Week7_6_NLP_preprocessing

May 21, 2021

NLP text preprocessing

- **In NLP text preprocessing is more essential, and NLP is an art to extract some information from the text documents.**
- ***First step is the preprocessing, and how to process the data are given in detail***
- **Import Libraries and download NLTK tools**

[1]:

```python
import pandas as pd
import numpy as np
import nltk
import matplotlib.pyplot as plt

from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))
wordnet_lemmatizer = WordNetLemmatizer()
from termcolor import colored

nltk.download('wordnet')
nltk.download('averaged_perceptron_tagger')
nltk.download('punkt')
nltk.download('stop')
nltk.download('stopwords')
# nltk.download('all')

from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences

import warnings
warnings.filterwarnings("ignore")
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     /home/jayanthikishore/nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /home/jayanthikishore/nltk_data…
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     /home/jayanthikishore/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Error loading stop: Package 'stop' not found in index
[nltk_data] Downloading package stopwords to
[nltk_data]     /home/jayanthikishore/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
```

Text datasets reading

```
[2]: tdata1 = pd.read_csv("~/Downloads/ML_classwork/Week7_srrt/tweets_labels.csv")
```

```
[3]: tdata1
```

```
[3]:                    *screams in 25 different languages*  0.6
     0     Families to sue over Legionnaires: More than 4…  0.1
     1     Pandemonium In Aba As Woman Delivers Baby With…  0.4
     2     My emotions are a train wreck. My body is a tr…  0.2
     3     Alton brown just did a livestream and he burne…  0.5
     4     @TinyJecht Are you another Stand-user? If you …  0.5
     …                                                  …    …
     1858  @Trollkrattos Juan Carlos Salvador The Secret …  0.5
     1859  @devon_breneman hopefully it doesn't electrocu…  0.5
     1860  Businesses are deluged with invokces. Make you…  0.5
     1861  #BREAKING411 4 police officers arrested for ab…  0.1
     1862  @News@ Refugio oil spill may have been costlie…  0.2

     [1863 rows x 2 columns]
```

Text dataset information and shape

```
[4]: #Total information of the data file
     tdata1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1863 entries, 0 to 1862
Data columns (total 2 columns):
 #   Column                               Non-Null Count  Dtype
---  ------                               --------------  -----
 0   *screams in 25 different languages*  1863 non-null   object
 1   0.6                                  1863 non-null   float64
dtypes: float64(1), object(1)
memory usage: 29.2+ KB
```

```
[5]: #Data shape
     tdata1.shape
```

```
[5]: (1863, 2)
```

Rename column name

[6]:
```
# renamming the column names for the text dataset
tdata = tdata1.rename(columns={"*screams in 25 different languages*":"tweet",␣
 ↪"0.6":"lbel"})
tdata.tweet
```

[6]:
```
0       Families to sue over Legionnaires: More than 4…
1       Pandemonium In Aba As Woman Delivers Baby With…
2       My emotions are a train wreck. My body is a tr…
3       Alton brown just did a livestream and he burne…
4       @TinyJecht Are you another Stand-user? If you …
                              …
1858    @Trollkrattos Juan Carlos Salvador The Secret …
1859    @devon_breneman hopefully it doesn't electrocu…
1860    Businesses are deluged with invokces. Make you…
1861    #BREAKING411 4 police officers arrested for ab…
1862    @News@ Refugio oil spill may have been costlie…
Name: tweet, Length: 1863, dtype: object
```

Display the first document from the text document

[7]:
```
#Displaying the first document
tdata.tweet[0]
```

[7]:
```
"Families to sue over Legionnaires: More than 40 families affected by the fatal
outbreak of Legionnaires' disea… http://t.co/ZA4AXFJSVB"
```

Binning the y column

[8]:
```
# Binning the y values to integers
bins = np.array([0.0,0.25,0.5,0.75])
# bins = np.array([0.0,0.2,0.4,0.6,0.8])
tdata['label'] = np.digitize(tdata['lbel'],bins)
# tdata['label'] = np.digitize(tdata['lble'],bins=[0.5])
tdata.head(5)
```

