PROJECT REPORT

on

“News Text Classification”

(CSE CC VI Semester Mini Project)

June 2022

Submitted to:

Dr. Surender Singh Samant

Assistant Professor

Submitted by:

Girija Rathore

Roll. No.: 2015016

CSE-CC-CCIS-VI-Sem

Graphic Era

Deemed to be University

1. ABOUT PROJECT

1.1. Introduction

Every news website categorizes news items before releasing them so that readers may quickly click on the sort of news that interests them every time, they visit their website. Popular categories on practically every news website include technology, entertainment, and sports.

1.2. Problem Statement

As the number of English news items grows, so do the need of new technologies for organizing textual material. To properly engage with raw text, these technologies should pre-process, analyze, and classify it.

The problem statement highlights the need of automated text classification, specifically news in this case, that too on the back-end server. Since this is in trend for quite some time in computing arena, some Python libraries support solutions for this, via the means of Machine Learning & NLP.

1.3. Getting Started with Machine Learning

Machine Learning (ML) is a subset of Artificial Intelligence (AI) that allows computers to automatically learn and improve from experience without being explicitly designed. Machine learning is not a new concept. In truth, the notion dates back more than 60 years, when it was said that " To be called intelligent, a machine must produce responses that are indistinguishable from those of a human.".

ML is made up of algorithms that educate computers to accomplish things that humans do intuitively on a regular basis. The initial attempts at artificial intelligence were establishing a rule/set of rules to train a computer.

Depending on the context of the problem, they can be classified into three major categories: Unsupervised Learning, Supervised Learning & Reinforcement Learning, each differing in their applicable algorithms and tools.

1.4. What NLP really is?

In the Netizen era, information circulates through numerous social media platforms and e-newspapers in many languages. With the help of natural language processing, it is now feasible to gather this unstructured data and analyze its many meanings.

Natural Language Processing is the branch of Artificial Intelligence concerned with the interaction of machines and human languages. To put it another way, NLP assists machines in deriving the meaning of human (natural) languages.

1.5. Need for the System

Every second, fresh information becomes available to the public: reports, books, articles, and news are published in several languages. The content administrators of news websites now classify news pieces by themselves, manually. To save time, they may also deploy a machine learning model on their websites that reads the news title or content and categorizes the news. Automatic classification will allow it to be processed and used more effectively for decision-making.

2. RESOURCES USED

2.1. System Requirements

- Central Processing Unit (CPU) – Intel Core i5 6th Generation Processor or higher
- RAM – 8GB Minimum
- Operating System – Ubuntu or Microsoft Windows 10
- Executable Google Colaboratory Notebook

2.2. Computer Languages

- Python versions 2.7 or 3.4+

2.3. Libraries

- RE

The re module includes a set of powerful regular expression facilities that allow you to quickly determine if a given text matches or contains a particular pattern (through the match function or by using the search function). A regular expression is a string pattern defined in a short (and rather obscure) syntax.

- CSV

The most common import and export format for spreadsheets and databases is CSV (Comma Separated Values). The csv module includes classes for reading and writing CSV tabular data. Programmers can also specify the CSV formats that other apps understand or create their own special-purpose CSV formats.

- OPERATOR

The operator module in Python offers a "functional" interface to the standard operators. When processing data with functions like map and filter, the functions in this module can be used instead of various lambda constructions.

- NLTK

NLTK is a standard Python package that includes a variety of NLP methods. It is one of the most used NLP and Computational Linguistics libraries.

2.4. Algorithms

- TOKENIZATION

Tokenization is the process of dividing raw text into tiny parts. Tokenization divides raw text into words and phrases known as tokens. These tokens aid in understanding the context or constructing the NLP model. Tokenization aids in determining the meaning of the text by evaluating the word sequence.

- STEMMING

Stemming is a word normalization approach used in Natural Language Processing. It is a method that converts a collection of words in a phrase into a sequence in order to decrease the lookup time. This approach normalizes words that have the same meaning but differ slightly depending on the context or phrase.

The stemming algorithm operates by removing the suffix from the word. In a larger sense, it removes either the beginning or the end of a word.

- LEMMATIZATION

In NLTK, lemmatization is the computational process of determining a word's lemma based on its meaning and context. Lemmatization is commonly used to describe the morphological study of words with the goal of removing inflectional ends. It aids in retrieving the lemma, or basic or dictionary form, of a word.

Lemmatization is a more powerful process that takes morphological analysis of the words into account. It returns the basic form of all its inflectional forms, the lemma.

