Marathwada Shikshan Prasarak Mandal’s
Deogiri Institute of Engineering and Management Studies,
Aurangabad

Project Report

on

News Summarization

Submitted By

Sakshi Sagar Kherdekar (36148)
Sakshi Rajesh Kakde (36157)

for

Continuous Assessment of

Machine Learning (TY CSE)

Dr. Babasaheb Ambedkar Technological University
Lonere (M.S.)

MSPM's Let us Rise Above The Rest
DIEMS

Department of Computer Science and Engineering
Deogiri Institute of Engineering and Management Studies,
Aurangabad
(2021- 2022)

Project Report
on

News Summarization

Submitted By
Sakshi Sagar Kherdekar (36148)
Sakshi Rajesh Kakde (36157)

In partial fulfillment of
Bachelor of Technology
(Computer Science & Engineering)

Guided By
Dr. Padmapani P. Tribhuvan

Department of Computer Science & Engineering
Deogiri Institute of Engineering and Management Studies,
Aurangabad
(2021- 2022)

CERTIFICATE

This is to certify that, the Project entitled “**News Summarization**” submitted by **Sakshi Sagar Kherdekar (36148), Sakshi Rajesh Kakde (36157)** ,is a bonafide work completed under my supervision and guidance in partial fulfillment for award of Bachelor of Technology(Computer Science and Engineering) Degree of Dr. Babasaheb Ambedkar Technological University, Lonere.

Place: Aurangabad
Date: 12-1-2022

Dr. Padmapani P. Tribhuvan
Guide

Mr. Sanjay B. Kalyankar
Head

Dr. Ulhas D. Shiurkar
Director,
Deogiri Institute of Engineering and Management Studies,
Aurangabad

Contents

List of Screen Shots i

1. INTRODUCTION 4

1.1 Introduction

1.2 Problem Statement

1.3 Objectives

2. DATA COLLECTION AND ANALYSIS 6

3. FINAL DESIGN AND IMPLEMENTATION 7

4. PERFORMANCE ANALYSIS 8

5. CONCLUSION AND FUTURE SCOPE Project Code 15

Acknowledgement

1. Introduction

1.1 Introduction

News Summarization comes under Unsupervised Machine Learning. Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision. News Summarization is the problem of creating a short, accurate, and fluent summary of a long text document.

Automatic news summarization is the process of creating a short and coherent version of a longer document. The idea of automatic summarization work is to develop techniques by which a machine can generate summarize that successfully imitate summaries generated by human beings. News summarization in NLP is the process of summarizing the information in large texts for quicker consumption The method of extracting these summaries from the original huge text without losing vital information is called as Text Summarization. It is essential for the summary to be a fluent, continuous and depict the significant. In fact, the Google news, the inshorts app and various other news aggregator apps take advantage of text summarization algorithms.

Text summarization methods can be grouped into two main categories: Extractive and Abstractive methods Extractive Text Summarization. It is the traditional method developed first. The main objective is to identify the significant sentences of the text and add them to the summary. Abstractive Text Summarization: The approach is to identify the important sections, interpret the context and reproduce in a new way. This ensures that the core information is conveyed through shortest text possible. Here, the sentences in summary are generated, not just extracted from original text.

1.2 Problem Statement

Preparing A Web application and a ML Model which will Summarize the news using NLP(Natural Language Processing) with spacy library an unsupervised Machine Learning Approach, model using Python Programming language in which the Web Page will be integrated with our News Summarization ML Model. The Web page will contain a text box, in which the user will input the long news and after clicking the summarize button it will display the Summarized news.

1.3 Objectives

- Automatic News summarization methods are greatly needed to address the ever-growing amount of text data available online to both better help discover relevant information, faster.

- We cannot possibly create summaries of all the text manually
- News Summarization reduces reading time.
- Automatic Summarization algorithms are less biased than human summarizers.
- The main objective is to identify the significant sentences of the text and add them to the summary

2. DATA COLLECTION

For our project we are not having any dataset, because it comes under unsupervised machine learning. We can give the input as long news and the model will summarize it. So, basically our dataset will look something like this, as shown below in the table. It will have input as long news and output as summarized news.

