PRACTICAL – 4 EXTRACTION OF LATENT VARIABLES

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

		pandas as pd										
	: dataset	t = pd.read_csv('/home/pr	anay/D	esktop/rediff_	realtime_	news_201704_20	1706.csv',delimit	ter = '\t'	,nrows=20)	
In [4]	: dataset.head()											
Out[4]	:		url s	ource	crawl tin	ne	title	trimmed	description		summ	ary
	o http	p://feeds.reuters.com/~r /INtop	droutoro.		2017-04-01T00:00:0	27	Fed signals it could y start shedding b	The Federal Reserve		NEW YORK Federal F	Co. 1. (4) 24 27	The
	1	http://www. /news:2535003	nerve.in	Nerve	2017-04-01T00:00:0	F	lasten transmission cts to meet summer		lay, 31 March			VaN
	2	http://www. /news:2535003	nerve.in 905777	Nerve	2017-04-01T00:00:0	3Z Armed F	orces short of more than 9,000 officers	New Delhi, March 31	India's three armed serv		r	NaN
	3	http://www.proker /news/articles/a7	ala.com 29599 Pro	kerala	2017-04-01T00:00:0	3Z Nea	rly 30,000 migrants reach Italy in 2017	Nearly 30,000 migrar	nts reach Italy in 2017 Syn	A total 27,85 refug	0 migrants lees landed	and
	4	http://www. /news:2535003	nerve.in 905778	Nerve	2017-04-01T00:00:0	3Z Kudan beg	kulam n-plant unit 2 gins warranty oper	March 2017 http:// /news:2	www.nerve.in 535003905		V	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]	<pre>i: def StopWordsMixed(str): str = re.sub(r'[^a-zA-Z0-9_:]+', ' ', str) stop_words = set(stopwords.words('english')) word_tokens = word_token!ze(str) filtered_sentence = [w for w in word_tokens if not w in stop_words] return ' '.join(filtered_sentence)</pre>											
In [7]	wor	<pre>ked(str): r = re.sub(r'[^a rd_tokens = word turn ' '.join(wo</pre>	tokenize	(str)	'', str)							
In [8]	str	tractCities(str) r = StopWordsMix rd_tokens = word r word in word_t for city in ci if(word.lo return	ed(str) _tokenize okens: ties_list wer()==ci	:	/er()):							
In [9]	<pre>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</pre>											
		return match										
In [25]:	cities_d	return match df = pd.read_csv	/('Cities	Data.c	sv',usecols =	['cities'	1)					
						['cities'	1)					
In [26]:	cities_l	df = pd.read_csv list = cities_df	['cities].tol	ist()			scription']+" "+	row['summa	ry']),axi	s = 1)	
In [26]:	cities_l	df = pd.read_csv list = cities_df	['cities].tol	ist()				row['summa		s = 1)	
In [26]: In [27]: In [29]:	cities_l dataset[dataset	df = pd.read_csv list = cities_df ['city'] = datas	f['cities set.apply	'].tol	a row: extract crawl_time	Cities(ro	w['trimmed_de	ription summary NEW YORK (Reuters) -			ear month	48
In [26]: In [27]: In [29]:	cities_l dataset[dataset o http://orth	df = pd.read_csv list = cities_df ['city'] = datas url p://eeds.reuters.com	set.apply source Reuters	'].tol (lambd	crawl_time -01T00:00:03Z m	cities (ro title ed signals it uld promptly art shedding b Hasten ransmission projects to eet summer	trimmed_desc The Federal R-could begin shrini Friday, 31 2017 http://www.n-	ription summary NEW YORK (Reuters)- river Federal Reserve could March	date	city ye	ear month	1
In [26]: In [27]: In [29]:	cities_l dataset[dataset o http://orth	df = pd.read_csv list = cities_df ['city'] = datas url p://Reeds.reuters.com reuters/iNtopNews/	set.apply source Reuters	'].tol (lambd 2017-04-	crawl_time -01T00:00:032 tm	cities (ro title ed signals it uld promptly art shedding b Hasten ransmission projects to eet summer	trimmed_desc The Federal R-could begin shrini Friday, 31 2017 http://www.n-	ription summary NEW YORK (Reuters)- The Federal Reserve could March erve.in news ch 31 - serv	date 2017-04-01 00:00:03	city ye	ear month	1
In [26]: In [27]: In [29]:	cities_l dataset[dataset http://	df = pd.read_csv list = cities_df ['city'] = datas url p://leeds.reuters.com reuters/INtopNews/ http://www.nerve.in pws.2535003905774 http://www.nerve.in pws.2535003905777	set.apply source Reuters : Nerve :	[lambd]	crawl_time -01T00:00:03Z	cities (ro title ed signals it ald promptly art shedding b Haster ransmission projects to pocus eet summer po med Forces hort of more than 9,000	trimmed_desc The Federal R-could begin shrini Friday, 31 2017 http://www.n-/h	ription summary NEW YORK (Reuters) The Federal Reserve could March erve.in news A total igrants	date 2017-04-01 00:00:03 2017-04-01 00:00:03	None 20	ear month	1