    ('\n', 'SPACE', '\n'),
    ('20', 'NUM', '20'),
    ('   ', 'SPACE', '   '),
    ('Aagaard', 'PROPN', 'Aagaard'),
    ('convinced', 'VERB', 'convince'),
    ('small', 'ADJ', 'small'),
    ('amounts', 'NOUN', 'amount'),
    ('o', 'X', 'o'),
    ('\n', 'SPACE', '\n'),
    ('21', 'NUM', '21'),
    ('   ', 'SPACE', '   '),
    ('3-D', 'NUM', '3-d'),
    ('VISIONÂ', 'PROPN', 'VISIONÂ'),
    ('\xa0 ', 'SPACE', '\xa0 '),
    ('Milky', 'PROPN', 'Milky'),
    ('Wayâ\x80\x99s', 'PROPN', 'Wayâ\x80\x99s'),
    ('Cepheid', 'PROPN', 'Cepheid'),
    ('stars', 'VERB', 'star'),
    ('\n', 'SPACE', '\n'),
    ('22', 'NUM', '22'),
    ('   ', 'SPACE', '   '),
    ('intimidating', 'VERB', 'intimidate'),
    ('body', 'NOUN', 'body'),
    ('catalog', 'NOUN', 'catalog'),
    ('pufferf', 'NOUN', 'pufferf'),
    ('\n', 'SPACE', '\n'),
    ('23', 'NUM', '23'),
    ('   ', 'SPACE', '   '),
```

```
('Thatâ\x80\x99s', 'NUM', 'thatâ\x80\x99s'),
('probably', 'ADV', 'probably'),
('valuable', 'ADJ', 'valuable'),
('ability', 'NOUN', 'ability'),
('wi', 'NOUN', 'wi'),
('\n', 'SPACE', '\n'),
('Summary', 'PROPN', 'Summary'),
('dtype', 'NOUN', 'dtype'),
('object', 'NOUN', 'object')]
```

In [126]:

```
dict([(x, x.label_) for x in nlp(str(new)).ents]) # list of all named entity extracted
```

Out[126]:

```
{0: 'CARDINAL',
 1: 'CARDINAL',
 Poppinga: 'ORG',
 2: 'CARDINAL',
 3: 'CARDINAL',
 the August 17, 2019: 'EVENT',
 4: 'CARDINAL',
 NAL: 'ORG',
 Fermilab: 'PERSON',
 1974: 'DATE',
 5: 'CARDINAL',
 the August 17, 2019: 'EVENT',
 6: 'CARDINAL',
 The United States: 'GPE',
 7: 'CARDINAL',
 Knowing: 'GPE',
 8: 'CARDINAL',
 Wang: 'ORG',
 9: 'CARDINAL',
 10: 'CARDINAL',
 11: 'CARDINAL',
 12: 'CARDINAL',
 800: 'CARDINAL',
 Maya: 'NORP',
 13: 'CARDINAL',
 2021: 'CARDINAL',
 NASA: 'ORG',
 the Indian Space Research Or: 'ORG',
 14: 'CARDINAL',
 Soils: 'PERSON',
 15: 'CARDINAL',
 Amazon.com: 'PRODUCT',
 16: 'CARDINAL',
 17: 'CARDINAL',
 Wisconsin: 'GPE',
 18: 'CARDINAL',
 19: 'CARDINAL',
 20: 'CARDINAL',
 21: 'CARDINAL',
 The Milky: 'ORG',
 22: 'CARDINAL',
 23: 'CARDINAL'}
```

In [21]:

```
#.....................................TOPIC MODELLIN
G.....................................#####
import pandas as pd
df = pd.read_csv('sciencenews_all_data.csv')
data_text = df
data_text['index'] = data_text.index
documents = data_text
print(len(documents))
print(documents[:5])
```

```
24
                                          Topic Name  \
0  Exploding stars scattered traces of iron over ...
1     How these tiny insect larvae leap without legs
2  The worst wildfires can send smoke high enough...
3  How pieces of live human brain are helping sci...
4           50 years ago, Fermilab turned to bubbles