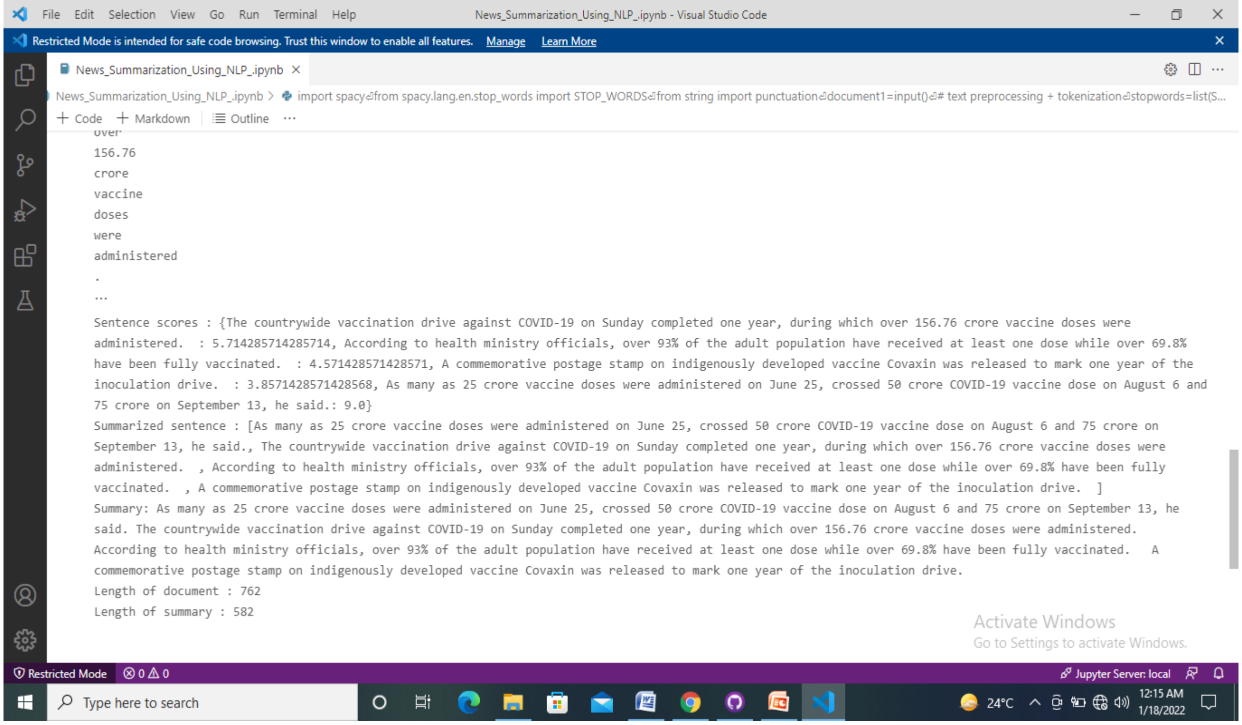
News	Summary
The State government has stopped sending samples for Omicron confirmation to the National Institute of Virology (NIV), as the infection was mild and patients were getting discharged even before the report was received, said Health Minister Ma. Subramanian. Speaking to media persons on Tuesday, Mr. Subramanian said samples from cluster areas only will be sent for testing to NIV, so that new variants, if any, could be identified.“	Government employees need a break during January 14-16,” said Mr. Subramanian. With the cooperation of residents, the number of cases can be controlled. “The Chief Minister is determined to control the spread of COVID-19 without causing any adverse impact on the economy. So there is no need for a major lockdown. We are also planning to postpone the lockdown this Sunday because of Pongal. Subramanian.

The Chief Minister is determined to control the spread of COVID-19 without causing any adverse impact on the economy. So there is no need for a major lockdown. With the cooperation of residents, the number of cases can be controlled. We are also planning to postpone the lockdown this Sunday because of Pongal. Government employees need a break during January 14-16," said Mr. Subramanian.	

3. FINAL DESIGN AND IMPLEMENTATION

News Summarization comes under unsupervised machine learning. We have implemented the project using NLP that is Natural Language Processing, in NLP we have used spaCy library. SpaCy is a free, open-source library for advanced Natural Language Processing (NLP) in Python. SpaCy is designed specifically for production use and helps you build applications that process and "understand" large volumes of text. It can be used to build information extraction or natural language understanding systems, or to pre-process text for deep learning.

