Abstractive and Extractive text summarization using CNN/DailyMail dataset.

Github link: https://github.com/pahembha/NLP_PROJECT_G3

Group-No: 3

Team Members

A, Bharath Chowdary - 11712655

M. Hemanth Kumar - 11637925

P. Pardha Saradhi - 11648100

1. Motivation

The prime motivation for this project is deep dive the machine learning, which focuses mainly on the natural language processing . we are looking to handle the situation of information overloading. here we accept and recognize the increase in difficulty which people are facing to stay themselves updated from all the possible sources of information. To mention details, we must speak about NLP applications which summarizes large datasets, particularly new pieces from SNN and daily Mail datasets. Also, we are working on this project to get a summary of news quickly and easily. The motive involves the power and speed of the technology to simplify the process and built the model. Our goal is to be a tool that people can read news and get an idea about it in a quick time.

2. Objectives

the ultimate objective is to build an effective tool which is sufficient to read long articles and make them effectively short using the different NLP techniques and ML algo which in return generate the accurate and concise large text. Next develop a platform where users can easily request and get the text easily to process them and this platform should also be able to accommodate various user preferences and provide a better user experience for accessing news. To use NLP techniques to better understand the way language is used in news articles. And our project also involves the NLP technique to gain a deeper and inner understanding of the language used in news and articles. Here our model aims to recognize and generate the main idea and key points behind. This model making the summaries more accurate and easier to understand.

3. Significance

In this world info is available in plenty of sources like in databases, articles, web pages and many other ways and they are stored in desired notations, but they can be considered as readily available state. Everywhere the time is always limited, and the winner comes out of the effective user of the time in any segment. The range of content and the volume of content from the global headlines to the topics can always be appreciated by the people to stay informed. I strongly believe that this project plays a vital role because it helps people to stay up to date in a quick time so that to not feel stressed. Present day the availability of news is on a large scale from Newspapers to website articles it is tough for people to cover all this which is stressed for them. To make it easier just make the entire article into paragraphs by summarizing the news. Our tool will help to solve this problem by making short summaries of long articles. This means people can quickly understand the main points without having to read the whole article. This will save people time and make it easier for them to know what's going on in the world.

4. Features

Deliverables and Technical features

1.Data summarisation tool

2.Ml model and NLP techniques

3.User interface

4.Scalability

5.Test and feedback

6.Feedback implementations

7.Performance measure and suggestions

8.Training materials and their updates

9.Tepository handling and creation

10.Project report

Uniqueness:

- Here we are using dual summarisation technique approach(both abstractive and extractive methods are used here)
- we are planning to make the user friendly interface to handle the model in best way
- we mainly concentrate to use the latest publication or future publications available to make the model an effective model.
- one of the step is to conduct a comparison analysis with the existing solutions of the similar or near to similar ideas
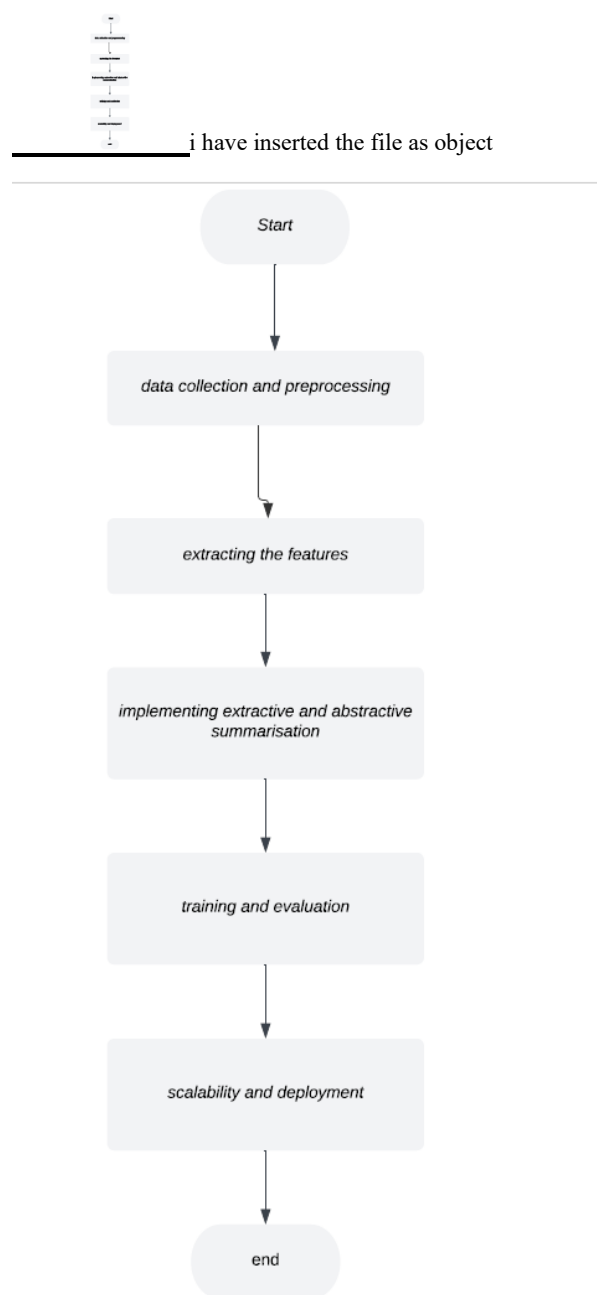
Milestone

