Phase 5 Project - Fast Tracking Media Intelligence

Natural Language Processing Model for Text Summarization of CNN Daily News articles

Name: Andrew Levinton Student Pace: Self Pace Instructor name: Ahbhineet Kukarni

Business Problem

In this study, we will be analyzing CNN-Daily Mail News Articles. Many companies want to keep track of the latest news trends and reading full length articles can be incredibly time consuming. With the way news headlines affect major business decisions such as legal matters and stock prices, its important that readers are able to obtain all the information and reasoning behind the purpose of these articles and many of them can have a large number of distractors. It's essential that companies that get this information have the ability to quickly analyze these articles and develop counter intel or reports of their own in order to stay current and relevant with today's media. In this study we take the first steps towards utilizing text summarization to fast track the information gathering and counter-reporting process.

Business Understanding

Description:

In today's information age, staying updated with news and information is crucial for both individuals and businesses. However, the sheer volume of news articles published daily can be overwhelming. Reading and processing these articles is time-consuming, especially when trying to gather insights from multiple sources.

To address this challenge, we propose the development of a text summarization model using natural language processing (NLP) techniques. This model will automatically generate concise and coherent summaries of news articles. Here's how it can benefit the public and support business decisions:

1. `Time Efficiency` : People often struggle to find the time to read lengthy news articles. Automated summarization allows individuals to quickly grasp the key points of an article, saving them time while keeping them informed.
2. `Enhanced Understanding` : Summaries provide a condensed version of the article, making complex topics more accessible to a wider audience. This can help people better understand important news and events.
3. `Multi-source Insights` : Readers can efficiently scan summaries from multiple sources to get a well-rounded view of a topic, fostering critical thinking and a broader perspective.
4. `Competitive Intelligence` : Businesses can use automated summarization to track news and developments in their industry, enabling them to stay ahead of competitors and adapt to market changes more effectively.
5. `Market Research` : Summarization can assist in analyzing customer sentiment, emerging trends, and competitor strategies by summarizing customer reviews, news, and social media posts.
6. `Content Curation` : Media companies and content aggregators can use summarization to curate content for their audience, improving user engagement and retention.

Goals of this study:

1. Streamline and automate the understanding of complex news topics.
2. Help track news developments to enhance market research.

Business Questions to consider:

- What specific NLP techniques or models can be employed to automate the process of summarizing complex news articles effectively?
- How can automation be leveraged to provide real-time updates on important news developments to users or clients?
- Are there opportunities to use automation to personalize news content delivery based on individual user preferences and interests?
- How can automated summarization and analysis be integrated into other tools or platforms, such as news aggregators or chatbots?
- What are the critical news sources, categories, or keywords that have the most significant impact on market trends and consumer sentiment?

We will be using the CNN Daily Mail News dataset to conduct our study. You can download the dataset from here

In [1]:

```
#dataframe manipulation
import numpy as np
import pandas as pd
import warnings

#plotting
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('ggplot')

#text preprocessing
import re

from bs4 import BeautifulSoup
from keras.preprocessing.text import Tokenizer
from keras_preprocessing.sequence import pad_sequences
from nltk.corpus import stopwords
from nltk.tokenize import sent_tokenize
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from gensim.parsing import strip_tags, strip_numeric, strip_multiple_whitespaces, stem_text, strip_punc
from gensim.parsing import preprocess_string
from gensim import parsing
import math

#modeling
from tensorflow.keras.layers import Input, LSTM, Embedding, Dense, Concatenate, TimeDistributed
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer, TfidfTransformer
from tensorflow.keras.models import Model
from tensorflow.keras.callbacks import EarlyStopping
from tensorflow.keras.layers import Attention
import warnings
pd.set_option("display.max_colwidth", 200)

#model performance
from rouge import Rouge
from collections import Counter

warnings.filterwarnings("ignore")
```

I. Data Understanding

The CNN / DailyMail Dataset is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail. The current version supports both extractive and abstractive summarization, though the original version was created for machine reading and comprehension and abstractive question answering.

