Multi-Lingual Category Tree Classification

August 19, 2020

1 Packages & Import Functions

[1]:
```
from sklearn.tree import DecisionTreeClassifier
from collections import Counter
from sklearn.metrics import confusion_matrix,accuracy_score
import spacy, gensim
import nltk
import re
import matplotlib.pyplot as plt
import sklearn
import pandas as pd
import numpy as np
import seaborn as sns
from tqdm import tqdm
import spacy, gensim
from sklearn.metrics import f1_score
from sklearn.preprocessing import MultiLabelBinarizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import LatentDirichletAllocation, TruncatedSVD
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
# nlp = spacy.load('en')
import string
from plotly.offline import plot
nltk.download('wordnet')
from sklearn.decomposition.online_lda import LatentDirichletAllocation
from nltk.tokenize import RegexpTokenizer
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from gensim import corpora, models
from collections import Counter
import sklearn.feature_extraction.text as text
from sklearn import decomposition
from sklearn.decomposition import NMF
import gensim
from sklearn.metrics.pairwise import cosine_similarity
```

```
from textblob import TextBlob
from nltk.tag import pos_tag
from sklearn.svm import SVC
import heapq
import warnings
from os import listdir
from os.path import isfile, join
from sklearn.preprocessing import label_binarize
from sklearn.multiclass import OneVsRestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.decomposition import NMF
from sklearn.model_selection import train_test_split
pd.set_option('display.max_colwidth', 300)
```

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

2 Loading Competition Dataset

[2]:
```
data = pd.read_csv('ID_Desc_URL.csv',error_bad_lines=False,sep=',')
```

[3]:
```
data.head(5)
```

[3]:
```
     id  \
0  3874
1  3260
2  1867
3  4988
4  2882

                                                long_description  \
0
B Sandipan awakened the play with consciousness giving direction to the artists.
This was said by senior painter and actor Rakesh Pandey, Mathura color worker
and Braj-speaking ...
1
Samajwadi Party founder Mulayam Singh has been admitted to a private hospital in
Mumbai due to a stomach problem. Samajwadi Party founder Mulayam Singh (80) was
abruptly ...
2
The Union Minister of State for the establishment of a new Technology Center in
Chennai to promote battery-powered vehicle manufacturing. The use of vehicles to
suit the population of the country ...
3
```

```
PANCARD is to be canceled within three days. Don't be surprised. If you do not
link a PAN card with Aadhaar, the PAN card is canceled. They have only 3 days to
expire ...
4  Here's Presenting popular Children Nursery Story 'Krishna And Kaliya | Sri
Krishna'. For popular children Stories, kids songs, children songs, children
poems, baby songs, baby rhymes, kids nursery rhymes, nursery poems in Tamil
visit Etimes Tamil kids sections. Check out Etimes Kids videos secti...

                          link
0
https://navbharattimes.indiatimes.com/metro/lucknow/other-news/sandeepan-
vimalkant-nagar-was-a-struggling-and-combative-cultural-
worker/articleshow/73022935.cms
1
https://www.seithisolai.com/mulayam-singh-admitted-in-hospital.php
2                                      https://tamil.webdunia.com/article
/regional-tamil-news/central-minister-put-stone-for-new-battery-manufacturing-
building-in-chennai-119123000004_1.html
3
https://www.newspointapp.com/telugu-news/publisher-webdunia-telugu/top-
news/articleshow/1450482007046259cf0a35c4c54156d3cd66c1fe
4  https://timesofindia.indiatimes.com/videos/entertainment/kids/tamil/kids-
stories-nursery-rhymes-baby-songs-krishna-and-kaliya-sri-krishna-kids-nursery-
story-in-tamil/videoshow/73010220.cms
```

3 Data Preprocessing

[4]:
```
# Preprocessing steps

data2Cat = data[data.id.notnull()]
data2Cat["id"] = data2Cat['id'].astype('int64')
data2Cat.dropna(subset=['link','long_description'],how='all',inplace=True)
data2Cat.reset_index(drop=True, inplace=True)
data2Cat
```

[4]:
```
             id  \
0          3874
1          3260
2          1867
3          4988
4          2882
5          2675
6          2058
7          3875
8          2481
```

9	4615
10	4616
11	4229
12	629
13	3677
14	6523
15	3476
16	3063
17	630
18	6329
19	5766
20	4617
21	3064
22	3678
23	4811
24	6893
25	2253
26	4989
27	6524
28	4413
29	6894
...	...
1093978	169330
1093979	165935
1093980	163674
1093981	168140
1093982	164267
1093983	165699
1093984	164675
1093985	164268
1093986	165310
1093987	168544
1093988	165936
1093989	163108
1093990	163878
1093991	162917
1093992	166909
1093993	168357
1093994	168939
1093995	167732
1093996	166705
1093997	168358
1093998	168754
1093999	167107
1094000	162547
1094001	162723
1094002	163675

```
1094003  167108
1094004  166121
1094005  164483
1094006  162919
1094007  168359

                                                long_description  \
0
B Sandipan awakened the play with consciousness giving direction to the artists.
This was said by senior painter and actor Rakesh Pandey, Mathura color worker
and Braj-speaking ...
1
Samajwadi Party founder Mulayam Singh has been admitted to a private hospital in
Mumbai due to a stomach problem. Samajwadi Party founder Mulayam Singh (80) was
abruptly ...
2
The Union Minister of State for the establishment of a new Technology Center in
Chennai to promote battery-powered vehicle manufacturing. The use of vehicles to
suit the population of the country ...
3
PANCARD is to be canceled within three days. Don't be surprised. If you do not
link a PAN card with Aadhaar, the PAN card is canceled. They have only 3 days to
expire ...
4       Here's Presenting popular Children Nursery Story 'Krishna And Kaliya |
Sri Krishna'. For popular children Stories, kids songs, children songs, children
poems, baby songs, baby rhymes, kids nursery rhymes, nursery poems in Tamil
visit Etimes Tamil kids sections. Check out Etimes Kids videos secti...
5       Norwich City v Tottenham Hotspur - Premier League Hello and welcome to
the EPL transfer news roundup for the day! Here are the top stories of the day
surrounding the ! Erling Braut Haaland joins Dortmund Borussia Dortmund have
confirmed the signing of Erling Braut Haaland from Red Bull Salzburg...
6
New Delhi is good news for Reliance JioFiber users. The company is now offering
1TB (1000GB) of data on its Rs 199 top-up voucher. Earlier only 100GB data was
available in this plan. In plan's validity ...
7       The Bollywood fraternity on Friday expressed pain and shock on the news
of TV actor Kushal Punjabi's untimely death. Kushal committed suicide at his
home in Mumbai's Bandra area late on Thursday night. As per a statement from the
police, the actor \committed suicide by hanging himself from a fan...
8
Hemant Soren, the working president of the Jharkhand Mukti Morcha (JMM), was
sworn in as the Chief Minister of the state here on Sunday and said that his
government will continue to emphasize on education and health. Soren's main ...
9
A traffic police constable helped an underprivileged boy recover a bicycle,
which was gifted to him, in Bandra (west) on Saturday.
10      Recruitment Scam: 417 candidates booked for getting proxies to write
```

exam Top Searches: Recruitment Scam: 417 candidates booked for getting proxies to write exam TNN | Updated: Dec 30, 2019, 10:40 IST In yet another recruitment scam in the state, the Haryana police have booked 417 candidates who...
11
Leading online shopping giant Amazon runs a quiz every day on its app. The quiz involves selecting some of the answers and giving them presents. Today (Dec ...
12
Melbourne: Australian pacer Peter Siddle has retired from international cricket. Siddell announced his retirement following the Melbourne Test against New Zealand. Friends ...
13
Mum bye: Prohibition of social media in the Navy. The bans are Facebook WhatsApp and Instagram-based social media apps. NON-SECURE-NON-TECHNICAL CHARACTERISTICS ...
14
TANZANIA: Fosta, 57, the oldest rhinoceros in the world, has left for Tanzania. The female rhinoceros is in the Ngorongoro protected area of Tanzania.
15
Kuwait City: The 5th Islamic Seminar Propaganda Seminar, organized by Kerala Kuwait Islami Center, Kuwait ...
16 Haryana chief minister Manohar Lal Khattar on Sunday inaugurated a canteen in Karnal that will cater to the poor, especially farmers and labourers. Full meal is priced at Rs 10 at the canteen. CM said more such canteens will be opened in the state.
17
DMK leader Stalin and former chief minister Karunanidhi have protested against the amendment to the citizenship law. The Central BJP government has ...
18
New Delhi (Uttam Hindu News): Pakistan, which has been plotting against India again and again, has once again lost its face. This time Pakistan has made its gritty in the UN Security Council. Media Reports ...
19 Bigg Boss 13 housemates gave a tough time to the new captain Shehnaaz Gill by avoiding their duties. This led to host and Bollywood superstar SalmanKhan entering the house and cleaning the kitchen and bathroom on his own as the contestants watched in embarrassment.
20
Pakistani Danish Kaneria's statement has been the subject of a lot of talk about discrimination against Hindus in Pakistan. Former Pakistani leader Shahid Afridi has now revealed another issue. Interview ...
21
The Transport Corporation has begun its investigation after a complaint over state-owned buses was brought to light in Chennai. Trend ...
22
It is quite natural for field umpires to make decisions on the field. Most importantly, umpires often make mistakes during the match. Fielders on the ground loudly ...
23

The criminals in Uttar Pradesh are fearless in Yogi Raj. The police are also under question. Alam is that along with the leaders of the opposition party, BJP MPs and MLAs also question the law and order of Uttar Pradesh.
24
Surprisingly, the incident in which the beanbirds in Russia have been admitted to a veterinary hospital in due time is a serious health condition.
25
Bangla Hunt Desk: Why did you file for citizenship law in 28, did you change colors? Your character is at fault. So you keep the Bengali brothers confused. Banda's Onda. BJP Citizens ...
26
From time to time, we know how important vitamin D is to the body and health. In our tropical country, India has suffered from this shortage in recent years.
27 Banga, Dec 30 (Jasbir Singh Nurpur) - Satvir Singh Palli Fikki, in-charge of the Banga constituency has been appointed by the Punjab Government as Chairman of the District Planning Board. Satvir Singh Palli Fikki appointed Chief after
28
Sex in wet shower. If you want to lick each other's bodies and get wet, you can have sex with each other. The experience is different. Manassu ... Close Button
29
Hyderabad Senior Congress leader V Hanumanta Rao has filed a complaint against Rashtriya Swayamsevak Sangh (RSS) chief Mohan Bhagwat. In his complaint, Rao said that Bhagwat gave 130 crore countrymen a Hindu ...
...
...
1093978
Highlights Jungipur Municipal Board (PBB) indirectly alleges financial corruption, State Minister of State for Legislative Assembly Zakir Hossain launched the central government to solve the drinking water crisis in every state.
1093979
BNP chairperson Khaleda Zia's bail in the corruption case has been rejected by Bangladesh court. The court has said that Khaleda Zia will not get all the facilities like a common man. Zia's diabetic, hypertension ...
1093980 CHENNAI: The sixth offshore patrol vessel 'Yard 45006 VAJRA' to enhance coastal security was formally launched in the presence of Union Shipping Minister Mansukh Mandaviya and senior government officials here on Thursday. The vessel, built by Larsen and Toubro under the Centre's 'Make in India' ...
1093981
Bangalore: The state government has imposed a condition that Kannadigas should be given jobs when the state government gives space to Kempegowda International Airport. Accordingly, the report will be published in Kannada ...
1093982
State Chief Electoral Officer Teekaram Meena said college politics was unnecessary. Completely agree with the High Court ruling. Campuses cannot be riotous with political games. Rash on campus ...

