```
# Import package
import numpy as np
import pandas as pd
import nltk
import re
import string
import scipy.sparse as sp
import matplotlib.pyplot as plt
from sklearn.preprocessing import normalize
from sklearn.decomposition import TruncatedSVD
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.metrics import pairwise_distances_argmin_min
from sklearn.metrics.pairwise import cosine_similarity
```

```
from google.colab import drive

drive.mount('/content/gdrive')
```

```
     Mounted at /content/gdrive
```

```
raw_articles_data = pd.read_excel('/content/gdrive/MyDrive/data/news_dataset.xlsx')
```

```
raw_articles_data
```

	Unnamed: 0	source	author	title	description
0	0	{'id': 'the-verge', 'name': 'The Verge'}	Jon Porter	Twitter's voice DMs arrive in India	Twitter has rolled out support for voice DMs o...
1	1	{'id': 'engadget', 'name': 'Engadget'}	Daniel Cooper	Amazon follows Netflix with mobile-only video ...	Amazon Prime Video and Bharti Airtel, India's ...
2	2	{'id': 'techcrunch', 'name': 'TechCrunch'}	Manish Singh	India bans PUBG and over 100 additional Chines...	India has banned more than 100 additional Chin...
3	3	{'id': 'engadget', 'name': 'Engadget'}	Steve Dent	Samsung begins offering support requests via W...	With the COVID-19 crisis continuing unabated i...
4	4	{'id': 'engadget', 'name': 'Engadget'}	Mariella Moon	Sony is launching the PS5 in India on February...	PlayStation gamers in India will finally have ...
...	...	...	...	...	...
				A Cat Is	The last cat

```
titles=[]
dates=[]
descriptions=[]
contents=[]
for index,item in raw_articles_data.iterrows():
  titles.append(item['title'])
  dates.append(item['publishedAt'])
  descriptions.append(item['description'])
  contents.append(item['content'])
```

weeks

```
dataset=pd.DataFrame({'title': titles, 'date': dates, 'desc': descriptions, 'content': con
dataset=dataset.drop_duplicates(subset='title').reset_index(drop=True)
dataset=dataset.dropna()
```

News'} braces for poores

```
dataset.head()
```

	title	date	desc	
0	Twitter’s voice DMs arrive in India	2021-02-17T13:18:32Z	Twitter has rolled out support for voice DMs o...	For whe
1	Amazon follows Netflix with mobile-only video ...	2021-01-13T11:15:31Z	Amazon Prime Video and Bharti Airtel, India’s ...	Amazon Pr
2	India bans PUBG and over 100 additional Chines...	2020-09-02T12:02:29Z	India has banned more than 100 additional Chin...	India has ba

```
dataset.shape
```

```
    (100, 4)
```

```
# Create function to process and tokenize raw texts
def preprocess(text, stopwords={}, lemmatizer=nltk.stem.wordnet.WordNetLemmatizer()):
    # Lower case
    text = text.lower()
    # Handle URL
    text = re.sub(r"https?://t.co/\w{10}",' ', text)
    # Deal with "'s"
    text = re.sub(r"'s", "", text)
    # Deal with "'"
    translator2 = str.maketrans({key: None for key in string.punctuation[6]})
    text = text.translate(translator2)
    # Deal with the rest of punctuations
    translator3 = str.maketrans(string.punctuation, ' '*len(string.punctuation))
    text = text.translate(translator3)
    # Handle unicode
    text = re.sub(r'[^\x00-\x7F]+',' ', text)
    # Split the text
    r1 = nltk.word_tokenize(text)
    # Lemmatize the text
    r2 = [lemmatizer.lemmatize(word) for word in r1]
    # Remove the stopwords
    r3 = [word for word in r2 if not word in stopwords]
    # Remove digits
    r4 = [word for word in r3 if word.isalpha()]
    return r4
```

```
# Import NLTK stopwords
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('punkt')
extra_stopwords = set()
stopwords = set(nltk.corpus.stopwords.words('english')) | extra_stopwords
```

```
    [nltk_data] Downloading package stopwords to /root/nltk_data...
    [nltk_data]   Unzipping corpora/stopwords.zip.
    [nltk_data] Downloading package wordnet to /root/nltk_data...
    [nltk_data]   Unzipping corpora/wordnet.zip.
    [nltk_data] Downloading package punkt to /root/nltk_data...
    [nltk_data]   Unzipping tokenizers/punkt.zip.
```

```
# Put the preprocessed texts into a list
articles = []

for i in range(0,dataset.shape[0]):
    tokenized_text = preprocess(dataset['content'][i], stopwords)
    articles.append(' '.join(tokenized_text))


from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer


doc = CountVectorizer()
word_count=doc.fit_transform(articles)
word_count.shape
print(word_count)
```

```
  (0, 1030)	1
  (0, 622)	1
  (0, 992)	1
  (0, 471)	1
  (0, 29)	1
  (0, 147)	1
  (0, 1013)	1
  (0, 990)	1
  (0, 437)	2
  (0, 816)	1
  (0, 927)	1
  (0, 1024)	1
  (0, 279)	1
  (0, 597)	1
  (0, 508)	1
  (0, 45)	1
  (0, 484)	1
  (0, 898)	1
  (0, 968)	1
  (0, 191)	1
  (0, 157)	1
  (1, 484)	1
  (1, 157)	1
  (1, 39)	1
  (1, 740)	2
  :	:
  (98, 163)	1
  (98, 676)	1
  (98, 427)	1
  (98, 1049)	1
  (99, 437)	1
  (99, 484)	1
  (99, 191)	1
  (99, 157)	1
  (99, 47)	1
  (99, 348)	1
  (99, 1037)	2
  (99, 53)	1
  (99, 726)	1
  (99, 952)	1
  (99, 692)	2
  (99, 117)	1
  (99, 612)     1
  (99, 1015)    1
  (99, 691)     1
  (99, 540)     1
  (99, 317)     1
  (99, 98)      1
  (99, 110)     1
  (99, 1017)    1
  (99, 941)     1
```

```
print(word_count.toarray())
```

```
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