[8]:
```
                                              tweet  lbel  label
0  Families to sue over Legionnaires: More than 4…   0.1      1
1  Pandemonium In Aba As Woman Delivers Baby With…   0.4      2
2  My emotions are a train wreck. My body is a tr…   0.2      1
3  Alton brown just did a livestream and he burne…   0.5      3
4  @TinyJecht Are you another Stand-user? If you …   0.5      3
```

[9]:
```
ptext = tdata.tweet
ydata = tdata.label
ptext.shape
```

```
[9]: (1863,)
```

```
[10]: print(ptext[0])
```

```
Families to sue over Legionnaires: More than 40 families affected by the fatal
outbreak of Legionnaires' disea… http://t.co/ZA4AXFJSVB
```

Text preprocessing

```
[11]: import re
      import string
      # temp = []
      snow = nltk.stem.SnowballStemmer('english')
      def preprocess(data):
          temp = []
          for sentence in data:
              sentence = sentence.lower()      #converting lowercase

              #Removal of HTTP links/URLs mixed up in any text:
              sentence = re.sub('http://\S+|https://\S+', '', sentence)
              #OR
              #sentence = re.sub('http[s]?://\S+', '', sentence)
              #sentence = re.sub(r'https?:\/\/.*[\r\n]*','',sentence, flags=re.
       ↪MULTILINE) #remove the URLs

              #punctuations:'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
              sentence = "".join([char for char in sentence if char not in string.
       ↪punctuation])
              sentence = re.sub(r'\d+', '', sentence)     # removing numbers
              sentence = sentence.strip()         # removing extra white spaces
              tokens = nltk.word_tokenize(sentence)[2 :]
          #     tokens = nltk.word_tokenize(sentence)
              sentence = [l.lower() for l in tokens]
              filtered_result = list(filter(lambda l: l not in stop_words, sentence))
              lemmas = [wordnet_lemmatizer.lemmatize(t) for t in filtered_result]

              #(or)
          #     sentence = re.sub(r'[?|!|\'|"|#!$%&()*+-@]',r'',sentence)
          #     sentence = re.sub(r'[.|,|)|(|\|/]',r' ',sentence)

              temp.append(lemmas)
          return temp
```

```
[12]: # Xdata_cleaned = preprocess(ptext)
      # Xtest_cleaned = preprocess(Xtest)
      ptext_cleaned = preprocess(ptext)
```

```
[13]: # print(Xdata_cleaned[0])
      ptext_cleaned[:2]
```

```
[13]: [['sue',
        'legionnaire',
        'family',
        'affected',
        'fatal',
        'outbreak',
        'legionnaire',
        'disea'],
       ['aba', 'woman', 'delivers', 'baby', 'without', 'face', 'photo']]
```

```
[14]: # Normalized or cleaning dataset adding into original datafile for chcking
      tdata['Normalized_tweet'] = ptext_cleaned
      tdata.shape
```

```
[14]: (1863, 4)
```

Original text and cleaned/normalized text

```
[15]: tdata[['tweet','Normalized_tweet']].head(5)
```

```
[15]:                                                tweet  \
      0  Families to sue over Legionnaires: More than 4…
      1  Pandemonium In Aba As Woman Delivers Baby With…
      2  My emotions are a train wreck. My body is a tr…
      3  Alton brown just did a livestream and he burne…
      4  @TinyJecht Are you another Stand-user? If you …

                                        Normalized_tweet
      0  [sue, legionnaire, family, affected, fatal, ou…
      1  [aba, woman, delivers, baby, without, face, ph…
      2      [train, wreck, body, train, wreck, im, wreck]
      3  [livestream, burned, butter, touched, hot, pla…
      4      [another, standuser, detonate, killer, queen]
```