1093983
New Delhi: The Delhi High Court has issued notice to the Delhi government, police and others on Thursday in the North East Delhi violence case and the next hearing of the matter will be on April 13. Chief Justice DN Patel and ...
1093984
Riyadh: The Riyadh-based Sarangi Art and Cultural Forum, Ji. Member of Parliament for Thrissur TN Prathapasana received the Kartikeyan Memorial Voice of Democracy.
1093985
Kasargod: (www.kasaragodvartha.com)
1093986 Delhi chief minister Arvind Kejriwal said that people responsible for violence in northeast Delhi should not be spared, even if they are from the Aam Aadmi Party. Delhi CM Kejriwal said, \Any person found guilty should be given stringent punishment. If any Aam Aadmi Party person is found guilty,...
1093987 Moxley is leaving nothing to chance (Pic Source: AEW) It's still hard to believe that Jon Moxley shocked the world last year when he joined AEW at Double or Nothing and made a statement by attacking and . His feud with Jericho has been building over the last couple of months as Moxley has withst...
1093988
New Delhi: The NIA has denied the bail plea of the accused in the Pulwama case. NIA has not filed a charge sheet against the accused in the case.
1093989
Kapil Mishra said there was nothing provocative in his speech in support of the Amendment of Citizenship Law at Maujpur Chowk in Delhi. Ask the police officer to qualify for the road ...
1093990 Bollywood actress Bhumi Pednekar has collaborated with feminine care brand Whisper for its new campaign #KeepGirlsInSchool that aims to create awareness on how even today, girls across India drop out of school on hitting puberty. As part of this campaign, Whisper has launched its new film that b...
1093991
In the Indian capital, Delhi, the revised Citizenship Act (CAA) is being targeted with violence aimed at Muslims. Most of the incidents were reported in Muslim homes and shops. Claims to have ...
1093992
New Delhi, 27 February (Language) Government asked banks to dispose of about 1.18 lakh pending applications received for loans under self-employment scheme 'Prime Minister Employment Generation Program (PMEGP)' by 15 March.
1093993
Responding to a question in the Legislative Assembly regarding the primary and collective health centers, Health Minister Nitin Patel said that the best healthcare facilities in remote areas of the state are also ...
1093994 NEW DELHI: A new hostel will soon come up at the Jawaharlal Nehru University (JNU) for the students belonging to the Northeast besides a regional convention centre in Delhi, Union minister Jitendra Singh said on Thursday. Singh said new bamboo clusters and technology centres will also come up ...
1093995

Now various boards are under examination. There are various boards being tested across the country. The big test of life. Long tireless work But after the Delhi violence, the Delhi Chest Board Exam ...
1093996 PARIS: Slovenian Tadej Pogacar produced a stunning final sprint on the top of Jebel Hafeet to win the 162km fifth stage of the UAE Tour on Thursday. Pogacar, 21, edged Kazakh Alexey Lutsenko in a photo finish for the victory, in what was the Tour's second ascent of the mountain overlooking the d...
1093997
The Tamil Nadu All Stripped Drinking Water Supply Association (GTU) has announced an indefinite strike from 6 pm onwards. President of the Association of Drinking Water Manufacturers
1093998
Friends of Pudukkottai, Bhagarla and Riyaz are very welcome to the videos posted on Kalatu Diktak. Riaz Dicta says Bharla died in an accident ...
1093999
Telangana Chief Minister KCR once again expressed his humanity. The problem of a disabled old man was solved on the road. CM going to participate in a private event on Thursday ...
1094000 Final Tax liability with cess @ 4% Rs 15,080 As shown in the example above, even a marginal increase in income above Rs 5 lakh has resulted in more tax outgo than the incremental income. Increase in income beyond Rs 5 lakh is marginal - only by Rs 10,000 but the tax liability due to such increme...
1094001
Pune: 'Rashtriya Swayamsevak Sangh has made a big contribution to the work of various educational and social institutions in Pune city. Dr. A book on the glory of Shripati Shastri will be published soon. You ...
1094002
Aligarh (Uttar Pradesh) [India], Feb 27 (ANI): Suspension of mobile internet services on Thursday was extended till February 28 in Aligarh district following recent clashes between police and anti-Cit...
1094003
New Delhi, Fr. NEW DELHI: Delhi Chief Minister Arvind Kejriwal on Thursday announced that he will provide relief to the family of the victims of the violence that broke out in north-eastern Delhi. Injury in violence ...
1094004
Recalling his long association with RSS ideologue P Parameswaran who passed away last week, RSS sarsangh chalak Mohan Bhagwat here, on Wednesday, said Parameswaran was a model swayam sewak who very mu...
1094005
The Kerala Administrative Service (KAS) has decided to cancel the examination on the basis of serious allegations leveled against it.
1094006 Rome [Italy]. Feb 28 (Sputnik/ANI): The death toll from coronavirus infection in Italy has risen to 17 with the addition of three more victims in the northwest of the country on Thursday, Angelo Borrelli, head of National Civil Protection Service, said. \The number of infected reached 650, while...
1094007
By keeping traffic management its top priority, Ahmedabad Urban Development

Authority (AUDA) aims to finish construction of four flyovers by 2021. With a project cost of Rs 105 cr the flyovers on SP R...

	link
0	https://navbharattimes.indiatimes.com/metro/lucknow/other-news/sandeepan-vimalkant-nagar-was-a-struggling-and-combative-cultural-worker/articleshow/73022935.cms
1	https://www.seithisolai.com/mulayam-singh-admitted-in-hospital.php
2	https://tamil.webdunia.com/article/regional-tamil-news/central-minister-put-stone-for-new-battery-manufacturing-building-in-chennai-119123000004_1.html
3	https://www.newspointapp.com/telugu-news/publisher-webdunia-telugu/top-news/articleshow/1450482007046259cf0a35c4c54156d3cd66c1fe
4	https://timesofindia.indiatimes.com/videos/entertainment/kids/tamil/kids-stories-nursery-rhymes-baby-songs-krishna-and-kaliya-sri-krishna-kids-nursery-story-in-tamil/videoshow/73010220.cms
5	https://www.newspointapp.com/english-news/publisher-sportskeeda/top-news/tottenham-willing-offer-christian-eriksen-real-madrid-swap-deal-more-epl-transfer-news-roundup-30th-december-2019/articleshow/14504820cbdd450bc9083874f53f83d52aeb1ece
6	https://navbharattimes.indiatimes.com/tech/gadgets-news/reliance-jio-rupees-199-top-up-voucher-offering-1tb-data-with-7-days-of-validity/articleshow/73024124.cms
7	https://timesofindia.indiatimes.com/videos/entertainment/hindi/from-farhan-akhtar-to-ranvir-shorey-bollywood-mournskushal-punjabis-untimely-death/videoshow/73024843.cms
8	https://www.newspointapp.com/hindi-news/publisher-vishvatimes-hindi/top-news/articleshow/14504820806e7880bf0c81749e1c135de9b397ef
9	https://timesofindia.indiatimes.com/city/mumbai/mumbai-cop-stops-bicycle-thief-saves-the-day-for-househelps-son-who-had-received-a-gift/articleshow/73025191.cms
10	https://timesofindia.indiatimes.com/home/education/news/recruitment-scam-417-candidates-booked-for-getting-proxies-to-write-exam/articleshow/73025288.cms
11	https://telugu.samayam.com/tech/news/these-are-the-answers-of-amazon-app-quiz-today-to-win-rs-5000/articleshow/73025096.cms
12	https://www.newspointapp.com/malayalam-news/publisher-mangalam-malayalam/top-

news/articleshow/14504820e25e717090a7f6b6a6dbdce6c4d58dd0
13
https://www.malayalamexpress.in/archives/991204/
14
https://www.malayalamexpress.in/archives/991231/
15
https://www.malayalamexpress.in/archives/991264/
16
https://timesofindia.indiatimes.com/city/gurgaon/full-meal-for-only-rs-10-at-karnal-grain-mkt/articleshow/73022866.cms
17
https://www.newspointapp.com/tamil-news/publisher-seithisolai-tamil/politics/articleshow/14504820209eacca43721bb50d36bfa9b73bab5b
18
https://www.newspointapp.com/hindi-news/publisher-uttamhindu-hindi/world/articleshow/14504820170b37788285aa75b0995201ec4a8742
19
https://timesofindia.indiatimes.com/videos/tv/hindi/bigg-boss-13-salman-khan-cleans-bathroom-washes-utensils/videoshow/73026487.cms
20
https://www.newspointapp.com/kannada-news/publisher-kannadadunia-kannada/top-news/articleshow/14504820688fb2101c56d229f5e191a82a0768f9
21
https://www.newspointapp.com/tamil-news/publisher-tamilsamayam/top-news/articleshow/145048200fe2fc0e89fb3112cea10b2ee992e509
22
https://www.newspointapp.com/telugu-news/publisher-telugusamayam/top-news/articleshow/145048209f13a386ebd667eb0a852c5ec19761d9
23
https://www.navjivanindia.com/news/bjp-leader-kaushal-kishore-attacks-up-police-over-law-and-order-in-up
24 https://tamil.samayam.com/latest-news/international-news/russian-veterinarian-successfully-performed-nanosurgery-on-cockroach/articleshow/73028127.cms?utm_source=colombia&utm_campaign=colombiaorganic&utm_medium=webwap
25
https://banglahunt.com/didi-character-loose-said-rajkumari-keshori/
26
https://tamil.samayam.com/lifestyle/health/foods-to-be-taken-to-boost-up-vitamin-d-in-your-body/articleshow/73028788.cms?utm_source=colombia&utm_campaign=colombiaorganic&utm_medium=webwap
27
https://www.newspointapp.com/punjabi-news/publisher-ajitjalandhar-punjabi/top-news/articleshow/145048208c0e8ddc261d4a5cfa1a19d57fb20aae
28
http://timeskerala.com/archives/169049
29

https://navbharattimes.indiatimes.com/state/other-states/hyderabad/congress-leader-files-complaint-against-rss-chief-mohan-bhagwat-in-hyderabad/articleshow/73030977.cms
...
...
1093978
https://bengali.mahanagar24x7.com/state-minister-zakir-hussain-alleges-against-jangipur-municipality/
1093979
https://eisamay.indiatimes.com/bangladesh-news/ex-bangladesh-pm-khaleda-zias-bail-plea-rejected/articleshow/74336937.cms
1093980
https://www.newspointapp.com/english-news/publisher-et/india/offshore-patrol-vessel-vajra-to-enhance-coastal-security-launched/articleshow/1450482061af4c7f0e8c0fc1d9cfa03b556cf1f0
1093981
https://kannadanewsnow.com/kannada/ts-nagabharana-speech-on-bangalore-international-airport-kannadigas-job/
1093982
https://www.newspointapp.com/malayalam-news/publisher-mediaonetv-malayalam/top-news/articleshow/14504820bef6a45797559ef0e8ad9cc79316d9d1
1093983
https://www.newspointapp.com/urdu-news/publisher-qaumiawaz-urdu/top-news/articleshow/14504820b70ec8ecd8bda229db78662cb64fae58
1093984
https://www.sathyamonline.com/tn-prathapan-mp-who-arrived-in-riyadh-sarangi-activists-gave-a-warm-reception/
1093985
http://feedproxy.google.com/~r/Kasargodvartha/~3/v8sa76hx3Wc/sabu-kottarakkara-on-delhi-clash.html
1093986
https://timesofindia.indiatimes.com/videos/city/delhi/people-responsible-for-riots-will-be-punished-cm-arvind-kejriwal/videoshow/74336973.cms
1093987 https://www.newspointapp.com/english-news/publisher-sportskeeda/top-news/jon-moxley-training-with-former-ufc-champion-for-his-match-with-chris-jericho-at-aew-revolution/articleshow/1450482022fbe34f7a4b1f0934fdc536ec5de456
1093988
http://timeskerala.com/archives/201274
1093989
https://www.newspointapp.com/malayalam-news/publisher-pradhanavartha-malayalam/top-news/articleshow/14504820a822e97f5c27ec1d9d94773026ed2a58
1093990
https://www.newspointapp.com/english-news/publisher-nationalherald/top-news/bhumi-padnekar-joins-whispers-keepgirlsinschool-campaign-/articleshow/145048208e750196095e2b9a618c9c6ef3dbb4a3
1093991

