ROLL NO – 162010022

PRACTICAL – 3
STUDY AND CLEANING OF DATASET

AIM: Study and cleaning of dataset

Description of dataset:

The Redif dataset contains more than 14 lakh news articles from various newspapers in India for the year 2017, crawled by the Rediff Realtime News platform. Most of the major newspapers are covered in this dataset, and top news from most of these sites should be a part of the dataset.

1) Trending Topics
2) Who can win the election based on treands
3) Who were the personalities most talked about
4) Which topic were covered most in newspaper.
5) Bias in terms of their political views.

The dataset contains the News source name, the time of crawl ,the title of the news article, a summary and a trimmed description of the news article .

Attributes :

1) URL: The URL gives direct link to see the resource .
2)Source : The Source Contains the details about the publishing news organization. 3)Crawl time : The time of crawl is time when the news was updated over web
4)Title: Main heading of the news.
5)Trimmed Description : Derived description.
6)Summary: Summary of the news article.

Advantages of using Python:

• **Easy Syntax** Python's syntax is easy to learn, so both non-programmers and programmers can start programming right away.

• **Readability** Python's syntax is very clear, so it is easy to understand program code. (Python is often referred to as "executable pseudo-code" because its syntax mostly follows the conventions used by programmers to outline their ideas without the formal verbosity of code in most programming languages; in other words syntax of Python is almost identical to the simplified "pseudo-code" used by many programmers to prototype and describe their solution to other
programmers. Thus Python can be used to prototype and test code which is later to be implemented in other programming languages).[citation needed]

• **High-Level Language** Python looks more like a readable, human language than like a low-level language. This gives you the ability to program at a faster rate than a low-level language will allow you.

• **Object oriented programming** Object-oriented programming allows you to create data structures that can be re-used, which reduces the amount of repetitive work that you'll need to do. Programming languages usually define objects with namespaces, like class or def, and objects can edit themselves by using keyword, like this or self. Most modern programming languages are object-oriented (such as Java, C++, and C#) or have support for OOP features (such as Perl version 5 and later). Additionally object-oriented techniques can be used in the design of almost any nontrivial software and implemented in almost any programming or scripting language. (For example a number of Linux kernel features are "objects" which

implement their own encapsulation of behavior and data structive via pointers, specifically pointers to functions, in the C programming language).[citation needed] Python's support for object-oriented programming is one of its greatest benefits to new programmers because they will be encountering the same concepts and terminology in their work environment. If you ever decide to switch languages, or use any other for that fact, you'll have a significant chance that you'll be working with object-oriented programming.

• **It's Free** Python is both free and open-source. The Python Software Foundation distributes pre-made binaries that are freely available for use on all major operating systems called CPython. You can get CPython's source-code, too. Plus, you can modify the source code and distribute as allowed by CPython's license. [2] (Luckily, CPython has a permissive free software license attitude.)

• **Cross-platform** Python runs on all major operating systems like Microsoft Windows, Linux, and Mac OS X.

• **Widely Supported** Python has an active support community with many web sites, mailing lists, and USENET "netnews" groups that attract a large number of knowledgeable and helpful contributes.

• **It's Safe** Python doesn't have pointers like other C-based languages, making it much more reliable. Along with that, errors never pass silently unless they're explicitly silenced. This allows you to see and read why the program crashed and where to correct your error.

Code:-

```
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 [24]:
```
dataset['trimmed_description'].fillna('',inplace=True)

dataset['summary'].fillna('',inplace=True)

dataset['date'] = dataset.apply(lambda row: extractDate(row['trimmed_description']+" "+row['summary']),axis = 1)

dataset['date'] = dataset['date'].fillna(dataset['crawl_time'])

dataset['date'] = pd.to_datetime(dataset['date'])

dataset['year'],dataset['month'],dataset['day'] = dataset['date'].dt.year,dataset['date'].dt.month,dataset['date'].dt.day
```

```
The Federal Reserve could begin shrinking its  balance sheet as soon as this year, earlier than most economists expect,
New York Fed President William Dudley said on Friday in the central  most definitive comments on the question that loom
s over financial  The  assertion temporarily pushed the dollar lower and raised yields on  bonds, and added  influentia
l voice to at least three other officials at the Fed eyeing a prompt end to a c NEW YORK   The Federal Reserve could be
gin shrinking its  balance sheet as soon as this year, earlier than most economists expect, New York Fed President Will
iam Dudley said on Friday in the central  most definitive comments on the question that looms over financial



Friday, 31 March    New Delhi, March 31  The Central government on Friday urged the state utilities to hasten the proce
ss of completion of transmission projects in the pipeline in order to meet the power demand in the coming summer, an of
ficial statement  According to the Power Ministry, the all India peak demand during the upcoming summer is expected to
be of the order of 165  Certain states may experience constraints due
```