Bag of Words

```
[16]: from nltk import word_tokenize
      word2count = {}
      for data in ptext_cleaned:
          words = data
          for word in words:
              if word not in word2count.keys():
                  word2count[word] = 1
              else:
                  word2count[word] += 1
```

```
word2count
```

```
[16]: {'sue': 5,
       'legionnaire': 12,
       'family': 24,
       'affected': 7,
       'fatal': 11,
       'outbreak': 6,
       'disea': 5,
       'aba': 3,
       'woman': 11,
       'delivers': 2,
       'baby': 7,
       'without': 15,
       'face': 14,
       'photo': 11,
       'train': 26,
       'wreck': 14,
       'body': 34,
       'im': 47,
       'livestream': 1,
       'burned': 9,
       'butter': 4,
       'touched': 1,
       'hot': 9,
       'plate': 2,
       'soon': 2,
       'made': 10,
       'nut': 1,
       'joke': 3,
       'another': 12,
       'standuser': 2,
       'detonate': 9,
       'killer': 3,
       'queen': 3,
       'timed': 1,
       'concert': 2,
       'screamed': 11,
       'minute': 17,
       'straight': 2,
       'florida': 3,
       'forest': 18,
       'service': 19,
       'firefighter': 2,
       'could': 17,
       'deployed': 1,
       'california': 24,
       'help': 13,
       'contain': 2,
       'fire': 73,
       'detail': 1,
       'atomic': 9,
       'bomb': 22,
       'japan': 8,
       'still': 32,
       'struggle': 3,
       'war': 20,
       'past': 6,
       'anniversary': 6,
       'devastation': 9,
       'wrought': 4,
       'b': 6,
       'killed': 20,
       'civilian': 3,
       'ground': 4,
       'jet': 3,
       'first': 24,
       'bombed': 10,
       'city': 13,
       'main': 3,
       'street': 8,
       'dramatically': 1,
       'plummeted': 1,
       'burn': 4,
       'whole': 11,
       'gotham': 1,
       'bcs': 1,
       'gon': 10,
       'na': 16,
       'laugh': 2,
       'everyone': 15,
       'theyre': 8,
       'panicking': 7,
       'phone': 8,
       'worstoverdose': 1,
       'scream': 9,
       'jaileens': 1,
       'caked': 1,
       'look': 29,
       'email': 2,
       'people': 44,
       'use': 7,
       'report': 13,
       'sthing': 1,
       'think': 18,
       'hazarddangerous': 1,
       'old': 16,
       'lady': 7,
       'went': 15,
       'something': 9,
       'ask': 1,
       'alone': 4,
       'coming': 5,
       'tsunami': 9,
       'showing': 1,
       'storm': 16,
       'ne': 2,
       'edmond': 1,
       'severe': 8,
       'hail': 11,
       'mph': 14,
       'wind': 16,
       'possible': 8,
       'okwx': 3,
       'attack': 22,
       'gun': 8,
       'grabber': 2,
       'demand': 2,
       'life': 39,
       'tragedy': 6,
       'monicas': 1,
       'hair': 3,
       'given': 2,
       'season': 4,
       'friend': 10,
       'case': 7,
       'drowning': 4,
       'sweatfyi': 1,
       'mile': 4,
       'east': 4,
       'pickens': 1,
       'moving': 8,
       'pea': 1,
       'size': 1,
       'gust': 4,
       'scwx': 1,
       'dude': 5,
       'obliterated': 5,
       'meek': 2,
       'mill': 2,
       'dont': 31,
       'judge': 1,
       'book': 11,
       'cover': 5,
       'explode': 7,
       'sewage': 1,
       'open': 7,
       'housing': 2,
       'estate': 1,
       'irish': 4,
       'independent': 2,
       'aug': 6,
       'team': 9,
       'usagi': 1,
       'even': 26,
       'blew': 9,
       'entire': 8,
       'solar': 2,
       'system': 4,
       'airhead': 1,
       'misstep': 1,
       'good': 24,
       'chunk': 1,
       'running': 4,