- BAG OF WORDS

A bag of words is a text representation that defines the appearance of words in a document. We only keep track of word counts, ignoring grammatical intricacies and word order. Because all information about the sequence or structure of words in the text is deleted, it is referred to as a "bag" of words.

The model is solely concerned with whether recognized terms appear in the document, not with where they appear.

2.5. Description of Dataset

The dataset primarily consists of ‘n’ categories, depending upon the range or articles and programmer. Here, we are working with three easily distinguishable categories, namely Political, Sports & Entertainment with 100 words each. The dataset can be refined to produce more efficient and accurate results.

3. MODULES AT A GLANCE

3.1. Working Components

The entire project can simply be broken down into following modules:

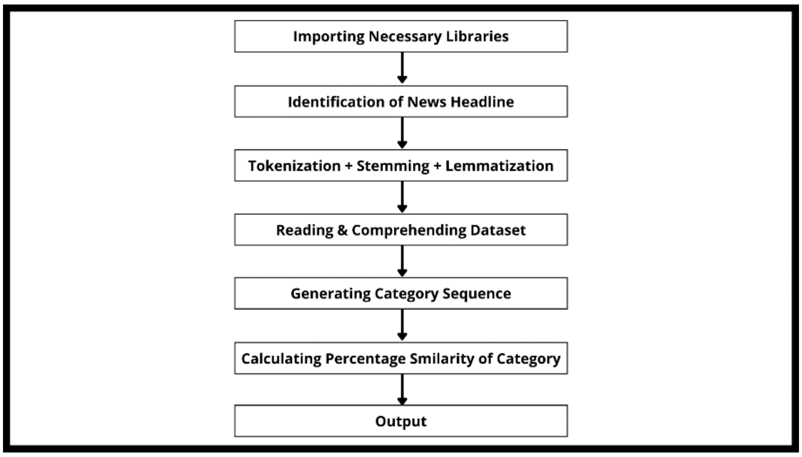
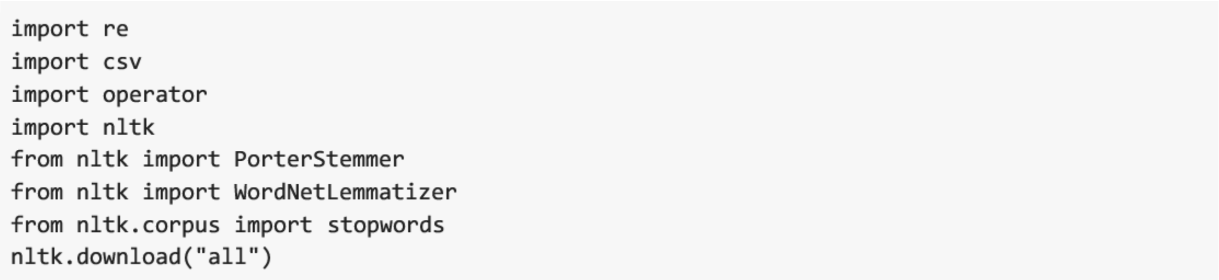

Importing Necessary Libraries
Identification of News Headline
Tokenization + Stemming + Lemmatization
Reading & Comprehending Dataset
Generating Category Sequence
Calculating Percentage Smilarity of Category
Output

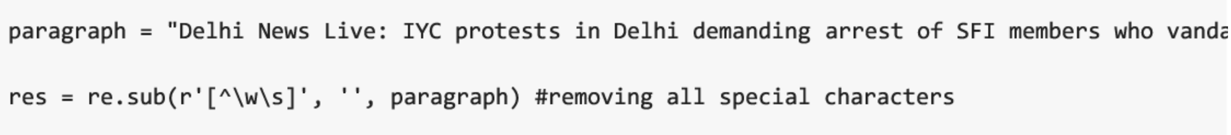
Fig. Working Components of Project

- Importing Necessary Libraries



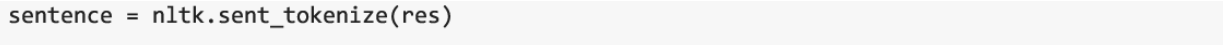
```
import re
import csv
import operator
import nltk
from nltk import PorterStemmer
from nltk import WordNetLemmatizer
from nltk.corpus import stopwords
nltk.download("all")
```