https://www.majaru.com/2020/02/Majaru-entertainment_2524.html
1093992
https://navbharattimes.indiatimes.com/business/business-news/banks-to-clear-118-lakh-pending-applications-of-prime-minister-employment-generation-scheme-by-march-15/articleshow/74340967.cms
1093993
https://www.newspointapp.com/gujarati-news/publisher-khabarchhe-gujarati/top-news/articleshow/14504820fc2158547ac1079022f6a9d56ddff4bb
1093994
https://timesofindia.indiatimes.com/city/delhi/new-hostel-to-come-up-in-jnu-for-northeast-students-convention-centre-at-dwarka/articleshow/74340274.cms
1093995
https://www.newspointapp.com/bengali-news/publisher-nilkantho-bengali/top-news/articleshow/14504820fbec033dcc03d6c8d847294fe94d53e6
1093996
https://www.newspointapp.com/english-news/publisher-toi/sports/yates-retains-lead-as-pogacar-wins-uae-tour-fifth-stage/articleshow/14504820b9a1e519b7f2889d1dcb4b213da5f42c
1093997
https://www.newspointapp.com/tamil-news/publisher-seithipunal-tamil/top-news/articleshow/1450482012cf63eb87ce0d17d89a5892edaf18be
1093998
https://www.newstm.in/news/tamilnadu/77767-pudhukottai-obituary-titok-video-goes-viral.html
1093999
https://telugu.asianetnews.com/telangana/telangana-cm-kcr-helping-hand-to-physically-challenged-old-man-q6dcaq
1094000
https://m.economictimes.com/wealth/tax/post-tax-rs-5-lakh-income-will-be-higher-than-rs-5-16-lakh-heres-why/articleshow/74311201.cms
1094001
https://maharashtratimes.indiatimes.com/maharashtra/pune-news/sripati-shastri-is-the-only-book-of-glory/articleshow/74344723.cms
1094002
https://www.newspointapp.com/english-news/publisher-aninews/health/internet-suspension-extended-till-feb-28-in-aligarh/articleshow/14504820728c3e117e176244008af55065e9085e
1094003
https://www.newspointapp.com/kannada-news/publisher-varthabharati-kannada/top-news/articleshow/14504820bc6078ac926ded4f95b0b95aedcdbedb
1094004
https://timesofindia.indiatimes.com/city/thiruvananthapuram/parameswaran-very-much-like-hedgewar-bhagwat/articleshow/74325730.cms
1094005
https://www.newspointapp.com/malayalam-news/publisher-keralaonlinenews-malayalam/top-news/articleshow/14504820d7979e2c81ba3da1dd66eda771c29d8a
1094006

```
https://www.aninews.in/news//world/asia/toll-from-coronavirus-rises-to-17-in-
italy20200228011759
1094007
https://timesofindia.indiatimes.com/city/ahmedabad/auda-focuses-on-roads-
flyovers-and-drainage/articleshow/74346692.cms

[1094008 rows x 3 columns]
```

4 Loading Google Word2Vec Model

Google's pre-trained model which includes word vectors for a vocabulary of 3 million words and phrases that they trained on roughly 100 billion words from a Google News dataset. The vector length is 300 features.

[5]:
```
# from gensim.models import Word2Vec
from gensim.models.word2vec import Word2Vec
from gensim.models import KeyedVectors

model = KeyedVectors.load_word2vec_format('GoogleNews-vectors-negative300.bin.
 ↪gz', binary=True)
```

5 Constructing Category Tree

We need to construct category tree using any python data structure. Here we use Anytree package as below.

[6]:
```
from anytree import Node, RenderTree, PreOrderIter

# Root node
root = Node("Categories")

# First level nodes
politics = Node("politics", parent=root)
sports = Node("sports", parent=root)
news = Node("news", parent=root)
education = Node("education", parent=root)
finance = Node("finance", parent=root)
entertainment = Node("entertainment", parent=root)
health = Node("health", parent=root)
environment = Node("environment", parent=root)

# Second level nodes
humor = Node("humor", parent=entertainment)
game = Node("game", parent=entertainment)
```

```
movie = Node("movie", parent=entertainment)
drama = Node("drama", parent=entertainment)
puzzle = Node("puzzle", parent=entertainment)
hobby = Node("hobby", parent=entertainment)
music = Node("music", parent=entertainment)
dance = Node("dance", parent=entertainment)

# Second level nodes
world = Node("global", parent=news)
national = Node("national", parent=news)
regional = Node("regional", parent=news)

# Second level nodes
pol_global = Node("global", parent=politics)
pol_national = Node("national", parent=politics)
pol_regional = Node("regional", parent=politics)

# Second level nodes
emergency = Node("emergency", parent=health)
casual = Node("casual", parent=health)

# Second level nodes
investment = Node("investment", parent=finance)
banking = Node("banking", parent=finance)

# Second level nodes
climate = Node("climate", parent=environment)
forest = Node("forest", parent=environment)


for pre, fill, node in RenderTree(root):
    print("%s%s" % (pre, node.name))
```

```
Categories
âŤŰâŤĂâŤĂ politics
âŤĆ   âŤŰâŤĂâŤĂ global
âŤĆ   âŤŰâŤĂâŤĂ national
âŤĆ   âŤŤâŤĂâŤĂ regional
âŤŰâŤĂâŤĂ sports
âŤŰâŤĂâŤĂ news
âŤĆ   âŤŰâŤĂâŤĂ global
âŤĆ   âŤŰâŤĂâŤĂ national
âŤĆ   âŤŤâŤĂâŤĂ regional
âŤŰâŤĂâŤĂ education
âŤŰâŤĂâŤĂ finance
âŤĆ   âŤŰâŤĂâŤĂ investment
âŤĆ   âŤŤâŤĂâŤĂ banking
```

```
âĬŰâĬĂâĬĂ entertainment
âĬĆ   âĬŰâĬĂâĬĂ humor
âĬĆ   âĬŰâĬĂâĬĂ game
âĬĆ   âĬŰâĬĂâĬĂ movie
âĬĆ   âĬŰâĬĂâĬĂ drama
âĬĆ   âĬŰâĬĂâĬĂ puzzle
âĬĆ   âĬŰâĬĂâĬĂ hobby
âĬĆ   âĬŰâĬĂâĬĂ music
âĬĆ   âĬĬâĬĂâĬĂ dance
âĬŰâĬĂâĬĂ health
âĬĆ   âĬŰâĬĂâĬĂ emergency
âĬĆ   âĬĬâĬĂâĬĂ casual
âĬĬâĬĂâĬĂ environment
    âĬŰâĬĂâĬĂ climate
    âĬĬâĬĂâĬĂ forest
```

6 Reading given Category Tree HDFS

[7]:
```
from anytree import Node, RenderTree, PreOrderIter, find_by_attr

catData = pd.read_csv('cat_tree.csv',error_bad_lines=False,sep=',')

catData.head(5)
```

[7]:

	cm_name	category_tree_text
0	Industry	Industry
1	International	Industry^Travel^Air Travel^International
2	Hotel	Industry^Travel^Hotel
3	Travel Booking Services	Industry^Travel^Travel Booking Services
4	Automobiles	Industry^Automobiles

7 Constructing given Category Tree

[8]:
```
# Creating the Category Tree as Python Data structure - Anytree

with open('cat_tree.csv', 'r') as f:
    lines = f.readlines()[1:]

    root = Node(lines[0].split(",")[1].strip())

    for line in lines:

        line = line.split(",")[1]
```

```
        for idx, cats in enumerate(line.split("^")[1:]):

            catNode = find_by_attr(root, cats.strip())

            if (catNode is None):
                parentNode = find_by_attr(root, line.split("^")[idx])

                if parentNode is None:
                    Node(cats.strip(), parent=root)
                else:
                    Node(cats.strip(), parent=parentNode)


for pre, _, node in RenderTree(root):
                print("%s%s" % (pre, node.name))
```

```
Industry
âŤŴâŤĂâŤĂ Travel
âŤĆ   âŤŴâŤĂâŤĂ Air Travel
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ International
âŤĆ   âŤŴâŤĂâŤĂ Hotel
âŤĆ   âŤŴâŤĂâŤĂ Travel Booking Services
âŤĆ   âŤŴâŤĂâŤĂ Auto Rentals
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Bike
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ Car
âŤĆ   âŤŴâŤĂâŤĂ Cabs
âŤĆ   âŤŴâŤĂâŤĂ Tourism
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ Domestic
âŤĆ   âŤŤâŤĂâŤĂ Vacation
âŤŴâŤĂâŤĂ Automobiles
âŤĆ   âŤŴâŤĂâŤĂ Four Wheeler
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Hatchback
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ SUV
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Sedan
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Budget Cars
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Certified Pre-Owned Cars
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Luxury Cars
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Electric Cars
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Large Vehicles
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ Premium Cars
âŤĆ   âŤŴâŤĂâŤĂ Auto Parts
âŤĆ   âŤŤâŤĂâŤĂ Two Wheeler
âŤĆ       âŤŴâŤĂâŤĂ Bikes
âŤĆ       âŤŴâŤĂâŤĂ Scooters
âŤĆ       âŤŴâŤĂâŤĂ Electric Bikes
âŤĆ       âŤŤâŤĂâŤĂ Premium Bikes
```

âĹŰâĹĂâĹĂ Shopping
âĹĆ   âĹŰâĹĂâĹĂ Books & Audible
âĹĆ   âĹĆ   âĹŤâĹĂâĹĂ Kindle E-Reader & eBooks
âĹĆ   âĹŰâĹĂâĹĂ Malls & Shopping Centers
âĹĆ   âĹŰâĹĂâĹĂ Electronics
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ Home Appliances
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ Kitchen Appliances
âĹĆ   âĹĆ   âĹŤâĹĂâĹĂ Entertainment Appliances
âĹĆ   âĹŰâĹĂâĹĂ Coupons and Discounts
âĹĆ   âĹŰâĹĂâĹĂ Fashion
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ Men
âĹĆ   âĹĆ   âĹŤâĹĂâĹĂ Women
âĹĆ   âĹŰâĹĂâĹĂ Health & Beauty
âĹĆ   âĹŰâĹĂâĹĂ Makeup and Accessories
âĹĆ   âĹŰâĹĂâĹĂ Watches
âĹĆ   âĹŰâĹĂâĹĂ Footwear
âĹĆ   âĹŰâĹĂâĹĂ Interior Decorating
âĹĆ   âĹŰâĹĂâĹĂ Outdoor Decorating
âĹĆ   âĹŰâĹĂâĹĂ Baby Products
âĹĆ   âĹŰâĹĂâĹĂ Gift Cards & Mobile Recharges
âĹĆ   âĹŰâĹĂâĹĂ Grocery
âĹĆ   âĹŰâĹĂâĹĂ Jewellery
âĹĆ   âĹŰâĹĂâĹĂ Luggage
âĹĆ   âĹŰâĹĂâĹĂ Mobile & Computers
âĹĆ   âĹŤâĹĂâĹĂ Sports & Fitness
âĹŰâĹĂâĹĂ Banking & Finance
âĹĆ   âĹŰâĹĂâĹĂ Banking
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ Credit Cards
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ Debit Cards
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ Demat Account
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ NRI Acoount
âĹĆ   âĹĆ   âĹŤâĹĂâĹĂ Online Banking
âĹĆ   âĹŰâĹĂâĹĂ Investment
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ SIPs
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ Mutual Funds
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ Shares and Stocks
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ Credit Checks
âĹĆ   âĹĆ   âĹŤâĹĂâĹĂ Foreign exchange
âĹĆ   âĹŰâĹĂâĹĂ Loans
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ Business Loans
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ Personal Loans
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ Education Loan
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ Car Loan
âĹĆ   âĹĆ   âĹŰâĹĂâĹĂ Heavy Vehicle Loan
âĹĆ   âĹĆ   âĹŤâĹĂâĹĂ Two Wheeler Loan
âĹĆ   âĹŤâĹĂâĹĂ Insurance
âĹĆ   âĹŰâĹĂâĹĂ Health Insurance
âĹĆ   âĹŰâĹĂâĹĂ Home Insurance