The columns of the dataframe are:

1. `ID` - Unique ID of the article
2. `article` - The raw content of the article
3. `Highlights` - A summary of the article

Preview of dataframe

In [2]:
```
df = pd.read_csv('./data/test.csv')

df.head(3)
```

Out[2]:

	id	article	highlights
0	92c514c913c0bdfe25341af9fd72b29db544099b	Ever noticed how plane seats appear to be getting smaller and smaller? With increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putti...	Experts question if packed out planes are putting passengers at risk .\nU.S consumer advisory group says minimum space must be stipulated .\nSafety tests conducted on planes with more leg room th...
1	2003841c7dc0e7c5b1a248f9cd536d727f27a45a	A drunk teenage boy had to be rescued by security after jumping into a lions' enclosure at a zoo in western India. Rahul Kumar, 17, clambered over the enclosure fence at the Kamla Nehru Zoological...	Drunk teenage boy climbed into lion enclosure at zoo in west India .\nRahul Kumar, 17, ran towards animals shouting 'Today I kill a lion!'\nFortunately he fell into a moat before reaching lions an...
2	91b7d2311527f5c2b63a65ca98d21d9c92485149	Dougie Freedman is on the verge of agreeing a new two-year deal to remain at Nottingham Forest. Freedman has stabilised Forest since he replaced cult hero Stuart Pearce and the club's owners are p...	Nottingham Forest are close to extending Dougie Freedman's contract .\nThe Forest boss took over from former manager Stuart Pearce in February .\nFreedman has since lead the club to ninth in the C...

Sample Article and reference summary

To start, we look at a sample article from the dataframe of CNN articles, as well as its reference summary, labeled "highlights".

The purpose of the reference summary is to use it in helping determine how well our models perform in generating an effective summary.

In this business case, the following objectives are important:

- `Content Extraction` : The model should accurately extract and prioritize key information from news articles, such as goal tallies, player performances, club interests, and tournament status, similar to what is done in the reference summary.
- `Conciseness` : The generated summaries should be concise, avoiding unnecessary details and providing the most important facts. This is crucial for creating efficient news feed content and trend analysis.
- `Relevance` : The model should focus on relevant and impactful information to meet the needs of the business case, ensuring that the summaries provide insights that matter.
- `Structure` : The generated summaries should be well-structured, possibly using bullet points or other formatting techniques to enhance readability and quick comprehension.
- `Clarity` : The language used in the summaries should be clear and understandable, making the information accessible to a wide audience, including those who may not be experts in the field.
- `Performance Measurement` : The success of the model can be measured by comparing the generated summaries to reference summaries, just as in your current approach, to assess how effectively it condenses the content and provides valuable insights.

Ultimately, the desired outcome is to have a summarization model that can efficiently and accurately produce news feed content and trend analysis summaries from public news articles, in a manner that aligns with the characteristics and quality of the reference summary you've provided.

In [3]:
```
df['article'][1000]
```

Out[3]:
```
"Cristiano Ronaldo and Lionel Messi will go head-to-head once more in the race to be this season's top
scorer in the Champions League – although Luiz Adriano threatens to spoil the party. Both Barcelona an
d Real Madrid booked their spots in the semi-finals this week with victories over Paris Saint-Germain
and Atletico Madrid respectively. The planet's best footballers have scored eight times in Europe this
season. But Shakhtar Donetsk's Adriano, courted by Arsenal and Liverpool, has netted on nine occasions
this term. Cristiano Ronaldo, in action against Atletico Madrid on Wednesday evening, has scored eight
goals in Europe . Lionel Messi also has eight goals in the Champions League this term; one fewer than
Luiz Adriano . Ronaldo and Messi will both play at least two more times after Real Madrid and Barcelon
a reached the last four . Adriano, who moved to Donetsk in 2007, scored five against BATE Borsiov in t
he group stages. His performance that night made history, with the 27-year-old becoming only the secon
d player to score five times in a Champions League game. The other was Messi for Barcelona against Bay
er Leverkusen in 2012. He also scored the third quickest hat-trick in the competition's history (12 mi
nutes) as the Ukrainian side, knocked out by Bayern Munich in the round of 16, racked up the biggest-e
ver half-time lead (6-0) in Europe's premier tournament. 'I am in a good moment of my career and we'll
do what will be best for me and for the club,' said Adriano last month when quizzed over his future. A
driano, who netted five times against BATE Borisov in the group, has scored more goals than any other
player in the Champions League... he is out of contract in December and could move to the Premier Leag
ue . 'With my contract set to expire and many good performances, it'll be difficult to stay in Ukrain
e.' Arsenal have sent scouts to watch Adriano in recent months, while Liverpool are also keen on the B
razilian. His contract with Shakhtar Donetsk runs out at the end of the year. Ronaldo and Messi howeve
r, remain in pole-position to top the scoring charts with Barcelona and Real Madrid both in the hat fo
r the two-legged semi-finals to be played next month. Of the teams still in the pot, Neymar and Luis S
uarez of Barcelona, Real Madrid's Karim Benzema and former Manchester United and City striker Carlos T
evez, now plying his trade for Juventus, each have six goals. The draw for the last four will take pla
ce on Friday."
```