- Identification of News Headline

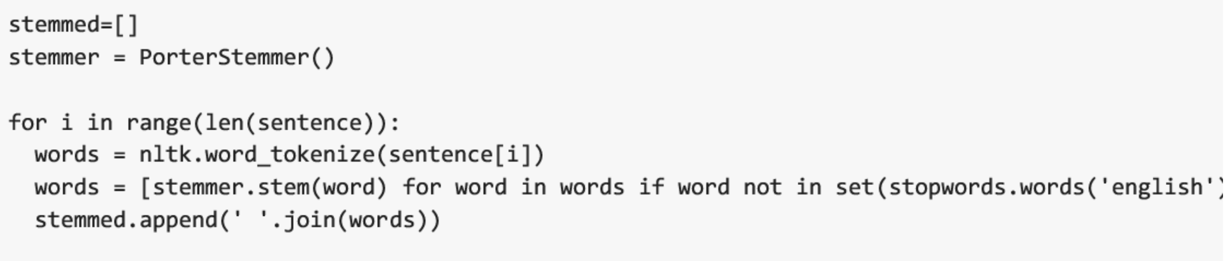


```
paragraph = "Delhi News Live: IYC protests in Delhi demanding arrest of SFI members who vanda

res = re.sub(r'[^\w\s]', '', paragraph) #removing all special characters
```

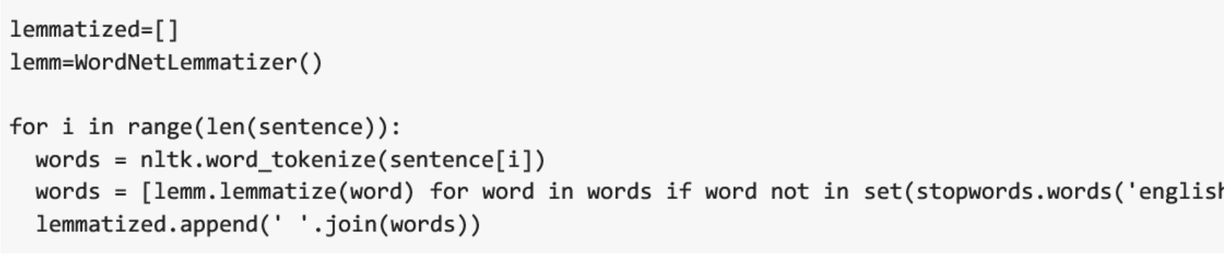
- Tokenization, Stemming & Lemmatization



```
sentence = nltk.sent_tokenize(res)
```

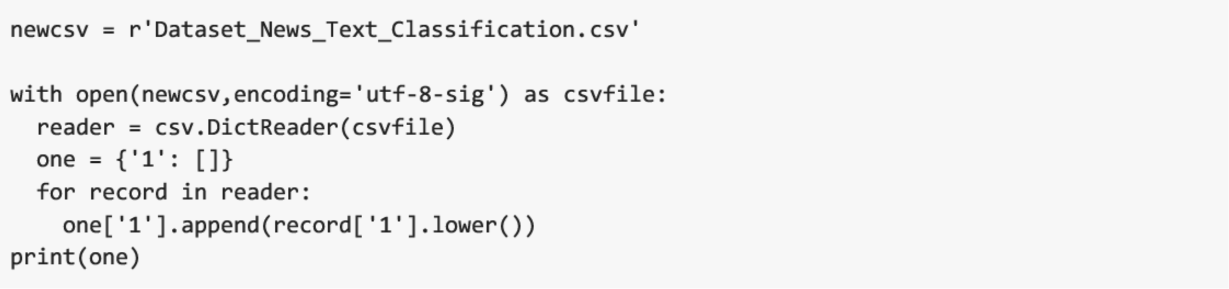


```
stemmed=[]
stemmer = PorterStemmer()

for i in range(len(sentence)):
  words = nltk.word_tokenize(sentence[i])
  words = [stemmer.stem(word) for word in words if word not in set(stopwords.words('english')
  stemmed.append(' '.join(words))
```