       'time': 37,
       'trapped': 8,
       'effective': 3,
       'claustrophobic': 1,
       'thriller': 3,
       'try': 9,
       'always': 17,
       'end': 11,
       'sinking': 14,
       'way': 28,
       'feel': 20,
       'drank': 1,
       'vodka': 1,
       'ice': 2,
       'would': 30,
       'bag': 26,
       'may': 24,
       'logically': 1,
       'right': 19,
       'call': 6,
       'maybe': 4,
       'act': 8,
       'mass': 17,
       'murder': 8,
       'cant': 15,
       'sanction': 2,
       'delivered': 1,
       'big': 13,
       'screen': 10,
       'safe': 3,
       'sound': 12,
       'like': 87,
       'yearold': 5,
       'superstar': 1,
       'girl': 16,
       'travel': 4,
       'fame': 1,
       'freeway': 2,
       'weakening': 1,
       'move': 4,
       'se': 4,
       'towards': 2,
       'lubbock': 1,
       'area': 12,
       'outflow': 1,
       'boundary': 2,
       'create': 2,
       'dust': 8,
       'quarantine': 7,
       'offensive\x89û': 1,
       'onlinecommunities': 1,
       'reddit': 1,
       'amageddon': 1,
       'freespeech': 1,
       'day': 48,
       'disaster': 26,
       'emotion': 2,
       'happening': 4,
       'lie': 5,
       'drift': 1,
       'away': 10,
       'storming': 1,
       'around': 10,
       'little': 9,
       'bebacksoon': 1,
       'four': 4,
       'year': 32,
       'making': 10,
       'home': 25,
       'razed': 7,
       'northern': 14,
       'wildfire': 22,
       'abc': 7,
       'news': 27,
       'second': 4,
       'perfect': 4,
       'heart': 17,
       'place': 11,
       'innocent': 3,
       'wellmeaning': 1,
       'wrecked': 7,
       'car': 21,
       'costing': 1,
       'apiece': 1,
       'purchased': 1,
       'film': 36,
       'anybody': 2,
       'else': 9,
       'problem': 3,
       'circle': 2,
       'epicentre': 2,
       'ride': 4,
       'quiet': 2,
       'cadence': 1,
       'pure': 2,
       'finesse': 1,
       'far': 7,
       'shortage': 1,
       'dilutes': 1,
       'potency': 1,
       'otherwise': 2,
       'respectable': 1,
       'action': 8,
       'bride': 3,
       'halfhour': 1,
       'long': 12,
       'come': 14,
       'replete': 2,
       'flattering': 1,
       'sense': 5,
       'mystery': 4,
       'quietness': 1,
       'birthday': 2,
       'bruh': 2,
       'windstorm': 9,
       'sheer': 1,
       'recovery': 2,
       'update': 8,
       'leelanau': 1,
       'amp': 58,
       'grand': 1,
       'traverse': 2,
       'state': 13,
       'emergency': 22,
       'extended': 1,
       'besafe': 1,
       'imax': 2,
       'strap': 1,
       'pair': 2,
       'goggles': 1,
       'shut': 1,
       'real': 9,
       'world': 31,
       'take': 28,
       'vicarious': 1,
       'voyage': 1,
       'last': 20,
       'frontier': 1,
       'space': 4,
       'airport': 6,
       'get': 61,
       'swallowed': 6,
       'sandstorm': 9,
       'though': 7,
       'refreshingly': 2,
       'novel': 4,
       'don\x89ûªt': 1,
       'want': 17,
       'mention': 3,
       'let\x89ûªs': 1,
       'anything': 5,
       'lead': 4,
       'best': 27,
       'american': 9,
       'movie': 31,
       'troubled': 1,
       'teen': 2,
       'since': 14,
       'whatever': 1,
       'win': 8,
       'kerry': 1,
       'obliteration': 7,
       'profit': 2,
       'rise': 7,
       'billion': 1,
       'worst': 4,
       'natural': 11,
       'claim': 7,
       'language': 4,
       'story': 25,
       'lively': 1,
       'script': 2,
       'sharp': 1,
       'acting': 3,