For our project we have implemented it using python programming language and unsupervised machine learning approach. We have used html for designing the web page on brackets platform.

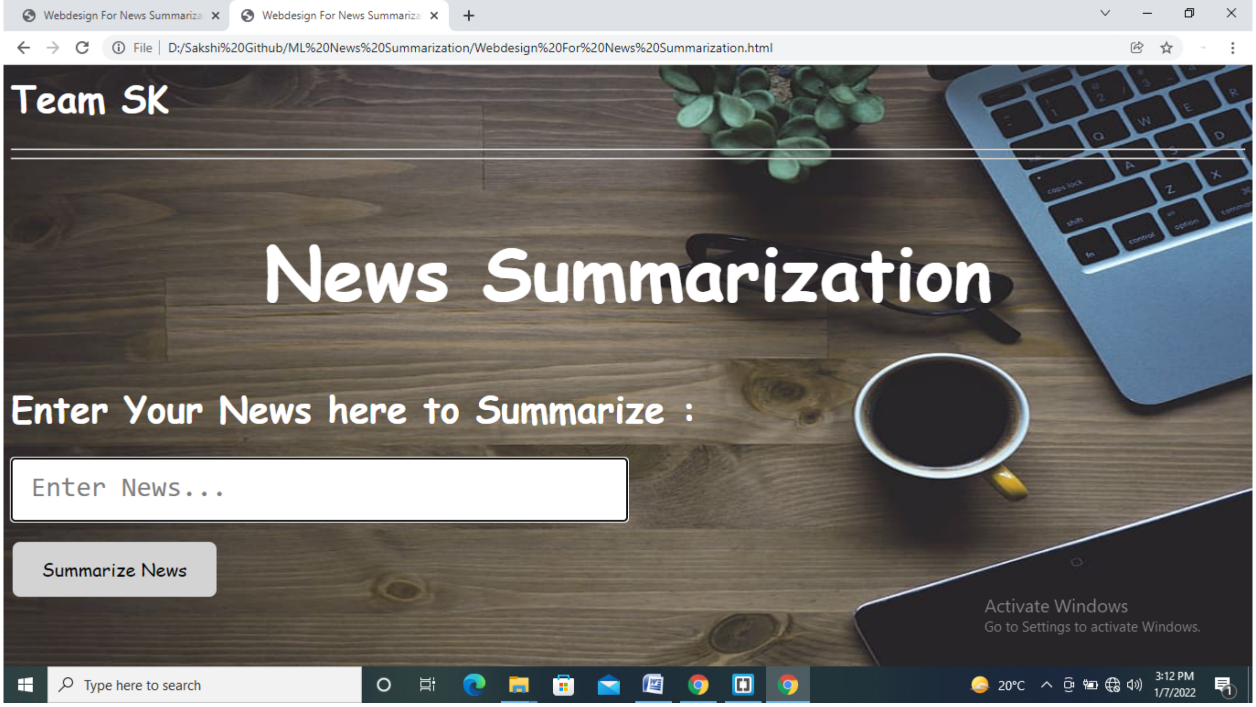


File Edit Selection View Go Run Terminal Help
News_Summarization_Using_NLP_.ipynb - Visual Studio Code
Restricted Mode is intended for safe code browsing. Trust this window to enable all features. Manage Learn More
News_Summarization_Using_NLP_.ipynb
News_Summarization_Using_NLP_.ipynb > import spacy⏎from spacy.lang.en.stop_words import STOP_WORDS⏎from string import punctuation⏎document1=input()⏎# text preprocessing + tokenization⏎stopwords=list(S...
+ Code + Markdown Outline
Activate Windows
Go to Settings to activate Windows.
Restricted Mode ⊗ 0 ⚠ 0
Jupyter Server: local
Type here to search
24°C
12:15 AM
1/18/2022

```
over
156.76
crore
vaccine
doses
were
administered
.
...
Sentence scores : {The countrywide vaccination drive against COVID-19 on Sunday completed one year, during which over 156.76 crore vaccine doses were
administered.  : 5.714285714285714, According to health ministry officials, over 93% of the adult population have received at least one dose while over 69.8%
have been fully vaccinated.  : 4.571428571428571, A commemorative postage stamp on indigenously developed vaccine Covaxin was released to mark one year of the
inoculation drive.  : 3.8571428571428568, As many as 25 crore vaccine doses were administered on June 25, crossed 50 crore COVID-19 vaccine dose on August 6 and
75 crore on September 13, he said.: 9.0}
Summarized sentence : [As many as 25 crore vaccine doses were administered on June 25, crossed 50 crore COVID-19 vaccine dose on August 6 and 75 crore on
September 13, he said., The countrywide vaccination drive against COVID-19 on Sunday completed one year, during which over 156.76 crore vaccine doses were
administered.  , According to health ministry officials, over 93% of the adult population have received at least one dose while over 69.8% have been fully
vaccinated.  , A commemorative postage stamp on indigenously developed vaccine Covaxin was released to mark one year of the inoculation drive.  ]
Summary: As many as 25 crore vaccine doses were administered on June 25, crossed 50 crore COVID-19 vaccine dose on August 6 and 75 crore on September 13, he
said. The countrywide vaccination drive against COVID-19 on Sunday completed one year, during which over 156.76 crore vaccine doses were administered.
According to health ministry officials, over 93% of the adult population have received at least one dose while over 69.8% have been fully vaccinated.   A
commemorative postage stamp on indigenously developed vaccine Covaxin was released to mark one year of the inoculation drive.