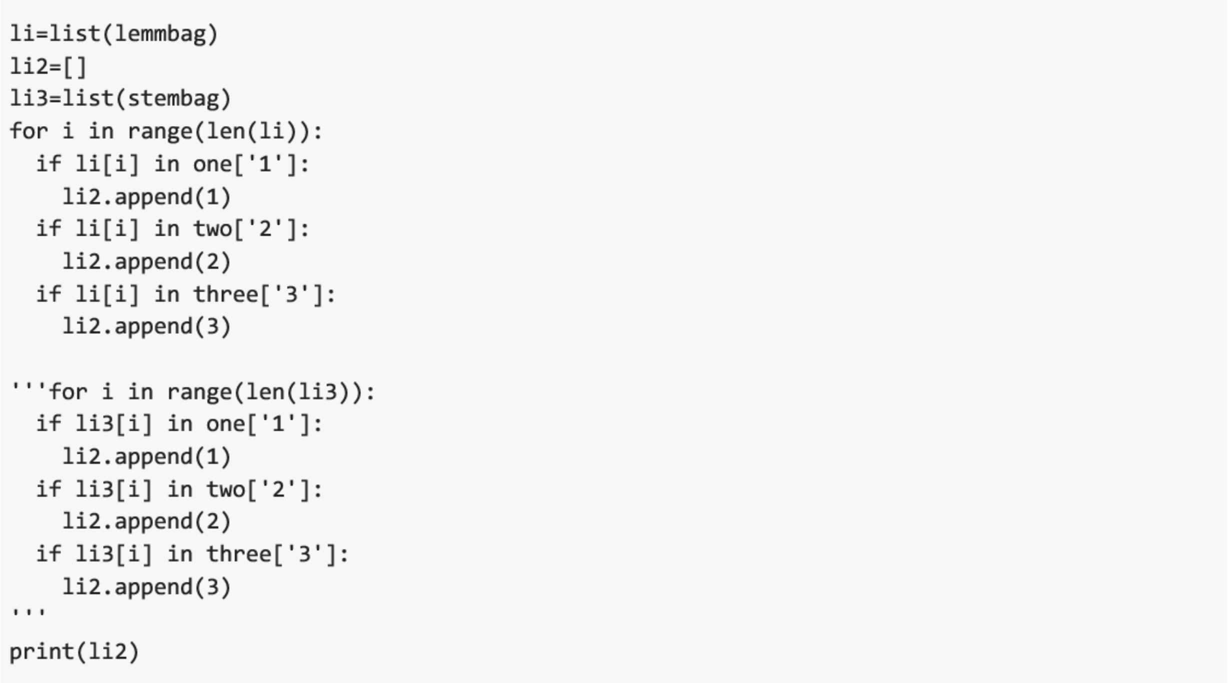
```
lemmatized=[]
lemm=WordNetLemmatizer()

for i in range(len(sentence)):
  words = nltk.word_tokenize(sentence[i])
  words = [lemm.lemmatize(word) for word in words if word not in set(stopwords.words('englis
  lemmatized.append(' '.join(words))
```

- Reading & Comprehending Dataset



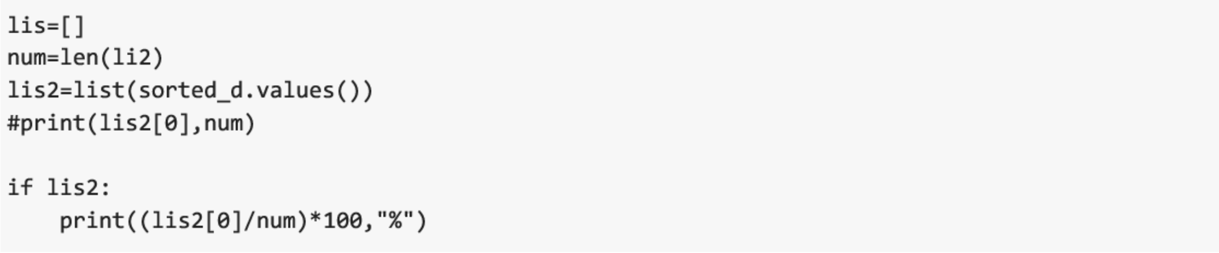
```
newcsv = r'Dataset_News_Text_Classification.csv'

with open(newcsv,encoding='utf-8-sig') as csvfile:
  reader = csv.DictReader(csvfile)
  one = {'1': []}
  for record in reader:
    one['1'].append(record['1'].lower())
print(one)
```