                                               Link         Author name
\
0  https://www.sciencenews.org//article/exploding...       Emily Conover
1  https://www.sciencenews.org//article/how-these...        Susan Milius
2  https://www.sciencenews.org//article/worst-wil...         Megan Sever
3  https://www.sciencenews.org//article/experimen...       Laura Sanders
4  https://www.sciencenews.org//article/50-years-...  Bethany Brookshire

          Date of Posting                                                Summa
ry  \
0  6:00am, August 9, 2019  “This is actually quite a profound thing,” sa
y...
1  6:20pm, August 8, 2019  Poppinga and colleagues recently showed that
C...
2  2:00pm, August 8, 2019  Given that climate change is increasing fire
f...
3  6:00am, August 7, 2019  This article appears in the August 17, 2019 i
s...
4  8:00am, August 8, 2019  NAL was renamed Fermilab in 1974 for physicis
t...

   index
0      0
1      1
2      2
3      3
4      4
```

In [22]:

```
import gensim
from gensim.utils import simple_preprocess
from gensim.parsing.preprocessing import STOPWORDS
from nltk.stem import WordNetLemmatizer, SnowballStemmer
from nltk.stem.porter import *
import numpy as np
np.random.seed(2018)
import nltk
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\AKSHIT\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Out[22]:

```
True
```

In [23]:

```
from nltk import PorterStemmer

# Lemmatize and stem preprocessing

PorterStemmer().stem('complications')
stemmer = PorterStemmer()
def lemmatize_stemming(text):
    return stemmer.stem(WordNetLemmatizer().lemmatize(text, pos='v'))
def preprocess(text):
    result = []
    for token in gensim.utils.simple_preprocess(text):
        if token not in gensim.parsing.preprocessing.STOPWORDS and len(token) > 3:
            result.append(lemmatize_stemming(token))
    return result
```

In [24]:

```
doc_sample = documents[documents['index'] == 12].values[0][0]
print('original sentence: ')
words = []
for word in doc_sample.split(' '):
    words.append(word)
print(words)
print('\n\n tokenized and lemmatized sentence: ')
print(preprocess(doc_sample))

# can run for different sentences within data range
```

```
original sentence:
['Ancient', 'Maya', 'warfare', 'flared', 'up', 'surprisingly', 'early']


 tokenized and lemmatized sentence:
['ancient', 'maya', 'warfar', 'flare', 'surprisingli', 'earli']
```

In [25]:

```
processed_docs = documents['Summary'].map(preprocess)
processed_docs[:20]
```

Out[25]:

```
0     [actual, profound, thing, say, astrophysicist,...
1     [poppinga, colleagu, recent, show, chines, wit...
2     [give, climat, chang, increas, frequenc, inten...
3     [articl, appear, august, issu, scienc, news, h...
4     [renam, fermilab, physicist, enrico, fermi, ac...
5     [articl, appear, august, issu, scienc, news, h...
6     [unit, state, consid, rel, risk, overal, use, ...
7     [know, certain, cell, cytokin, particular, cau...
8     [wang, colleagu, plan, larger, censu, ancient,...
9     [trent, pediatrician, help, famili, discuss, d...
10    [add, common, cosmet, ingredi, speed, pericosi...
11    [neighborhood, level, project, futur, level, r...
12    [prior, maya, peopl, consid, dishonor, kill, w...
13    [nasa, indian, space, research, organ, plan, l...
14    [soil, typic, slow, recov, calcium, lose, stud...
15    [archaeolog, space, amazon, scienc, news, part...
16    [increasingli, frequent, winter, warm, spell, ...
17    [investig, wisconsin, teen, provid, answer, re...
18    [unfortun, result, mean, astronom, star, spin,...
19    [trust, import, legitimaci, credibl, effect, b...
Name: Summary, dtype: object
```

In [26]:

```
# BAG OF WORDS
dictionary = gensim.corpora.Dictionary(processed_docs)
count = 0
for k, v in dictionary.iteritems():
    print(k, v)
    count += 1
    if count > 20:
        break

# OUTPUT: containing the number of times a word appears; I executed for 20 times
```

```
0 actual
1 astrophysicist
2 brian
3 champaign
4 death
5 field
6 galaxi
7 histori
8 illinoi
9 involv
10 live
11 massiv
12 neighborhood
13 profound
14 recent
15 research
16 say
17 star
18 tell
19 thing
20 univers
```