âŤĆ âŤŴâŤĂâŤĂ Life Insurance
âŤĆ âŤŴâŤĂâŤĂ Car Insurance
âŤĆ âŤŴâŤĂâŤĂ Bike Insurance
âŤĆ âŤŤâŤĂâŤĂ Term Insurance
âŤŴâŤĂâŤĂ Others
âŤĆ âŤŴâŤĂâŤĂ Logistics
âŤĆ âŤŴâŤĂâŤĂ Manufacturing
âŤĆ âŤŴâŤĂâŤĂ Non-Profit Organizations
âŤĆ âŤŴâŤĂâŤĂ Retail
âŤĆ âŤŴâŤĂâŤĂ Government Projects
âŤĆ âŤŴâŤĂâŤĂ Industrial B2B
âŤĆ âŤŤâŤĂâŤĂ Solar Energy
âŤŴâŤĂâŤĂ Career & Job
âŤĆ âŤŴâŤĂâŤĂ Job
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Job Search
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Job Fairs
âŤĆ âŤŤâŤĂâŤĂ Career Advice & Counselling
âŤŴâŤĂâŤĂ Entertainment
âŤĆ âŤŴâŤĂâŤĂ Offline Entertainment
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Events and Attractions
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Offline Movies
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Dish TV & D2H
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Offline TV Show & Serial
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Outdoor Activities
âŤĆ âŤŴâŤĂâŤĂ Music Streaming Services
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Airtel Wynk
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Amazon Prime Music
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Gaana
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Jio Saavn
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Spotify
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Youtube Music
âŤĆ âŤŤâŤĂâŤĂ OTT Apps
âŤĆ âŤŴâŤĂâŤĂ Airtel Xstream
âŤĆ âŤŴâŤĂâŤĂ Amazon Prime Video
âŤĆ âŤŴâŤĂâŤĂ Hotstar
âŤĆ âŤŴâŤĂâŤĂ MX Player
âŤĆ âŤŴâŤĂâŤĂ Netflix
âŤĆ âŤŴâŤĂâŤĂ Voot
âŤĆ âŤŤâŤĂâŤĂ Zee5
âŤŴâŤĂâŤĂ Food & Dining
âŤĆ âŤŴâŤĂâŤĂ Food delivery
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Offer & Deals
âŤĆ âŤŤâŤĂâŤĂ Restaurant table booking
âŤŴâŤĂâŤĂ Health & Pharmaceutical
âŤĆ âŤŴâŤĂâŤĂ Pharmaceutical
âŤĆ âŤŴâŤĂâŤĂ Weight Loss
âŤĆ âŤŴâŤĂâŤĂ Wellness
âŤĆ âŤŴâŤĂâŤĂ Health Supplements

âŤĆ âŤĆ âŤŤâŤĂâŤĂ Ayurveda
âŤĆ âŤWâŤĂâŤĂ Physical Therapy
âŤĆ âŤWâŤĂâŤĂ Medical Health
âŤĆ âŤĆ âŤWâŤĂâŤĂ Diseases and Conditions
âŤĆ âŤĆ âŤWâŤĂâŤĂ Dental Health
âŤĆ âŤĆ âŤWâŤĂâŤĂ Diabetes
âŤĆ âŤĆ âŤWâŤĂâŤĂ Eye and Vision Conditions
âŤĆ âŤĆ âŤWâŤĂâŤĂ Heart and Cardiovascular Diseases
âŤĆ âŤĆ âŤWâŤĂâŤĂ Reproductive Health
âŤĆ âŤĆ âŤWâŤĂâŤĂ Infertility
âŤĆ âŤĆ âŤWâŤĂâŤĂ Pregnancy
âŤĆ âŤĆ âŤWâŤĂâŤĂ Sexual Health & Conditions
âŤĆ âŤĆ âŤWâŤĂâŤĂ Skin and Dermatology
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Sleep Disorders
âŤĆ âŤWâŤĂâŤĂ Gym
âŤĆ âŤŤâŤĂâŤĂ Weight Gain
âŤWâŤĂâŤĂ Utility
âŤĆ âŤŤâŤĂâŤĂ Telecom
âŤWâŤĂâŤĂ Education
âŤĆ âŤWâŤĂâŤĂ College Education
âŤĆ âŤWâŤĂâŤĂ Professional School
âŤĆ âŤWâŤĂâŤĂ Kids Schooling
âŤĆ âŤWâŤĂâŤĂ Online Courses
âŤĆ âŤWâŤĂâŤĂ Online Admission
âŤĆ âŤWâŤĂâŤĂ Quiz
âŤĆ âŤŤâŤĂâŤĂ Training
âŤWâŤĂâŤĂ Family and Relationships
âŤĆ âŤWâŤĂâŤĂ Parenting
âŤĆ âŤĆ âŤWâŤĂâŤĂ Daycare and Pre-School
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Special Needs Kids
âŤĆ âŤŤâŤĂâŤĂ Matrimony
âŤĆ âŤWâŤĂâŤĂ Bride
âŤĆ âŤWâŤĂâŤĂ Dating
âŤĆ âŤŤâŤĂâŤĂ Groom
âŤWâŤĂâŤĂ News & Politics
âŤĆ âŤWâŤĂâŤĂ News
âŤĆ âŤĆ âŤWâŤĂâŤĂ International News
âŤĆ âŤĆ âŤWâŤĂâŤĂ Local News
âŤĆ âŤĆ âŤWâŤĂâŤĂ National News
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Elections News
âŤĆ âŤWâŤĂâŤĂ Politics
âŤĆ âŤĆ âŤWâŤĂâŤĂ AAP Followers
âŤĆ âŤĆ âŤWâŤĂâŤĂ BJP Followers
âŤĆ âŤĆ âŤWâŤĂâŤĂ Congress Followers
âŤĆ âŤĆ âŤŤâŤĂâŤĂ TMC Followers
âŤĆ âŤWâŤĂâŤĂ National Election
âŤĆ âŤŤâŤĂâŤĂ State Election
âŤWâŤĂâŤĂ Online Services

```
âŤĆ   âŤŤâŤĂâŤĂ Internet
âŤĆ       âŤŴâŤĂâŤĂ Internet Service Providers
âŤĆ       âŤŴâŤĂâŤĂ Cloud Computing
âŤĆ       âŤŴâŤĂâŤĂ Software and Applications
âŤĆ       âŤŴâŤĂâŤĂ Domain and Hosting
âŤĆ       âŤŴâŤĂâŤĂ Information and Network Security
âŤĆ       âŤŤâŤĂâŤĂ Social Networking
âŤŴâŤĂâŤĂ Real Estate
âŤĆ   âŤŴâŤĂâŤĂ Property Type
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ 2bhk Flat
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ 3bhk Flat
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ 4bhk+ Flat
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Malls
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Penthouse
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Shops
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Showrooms
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ Villas
âŤĆ   âŤŤâŤĂâŤĂ Location
âŤĆ       âŤŴâŤĂâŤĂ Ahmedabad
âŤĆ       âŤŴâŤĂâŤĂ Benguluru
âŤĆ       âŤŴâŤĂâŤĂ Chennai
âŤĆ       âŤŴâŤĂâŤĂ Delhi
âŤĆ       âŤŴâŤĂâŤĂ Faridabad
âŤĆ       âŤŴâŤĂâŤĂ Ghaziabad
âŤĆ       âŤŴâŤĂâŤĂ Greater Noida
âŤĆ       âŤŴâŤĂâŤĂ Gurgaon
âŤĆ       âŤŴâŤĂâŤĂ Hyderabad
âŤĆ       âŤŴâŤĂâŤĂ Jaipur
âŤĆ       âŤŴâŤĂâŤĂ Kanpur
âŤĆ       âŤŴâŤĂâŤĂ Kolkata
âŤĆ       âŤŴâŤĂâŤĂ Lucknow
âŤĆ       âŤŴâŤĂâŤĂ Mumbai
âŤĆ       âŤŴâŤĂâŤĂ Navi Mumbai
âŤĆ       âŤŴâŤĂâŤĂ Noida
âŤĆ       âŤŤâŤĂâŤĂ Thane
âŤŴâŤĂâŤĂ Sports & Games
âŤĆ   âŤŴâŤĂâŤĂ Outdoor Games
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Football
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Cricket
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Poker and Gambling
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Badminton
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Golf
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Kabbdi
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Tennis
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ Wrestling
âŤĆ   âŤŤâŤĂâŤĂ Online Games
âŤĆ       âŤŴâŤĂâŤĂ Action and Adventure
âŤĆ       âŤŴâŤĂâŤĂ Arcade
```

```
âŤĆ       âŤŴâŤĂâŤĂ Board
âŤĆ       âŤŴâŤĂâŤĂ Casino & Cards
âŤĆ       âŤŴâŤĂâŤĂ Simulation
âŤĆ       âŤŴâŤĂâŤĂ Sport
âŤĆ       âŤŤâŤĂâŤĂ Strategy
âŤŴâŤĂâŤĂ FMCG
âŤĆ   âŤŴâŤĂâŤĂ Hair Care
âŤĆ   âŤŴâŤĂâŤĂ Skin Care
âŤĆ   âŤŴâŤĂâŤĂ Cosmetics
âŤĆ   âŤŴâŤĂâŤĂ Foods & Beverages
âŤĆ   âŤŴâŤĂâŤĂ Household Products
âŤĆ   âŤŴâŤĂâŤĂ Lifestyle
âŤĆ   âŤŴâŤĂâŤĂ Medicines & Office Supplies
âŤĆ   âŤŤâŤĂâŤĂ Oral Care
âŤŴâŤĂâŤĂ Gadgets & Technology
âŤĆ   âŤŴâŤĂâŤĂ Laptops
âŤĆ   âŤŴâŤĂâŤĂ Cameras and Camcorders
âŤĆ   âŤŴâŤĂâŤĂ Home Entertainment Systems
âŤĆ   âŤŴâŤĂâŤĂ Smartphones
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Android Phones
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ IOS
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Less than 10k
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Less than 20k
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Less than 30k
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ More than 30K
âŤĆ   âŤŴâŤĂâŤĂ Tablets
âŤĆ   âŤŴâŤĂâŤĂ Smartwatch
âŤĆ   âŤŴâŤĂâŤĂ IOT Devices
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Echo & Alexa
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Fire TV Stick
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ Google Home
âŤĆ   âŤŴâŤĂâŤĂ Computer Accessories
âŤĆ   âŤŤâŤĂâŤĂ Mobile Accessories
âŤŤâŤĂâŤĂ Religion & Spirituality
    âŤŴâŤĂâŤĂ Astrology
    âŤŴâŤĂâŤĂ Religion
    âŤŤâŤĂâŤĂ Spirituality
```

8 Tweaking Category Tree

Multi-words categories having different meanings are combined into similar meaning words so that distance metrics won't go awry.