```
tfidf_transformer=TfidfTransformer(smooth_idf=True,use_idf=True)
tfidf_transformer.fit(word_count)
df_idf = pd.DataFrame(tfidf_transformer.idf_, index=doc.get_feature_names(),columns=["idf_
df_idf.sort_values(by=['idf_weights']).head(40)
```

char	1.009950
india	1.258412
ha	1.853920
new	2.619388
largest	2.670682
world	2.724749
market	2.781907
country	2.781907
second	2.907070
service	2.976063
said	2.976063
year	2.976063
facebook	3.050171
million	3.050171
announced	3.130214
indian	3.130214
wa	3.130214
time	3.217225
one	3.217225
amazon	3.312535
internet	3.312535
coronavirus	3.312535

```
tf_idf_vector=tfidf_transformer.transform(word_count)
feature_names = doc.get_feature_names()
first_document_vector=tf_idf_vector[1]
df = pd.DataFrame(first_document_vector.T.todense(), index=feature_names, columns=["tfidf"
df.sort_values(by=["tfidf"],ascending=False).head(45)
```

variety	0.201204
bharti	0.201204
bolt	0.201204
nu	0.201204
airtel	0.201204
edition	0.201204
report	0.156294
launch	0.149993
amazon	0.135412
service	0.121658
second	0.118837
largest	0.109174
india	0.051442
char	0.041285
pleading	0.000000
playstation	0.000000
paper	0.000000
played	0.000000
plunge	0.000000
outdoors	0.000000
platform	0.000000
plant	0.000000
overseas	0.000000
plenty	0.000000
poetic	0.000000
podcast	0.000000
plan	0.000000
policy	0.000000
political	0.000000
politics	0.000000
pollution	0.000000
pop	0.000000
popular	0.000000
popularity	0.000000

populated	0.000000
pose	0.000000
position	0.000000