Sample Highlight

In [4]:
```
df['highlights'][1000]
```

Out[4]:
```
"Luiz Adriano scored nine times for Shakhtar Donetsk in Europe this season .\nThe Brazilian is out of
contract at the end of the year... both Arsenal and Liverpool are interested in signing the 27-year-ol
d .\nCristiano Ronaldo and Lionel Messi have netted eight goals this season .\nReal Madrid and Barcelo
na both in the Champions League semi-finals .\nREAD: Our reporters have their say on who will win the
Champions League .\nCLICK HERE for Sportsmail's guide to the Champions League final four ."
```

-The reference summary provided for the sample article appears to be a good representation of the article from a business case perspective for a text summarization model **aiming to extract key information from news articles**. Here's why its a good benchmark for achievement of our model:

- `Key Information Extraction` : The reference summary effectively extracts and condenses the most important information from the article. It mentions the number of goals scored by Luiz Adriano and the interest from Arsenal and Liverpool. It also highlights the goal tally of Cristiano Ronaldo and Lionel Messi, as well as the fact that both Real Madrid and Barcelona are in the Champions League semi-finals. This information encapsulates the main points of interest in the article.
- `Conciseness` : The reference summary is concise and to the point. It does not contain unnecessary details or filler content, which is crucial for generating news feed content and trends analysis where brevity is key.
- `Relevance` : The reference summary includes information that is directly relevant to the business case, such as the goals scored, contract status, and the interest of major football clubs. This relevance is essential for providing valuable insights.
- `Structure` : The summary is well-structured, with bullet points that make it easy to skim and understand the main points quickly. This is important for generating news feed content that needs to be quickly digestible.
- `Clarity` : The language used in the reference summary is clear and easy to understand, ensuring that the extracted information is accessible to a broad audience.

Checking for length and nulls

Checking the length of a DataFrame is a fundamental step in data analysis and manipulation. It provides valuable information about the data's size, which can be used for data validation, sampling, indexing, transformation, and various analytical tasks. Understanding the length of your data is a crucial aspect of effective data handling and analysis.

Checking for nulls in data is crucial for maintaining data quality, ensuring accurate analyses, supporting machine learning models,

In [5]:
```
len(df)
```

Out[5]:
```
11490
```

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
id            0
article       0
highlights    0
dtype: int64
```

In [7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11490 entries, 0 to 11489
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   id          11490 non-null  object
 1   article     11490 non-null  object
 2   highlights  11490 non-null  object
dtypes: object(3)
memory usage: 269.4+ KB
```

Getting the word count

The purpose of the code below is to provide a quantitative measure of the word count for each article in the DataFrame. Knowing the word count of each article can be valuable for various natural language processing tasks, such as text summarization, where you might want to generate a concise summary of the content while preserving essential information or meeting a specific length requirement. By having the word count readily available in the 'WordCount' column, you can make informed decisions and apply algorithms or methods that take the length of the text into account during the summarization process, ensuring that the generated summaries are of an appropriate length.

After we generate the word count we will:

- Look at distribution
- Look at general statistics like the mean and median to analyze the general text length

In [8]:

```python
# Function to count words in a text column
def count_words(text):
    words = text.split()
    return len(words)

# Apply the function to the DataFrame column
df['articleWordCount'] = df['article'].apply(count_words)
df['highlightsWordCount'] = df['highlights'].apply(count_words)
df.head()
```

Out[8]:

	id	article	highlights	articleWordCount	highlightsWordCount
0	92c514c913c0bdfe25341af9fd72b29db544099b	Ever noticed how plane seats appear to be getting smaller and smaller? With increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putti...	Experts question if packed out planes are putting passengers at risk .\nU.S consumer advisory group says minimum space must be stipulated .\nSafety tests conducted on planes with more leg room th...	370	36
1	2003841c7dc0e7c5b1a248f9cd536d727f27a45a	A drunk teenage boy had to be rescued by security after jumping into a lions' enclosure at a zoo in western India. Rahul Kumar, 17, clambered over the enclosure fence at the Kamla Nehru Zoological...	Drunk teenage boy climbed into lion enclosure at zoo in west India .\nRahul Kumar, 17, ran towards animals shouting 'Today I kill a lion!'\nFortunately he fell into a moat before reaching lions an...	311	38
2	91b7d2311527f5c2b63a65ca98d21d9c92485149	Dougie Freedman is on the verge of agreeing a new two-year deal to remain at Nottingham Forest. Freedman has stabilised Forest since he replaced cult hero Stuart Pearce and the club's owners are p...	Nottingham Forest are close to extending Dougie Freedman's contract .\nThe Forest boss took over from former manager Stuart Pearce in February .\nFreedman has since lead the club to ninth in the C...	110	35
3	caabf9cbdf96eb1410295a673e953d304391bfbb	Liverpool target Neto is also wanted by PSG and clubs in Spain as Brendan Rodgers faces stiff competition to land the Fiorentina goalkeeper, according to the Brazilian's agent Stefano Castagna. Th...	Fiorentina goalkeeper Neto has been linked with Liverpool and Arsenal .\nNeto joined Firoentina from Brazilian outfit Atletico Paranaense in 2011 .\nHe is also wanted by PSG and Spanish clubs, acc...	308	44
4	3da746a7d9afcaa659088c8366ef6347fe6b53ea	Bruce Jenner will break his silence in a two-hour interview with Diane Sawyer later this month. The former Olympian and reality TV star, 65, will speak in a 'far-ranging' interview with Sawyer for...	Tell-all interview with the reality TV star, 69, will air on Friday April 24 .\nIt comes amid continuing speculation about his transition to a woman and following his involvement in a deadly car c...	749	61

In [9]:
```
df['articleWordCount'].describe()['mean']
```

Out[9]: 683.5115752828547

In [10]:
```
df['highlightsWordCount'].describe()['mean']
```

Out[10]: 55.00931244560488

The average token count for the articles and the highlights are provided below:

Feature	Mean Token Count
Article	683
Highlights	55

Understanding the distribution of the sequences

Here, we will analyze the length of the reviews and the summary to get an overall idea about the distribution of length of the text. This will help us fix the maximum length of the sequence:

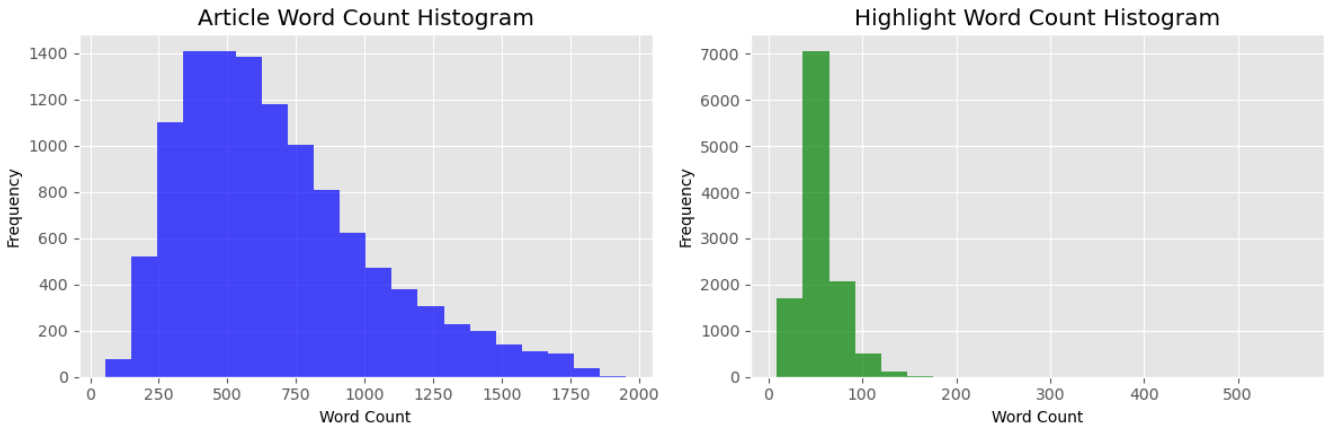
In [11]:
```
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))


ax1.hist(df['articleWordCount'], bins=20, color='blue', alpha=0.7)
ax1.set_title('Article Word Count Histogram')
ax1.set_xlabel('Word Count')
ax1.set_ylabel('Frequency')


ax2.hist(df['highlightsWordCount'], bins=20, color='green', alpha=0.7)
ax2.set_title('Highlight Word Count Histogram')
ax2.set_xlabel('Word Count')
ax2.set_ylabel('Frequency')


plt.tight_layout()


plt.show()
```



III. Data Preparation

Performing basic preprocessing steps is very important before we get to the model building part. Using messy and uncleaned text data is a potentially disastrous move. So in this step, we will drop all the unwanted symbols, characters, etc. from the text that do not affect the objective of our problem.