[9]:
```
len(root.children)

religionNode = find_by_attr(root, 'Religion & Spirituality')
```

```
religionNode.name = 'Religion'

onlineNode = find_by_attr(root, 'Online Services')
onlineNode.name = 'Online'

family = find_by_attr(root, 'Family and Relationships')
family.name = 'Family Relationships'

family = find_by_attr(root, 'Real Estate')
family.name = 'Realty'

education = find_by_attr(root, 'Education')
education.name = 'Schooling'

politics = find_by_attr(root, 'News & Politics')
politics.name = 'Politics'

banking = find_by_attr(root, 'Banking & Finance')
banking.name = 'Banking'

food = find_by_attr(root, 'Food & Dining')
food.name = 'Food'

career = find_by_attr(root, 'Career & Job')
career.name = 'Profession'

sports = find_by_attr(root, 'Sports & Games')
sports.name = 'Sports'


for pre, _, node in RenderTree(root):
                print("%s%s" % (pre, node.name))
```

```
Industry
âŤŃâŤĂâŤĂ Travel
âŤĆ   âŤŃâŤĂâŤĂ Air Travel
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ International
âŤĆ   âŤŃâŤĂâŤĂ Hotel
âŤĆ   âŤŃâŤĂâŤĂ Travel Booking Services
âŤĆ   âŤŃâŤĂâŤĂ Auto Rentals
âŤĆ   âŤĆ   âŤŃâŤĂâŤĂ Bike
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ Car
âŤĆ   âŤŃâŤĂâŤĂ Cabs
âŤĆ   âŤŃâŤĂâŤĂ Tourism
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ Domestic
âŤĆ   âŤŤâŤĂâŤĂ Vacation
âŤŃâŤĂâŤĂ Automobiles
```

```
âŤĆ   âŤŴâŤĂâŤĂ Four Wheeler
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Hatchback
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ SUV
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Sedan
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Budget Cars
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Certified Pre-Owned Cars
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Luxury Cars
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Electric Cars
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Large Vehicles
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ Premium Cars
âŤĆ   âŤŴâŤĂâŤĂ Auto Parts
âŤĆ   âŤŤâŤĂâŤĂ Two Wheeler
âŤĆ       âŤŴâŤĂâŤĂ Bikes
âŤĆ       âŤŴâŤĂâŤĂ Scooters
âŤĆ       âŤŴâŤĂâŤĂ Electric Bikes
âŤĆ       âŤŤâŤĂâŤĂ Premium Bikes
âŤŴâŤĂâŤĂ Shopping
âŤĆ   âŤŴâŤĂâŤĂ Books & Audible
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ Kindle E-Reader & eBooks
âŤĆ   âŤŴâŤĂâŤĂ Malls & Shopping Centers
âŤĆ   âŤŴâŤĂâŤĂ Electronics
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Home Appliances
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Kitchen Appliances
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ Entertainment Appliances
âŤĆ   âŤŴâŤĂâŤĂ Coupons and Discounts
âŤĆ   âŤŴâŤĂâŤĂ Fashion
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Men
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ Women
âŤĆ   âŤŴâŤĂâŤĂ Health & Beauty
âŤĆ   âŤŴâŤĂâŤĂ Makeup and Accessories
âŤĆ   âŤŴâŤĂâŤĂ Watches
âŤĆ   âŤŴâŤĂâŤĂ Footwear
âŤĆ   âŤŴâŤĂâŤĂ Interior Decorating
âŤĆ   âŤŴâŤĂâŤĂ Outdoor Decorating
âŤĆ   âŤŴâŤĂâŤĂ Baby Products
âŤĆ   âŤŴâŤĂâŤĂ Gift Cards & Mobile Recharges
âŤĆ   âŤŴâŤĂâŤĂ Grocery
âŤĆ   âŤŴâŤĂâŤĂ Jewellery
âŤĆ   âŤŴâŤĂâŤĂ Luggage
âŤĆ   âŤŴâŤĂâŤĂ Mobile & Computers
âŤĆ   âŤŤâŤĂâŤĂ Sports & Fitness
âŤŴâŤĂâŤĂ Banking
âŤĆ   âŤŴâŤĂâŤĂ Banking
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Credit Cards
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Debit Cards
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Demat Account
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ NRI Acoount
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ Online Banking
```

âŤĆ âŤŴâŤĂâŤĂ Investment
âŤĆ âŤĆ âŤŴâŤĂâŤĂ SIPs
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Mutual Funds
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Shares and Stocks
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Credit Checks
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Foreign exchange
âŤĆ âŤŴâŤĂâŤĂ Loans
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Business Loans
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Personal Loans
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Education Loan
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Car Loan
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Heavy Vehicle Loan
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Two Wheeler Loan
âŤĆ âŤŤâŤĂâŤĂ Insurance
âŤĆ   âŤŴâŤĂâŤĂ Health Insurance
âŤĆ   âŤŴâŤĂâŤĂ Home Insurance
âŤĆ   âŤŴâŤĂâŤĂ Life Insurance
âŤĆ   âŤŴâŤĂâŤĂ Car Insurance
âŤĆ   âŤŴâŤĂâŤĂ Bike Insurance
âŤĆ   âŤŤâŤĂâŤĂ Term Insurance
âŤŴâŤĂâŤĂ Others
âŤĆ âŤŴâŤĂâŤĂ Logistics
âŤĆ âŤŴâŤĂâŤĂ Manufacturing
âŤĆ âŤŴâŤĂâŤĂ Non-Profit Organizations
âŤĆ âŤŴâŤĂâŤĂ Retail
âŤĆ âŤŴâŤĂâŤĂ Government Projects
âŤĆ âŤŴâŤĂâŤĂ Industrial B2B
âŤĆ âŤŤâŤĂâŤĂ Solar Energy
âŤŴâŤĂâŤĂ Profession
âŤĆ âŤŴâŤĂâŤĂ Job
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Job Search
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Job Fairs
âŤĆ âŤŤâŤĂâŤĂ Career Advice & Counselling
âŤŴâŤĂâŤĂ Entertainment
âŤĆ âŤŴâŤĂâŤĂ Offline Entertainment
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Events and Attractions
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Offline Movies
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Dish TV & D2H
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Offline TV Show & Serial
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Outdoor Activities
âŤĆ âŤŴâŤĂâŤĂ Music Streaming Services
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Airtel Wynk
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Amazon Prime Music
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Gaana
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Jio Saavn
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Spotify
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Youtube Music
âŤĆ âŤŤâŤĂâŤĂ OTT Apps

âŤĆ âŤŴâŤĂâŤĂ Airtel Xstream
âŤĆ âŤŴâŤĂâŤĂ Amazon Prime Video
âŤĆ âŤŴâŤĂâŤĂ Hotstar
âŤĆ âŤŴâŤĂâŤĂ MX Player
âŤĆ âŤŴâŤĂâŤĂ Netflix
âŤĆ âŤŴâŤĂâŤĂ Voot
âŤĆ âŤŤâŤĂâŤĂ Zee5
âŤŴâŤĂâŤĂ Food
âŤĆ âŤŴâŤĂâŤĂ Food delivery
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Offer & Deals
âŤĆ âŤŤâŤĂâŤĂ Restaurant table booking
âŤŴâŤĂâŤĂ Health & Pharmaceutical
âŤĆ âŤŴâŤĂâŤĂ Pharmaceutical
âŤĆ âŤŴâŤĂâŤĂ Weight Loss
âŤĆ âŤŴâŤĂâŤĂ Wellness
âŤĆ âŤŴâŤĂâŤĂ Health Supplements
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Ayurveda
âŤĆ âŤŴâŤĂâŤĂ Physical Therapy
âŤĆ âŤŴâŤĂâŤĂ Medical Health
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Diseases and Conditions
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Dental Health
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Diabetes
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Eye and Vision Conditions
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Heart and Cardiovascular Diseases
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Reproductive Health
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Infertility
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Pregnancy
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Sexual Health & Conditions
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Skin and Dermatology
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Sleep Disorders
âŤĆ âŤŴâŤĂâŤĂ Gym
âŤĆ âŤŤâŤĂâŤĂ Weight Gain
âŤŴâŤĂâŤĂ Utility
âŤĆ âŤŤâŤĂâŤĂ Telecom
âŤŴâŤĂâŤĂ Schooling
âŤĆ âŤŴâŤĂâŤĂ College Education
âŤĆ âŤŴâŤĂâŤĂ Professional School
âŤĆ âŤŴâŤĂâŤĂ Kids Schooling
âŤĆ âŤŴâŤĂâŤĂ Online Courses
âŤĆ âŤŴâŤĂâŤĂ Online Admission
âŤĆ âŤŴâŤĂâŤĂ Quiz
âŤĆ âŤŤâŤĂâŤĂ Training
âŤŴâŤĂâŤĂ Family Relationships
âŤĆ âŤŴâŤĂâŤĂ Parenting
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Daycare and Pre-School
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Special Needs Kids
âŤĆ âŤŤâŤĂâŤĂ Matrimony
âŤĆ âŤŴâŤĂâŤĂ Bride

```
âŤĆ       âŤŴâŤĂâŤĂ Dating
âŤĆ       âŤŤâŤĂâŤĂ Groom
âŤŴâŤĂâŤĂ Politics
âŤĆ   âŤŴâŤĂâŤĂ News
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ International News
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Local News
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ National News
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ Elections News
âŤĆ   âŤŴâŤĂâŤĂ Politics
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ AAP Followers
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ BJP Followers
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Congress Followers
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ TMC Followers
âŤĆ   âŤŴâŤĂâŤĂ National Election
âŤĆ   âŤŤâŤĂâŤĂ State Election
âŤŴâŤĂâŤĂ Online
âŤĆ   âŤŤâŤĂâŤĂ Internet
âŤĆ       âŤŴâŤĂâŤĂ Internet Service Providers
âŤĆ       âŤŴâŤĂâŤĂ Cloud Computing
âŤĆ       âŤŴâŤĂâŤĂ Software and Applications
âŤĆ       âŤŴâŤĂâŤĂ Domain and Hosting
âŤĆ       âŤŴâŤĂâŤĂ Information and Network Security
âŤĆ       âŤŤâŤĂâŤĂ Social Networking
âŤŴâŤĂâŤĂ Realty
âŤĆ   âŤŴâŤĂâŤĂ Property Type
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ 2bhk Flat
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ 3bhk Flat
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ 4bhk+ Flat
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Malls
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Penthouse
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Shops
âŤĆ   âŤĆ   âŤŴâŤĂâŤĂ Showrooms
âŤĆ   âŤĆ   âŤŤâŤĂâŤĂ Villas
âŤĆ   âŤŤâŤĂâŤĂ Location
âŤĆ       âŤŴâŤĂâŤĂ Ahmedabad
âŤĆ       âŤŴâŤĂâŤĂ Benguluru
âŤĆ       âŤŴâŤĂâŤĂ Chennai
âŤĆ       âŤŴâŤĂâŤĂ Delhi
âŤĆ       âŤŴâŤĂâŤĂ Faridabad
âŤĆ       âŤŴâŤĂâŤĂ Ghaziabad
âŤĆ       âŤŴâŤĂâŤĂ Greater Noida
âŤĆ       âŤŴâŤĂâŤĂ Gurgaon
âŤĆ       âŤŴâŤĂâŤĂ Hyderabad
âŤĆ       âŤŴâŤĂâŤĂ Jaipur
âŤĆ       âŤŴâŤĂâŤĂ Kanpur
âŤĆ       âŤŴâŤĂâŤĂ Kolkata
âŤĆ       âŤŴâŤĂâŤĂ Lucknow
âŤĆ       âŤŴâŤĂâŤĂ Mumbai
```