```
tf_idf_vector
```

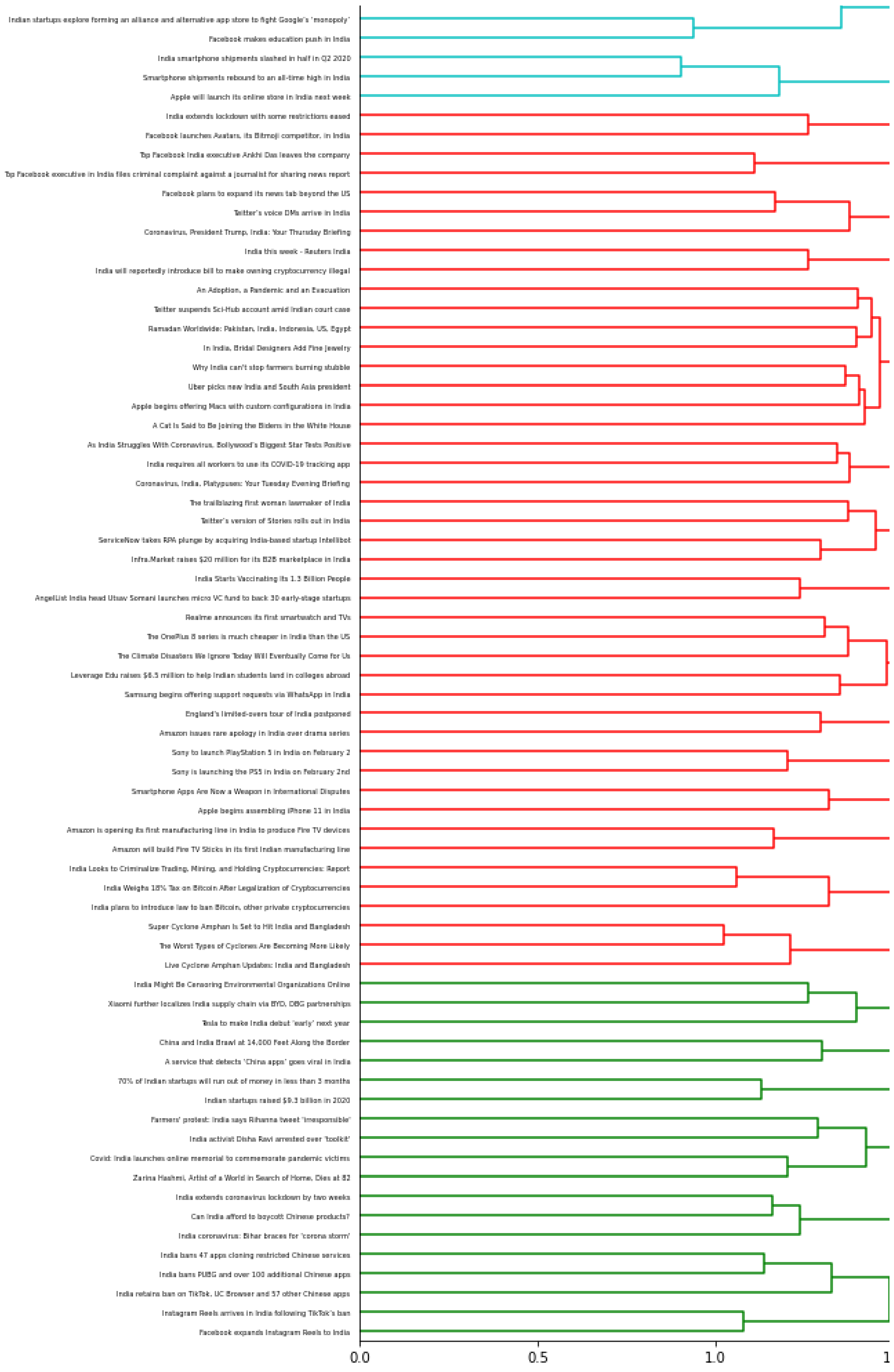
```
<100x1064 sparse matrix of type '<class 'numpy.float64'>'
        with 2028 stored elements in Compressed Sparse Row format>
```

```
print(tf_idf_vector.shape)
```

```
(100, 1064)
```

```
from sklearn.metrics.pairwise import cosine_similarity
dist = 1 - cosine_similarity(tf_idf_vector)
```