Here is the dictionary that we will use for expanding the contractions:

Contraction Mapping

A contraction mapping, also known as a "contraction operator" or "contraction function," is a concept from mathematics, specifically in the field of functional analysis. In the context of text summarization, it's not directly used, but related concepts and techniques from mathematics and natural language processing can be applied to improve the quality of text summarization models.

Contraction Mapping was utilized in our **2nd Iteration** of the modeling process, but not in this notebook. The concept of contraction mapping was used here just as a demonstration of how it was used in the actual model.

In [12]:
```
contraction_mapping = {"ain't": "is not", "aren't": "are not","can't": "cannot", "'cause": "because", "
                       "didn't": "did not",  "doesn't": "does not", "don't": "do not", "hadn't": "h
                       "he'd": "he would","he'll": "he will", "he's": "he is", "how'd": "how did",
                       "I'd": "I would", "I'd've": "I would have", "I'll": "I will", "I'll've": "I
                       "i'd've": "i would have", "i'll": "i will",  "i'll've": "i will have","i'm":
                       "it'd've": "it would have", "it'll": "it will", "it'll've": "it will have","
                       "mayn't": "may not", "might've": "might have","mightn't": "might not","might
                       "mustn't": "must not", "mustn't've": "must not have", "needn't": "need not",
                       "oughtn't": "ought not", "oughtn't've": "ought not have", "shan't": "shall n
                       "she'd": "she would", "she'd've": "she would have", "she'll": "she will", "s
                       "should've": "should have", "shouldn't": "should not", "shouldn't've": "shou
                       "this's": "this is","that'd": "that would", "that'd've": "that would have",
                       "there'd've": "there would have", "there's": "there is", "here's": "here is"
                       "they'll": "they will", "they'll've": "they will have", "they're": "they are
                       "wasn't": "was not", "we'd": "we would", "we'd've": "we would have", "we'll"
                       "we've": "we have", "weren't": "were not", "what'll": "what will", "what'll'
                       "what's": "what is", "what've": "what have", "when's": "when is", "when've":
                       "where've": "where have", "who'll": "who will", "who'll've": "who will have"
                       "why's": "why is", "why've": "why have", "will've": "will have", "won't": "w
                       "would've": "would have", "wouldn't": "would not", "wouldn't've": "would not
                       "y'all'd": "you all would","y'all'd've": "you all would have","y'all're": "y
                       "you'd": "you would", "you'd've": "you would have", "you'll": "you will", "y
                       "you're": "you are", "you've": "you have"}
```

We will perform the below preprocessing tasks for our data:

1.Convert everything to lowercase

2.Remove HTML tags

3.Contraction mapping

4.Remove (‘s)

5.Remove any text inside the parenthesis ( )

6.Eliminate punctuations and special characters

7.Remove stopwords

8.Remove short words

Let’s define the function:

Text Cleaner

This code block performs text preprocessing tasks such as lowercasing, HTML tag removal, parentheses and quotation marks removal, contraction expansion, possessive form removal, and non-alphabetic character removal. It also offers the option to remove common stopwords and short words, resulting in cleaner and more structured text data for analysis.

In [13]:
```
stop_words = set(stopwords.words('english'))
stop_words.update(['daily', 'news', 'mail'])
def text_cleaner(text,num):
    newString = text.lower()
    newString = BeautifulSoup(newString, "lxml").text
    newString = re.sub(r'\([^)]*\)', '', newString)
    newString = re.sub('"','', newString)
    newString = ' '.join([contraction_mapping[t] if t in contraction_mapping else t for t in newString.
    newString = re.sub(r"'s\b","",newString)
    newString = re.sub("[^a-zA-Z]", " ", newString)
    newString = re.sub('[m]{2,}', 'mm', newString)
    if(num==0):
        tokens = [w for w in newString.split() if not w in stop_words]
    else:
        tokens=newString.split()
    long_words=[]
    for i in tokens:
        if len(i)>1:                                                 #removing short word
            long_words.append(i)
    return (" ".join(long_words)).strip()
```

Applying the function to the dataframe articles

In [14]:
```
#call the function
cleaned_text = []
for t in df['article']:
    cleaned_text.append(text_cleaner(t,0))
```