M1. Data summarisation model and NLP techniques, user interface, scalability (on or before oct-15)

M2. Test and feedback, feedback implementations, performance measure and suggestions (on or before nov-05)

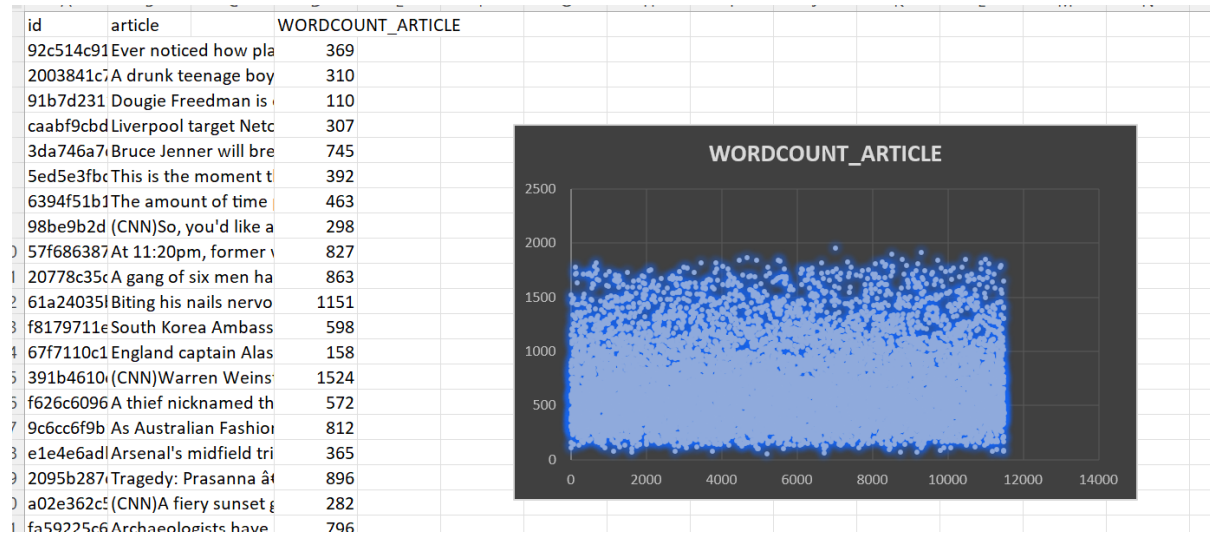M3. Repository handling and creation, Project report (on or before nov-20)

For any further changes or if we miss any of the segments to satisfy the rubric we can implement on or before dec-05

Workflow:

i have inserted the file as object

Start

data collection and preprocessing

extracting the features

implementing extractive and abstractive summarisation

training and evaluation

scalability and deployment

end

Visualisation:

Here we have plotted the wordcount which we are going to further process and analyse using the NLP techniques we just gave a overflow of the words which we Gona test and implement the model based on these.

id	article	WORDCOUNT_ARTICLE
92c514c91	Ever noticed how pla	369
2003841c7	A drunk teenage boy	310
91b7d231	Dougie Freedman is	110
caabf9cbd	Liverpool target Netc	307
3da746a7	Bruce Jenner will bre	745
5ed5e3fbc	This is the moment t	392
6394f51b1	The amount of time	463
98be9b2d	(CNN)So, you'd like a	298
57f686387	At 11:20pm, former v	827
20778c35c	A gang of six men ha	863
61a24035l	Biting his nails nervo	1151
f8179711e	South Korea Ambass	598
67f7110c1	England captain Alas	158
391b4610	(CNN)Warren Weins	1524
f626c6096	A thief nicknamed th	572
9c6cc6f9b	As Australian Fashior	812
e1e4e6adl	Arsenal's midfield tri	365
2095b287	Tragedy: Prasanna â	896
a02e362c5	(CNN)A fiery sunset g	282
fa59225c6	Archaeologists have	796

WORDCOUNT_ARTICLE

REFERENCES

1. "Natural Language Processing, Jacob Einstein"

2. https://arxiv.org/ftp/arxiv/papers/2204/2204.01849.pdf

3. "Speech and Language Processing, by Jurafsky and Martin"

4. https://github.com/pahembha/NLP_PROJECT_G3