       'partially': 1,
       'animated': 1,
       'interlude': 1,
       'make': 34,
       'kiss': 2,
       'seem': 1,
       'minty': 1,
       'fresh': 3,
       'setting': 3,
       'flame': 11,
       'upon': 7,
       'sunk': 10,
       'saw': 6,
       'johnny': 1,
       'marr': 1,
       'primal': 1,
       'hour': 9,
       'sunday': 1,
       'fully': 4,
       'aware': 3,
       'battle': 7,
       'support': 6,
       'fight': 5,
       'today': 14,
       'allegation': 1,
       'timeline': 2,
       'damn': 3,
       'fast': 4,
       'nw': 4,
       'wth': 1,
       'rotating': 1,
       'w': 11,
       'huge': 7,
       'massive': 2,
       'violent': 9,
       'tornado': 3,
       'great': 19,
       'role': 7,
       'never': 14,
       'hog': 1,
       'scene': 4,
       'fellow': 2,
       'cast': 4,
       'plenty': 1,
       'line': 7,
       'comedy': 10,
       'fell': 2,
       'rock': 2,
       'scraped': 1,
       'butt': 1,
       'nearly': 4,
       'drowned': 6,
       'summerk': 1,
       'record': 2,
       'hurricane': 8,
       'drought': 12,
       'blazing': 5,
       'weatherstay': 1,
       'std': 1,
       'yet': 12,
       'rejected': 1,
       'slogan': 1,
       'notification': 1,
       'thats': 8,
       'bad': 15,
       'haha': 5,
       'wouldve': 1,
       'tho': 2,
       'mudslide': 10,
       'aw': 1,
       'country': 10,
       'latin': 1,
       'america': 5,
       'next': 8,
       'argentina': 1,
       'one': 49,
       'week': 11,
       'ago': 3,
       'reported': 4,
       'economic': 2,
       'frequent': 1,
       'thunder': 6,
       'gusty': 1,
       'part': 14,
       'uptown': 1,
       'midtown': 1,
       'cn': 1,
       'paramedic': 2,
       'really': 18,
       'leave': 5,
       'someone': 7,
       'inside': 9,
       'building': 29,
       'collapseblow': 1,
       'halloikbenwill': 1,
       'motivator': 1,
       'hoot': 1,
       'half': 12,
       'see': 25,
       'candidate': 1,
       'giving': 6,
       'cent': 1,
       'stump': 1,
       'speech': 1,
       'destined': 1,
       'st': 9,
       'century': 2,
       'new': 32,
       'conan': 1,
       'going': 26,
       'splash': 1,
       'greater': 2,
       'arnold': 1,
       'schwarzenegger': 1,
       'jeanclaud': 1,
       'van': 3,
       'damme': 1,
       'steven': 2,
       'segal': 1,
       'moron': 1,
       'flag': 7,
       'man': 16,
       'brainless': 1,
       'muscle': 2,
       'dobut': 1,
       'youre': 10,
       'murderer': 9,
       'village': 8,
       'pugwash': 1,
       'every': 13,
       'truck': 9,
       'town': 4,
       'house': 21,
       'population': 6,
       'survive': 9,
       'tonight': 8,
       'wouldnt': 1,
       'change': 12,
       'thing': 12,
       'starve': 1,
       'death': 16,
       'wild': 7,
       'morel': 1,
       'ambleside': 1,
       'farmr': 1,
       'martsunmushroom': 1,
       'foragesecret': 1,
       'know': 29,
       'tree': 6,
       'grow': 4,
       'climbed': 1,
       'wheelsio': 1,
       'hawkhis': 1,
       'knee': 2,
       'injury': 21,
       'wheres': 1,
       'beltmr': 1,
       'srk': 1,
       'cook': 1,
       'ur': 5,
       'beautiful': 7,
       'as': 13,
       'punishment': 2,
       'bombedout': 1,
       'britain': 2,
       'art': 7,
       'undercover': 2,
       'brother': 3,
       'run': 11,
       'steam': 2,
       'find': 7,
       'surprise': 4,
       'amuse': 1,
       'earring': 3,
       'beach': 7,
       'jewelry': 1,
       'vacation': 1,
       'keep': 10,
