ROLL NO – 162010022

PRACTICAL – 4
EXTRACTION OF LATENT VARIABLES

AIM:- To clean the given dataset and extract date and cities to different columns

In [2]:
```
import numpy as np
import pandas as pd
```

In [3]:
```
dataset = pd.read_csv('/home/pranay/Desktop/rediff_realtime_news_201704_201706.csv',delimiter = '\t',nrows=20)
```

In [4]:
```
dataset.head()
```

Out[4]:

	url	source	crawl_time	title	trimmed_description	summary
0	http://feeds.reuters.com/~r/reuters/INtopNews/...	Reuters	2017-04-01T00:00:03Z	Fed signals it could promptly start shedding b...	The Federal Reserve could begin shrinking its ...	NEW YORK (Reuters) - The Federal Reserve could...
1	http://www.nerve.in/news:2535003905774	Nerve	2017-04-01T00:00:03Z	Hasten transmission projects to meet summer po...	Friday, 31 March 2017|http://www.nerve.in/news...	NaN
2	http://www.nerve.in/news:2535003905777	Nerve	2017-04-01T00:00:03Z	Armed Forces short of more than 9,000 officers	New Delhi, March 31 - India's three armed serv...	NaN
3	http://www.prokerala.com/news/articles/a729599...	Prokerala	2017-04-01T00:00:03Z	Nearly 30,000 migrants reach Italy in 2017	Nearly 30,000 migrants reach Italy in 2017 Syn...	A total 27,850 migrants and refugees landed in...
4	http://www.nerve.in/news:2535003905778	Nerve	2017-04-01T00:00:03Z	Kudankulam n-plant unit 2 begins warranty oper...	March 2017|http://www.nerve.in/news:2535003905...	NaN

In [5]:
```
import datefinder
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import re
import pandas as pd
from numpy import nan
import datetime
```

In [6]:
```
def StopWordsMixed(str):
    str = re.sub(r'[^a-zA-Z0-9_:]+', ' ', str)
    stop_words = set(stopwords.words('english'))
    word_tokens = word_tokenize(str)
    filtered_sentence = [w for w in word_tokens if not w in stop_words ]
    return ' '.join(filtered_sentence)
```

In [7]:
```
def Mixed(str):
    str = re.sub(r'[^a-zA-Z0-9_: ]+', '', str)
    word_tokens = word_tokenize(str)
    return ' '.join(word_tokens)
```

In [8]:
```
def extractCities(str):
    str = StopWordsMixed(str)
    word_tokens = word_tokenize(str)
    for word in word_tokens:
        for city in cities_list:
            if(word.lower()==city.lower()):
                return word
```

In [9]:
```
def extractDate(str):
    matches = None
    str = re.sub("\S*[^\w\s,]\S*|\S*\d,\d\S*","",str)
    print('\n')
    print(str)
    print('\n')
    matches = datefinder.find_dates(str,strict=True)
    for match in matches:
        if(match.year!=2017):
            continue
        return match
```

In [25]:
```
cities_df = pd.read_csv('CitiesData.csv',usecols = ['cities'])
```

In [26]:
```
cities_list = cities_df['cities'].tolist()
```

In [27]:
```
dataset['city'] = dataset.apply(lambda row: extractCities(row['trimmed_description']+" "+row['summary']),axis = 1)
```

In [29]:
```
dataset
```

Out[29]:

	url	source	crawl_time	title	trimmed_description	summary	date	city	year	month	day
0	http://feeds.reuters.com/~r/reuters/INtopNews/...	Reuters	2017-04-01T00:00:03Z	Fed signals it could promptly start shedding b...	The Federal Reserve could begin shrinking its ...	NEW YORK (Reuters) - The Federal Reserve could...	2017-04-01 00:00:03	None	2017	4	1
1	http://www.nerve.in/news:2535003905774	Nerve	2017-04-01T00:00:03Z	Hasten transmission projects to meet summer po...	Friday, 31 March 2017|http://www.nerve.in/news...		2017-04-01 00:00:03	Delhi	2017	4	1
2	http://www.nerve.in/news:2535003905777	Nerve	2017-04-01T00:00:03Z	Armed Forces short of more than 9,000 officers	New Delhi, March 31 - India's three armed serv...		2017-04-01 00:00:03	Delhi	2017	4	1
3	http://www.prokerala.com/news/articles/a729599...	Prokerala	2017-04-01T00:00:03Z	Nearly 30,000 migrants reach Italy in 2017	Nearly 30,000 migrants reach Italy in 2017 Syn...	A total 27,850 migrants and refugees landed in...	2017-04-01 00:00:03	None	2017	4	1
4	http://www.nerve.in/news:2535003905778	Nerve	2017-04-01T00:00:03Z	Kudankulam n-plant unit 2 begins warranty oper...	March 2017|http://www.nerve.in/news:2535003905...		2017-04-01 00:00:03	Delhi	2017	4	1