âŤĆ âŤŴâŤĂâŤĂ Navi Mumbai
âŤĆ âŤŴâŤĂâŤĂ Noida
âŤĆ âŤŤâŤĂâŤĂ Thane
âŤŴâŤĂâŤĂ Sports
âŤĆ âŤŴâŤĂâŤĂ Outdoor Games
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Football
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Cricket
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Poker and Gambling
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Badminton
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Golf
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Kabbdi
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Tennis
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Wrestling
âŤĆ âŤŤâŤĂâŤĂ Online Games
âŤĆ âŤŴâŤĂâŤĂ Action and Adventure
âŤĆ âŤŴâŤĂâŤĂ Arcade
âŤĆ âŤŴâŤĂâŤĂ Board
âŤĆ âŤŴâŤĂâŤĂ Casino & Cards
âŤĆ âŤŴâŤĂâŤĂ Simulation
âŤĆ âŤŴâŤĂâŤĂ Sport
âŤĆ âŤŤâŤĂâŤĂ Strategy
âŤŴâŤĂâŤĂ FMCG
âŤĆ âŤŴâŤĂâŤĂ Hair Care
âŤĆ âŤŴâŤĂâŤĂ Skin Care
âŤĆ âŤŴâŤĂâŤĂ Cosmetics
âŤĆ âŤŴâŤĂâŤĂ Foods & Beverages
âŤĆ âŤŴâŤĂâŤĂ Household Products
âŤĆ âŤŴâŤĂâŤĂ Lifestyle
âŤĆ âŤŴâŤĂâŤĂ Medicines & Office Supplies
âŤĆ âŤŤâŤĂâŤĂ Oral Care
âŤŴâŤĂâŤĂ Gadgets & Technology
âŤĆ âŤŴâŤĂâŤĂ Laptops
âŤĆ âŤŴâŤĂâŤĂ Cameras and Camcorders
âŤĆ âŤŴâŤĂâŤĂ Home Entertainment Systems
âŤĆ âŤŴâŤĂâŤĂ Smartphones
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Android Phones
âŤĆ âŤĆ âŤŴâŤĂâŤĂ IOS
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Less than 10k
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Less than 20k
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Less than 30k
âŤĆ âŤĆ âŤŤâŤĂâŤĂ More than 30K
âŤĆ âŤŴâŤĂâŤĂ Tablets
âŤĆ âŤŴâŤĂâŤĂ Smartwatch
âŤĆ âŤŴâŤĂâŤĂ IOT Devices
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Echo & Alexa
âŤĆ âŤĆ âŤŴâŤĂâŤĂ Fire TV Stick
âŤĆ âŤĆ âŤŤâŤĂâŤĂ Google Home
âŤĆ âŤŴâŤĂâŤĂ Computer Accessories

```
âŤĆ   âŤŤâŤĂâŤĂ Mobile Accessories
âŤŤâŤĂâŤĂ Religion
    âŤŰâŤĂâŤĂ Astrology
    âŤŰâŤĂâŤĂ Religion
    âŤŤâŤĂâŤĂ Spirituality
```

[10]:
```
from anytree.exporter import DotExporter
from IPython.display import Image

# from graphviz import Source
# from graphviz import render

import os
os.environ["PATH"] += os.pathsep + 'D:/graphviz-2.38/release/bin'

# graphviz needs to be installed for the next line!
DotExporter(root).to_picture("root.png")
Image(filename='root.png')
```

[10]:

9 URL based Category Tree Classifier

Here we use the tree information embedded in the URL links from database to find the category.

[11]:
```
# idURL = pd.read_csv('ID_URLs.csv',error_bad_lines=False,sep=',')
idURL = data2Cat[['id', 'link']]
idURL
```

[11]:
```
              id  \
0           3874
1           3260
2           1867
3           4988
4           2882
5           2675
6           2058
7           3875
8           2481
9           4615
10          4616
11          4229
12           629
```

```
13          3677
14          6523
15          3476
16          3063
17           630
18          6329
19          5766
20          4617
21          3064
22          3678
23          4811
24          6893
25          2253
26          4989
27          6524
28          4413
29          6894
...          ...
1093978  169330
1093979  165935
1093980  163674
1093981  168140
1093982  164267
1093983  165699
1093984  164675
1093985  164268
1093986  165310
1093987  168544
1093988  165936
1093989  163108
1093990  163878
1093991  162917
1093992  166909
1093993  168357
1093994  168939
1093995  167732
1093996  166705
1093997  168358
1093998  168754
1093999  167107
1094000  162547
1094001  162723
1094002  163675
1094003  167108
1094004  166121
1094005  164483
1094006  162919
```

```
1094007  168359

    link
0
https://navbharattimes.indiatimes.com/metro/lucknow/other-news/sandeepan-
vimalkant-nagar-was-a-struggling-and-combative-cultural-
worker/articleshow/73022935.cms
1
https://www.seithisolai.com/mulayam-singh-admitted-in-hospital.php
2
https://tamil.webdunia.com/article/regional-tamil-news/central-minister-put-
stone-for-new-battery-manufacturing-building-in-chennai-119123000004_1.html
3
https://www.newspointapp.com/telugu-news/publisher-webdunia-telugu/top-
news/articleshow/1450482007046259cf0a35c4c54156d3cd66c1fe
4
https://timesofindia.indiatimes.com/videos/entertainment/kids/tamil/kids-
stories-nursery-rhymes-baby-songs-krishna-and-kaliya-sri-krishna-kids-nursery-
story-in-tamil/videoshow/73010220.cms
5        https://www.newspointapp.com/english-news/publisher-sportskeeda/top-
news/tottenham-willing-offer-christian-eriksen-real-madrid-swap-deal-more-epl-
transfer-news-roundup-30th-
december-2019/articleshow/14504820cbdd450bc9083874f53f83d52aeb1ece
6
https://navbharattimes.indiatimes.com/tech/gadgets-news/reliance-jio-rupees-199
-top-up-voucher-offering-1tb-data-with-7-days-of-
validity/articleshow/73024124.cms
7
https://timesofindia.indiatimes.com/videos/entertainment/hindi/from-farhan-
akhtar-to-ranvir-shorey-bollywood-mournskushal-punjabis-untimely-
death/videoshow/73024843.cms
8
https://www.newspointapp.com/hindi-news/publisher-vishvatimes-hindi/top-
news/articleshow/14504820806e7880bf0c81749e1c135de9b397ef
9
https://timesofindia.indiatimes.com/city/mumbai/mumbai-cop-stops-bicycle-thief-
saves-the-day-for-househelps-son-who-had-
received-a-gift/articleshow/73025191.cms
10
https://timesofindia.indiatimes.com/home/education/news/recruitment-scam-417
-candidates-booked-for-getting-proxies-to-write-exam/articleshow/73025288.cms
11
https://telugu.samayam.com/tech/news/these-are-the-answers-of-amazon-app-quiz-
today-to-win-rs-5000/articleshow/73025096.cms
12
https://www.newspointapp.com/malayalam-news/publisher-mangalam-malayalam/top-
news/articleshow/14504820e25e717090a7f6b6a6dbdce6c4d58dd0
```

13
https://www.malayalamexpress.in/archives/991204/
14
https://www.malayalamexpress.in/archives/991231/
15
https://www.malayalamexpress.in/archives/991264/
16
https://timesofindia.indiatimes.com/city/gurgaon/full-meal-for-only-rs-10-at-karnal-grain-mkt/articleshow/73022866.cms
17
https://www.newspointapp.com/tamil-news/publisher-seithisolai-tamil/politics/articleshow/14504820209eacca43721bb50d36bfa9b73bab5b
18
https://www.newspointapp.com/hindi-news/publisher-uttamhindu-hindi/world/articleshow/14504820170b37788285aa75b0995201ec4a8742
19
https://timesofindia.indiatimes.com/videos/tv/hindi/bigg-boss-13-salman-khan-cleans-bathroom-washes-utensils/videoshow/73026487.cms
20
https://www.newspointapp.com/kannada-news/publisher-kannadadunia-kannada/top-news/articleshow/14504820688fb2101c56d229f5e191a82a0768f9
21
https://www.newspointapp.com/tamil-news/publisher-tamilsamayam/top-news/articleshow/145048200fe2fc0e89fb3112cea10b2ee992e509
22
https://www.newspointapp.com/telugu-news/publisher-telugusamayam/top-news/articleshow/145048209f13a386ebd667eb0a852c5ec19761d9
23
https://www.navjivanindia.com/news/bjp-leader-kaushal-kishore-attacks-up-police-over-law-and-order-in-up
24 https://tamil.samayam.com/latest-news/international-news/russian-veterinarian-successfully-performed-nanosurgery-on-cockroach/articleshow/73028127.cms?utm_source=colombia&utm_campaign=colombiaorganic&utm_medium=webwap
25
https://banglahunt.com/didi-character-loose-said-rajkumari-keshori/
26
https://tamil.samayam.com/lifestyle/health/foods-to-be-taken-to-boost-up-vitamin-d-in-your-body/articleshow/73028788.cms?utm_source=colombia&utm_campaign=colombiaorganic&utm_medium=webwap
27
https://www.newspointapp.com/punjabi-news/publisher-ajitjalandhar-punjabi/top-news/articleshow/145048208c0e8ddc261d4a5cfa1a19d57fb20aae
28
http://timeskerala.com/archives/169049
29
https://navbharattimes.indiatimes.com/state/other-states/hyderabad/congress-

leader-files-complaint-against-rss-chief-mohan-bhagwat-in-hyderabad/articleshow/73030977.cms
...
...
1093978
https://bengali.mahanagar24x7.com/state-minister-zakir-hussain-alleges-against-jangipur-municipality/
1093979
https://eisamay.indiatimes.com/bangladesh-news/ex-bangladesh-pm-khaleda-zias-bail-plea-rejected/articleshow/74336937.cms
1093980
https://www.newspointapp.com/english-news/publisher-et/india/offshore-patrol-vessel-vajra-to-enhance-coastal-security-launched/articleshow/1450482061af4c7f0e8c0fc1d9cfa03b556cf1f0
1093981
https://kannadanewsnow.com/kannada/ts-nagabharana-speech-on-bangalore-international-airport-kannadigas-job/
1093982
https://www.newspointapp.com/malayalam-news/publisher-mediaonetv-malayalam/top-news/articleshow/14504820bef6a45797559ef0e8ad9cc79316d9d1
1093983
https://www.newspointapp.com/urdu-news/publisher-qaumiawaz-urdu/top-news/articleshow/14504820b70ec8ecd8bda229db78662cb64fae58
1093984
https://www.sathyamonline.com/tn-prathapan-mp-who-arrived-in-riyadh-sarangi-activists-gave-a-warm-reception/
1093985
http://feedproxy.google.com/~r/Kasargodvartha/~3/v8sa76hx3Wc/sabu-kottarakkara-on-delhi-clash.html
1093986
https://timesofindia.indiatimes.com/videos/city/delhi/people-responsible-for-riots-will-be-punished-cm-arvind-kejriwal/videoshow/74336973.cms
1093987 https://www.newspointapp.com/english-news/publisher-sportskeeda/top-news/jon-moxley-training-with-former-ufc-champion-for-his-match-with-chris-jericho-at-aew-revolution/articleshow/1450482022fbe34f7a4b1f0934fdc536ec5de456
1093988
http://timeskerala.com/archives/201274
1093989
https://www.newspointapp.com/malayalam-news/publisher-pradhanavartha-malayalam/top-news/articleshow/14504820a822e97f5c27ec1d9d94773026ed2a58
1093990
https://www.newspointapp.com/english-news/publisher-nationalherald/top-news/bhumi-padnekar-joins-whispers-keepgirlsinschool-campaign-/articleshow/145048208e750196095e2b9a618c9c6ef3dbb4a3
1093991
https://www.majaru.com/2020/02/Majaru-entertainment_2524.html