Length of document : 762
Length of summary : 582
```

i

So, our webpage is looking something like this. Wherein the user needs to copy/ paste the news in the given text box. After entering the news he/she needs to click on the summarize news button. The models will run after that and it will give the summarized news below it. So, this is how our model will work.


Webdesign For News Summariza
Webdesign For News Summariza
File
D:/Sakshi%20Github/ML%20News%20Summarization/Webdesign%20For%20News%20Summarization.html
Summarize News
Activate Windows
Go to Settings to activate Windows.
Type here to search
20°C
3:12 PM
1/7/2022

Team SK

News Summarization

Enter Your News here to Summarize :

Enter News...

ii

4. PERFORMANCE ANALYSIS

For performance analysis we have compared the summary of our model with the summary of another alternate design for our project. And we have measured the performance of our model. So we have checked for the word count, and also we have compared the summaries manually. So we come to know that npl i.e. with spacy we get more accurate results. The table shown below shows the long news then summarized news through our model and also summarized news of alternate design.

Even if you read both the news you come to know that Nlp gives the most accurate news.

NEWS	NLP SUMMARY	ALTERNATE DESIGN SUMMARY

	Using Spacy	
The State government has stopped sending samples for Omicron confirmation to the National Institute of Virology (NIV), as the infection was mild and patients were getting discharged even before the report was received, said Health Minister Ma. Subramanian. Speaking to media persons on Tuesday, Mr. Subramanian said samples from cluster areas only will be sent for testing to NIV, so that new variants, if any, could be identified.“ The Chief Minister is determined to control the spread of COVID-19 without causing any adverse impact on the economy. So there is no need for a major lockdown. With the cooperation of residents, the number of cases can be	Government employees need a break during January 14-16,” said Mr. Subramanian. With the cooperation of residents, the number of cases can be controlled. “The Chief Minister is determined to control the spread of COVID-19 without causing any adverse impact on the economy. So, there is no need for a major lockdown. We are also planning to postpone the lockdown this Sunday because of Pongal.	Speaking to media persons on Tuesday, Mr. Subramanian said samples from cluster areas only will be sent for testing to NIV, so that new variants, if any, could be identified.“ The Chief Minister is determined to control the spread of COVID-19 without causing any adverse impact on the economy.