       'calm': 2,
       'flattened': 7,
       'cold': 3,
       'nuke': 2,
       'ban': 5,
       'ocean': 1,
       'superiority': 1,
       'unconditional': 1,
       'surrender': 1,
       'putin': 1,
       'game': 9,
       'set': 11,
       'match': 3,
       'release': 4,
       'hostage': 15,
       'wearing': 2,
       'dead': 24,
       'black': 15,
       'flaming': 2,
       'red': 6,
       'stark': 2,
       'white': 6,
       'much': 26,
       'esp': 1,
       'debate': 1,
       'go': 34,
       'blue': 6,
       'gold': 5,
       'brown': 5,
       'shoe': 3,
       'result': 4,
       'hahahah': 1,
       'worrying': 1,
       'performance': 6,
       'incredible': 3,
       'effected': 1,
       'cali': 1,
       'effectively': 1,
       'teach': 1,
       'kid': 6,
       'danger': 6,
       'drug': 4,
       'project': 5,
       'lrb': 3,
       'unfortunately': 1,
       'rrated': 1,
       'rrb': 7,
       'paid': 1,
       'encaustic': 1,
       'cerography': 1,
       'portion': 2,
       'till': 3,
       'give': 13,
       'voice': 2,
       'deluge': 8,
       'byityf': 1,
       'hope': 8,
       'ok': 8,
       'warning': 15,
       'dry': 2,
       'thunderstorm': 11,
       'bay': 2,
       'weather': 9,
       'cawx': 1,
       'nwsbayarea': 1,
       'enough': 10,
       'held': 5,
       'together': 7,
       'skilled': 1,
       'ensemble': 1,
       'actor': 3,
       'drama': 6,
       'cube': 1,
       'stepped': 2,
       'broken': 3,
       'glass': 5,
       'pun': 2,
       'tak': 1,
       'sedar': 1,
       'pain': 3,
       'also': 9,
       'bleeding': 9,
       'shit': 14,
       'progress': 2,
       'shot': 11,
       'syncopated': 1,
       'style': 5,
       'mimicking': 1,
       'work': 30,
       'subject': 3,
       'pray': 3,
       'turn': 8,
       'idea': 11,
       'documentary': 6,
       'head': 10,
       'rousing': 1,
       'invigorating': 1,
       'fun': 11,
       'lacking': 1,
       'mtv': 1,
       'puffery': 1,
       'say': 24,
       'silas': 1,
       'sliced': 1,
       'headlinelike': 1,
       'chopped': 1,
       'piece': 5,
       'cabbage': 1,
       'gh': 1,
       'twister': 7,
       'found': 7,
       'reunion': 6,
       'island': 6,
       'flight': 3,
       'mh': 13,
       'behind': 3,
       'plane': 9,
       'disappearance': 1,
       'better': 9,
       'mixed': 2,
       'disapproval': 1,
       'justine': 1,
       'combined': 2,
       'tinge': 1,
       'understanding': 3,
       'power': 11,
       'jedi': 1,
       'collection': 1,
       'droid': 1,
       'hasbro': 1,
       'full': 24,
       'read': 12,
       'ebay': 8,
       'effect': 13,
       'hiroshima': 15,
       'nagasaki': 3,
       'bombing': 13,
       'felt': 2,
       'fan': 11,
       'angry': 3,
       'odeon': 2,
       'cinema': 8,
       'evacuated': 10,
       'following': 4,
       'false': 5,
       'alarm': 4,
       'doesnt': 8,
       'replace': 5,
       'eyewitness': 6,
       'video': 23,
       'ferguson': 2,
       'wounded': 9,
       'suspect': 7,
       'exchanging': 3,
       'richmond': 3,
       'police': 18,
       'officer': 6,
       'exchange': 1,
       'gunfire': 1,
       'incarnation': 1,
       'fizz': 1,
       'infectious': 2,
       'stop': 10,
       'terrorism': 7,
       'inviting': 1,
       'arsonist': 3,
       'join': 2,
       'brigade': 1,
       'telegraph': 1,
       'happen': 2,
       'anywhere': 2,
       'school': 14,
       'etc': 2,
       'learn': 7,
       'trauma': 6,