Sample cleaned articles

In [15]:
```
print('Here is a sample of what the articles look like after cleaning:')
print('\n')
print('Article 1: ')
print(cleaned_text[0])
print('\n')
print('Article 2: ')
print(cleaned_text[1])
```

```
Here is a sample of what the articles look like after cleaning:


Article 1:
ever noticed plane seats appear getting smaller smaller increasing numbers people taking skies experts
questioning packed planes putting passengers risk say shrinking space aeroplanes uncomfortable putting
health safety danger squabbling arm rest shrinking space planes putting health safety danger week cons
umer advisory group set department transportation said public hearing government happy set standards a
nimals flying planes stipulate minimum amount space humans world animals rights space food humans said
charlie leocha consumer representative committee time dot faa take stand humane treatment passengers c
ould crowding planes lead serious issues fighting space overhead lockers crashing elbows seat back kic
king tests conducted faa use planes inch pitch standard airlines decreased many economy seats united a
irlines inches room airlines offer little inches cynthia corbertt human factors researcher federal avi
ation administration conducts tests quickly passengers leave plane tests conducted using planes inches
row seats standard airlines decreased reported detroit distance two seats one point seat point seat be
hind known pitch airlines stick pitch inches fall united airlines inches space gulf air economy seats
inches air asia offers inches spirit airlines offers inches british airways seat pitch inches easyjet
inches thomson short haul seat pitch inches virgin atlantic


Article 2:
drunk teenage boy rescued security jumping lions enclosure zoo western india rahul kumar clambered enc
losure fence kamla nehru zoological park ahmedabad began running towards animals shouting would kill m
r kumar explained afterwards drunk thought would stand good chance predators next level drunk intoxica
ted rahul kumar climbed lions enclosure zoo ahmedabad began running towards animals shouting today kil
l lion mr kumar sitting near enclosure suddenly made dash lions surprising zoo security intoxicated te
enager ran towards lions shouting today kill lion lion kills zoo spokesman said guards earlier spotted
close enclosure idea planing enter fortunately eight moats cross getting lions usually fell second one
allowing guards catch take handed police brave fool fortunately mr kumar fell moat ran towards lions c
ould rescued zoo security staff reaching animals kumar later explained really know drunk thought would
stand good chance police spokesman said cautioned sent psychiatric evaluation fortunately lions asleep
zoo guards acted quickly enough prevent tragedy similar delhi last year year old man mauled death tige
r indian capital climbing enclosure city zoo
```

Applying the same function to the dataframe highlights:

In [16]:
```
#call the function
cleaned_summary = []
for t in df['highlights']:
    cleaned_summary.append(text_cleaner(t,1))
```

In [17]:
```
print('Here is a sample of what the summaries look like after cleaning:')
print('\n')
print('Highlight 1: ')
print(cleaned_summary[0])
print('\n')
print('Highlight 2: ')
print(cleaned_summary[1])
```

```
Here is a sample of what the summaries look like after cleaning:


Highlight 1:
experts question if packed out planes are putting passengers at risk consumer advisory group says mini
mum space must be stipulated safety tests conducted on planes with more leg room than airlines offer


Highlight 2:
drunk teenage boy climbed into lion enclosure at zoo in west india rahul kumar ran towards animals sho
uting today kill lion fortunately he fell into moat before reaching lions and was rescued
```

In [20]:
```
df['cleaned_text']=cleaned_text
df['cleaned_summary']=cleaned_summary
```

Drop empty rows

In [21]:
```
df.replace('', np.nan, inplace=True)
df.dropna(axis=0,inplace=True)
```

Analyzing the length of the highlights

The purpose the code below is to analyze and gain insights into the distribution of summary lengths within a dataset. By iterating through the 'cleaned_summary' column of the DataFrame and checking the number of words in each cleaned summary, the code helps identify how many of these summaries are concise, containing 75 words or less. This analysis is valuable in various natural language processing and text analysis tasks, such as text summarization or information retrieval, as it provides an understanding of the dataset's summarization patterns. Furthermore, it allows data practitioners to assess the suitability of the dataset for specific applications that may have constraints on summary length.

In [22]:
```
cnt=0
for i in df['cleaned_summary']:
    if(len(i.split())<=75):
        cnt=cnt+1
print(cnt/len(df['cleaned_summary']))
```

```
0.9079199303742385
```

We observe that 91% of the highlights have length below 75. So, we can fix maximum length of summary to 75.

Let us fix the maximum length of an article to 750

In [23]:
```
cnt=0
for i in df['cleaned_text']:
    if(len(i.split())<=750):
        cnt=cnt+1
print(cnt/len(df['cleaned_text']))
```

```
0.9485639686684073
```

We observe that 95% of the articles have length below 750. So, we can fix maximum length of summary to 750.

The purpose of the code below is to filter and create a new dataset (data) that contains text and summary pairs where both the text and summary have lengths within certain limits (max_summary_len and max_text_len). This can be useful when working with text summarization tasks or other natural language processing applications where controlling the length of input and output sequences

is important. The resulting data DataFrame can be used for training, validation, or testing in machine learning models or other text analysis tasks.