controlled. We are also planning to postpone the lockdown this Sunday because of Pongal. Government employees need a break during January 14-16," said Mr. Subramanian.		
New Delhi: After some decent gains, the cryptocurrency market took a breather as major crypto tokens were trading flat. Trading was light as investors looked for signs that Bitcoin's downward spiral has reached an endpoint and that the largest cryptocurrency by market capitalization is ready to enter a new bull cycle. Barring Solana and Polkadot, all other eight out of the top-10 digital tokens were trading with a positive bias during the early trade on Monday. Cardano,	After some decent gains, the cryptocurrency market took a breather as major crypto tokens were trading flat. Cardano, meanwhile, zoomed over 12 per cent to become the fifth largest token by market cap. However, the total crypto market volume increased by 5 per cent to $64.30 billion. The global crypto market cap was largely unchanged at $2.07 trillion compared to the last day. Barring Solana and Polkadot, all other eight out of the top-10 digital tokens were trading with a positive bias during the	New Delhi: After some decent gains, the cryptocurrency market took a breather as major crypto tokens were trading flat.

meanwhile, zoomed over 12 per cent to become the fifth largest token by market cap. The global crypto market cap was largely unchanged at $2.07 trillion compared to the last day. However, the total crypto market volume increased by 5 per cent to $64.30 billion.	early trade on Monday.	

5. CONCLUSION AND FUTURE SCOPE

Here, we can conclude that news summarization using npl with spacy library gives us the best results. Also this project can be used for different type of text, other than news also. In that case also it gives us the best results.

Project Code

```
import spacy
from spacy.lang.en.stop_words import STOP_WORDS
from string import punctuation
document1=input()
# text preprocessing + tokenization
stopwords=list(STOP_WORDS)
print('Length of stopwords is :',len(stopwords))
```

```
nlp=spacy.load('en')
docx=nlp(document1)
for token in docx:
  print(token.text)
# word frequency table
# dictionary of words and their counts
# using non-stopwords
word_frequencies={}
for word in docx:
  if word.text not in stopwords:
    if word.text not in word_frequencies.keys():
      word_frequencies[word.text]=1
    else:
      word_frequencies[word.text]+=1
print('Word frquencies :',word_frequencies)
# maximum frequency
#find the weighted frequency,
#each word over most occuring word
#long sentence over short sentence
maximum_frequency=max(word_frequencies.values())
print('Maximum frequency :',maximum_frequency)
for word in word_frequencies.keys():
  word_frequencies[word]=(word_frequencies[word]/maximum_frequency)
```

```
print('Word frequencies :',word_frequencies)

# sentence tokenization

# scoring every sentence based on number of words

#(non-stopwords in our word frequency table)

sentence_list=[sentence for sentence in docx.sents]

sentence_scores={}

for sent in sentence_list:

 for word in sent:

  if word.text.lower() in word_frequencies.keys():

   if len(sent.text.split(' '))<30:

    if sent not in sentence_scores.keys():

     sentence_scores[sent]=word_frequencies[word.text.lower()]

    else:

     sentence_scores[sent]+=word_frequencies[word.text.lower()]

print('Sentence scores :',sentence_scores)

#find top n sentences with largest score

from heapq import nlargest

summarized_sentences=nlargest(7,sentence_scores,key=sentence_scores.get)

print('Summarized sentence :',summarized_sentences)

#convert from spacy span to string

final_sentences=[w.text for w in summarized_sentences]

#join sentences
```

```
summary=' '.join(final_sentences)

print('Summary:',summary)

print('Length of document :',len(document1))

print('Length of summary :',len(summary))
```

ACKNOWLEDGEMENT

We would like to place on record our deep sense of gratitude to Mr. Sanjay Kalyankar, Head of Department Computer Science and Engineering, Deogiri Institute of Engineering and management Studies Aurangabad, for his generous guidance, help and useful suggestions.

We express our sincere gratitude to Dr. Padmapani P. Tribhuvan, Dept. of Computer Science and Engineering, Deogiri Institute of Engineering and management Studies Aurangabad, for her stimulating guidance, continuous encouragement and supervision throughout the course of present work.

We are extremely thankful to Dr. Ulhas Shiurkar, Director, Deogiri Institute of Engineering and management Studies Aurangabad, for providing me infrastructural facilities to work in, without which this work would not have been possible.

Signature(s) of Students

Sakshi Sagar Kherdekar (36148) Sakshi Kherdekar

Sakshi Rajesh Kakde (36157) SKakde