In [27]:

```
#.....................................................Topic Modelling.............usi
ng LSA-Latent Semantic Analysis
import seaborn as sns
```

In [28]:

```
from sklearn.decomposition import TruncatedSVD
from sklearn.feature_extraction.text import TfidfVectorizer
#.........sklearn's TfidfVectorizer to create a document-term matrix ..............
vectorizer = TfidfVectorizer(stop_words='english',
max_features= 1000,
max_df = 0.5,
smooth_idf=True)

X = vectorizer.fit_transform(df['Summary'])

X.shape
# SVD represent data in vectors
svd_model = TruncatedSVD(n_components=20, algorithm='randomized', n_iter=100, random_st
ate=122)

svd_model.fit(X)

len(svd_model.components_)
# used sklearn's TruncatedSVD to perform the task of matrix decomposition
```

Out[28]:

```
20
```

In [29]:

```
terms = vectorizer.get_feature_names()
# Now printing each of the 20 topics.
for i, comp in enumerate(svd_model.components_):
    terms_comp = zip(terms, comp)
    sorted_terms = sorted(terms_comp, key= lambda x:x[1], reverse=True)[:7]
    print("Topic "+str(i)+": ")
    for t in sorted_terms:
        print(t[0])
        print(" ")
```

Topic 0:
news

science

boston

2019

appears

article

august

Topic 1:
says

stars

galaxy

research

ability

use

involved

Topic 2:
18

sn

probably

ability

climate

like

large

Topic 3:
water

states

percent

risk

new

typically

considered

Topic 4:

stars

18

sn

families

galaxy

cepheids

warped

Topic 5:
changes

plan

space

colleagues

2021

allowing

antarctica

Topic 6:
amazon

teens

details

changes

space

archaeology

associates

Topic 7:
spines

pufferfishes

puffers

colleagues

amazon

seed

water

Topic 8:
bubble

chamber

study

particles

massive

fermilab

lab

Topic 9:
amounts

aagaard

bacteria

biological

convinced

microbes

placenta

Topic 10:
trust

spines

celiac

gluten

cells

scientists

boykoff

Topic 11:
trust

aagaard

bacteria

biological

convinced

microbes

placenta

Topic 12:
trust

boykoff

credibility

effectiveness

important

just

legitimacy

Topic 13:
celiac

gluten

cells

boston

seed

able

anderson

Topic 14:
bubble

chamber

spines

particles

trust

families

fermilab

Topic 15:
families

trust

maya

address

adolescents

adopt

advocacy

Topic 16:
spines

maya

amazon

pufferfishes

puffers

800

battle

Topic 17:
teens

trust

says

aid

alexander

answers

cigarette

Topic 18:
amazon

archaeology

associates

buy

com

faq

llc

Topic 19:
changes

ages

anymore

astronomers

bummer

curtis

guess

In [30]:

```
#.....................TOPIC MODELLING BY USING LDA- Latent Dirichlet Allocation
 ................#
from nltk.corpus import stopwords
from nltk.stem.wordnet import WordNetLemmatizer
import string
stop = set(stopwords.words('english'))
exclude = set(string.punctuation)
lemma = WordNetLemmatizer()
def clean(doc):
    stop_free = " ".join([i for i in doc.lower().split() if i not in stop])
    punc_free = ''.join(ch for ch in stop_free if ch not in exclude)
    normalized = " ".join(lemma.lemmatize(word) for word in punc_free.split())
    return normalized

doc_clean = [clean(doc).split() for doc in df]
```

In [47]:

```
import gensim
from gensim import corpora

Lda = gensim.models.ldamodel.LdaModel

# Running and Trainign LDA model on the document term matrix.
ldamodel = Lda(doc_term_matrix, num_topics=3, id2word = dictionary, passes=50)
```

In [51]:

```
print(ldamodel.print_topics(num_topics=1, num_words=4))
```

```
[(0, '0.002*"rain" + 0.002*"lose" + 0.002*"slow" + 0.002*"point"')]
```

In [ ]:

```
# Each line is a topic with individual topic terms and weights.
#Topic1 can be termed as "Rain Lose SLow Point" : like during rain commute becomes slo
w, getting lost in traffic,etc.
# eg : 0.168*health + 0.083*sugar + 0.072*bad can be termed as BAD HEALTH.
#based on ones view.
```