In [24]:
```
max_text_len = 750
max_summary_len = 75
```

In [25]:
```
cleaned_text =np.array(df['cleaned_text'])
cleaned_summary=np.array(df['cleaned_summary'])

short_text=[]
short_summary=[]

for i in range(len(cleaned_text)):
    if(len(cleaned_summary[i].split())<=max_summary_len and len(cleaned_text[i].split())<=max_text_len)
        short_text.append(cleaned_text[i])
        short_summary.append(cleaned_summary[i])

data=pd.DataFrame({'article':short_text,'highlights':short_summary})
```

In [26]:
```
data.head()
```

Out[26]:

	article	highlights
0	ever noticed plane seats appear getting smaller smaller increasing numbers people taking skies experts questioning packed planes putting passengers risk say shrinking space aeroplanes uncomfortabl...	experts question if packed out planes are putting passengers at risk consumer advisory group says minimum space must be stipulated safety tests conducted on planes with more leg room than airlines...
1	drunk teenage boy rescued security jumping lions enclosure zoo western india rahul kumar clambered enclosure fence kamla nehru zoological park ahmedabad began running towards animals shouting woul...	drunk teenage boy climbed into lion enclosure at zoo in west india rahul kumar ran towards animals shouting today kill lion fortunately he fell into moat before reaching lions and was rescued
2	dougie freedman verge agreeing new two year deal remain nottingham forest freedman stabilised forest since replaced cult hero stuart pearce club owners pleased job done city ground dougie freedman...	nottingham forest are close to extending dougie freedman contract the forest boss took over from former manager stuart pearce in february freedman has since lead the club to ninth in the championship
3	liverpool target neto also wanted psg clubs spain brendan rodgers faces stiff competition land fiorentina goalkeeper according brazilian agent stefano castagna reds linked move year old whose cont...	fiorentina goalkeeper neto has been linked with liverpool and arsenal neto joined firoentina from brazilian outfit atletico paranaense in he is also wanted by psg and spanish clubs according to hi...
4	bruce jenner break silence two hour interview diane sawyer later month former olympian reality tv star speak far ranging interview sawyer special edition friday april abc announced monday intervie...	tell all interview with the reality tv star will air on friday april it comes amid continuing speculation about his transition to woman and following his involvement in deadly car crash in februar...

In [27]:
```
data['article'][1000]
```

Out[27]:
```
'panama city handshake shook western hemisphere president obama briefly met cuban counterpart raul cas
tro friday night dinner dozens latin american leaders convening panama city summit americas historic t
wo nations barely speaking terms officially years meeting important bernadette meehan national securit
y council spokesperson issued statement summit americas evening president obama president castro greet
ed shook hands cuba united states endured half century enmity tension worsened two nations miles apart
key events years include traumatic modern history cuban missile crisis bay pigs mariel boatlift two le
aders building historic face face obama spoke phone wednesday cuban leader heading panama met friday d
inner expected spend lot time together saturday summit begins earnest obama arrived panama late thursd
ay conference years past tinged animosity cuba exclusion moments marine one obama helicopter touched p
anama city castro plane landed tarmac panamanian television carried arrivals live phone call wednesday
obama castro discussed ongoing process normalizing relations united states cuba according deputy natio
nal security adviser ben rhodes said made sense two leaders communicate anticipated interactions frida
y saturday run ins represent highest level talks united states cuba since meeting vice president richa
rd nixon prime minister fidel castro new territory rhodes said friday reason president strongly believ
es approach focused totally isolation focused totally seeking cut cuban people united states america f
ailed obama expecting warm welcome dozens countries represented conference announcing december seeking
engage havana talks reopening embassies removing barriers commerce travel panama obama expected announ
ce removing cuba united states list countries sponsor terrorism major advance building diplomatic ties
two countries state department delivered report designation white house wednesday obama said thursday
panel experts reviewing makes final determination white house ruling final decision obama leaves panam
a late saturday night remarks brief stopover jamaica thursday obama strongly hinted ready remove cuba
list also includes iran sudan syria throughout process emphasis facts obama said want make sure given
powerful tool isolate countries genuinely support terrorism make designations got strong evidence fact
case circumstances change list change well said inside cuba expressed dissatisfaction pace diplomatic
thaw officials say pleased progress toward establishing diplomatic ties white house argues helped impr
ove relations countries region obama said jamaica never foresaw immediately overnight everything would
transform overtures cuba universally popular united states lawmakers irate obama seeking engage regard
corrupt government even obama landed panama long standing tensions pro anti castro activists full disp
lay dissidents opposed castro regime violently accosted earlier week supporters cuban government rhode
s said white house expressed serious concerns violence would continue speak support human rights refor
ms island'
```