```
from scipy.cluster.hierarchy import ward, dendrogram

linkage_matrix = ward(dist) #define the linkage_matrix using ward clustering pre-computed

fig, ax = plt.subplots(figsize=(15, 20)) # set size
ax = dendrogram(linkage_matrix, orientation="right", labels=titles);

plt.tick_params(\
    axis= 'x',          # changes apply to the x-axis
    which='both',      # both major and minor ticks are affected
    bottom='off',      # ticks along the bottom edge are off
    top='off',         # ticks along the top edge are off
    labelbottom='off')

plt.tight_layout() #show plot with tight layout

#uncomment below to save figure
plt.savefig('ward_clusters.png', dpi=200) #save figure as ward_clusters
```


Indian startups explore forming an alliance and alternative app store to fight Google's 'monopoly'
Facebook makes education push in India
India smartphone shipments slashed in half in Q2 2020
Smartphone shipments rebound to an all-time high in India
Apple will launch its online store in India next week
India extends lockdown with some restrictions eased
Facebook launches Avatars, its Bitmoji competitor, in India
Top Facebook India executive Ankhi Das leaves the company
Top Facebook executive in India files criminal complaint against a journalist for sharing news report
Facebook plans to expand its news tab beyond the US
Twitter's voice DMs arrive in India
Coronavirus, President Trump, India: Your Thursday Briefing
India this week - Reuters India
India will reportedly introduce bill to make owning cryptocurrency illegal
An Adoption, a Pandemic and an Evacuation
Twitter suspends Sci-Hub account amid Indian court case
Ramadan Worldwide: Pakistan, India, Indonesia, US, Egypt
In India, Bridal Designers Add Fine Jewelry
Why India can't stop farmers burning stubble
Uber picks new India and South Asia president
Apple begins offering Macs with custom configurations in India
A Cat Is Said to Be Joining the Bidens in the White House
As India Struggles With Coronavirus, Bollywood's Biggest Star Tests Positive
India requires all workers to use its COVID-19 tracking app
Coronavirus, India, Platypuses: Your Tuesday Evening Briefing
The trailblazing first woman lawmaker of India
Twitter's version of Stories rolls out in India
ServiceNow takes RPA plunge by acquiring India-based startup Intellibot
Infra.Market raises $20 million for its B2B marketplace in India
India Starts Vaccinating Its 1.3 Billion People
AngelList India head Utsav Somani launches micro VC fund to back 30 early-stage startups
Realme announces its first smartwatch and TVs
The OnePlus 8 series is much cheaper in India than the US
The Climate Disasters We Ignore Today Will Eventually Come for Us
Leverage Edu raises $6.5 million to help Indian students land in colleges abroad
Samsung begins offering support requests via WhatsApp in India
England's limited-overs tour of India postponed
Amazon issues rare apology in India over drama series
Sony to launch PlayStation 5 in India on February 2
Sony is launching the PS5 in India on February 2nd
Smartphone Apps Are Now a Weapon in International Disputes
Apple begins assembling iPhone 11 in India
Amazon is opening its first manufacturing line in India to produce Fire TV devices
Amazon will build Fire TV Sticks in its first Indian manufacturing line
India Looks to Criminalize Trading, Mining, and Holding Cryptocurrencies: Report
India Weighs 18% Tax on Bitcoin After Legalization of Cryptocurrencies
India plans to introduce law to ban Bitcoin, other private cryptocurrencies
Super Cyclone Amphan Is Set to Hit India and Bangladesh
The Worst Types of Cyclones Are Becoming More Likely
Live Cyclone Amphan Updates: India and Bangladesh
India Might Be Censoring Environmental Organizations Online
Xiaomi further localizes India supply chain via BYD, DBG partnerships
Tesla to make India debut 'early' next year
China and India Brawl at 14,000 Feet Along the Border
A service that detects 'China apps' goes viral in India
70% of Indian startups will run out of money in less than 3 months
Indian startups raised $9.3 billion in 2020
Farmers' protest: India says Rihanna tweet 'irresponsible'
India activist Disha Ravi arrested over 'toolkit'
Covid: India launches online memorial to commemorate pandemic victims
Zarina Hashmi, Artist of a World in Search of Home, Dies at 82
India extends coronavirus lockdown by two weeks
Can India afford to boycott Chinese products?
India coronavirus: Bihar braces for 'corona storm'
India bans 47 apps cloning restricted Chinese services
India bans PUBG and over 100 additional Chinese apps
India retains ban on TikTok, UC Browser and 57 other Chinese apps
Instagram Reels arrives in India following TikTok's ban
Facebook expands Instagram Reels to India
0.0
0.5
1.0
1