- Generating Category Sequence



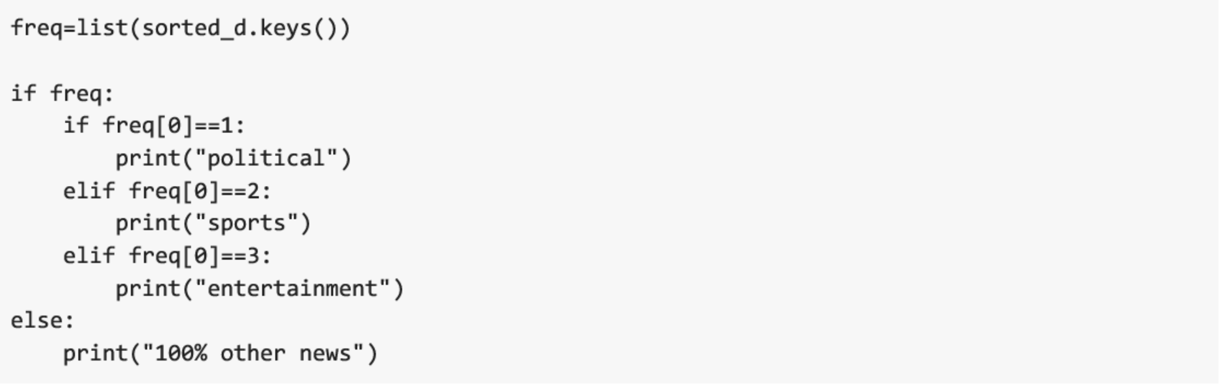
```
li=list(lemmbag)
li2=[]
li3=list(stembag)
for i in range(len(li)):
  if li[i] in one['1']:
    li2.append(1)
  if li[i] in two['2']:
    li2.append(2)
  if li[i] in three['3']:
    li2.append(3)

'''for i in range(len(li3)):
  if li3[i] in one['1']:
    li2.append(1)
  if li3[i] in two['2']:
    li2.append(2)
  if li3[i] in three['3']:
    li2.append(3)
'''
print(li2)
```

- Calculating Percentage Similarity of Category



```
lis=[]
num=len(li2)
lis2=list(sorted_d.values())
#print(lis2[0],num)

if lis2:
    print((lis2[0]/num)*100,"%")
```

- Output



```
freq=list(sorted_d.keys())

if freq:
    if freq[0]==1:
        print("political")
    elif freq[0]==2:
        print("sports")
    elif freq[0]==3:
        print("entertainment")
else:
    print("100% other news")
```

4. RESULTS & OUTCOME

The program exhibits output in terms of percentage associated with the category. Since the dataset used for beginner basis consists of only three categories, rest of the other categories are classified under “Other News”.

5. CONCLUSION

5.1. Future Scope

For future modifications under such program, we can enable a browser extension, which when enabled can propose an on-page category of every news scrolled over. Future research will use the recommended approach with multiple searches and a big collection of text documents. The study also lays the groundwork for a future study that will investigate the influence of words with similar meanings inside a phrase.

5.2. Challenges

5.2.1.Language Barriers

Varied languages have drastically different collections of vocabulary, as well as different forms of phrasing, ways of inflection, and cultural expectations. This problem can be solved by using "universal" models that can transfer at least some learning to other languages. You will, however, need to retrain your NLP system for each new language.

5.2.2. Phrasing Ambiguities

Even another human being might struggle to decipher what someone intends when they say something confusing. A rigorous study of their words may not reveal a clear, crisp meaning. To remedy this, an NLP system must be able to seek context that will assist it in understanding the wording.

5.2.3.Words with Multiple Meanings

No language is flawless, and many words in most languages can have numerous meanings depending on the context. With the aid of context, good NLP technologies should be able to distinguish between these sentences.

6. REFERENCES

- “Text Classification of News Articles”, https://www.analyticsvidhya.com/blog/2021/12/text-classification-of-news-articles/, May 2022

- “IRJET – News Classification using Natural Language Processing”, https://www.irjet.net/archives/V9/i4/IRJET-V9I4336.pdf, May 2022

- “Text Classification of English News Articles using Graph Mining Techniques”, https://www.scitepress.org/Papers/2022/109546/109546.pdf, May 2022

- “Natural Language Processing: A Beginner’s Guide Part – 1, https://towardsdatascience.com/natural-language-processing-a-beginners-guide-part-i-1a5880cc3bdc, May 2022

- “The Python Standard Library”, https://docs.python.org/3/library/, May 2022

- “Getting Started with NLP using NLTK Library”, https://www.analyticsvidhya.com/blog/2021/07/getting-started-with-nlp-using-nltk-library/, June 2022

- “Natural Language Toolkit (NLTK) Tutorial with Python”, https://www.mygreatlearning.com/blog/nltk-tutorial-with-python/, June 2022

- “The 10 Biggest Issues in Natural Language Processing (NLP)”, https://www.rosoka.com/blog/10-biggest-issues-natural-language-processing-nlp, June 2022

- “Stemming and Lemmatization in Python NLTK with Examples”, https://www.guru99.com/stemming-lemmatization-python-nltk.html, June 2022