1093992
https://navbharattimes.indiatimes.com/business/business-news/banks-to-clear-118-lakh-pending-applications-of-prime-minister-employment-generation-scheme-by-march-15/articleshow/74340967.cms
1093993
https://www.newspointapp.com/gujarati-news/publisher-khabarchhe-gujarati/top-news/articleshow/14504820fc2158547ac1079022f6a9d56ddff4bb
1093994
https://timesofindia.indiatimes.com/city/delhi/new-hostel-to-come-up-in-jnu-for-northeast-students-convention-centre-at-dwarka/articleshow/74340274.cms
1093995
https://www.newspointapp.com/bengali-news/publisher-nilkantho-bengali/top-news/articleshow/14504820fbec033dcc03d6c8d847294fe94d53e6
1093996
https://www.newspointapp.com/english-news/publisher-toi/sports/yates-retains-lead-as-pogacar-wins-uae-tour-fifth-stage/articleshow/14504820b9a1e519b7f2889d1dcb4b213da5f42c
1093997
https://www.newspointapp.com/tamil-news/publisher-seithipunal-tamil/top-news/articleshow/1450482012cf63eb87ce0d17d89a5892edaf18be
1093998
https://www.newstm.in/news/tamilnadu/77767-pudhukottai-obituary-titok-video-goes-viral.html
1093999
https://telugu.asianetnews.com/telangana/telangana-cm-kcr-helping-hand-to-physically-challenged-old-man-q6dcaq
1094000
https://m.economictimes.com/wealth/tax/post-tax-rs-5-lakh-income-will-be-higher-than-rs-5-16-lakh-heres-why/articleshow/74311201.cms
1094001
https://maharashtratimes.indiatimes.com/maharashtra/pune-news/sripati-shastri-is-the-only-book-of-glory/articleshow/74344723.cms
1094002
https://www.newspointapp.com/english-news/publisher-aninews/health/internet-suspension-extended-till-feb-28-in-aligarh/articleshow/14504820728c3e117e176244008af55065e9085e
1094003
https://www.newspointapp.com/kannada-news/publisher-varthabharati-kannada/top-news/articleshow/14504820bc6078ac926ded4f95b0b95aedcdbedb
1094004
https://timesofindia.indiatimes.com/city/thiruvananthapuram/parameswaran-very-much-like-hedgewar-bhagwat/articleshow/74325730.cms
1094005
https://www.newspointapp.com/malayalam-news/publisher-keralaonlinenews-malayalam/top-news/articleshow/14504820d7979e2c81ba3da1dd66eda771c29d8a
1094006
https://www.aninews.in/news//world/asia/toll-from-coronavirus-rises-to-17-in-

```
italy20200228011759
1094007
https://timesofindia.indiatimes.com/city/ahmedabad/auda-focuses-on-roads-
flyovers-and-drainage/articleshow/74346692.cms

[1094008 rows x 2 columns]
```

9.1 URL Parser Matcher with Category Tree

[12]:
```
# To find out category tree based on the URL links

stop_words = set(stopwords.words('english'))
cutoff = 0.25
urlMappings = []

for id, url in zip(idURL['id'], idURL ['link']):

    urlCat = url.split("/")[3:-2]
    urlCat

    totalSimScore = 0
    category_levels = [root]

    for word in urlCat:

        if ('-' in word):
            keywords = word.lower().split("-")
        else:
            keywords = [word.lower()]

        filtered_words = [w for w in keywords if not w in stop_words]

        for urlWords in filtered_words:
            level_similarity = []
            subCategories = category_levels[len(category_levels)-1].children

            for i in range(len(subCategories)):

                subCatWords = re.sub(r'[^A-Za-z0-9 ]+', '', subCategories[i].
↪name).lower().split()

                similarityScore = 0
                for catWords in subCatWords:
                    if urlWords in model and catWords in model:
                        similarityScore += model.similarity(urlWords, catWords)
```

```
                level_similarity.append(similarityScore/ len(subCatWords))

            maxsim_this_level = max(level_similarity, default=0)
            if (maxsim_this_level > cutoff):
                category_levels.append(subCategories[np.
↪argmax(level_similarity)])
                totalSimScore += maxsim_this_level


    if (len(category_levels) > 1):
        confidence = float(totalSimScore/(len(category_levels)-1))
    else:
        confidence = 0


    urlMappings.append([id, category_levels[len(category_levels)-1], confidence])

urlCatDetects = pd.DataFrame(urlMappings)
urlCatDetects.head(5)
```

[12]:

	0	1	2
0	3874	Node('/Industry/Sports')	0.313866
1	3260	Node('/Industry')	0.000000
2	1867	Node('/Industry')	0.000000
3	4988	Node('/Industry/Sports')	0.313866
4	2882	Node('/Industry/Online/Internet')	0.339198

[ ]:
```
urlCatDetects.to_csv('urlCatDetects.csv',index=False)
```

urlCatDetects.csv contains the category tree. Please ignore the above error (i just terminated the execution)

10 Category Tree Classifier using Article Description

[20]:
```
idesc = data2Cat[['id', 'long_description']]
idesc
```

[20]:

	id
0	3874
1	3260
2	1867
3	4988
4	2882
5	2675
6	2058

```
7           3875
8           2481
9           4615
10          4616
11          4229
12           629
13          3677
14          6523
15          3476
16          3063
17           630
18          6329
19          5766
20          4617
21          3064
22          3678
23          4811
24          6893
25          2253
26          4989
27          6524
28          4413
29          6894
...          ...
1093978  169330
1093979  165935
1093980  163674
1093981  168140
1093982  164267
1093983  165699
1093984  164675
1093985  164268
1093986  165310
1093987  168544
1093988  165936
1093989  163108
1093990  163878
1093991  162917
1093992  166909
1093993  168357
1093994  168939
1093995  167732
1093996  166705
1093997  168358
1093998  168754
1093999  167107
1094000  162547
```

```
1094001  162723
1094002  163675
1094003  167108
1094004  166121
1094005  164483
1094006  162919
1094007  168359
```

	long_description
0	B Sandipan awakened the play with consciousness giving direction to the artists. This was said by senior painter and actor Rakesh Pandey, Mathura color worker and Braj-speaking ...
1	Samajwadi Party founder Mulayam Singh has been admitted to a private hospital in Mumbai due to a stomach problem. Samajwadi Party founder Mulayam Singh (80) was abruptly ...
2	The Union Minister of State for the establishment of a new Technology Center in Chennai to promote battery-powered vehicle manufacturing. The use of vehicles to suit the population of the country ...
3	PANCARD is to be canceled within three days. Don't be surprised. If you do not link a PAN card with Aadhaar, the PAN card is canceled. They have only 3 days to expire ...
4	Here's Presenting popular Children Nursery Story 'Krishna And Kaliya | Sri Krishna'. For popular children Stories, kids songs, children songs, children poems, baby songs, baby rhymes, kids nursery rhymes, nursery poems in Tamil visit Etimes Tamil kids sections. Check out Etimes Kids videos secti...
5	Norwich City v Tottenham Hotspur - Premier League Hello and welcome to the EPL transfer news roundup for the day! Here are the top stories of the day surrounding the ! Erling Braut Haaland joins Dortmund Borussia Dortmund have confirmed the signing of Erling Braut Haaland from Red Bull Salzburg...
6	New Delhi is good news for Reliance JioFiber users. The company is now offering 1TB (1000GB) of data on its Rs 199 top-up voucher. Earlier only 100GB data was available in this plan. In plan's validity ...
7	The Bollywood fraternity on Friday expressed pain and shock on the news of TV actor Kushal Punjabi's untimely death. Kushal committed suicide at his home in Mumbai's Bandra area late on Thursday night. As per a statement from the police, the actor \committed suicide by hanging himself from a fan...
8	Hemant Soren, the working president of the Jharkhand Mukti Morcha (JMM), was sworn in as the Chief Minister of the state here on Sunday and said that his government will continue to emphasize on education and health. Soren's main ...
9	A traffic police constable helped an underprivileged boy recover a bicycle,

which was gifted to him, in Bandra (west) on Saturday.
10 Recruitment Scam: 417 candidates booked for getting proxies to write exam Top Searches: Recruitment Scam: 417 candidates booked for getting proxies to write exam TNN | Updated: Dec 30, 2019, 10:40 IST In yet another recruitment scam in the state, the Haryana police have booked 417 candidates who...
11
Leading online shopping giant Amazon runs a quiz every day on its app. The quiz involves selecting some of the answers and giving them presents. Today (Dec ...
12
Melbourne: Australian pacer Peter Siddle has retired from international cricket. Siddell announced his retirement following the Melbourne Test against New Zealand. Friends ...
13
Mum bye: Prohibition of social media in the Navy. The bans are Facebook WhatsApp and Instagram-based social media apps. NON-SECURE-NON-TECHNICAL CHARACTERISTICS ...
14
TANZANIA: Fosta, 57, the oldest rhinoceros in the world, has left for Tanzania. The female rhinoceros is in the Ngorongoro protected area of Tanzania.
15
Kuwait City: The 5th Islamic Seminar Propaganda Seminar, organized by Kerala Kuwait Islami Center, Kuwait ...
16 Haryana chief minister Manohar Lal Khattar on Sunday inaugurated a canteen in Karnal that will cater to the poor, especially farmers and labourers. Full meal is priced at Rs 10 at the canteen. CM said more such canteens will be opened in the state.
17
DMK leader Stalin and former chief minister Karunanidhi have protested against the amendment to the citizenship law. The Central BJP government has ...
18
New Delhi (Uttam Hindu News): Pakistan, which has been plotting against India again and again, has once again lost its face. This time Pakistan has made its gritty in the UN Security Council. Media Reports ...
19 Bigg Boss 13 housemates gave a tough time to the new captain Shehnaaz Gill by avoiding their duties. This led to host and Bollywood superstar SalmanKhan entering the house and cleaning the kitchen and bathroom on his own as the contestants watched in embarrassment.
20
Pakistani Danish Kaneria's statement has been the subject of a lot of talk about discrimination against Hindus in Pakistan. Former Pakistani leader Shahid Afridi has now revealed another issue. Interview ...
21
The Transport Corporation has begun its investigation after a complaint over state-owned buses was brought to light in Chennai. Trend ...
22
It is quite natural for field umpires to make decisions on the field. Most importantly, umpires often make mistakes during the match. Fielders on the

ground loudly ...
23
The criminals in Uttar Pradesh are fearless in Yogi Raj. The police are also under question. Alam is that along with the leaders of the opposition party, BJP MPs and MLAs also question the law and order of Uttar Pradesh.
24
Surprisingly, the incident in which the beanbirds in Russia have been admitted to a veterinary hospital in due time is a serious health condition.
25
Bangla Hunt Desk: Why did you file for citizenship law in 28, did you change colors? Your character is at fault. So you keep the Bengali brothers confused. Banda's Onda. BJP Citizens ...
26
From time to time, we know how important vitamin D is to the body and health. In our tropical country, India has suffered from this shortage in recent years.
27 Banga,
Dec 30 (Jasbir Singh Nurpur) - Satvir Singh Palli Fikki, in-charge of the Banga constituency has been appointed by the Punjab Government as Chairman of the District Planning Board. Satvir Singh Palli Fikki appointed Chief after
28
Sex in wet shower. If you want to lick each other's bodies and get wet, you can have sex with each other. The experience is different. Manassu ... Close Button
29
Hyderabad Senior Congress leader V Hanumanta Rao has filed a complaint against Rashtriya Swayamsevak Sangh (RSS) chief Mohan Bhagwat. In his complaint, Rao said that Bhagwat gave 130 crore countrymen a Hindu ...
...
...
1093978
Highlights Jungipur Municipal Board (PBB) indirectly alleges financial corruption, State Minister of State for Legislative Assembly Zakir Hossain launched the central government to solve the drinking water crisis in every state.
1093979
BNP chairperson Khaleda Zia's bail in the corruption case has been rejected by Bangladesh court. The court has said that Khaleda Zia will not get all the facilities like a common man. Zia's diabetic, hypertension ...
1093980 CHENNAI: The sixth offshore patrol vessel 'Yard 45006 VAJRA' to enhance coastal security was formally launched in the presence of Union Shipping Minister Mansukh Mandaviya and senior government officials here on Thursday. The vessel, built by Larsen and Toubro under the Centre's 'Make in India' ...
1093981
Bangalore: The state government has imposed a condition that Kannadigas should be given jobs when the state government gives space to Kempegowda International Airport. Accordingly, the report will be published in Kannada ...
1093982
State Chief Electoral Officer Teekaram Meena said college politics was