       'parent': 3,
       '\x89û': 4,
       'u': 53,
       'trip': 4,
       'n': 7,
       'fall': 13,
       'cliff': 6,
       'tweet': 5,
       'loses': 3,
       'bite': 1,
       'lastminute': 1,
       'happy': 6,
       'ending': 3,
       'le': 3,
       'plausible': 1,
       'rest': 4,
       'picture': 11,
       'owner': 2,
       'charged': 12,
       'swear': 1,
       'secret': 7,
       'well': 15,
       'uncover': 1,
       'god': 9,
       'slumber': 1,
       'there': 5,
       'blight': 4,
       'trying': 4,
       'racist': 1,
       'elitist': 1,
       'almost': 12,
       'spooky': 1,
       'sulky': 1,
       'calculating': 1,
       'lolita': 1,
       'catch': 4,
       'finally': 7,
       'monwabisi': 1,
       'lol': 16,
       'hlongwane': 1,
       'ryt': 1,
       'twin': 2,
       'r': 11,
       'destroy': 11,
       'ashestoashes': 1,
       'dad': 2,
       'survived': 8,
       'driving': 5,
       'missing': 4,
       'migrant': 5,
       'med': 1,
       'rescuer': 8,
       'search': 4,
       'survivor': 9,
       'boat': 7,
       'carrying': 2,
       'many': 25,
       'migrants\x89û': 1,
       'tom': 1,
       'clancy': 1,
       'military': 7,
       'paperback': 1,
       'tomclancy': 1,
       'anyone': 4,
       'answer': 2,
       'need': 18,
       'contact': 3,
       'flooding': 6,
       'vietnam': 2,
       'situation': 2,
       'night': 7,
       'panic': 11,
       'gem': 6,
       'obsession': 1,
       'hijacking': 5,
       'computer': 8,
       'send': 8,
       'data': 7,
       'wave': 9,
       'hat': 6,
       'prebreak': 10,
       'health': 3,
       'care': 6,
       'review': 5,
       'investigative': 1,
       'journalism': 1,
       'sick': 3,
       'injured': 11,
       'patient': 2,
       'local': 2,
       'er': 3,
       'entertainment': 3,
       'derives': 1,
       'sticking': 1,
       'fact': 4,
       'live': 13,
       'noaa': 1,
       'tracking': 1,
       'looping': 1,
       'wedaugth': 1,
       'electrocuted': 6,
       'morning': 5,
       'becomes': 4,
       'race': 3,
       'education': 2,
       'catastrophe': 9,
       'inundated': 6,
       'soggy': 1,
       'bottom': 2,
       'lashing': 1,
       'moist': 1,
       'rioting': 8,
       'valley': 2,
       'penn': 1,
       'storyline': 1,
       'interesting': 4,
       'entertaining': 4,
       'nt': 15,
       'magical': 1,
       'quality': 4,
       'beginning': 3,
       'later': 4,
       'crime': 6,
       'put': 9,
       'moscow': 1,
       'director': 2,
       'bourne': 1,
       'directs': 1,
       'traffic': 6,
       'nice': 3,
       'wintry': 1,
       'location': 2,
       'absorbs': 1,
       'spycraft': 1,
       'us': 2,
       'damon': 1,
       'ability': 2,
       'focused': 1,
       'sincere': 1,
       'surface': 2,
       'loversontherun': 1,
       'flick': 5,
       'lot': 10,
       'common': 2,
       'piesiewicz': 1,
       'kieslowski': 1,
       'earlier': 4,
       'double': 5,
       'veronique': 1,
       'soudelors': 2,
       'predicted': 3,
       'path': 3,
       'approach': 2,
       'taiwan': 7,
       'expected': 5,
       'landfall': 2,
       'southern': 3,
       'china': 4,
       's\x89û': 2,
       'north': 4,
       'japton': 2,
       'large': 6,
       'po': 4,
       'arwx': 2,
       'respond': 2,
       'chemical': 8,
       'spill': 14,
       'downtown': 5,
       'beaumont': 1,
       'benews': 1,
       'ready': 6,
       'client': 3,
       'outage': 2,
       'vet': 3,
       'design': 4,