unnecessary. Completely agree with the High Court ruling. Campuses cannot be riotous with political games. Rash on campus ...
1093983
New Delhi: The Delhi High Court has issued notice to the Delhi government, police and others on Thursday in the North East Delhi violence case and the next hearing of the matter will be on April 13. Chief Justice DN Patel and ...
1093984
Riyadh: The Riyadh-based Sarangi Art and Cultural Forum, Ji. Member of Parliament for Thrissur TN Prathapasana received the Kartikeyan Memorial Voice of Democracy.
1093985
Kasargod: (www.kasaragodvartha.com)
1093986 Delhi chief minister Arvind Kejriwal said that people responsible for violence in northeast Delhi should not be spared, even if they are from the Aam Aadmi Party. Delhi CM Kejriwal said, \Any person found guilty should be given stringent punishment. If any Aam Aadmi Party person is found guilty,...
1093987 Moxley is leaving nothing to chance (Pic Source: AEW) It's still hard to believe that Jon Moxley shocked the world last year when he joined AEW at Double or Nothing and made a statement by attacking and . His feud with Jericho has been building over the last couple of months as Moxley has withst...
1093988
New Delhi: The NIA has denied the bail plea of the accused in the Pulwama case. NIA has not filed a charge sheet against the accused in the case.
1093989
Kapil Mishra said there was nothing provocative in his speech in support of the Amendment of Citizenship Law at Maujpur Chowk in Delhi. Ask the police officer to qualify for the road ...
1093990 Bollywood actress Bhumi Pednekar has collaborated with feminine care brand Whisper for its new campaign #KeepGirlsInSchool that aims to create awareness on how even today, girls across India drop out of school on hitting puberty. As part of this campaign, Whisper has launched its new film that b...
1093991
In the Indian capital, Delhi, the revised Citizenship Act (CAA) is being targeted with violence aimed at Muslims. Most of the incidents were reported in Muslim homes and shops. Claims to have ...
1093992
New Delhi, 27 February (Language) Government asked banks to dispose of about 1.18 lakh pending applications received for loans under self-employment scheme 'Prime Minister Employment Generation Program (PMEGP)' by 15 March.
1093993
Responding to a question in the Legislative Assembly regarding the primary and collective health centers, Health Minister Nitin Patel said that the best healthcare facilities in remote areas of the state are also ...
1093994 NEW DELHI: A new hostel will soon come up at the Jawaharlal Nehru University (JNU) for the students belonging to the Northeast besides a regional convention centre in Delhi, Union minister Jitendra Singh said on Thursday. Singh said new bamboo clusters and technology centres will also come

up ...
1093995
Now various boards are under examination. There are various boards being tested across the country. The big test of life. Long tireless work But after the Delhi violence, the Delhi Chest Board Exam ...
1093996 PARIS: Slovenian Tadej Pogacar produced a stunning final sprint on the top of Jebel Hafeet to win the 162km fifth stage of the UAE Tour on Thursday. Pogacar, 21, edged Kazakh Alexey Lutsenko in a photo finish for the victory, in what was the Tour's second ascent of the mountain overlooking the d...
1093997
The Tamil Nadu All Stripped Drinking Water Supply Association (GTU) has announced an indefinite strike from 6 pm onwards. President of the Association of Drinking Water Manufacturers
1093998
Friends of Pudukkottai, Bhagarla and Riyaz are very welcome to the videos posted on Kalatu Diktak. Riaz Dicta says Bharla died in an accident ...
1093999
Telangana Chief Minister KCR once again expressed his humanity. The problem of a disabled old man was solved on the road. CM going to participate in a private event on Thursday ...
1094000 Final Tax liability with cess @ 4% Rs 15,080 As shown in the example above, even a marginal increase in income above Rs 5 lakh has resulted in more tax outgo than the incremental income. Increase in income beyond Rs 5 lakh is marginal - only by Rs 10,000 but the tax liability due to such increme...
1094001
Pune: 'Rashtriya Swayamsevak Sangh has made a big contribution to the work of various educational and social institutions in Pune city. Dr. A book on the glory of Shripati Shastri will be published soon. You ...
1094002
Aligarh (Uttar Pradesh) [India], Feb 27 (ANI): Suspension of mobile internet services on Thursday was extended till February 28 in Aligarh district following recent clashes between police and anti-Cit...
1094003
New Delhi, Fr. NEW DELHI: Delhi Chief Minister Arvind Kejriwal on Thursday announced that he will provide relief to the family of the victims of the violence that broke out in north-eastern Delhi. Injury in violence ...
1094004
Recalling his long association with RSS ideologue P Parameswaran who passed away last week, RSS sarsangh chalak Mohan Bhagwat here, on Wednesday, said Parameswaran was a model swayam sewak who very mu...
1094005
The Kerala Administrative Service (KAS) has decided to cancel the examination on the basis of serious allegations leveled against it.
1094006 Rome [Italy]. Feb 28 (Sputnik/ANI): The death toll from coronavirus infection in Italy has risen to 17 with the addition of three more victims in the northwest of the country on Thursday, Angelo Borrelli, head of National Civil Protection Service, said. \The number of infected reached 650, while...

```
1094007
By keeping traffic management its top priority, Ahmedabad Urban Development
Authority (AUDA) aims to finish construction of four flyovers by 2021. With a
project cost of Rs 105 cr the flyovers on SP R...

[1094008 rows x 2 columns]
```

[21]:
```
# To compute the confidence score of category prediction based on the numerical␣
 ↪distribution of values in the list
def computeConfidence(similarityList):

    similarScores = set(similarityList)
    highest = max(similarScores)

    similarScores.remove(highest)
    if (len(similarScores) == 0):
        return 0

    secondHighest = max(similarScores)

    return (highest - secondHighest)/ (highest)
```

[40]:
```
# To do multi level classification - category tree - using all the words
# in the article (stop words remove and cleaned) and also the words in the␣
 ↪category tree nodes.

stop_words = set(stopwords.words('english'))
similarity_cutoff = 0.3
confidence_cutoff = 0.05

categoryMappings = []

for id, fulltxt in zip(idesc['id'], idesc['long_description']):


    fulltxt_filtered = [w for w in str(fulltxt).lower().split() if not w in␣
 ↪stop_words]

    category_levels = [root]

    while True:

        tree_level = len(category_levels)

        subCategories = category_levels[tree_level - 1].children

        if len(subCategories) == 0:
```

```
            categoryMappings.append([id, category_levels[tree_level - 1],␣
↪confidenceRow])
            break

        level_similarity = [0] * len(subCategories)

        for word in fulltxt_filtered:
            # To exclude small words
            if len(word) < 3:
                continue

            if word in model:
                # To handle categories with multiple words. Eg: Banking & Finance
                for i in range(len(subCategories)):
                    subCatWords = re.sub(r'[^A-Za-z0-9 ]+', '', subCategories[i].
↪name).lower().split()

                    similarityCategory = []
                    for catWords in subCatWords:
                        if catWords in model:

                            similarityCategory.append(abs(model.similarity(word,␣
↪catWords)))

                    if (len(similarityCategory) > 0):
                        level_similarity[i] = level_similarity[i] +␣
↪similarityCategory[np.argmax(similarityCategory)]

        maxsim_this_level = max(level_similarity, default=0)
        confidence = computeConfidence(level_similarity)

        # For level 1 entry, there is no confidence cutoff.
        # If level > 1, then confidence score should be > cutoff
        if (maxsim_this_level > similarity_cutoff and
                (tree_level == 1 or confidence > confidence_cutoff)):

            if (tree_level == 1):
                confidenceRow = confidence
            category_levels.append(subCategories[np.argmax(level_similarity)])
        else:
            categoryMappings.append([id, category_levels[tree_level - 1],␣
↪confidenceRow])
            break

catDetects = pd.DataFrame(categoryMappings)
catDetects.head(5)
```

```
# computeConfidence < 0.25 then the w2v classification is considered ambiguous
```

[40]:
```
      0                                                                1  \
0  3874                    Node('/Industry/Politics/Politics/BJP Followers')
1  3260                Node('/Industry/Family Relationships/Matrimony/Groom')
2  1867     Node('/Industry/Gadgets & Technology/Home Entertainment Systems')
3  4988                                              Node('/Industry/Online')
4  2882  Node('/Industry/Family Relationships/Parenting/Special Needs Kids')

          2
0  0.090846
1  0.141484
2  0.173548
3  0.093610
4  0.037887
```

[41]:
```
catDetects.to_csv('textCatDetects.csv',index=False)
```

10.1 Standby Logic to do 1st Level Classification

[ ]:
```
# To do 1st level classification

# let category be defined as below
categories = ["politics", "sports", "news", "education", "finance",␣
 ↪"entertainment", "health", "environment"]

for id, article in zip(englishArticles['id'],␣
 ↪englishArticles['long_description']):

    similarity = [0] * len(categories)
    for word in article.split():

        if word in model:
            for i in range(len(categories)):
                similarity[i] = similarity[i] + model.similarity(word,␣
 ↪categories[i])

    print("Category: ")
    print(categories[np.argmax(similarity)], similarity, max(similarity),␣
 ↪computeConfidence(similarity))

    # computeConfidence < threshold then the w2v classification is ambiguous
```

[ ]:
```
```

11 Merging Article Description Classifier with URL & LDA-NMF Classifiers

The results of word to vector methods over article description and article URL need to be merged with LDA-NMF combination model, which is done separately.

[82]:
```
artClassifier = pd.read_csv('textCatDetects.csv',error_bad_lines=False,sep=',')
artClassifier.head(5)
```

[82]:

	id	category	confidence
0	3874	Node('/Industry/Politics/Politics/BJP Followers')	0.090846
1	3260	Node('/Industry/Family Relationships/Matrimony/Groom')	0.141484
2	1867	Node('/Industry/Gadgets & Technology/Home Entertainment Systems')	0.173548
3	4988	Node('/Industry/Online')	0.093610
4	2882	Node('/Industry/Family Relationships/Parenting/Special Needs Kids')	0.037887

[83]:
```
urlClassifier = pd.read_csv('urlCatDetects.csv',error_bad_lines=False,sep=',')
urlClassifier.head(5)
```

[83]:

	id	category	confidence
0	3874	Node('/Industry/News & Politics/Politics')	0.432521
1	3260	Node('/Industry')	0.000000
2	1867	Node('/Industry')	0.000000
3	4988	Node('/Industry/Entertainment Movie')	0.325226
4	2882	Node('/Industry/Online/Internet')	0.339198

[84]:
```
ldanmfClassifier = pd.read_csv('LDA-NMF-Result.
 ↪csv',error_bad_lines=False,sep=',')
ldanmfClassifier.head(5)
```

[84]:

	id	category_tree
0	5177	News & Politics
1	2882	Education/KidEducation
2	2675	News/Relationship
3	2676	News/
4	3875	Finance/OTT APPS/News/Business/Location

11.1 LDA, NMF, W2V Merger Logic

[89]:
```
result = []
for idx in artClassifier.index:

    if (artClassifier['confidence'][idx] > 0.05):
        result.append([artClassifier['id'][idx],␣
↪str(artClassifier['category'][idx])[6:-2].replace("/", "^")])

    elif (urlClassifier['confidence'][idx] > 0.4):
        result.append([urlClassifier['id'][idx],␣
↪str(urlClassifier['category'][idx])[6:-2].replace("/", "^")])
    else:
        row = ldanmfClassifier.loc[ldanmfClassifier['id'] ==␣
↪artClassifier['id'][idx]]

        if (len(row) == 0):
            result.append([artClassifier['id'][idx],␣
↪str(artClassifier['category'][idx])[6:-2].replace("/", "^")])
        else:
            result.append([artClassifier['id'][idx], row['category_tree'].
↪to_string(index=False)])
```

[90]:
```
resultDF = pd.DataFrame(result)
resultDF.to_csv('Result_Submission.csv',index=False)
```

The resuls are found in the file: Result_Submission.csv. It contains document id and corresponding category tree.