       'western': 3,
       'food': 3,
       'en': 1,
       'masse': 1,
       'causing': 4,
       'public': 3,
       'backlash': 1,
       'san': 2,
       'antonio': 1,
       'star': 4,
       'coach': 3,
       'dan': 2,
       'hughes': 1,
       'sideline': 4,
       'chair': 2,
       'onto': 5,
       'floor': 2,
       'stretcher': 4,
       'except': 4,
       'actually': 5,
       'colorado': 4,
       'tomorrow': 10,
       'dreaming': 1,
       'capture': 7,
       'complexity': 1,
       'trial': 3,
       'tribulation': 1,
       'gone': 4,
       'warped': 1,
       'tony': 1,
       'played': 6,
       'issue': 6,
       'showed': 3,
       'sleeping': 1,
       'siren': 6,
       'attila': 1,
       'hasnt': 2,
       'coat': 1,
       'hand': 5,
       'id': 8,
       'worn': 1,
       'certainty': 1,
       'armageddon': 7,
       'bear': 2,
       'occasion': 2,
       'electrocute': 3,
       'thanks': 5,
       'normal': 3,
       'sit': 4,
       'front': 6,
       'uber': 1,
       'driver': 4,
       'original': 5,
       'sensei': 1,
       'write': 1,
       'rhyme': 1,
       'attic': 1,
       'reviewing': 1,
       'policy': 8,
       'leaving': 3,
       'hundred': 4,
       'commuter': 1,
       'stranded': 2,
       'hail\x89û': 1,
       'patio': 1,
       'table': 1,
       'umbrella': 1,
       'flipped': 1,
       'foul': 2,
       'play': 8,
       'instead': 6,
       'suspense': 2,
       'writer': 2,
       'divided': 2,
       'headed': 2,
       'destruction': 6,
       'stand': 5,
       'matthew': 2,
       'anthropologically': 1,
       'detailed': 1,
       'realization': 1,
       'early': 5,
       'suburbia': 1,
       'significant': 1,
       'overstated': 1,
       'playful': 3,
       'constantly': 1,
       'frustrates': 1,
       'desire': 5,
       'truth': 1,
       'deconstructing': 1,
       'format': 1,
       'biography': 1,
       'manner': 1,
       'derrida': 3,
       'doubtless': 1,
       'blessing': 2,
       'close': 5,
       'floyd': 1,
       'mayweathers': 1,
       'money': 8,
       'bloody': 10,
       'elbow': 1,
       'boxing': 1,
       'kill': 11,
       'general': 4,
       'highestranking': 1,
       'fatality': 10,
       'wicked': 2,
       'wont': 8,
       'fbi': 1,
       'stole': 1,
       'married': 2,
       'honduran': 1,
       'minor': 2,
       'sex': 2,
       'latifah': 1,
       'offer': 2,
       'seemed': 3,
       'flaunting': 1,
       'gift': 1,
       'sexy': 1,
       'happiest': 1,
       'deal': 7,
       'surprising': 1,
       'infuriating': 1,
       'flaw': 1,
       'least': 8,
       'amy': 2,
       'selfabsorbed': 2,
       'personality': 1,
       'honesty': 4,
       'taking': 6,
       'extra': 1,
       'security': 6,
       'harrybecareful': 2,
       …}
```

[17]:

```
print("Word count :" ,colored(len(word2count),'blue'))
```

```
Word count : 5830
```

[18]:

```
# selecting best 100 features only
import heapq
ptext_words = heapq.nlargest(1200, word2count, key=word2count.get)
```

Vectorization

*Arbitarary words (units/tokens) into fixed length of vectors

[19]:

```
# Converting sentences into vectors: Xtrain data
x_final = []
for data in ptext_cleaned:
    vec = []
    for word in ptext_words:
        if word in data:
            vec.append(1)
        else:
            vec.append(0)
    x_final.append(vec)

X_final = np.array(x_final)
print(X_final)
```

```
[[0 0 0 … 0 0 0]
 [0 0 0 … 0 0 0]
 [0 0 0 … 0 0 0]
 …
 [0 0 0 … 0 0 0]
 [0 0 0 … 0 0 0]
 [0 0 0 … 0 0 0]]
```

[ ]:

```

```