In [28]:
```
data['highlights'][1000]
```

Out[28]:
```
'president obama cuban president raul castro meet in panama city the two nations only miles apart have
been at odds for more than years'
```

Analyzing n-grams - Articles

In [29]:
```
X = df['cleaned_text']
y = df.drop('cleaned_text', axis=1)
```

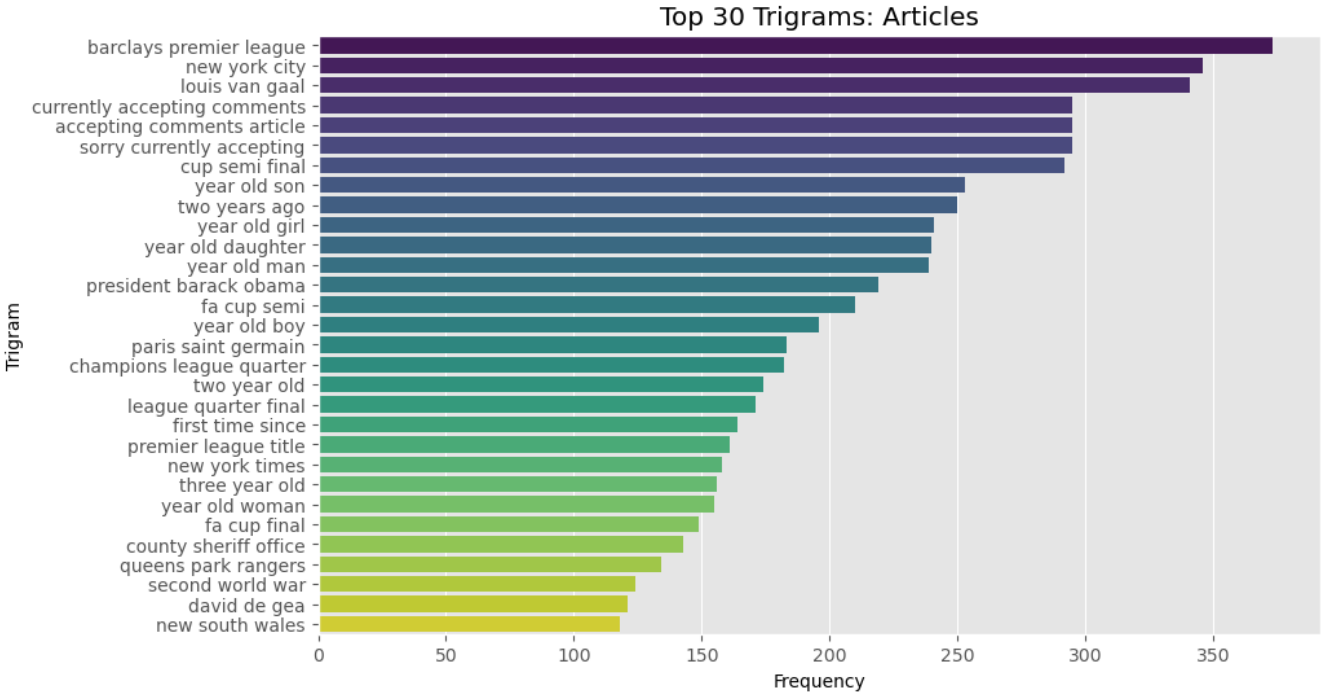
In [31]:
```
cv = CountVectorizer(ngram_range = (3,3))
X_count = cv.fit_transform(X)
X_count  = pd.DataFrame.sparse.from_spmatrix(X_count)
X_count.columns = sorted(cv.vocabulary_)
X_count.set_index(y.index, inplace=True)

all_tri_labels = X_count.sum().sort_values(ascending = False)[0:30]
```

In [32]:
```
# Create a bar plot
plt.figure(figsize=(10, 6))
sns.barplot(x=all_tri_labels.values, y=all_tri_labels.index, palette='viridis')

# Set plot labels and title
plt.xlabel('Frequency')
plt.ylabel('Trigram')
plt.title('Top 30 Trigrams: Articles')

# Display the plot
plt.show()
```



Analyzing n-grams - Highlights

In [35]:
```
X = df['cleaned_summary']
y = df.drop('cleaned_summary', axis=1)
```

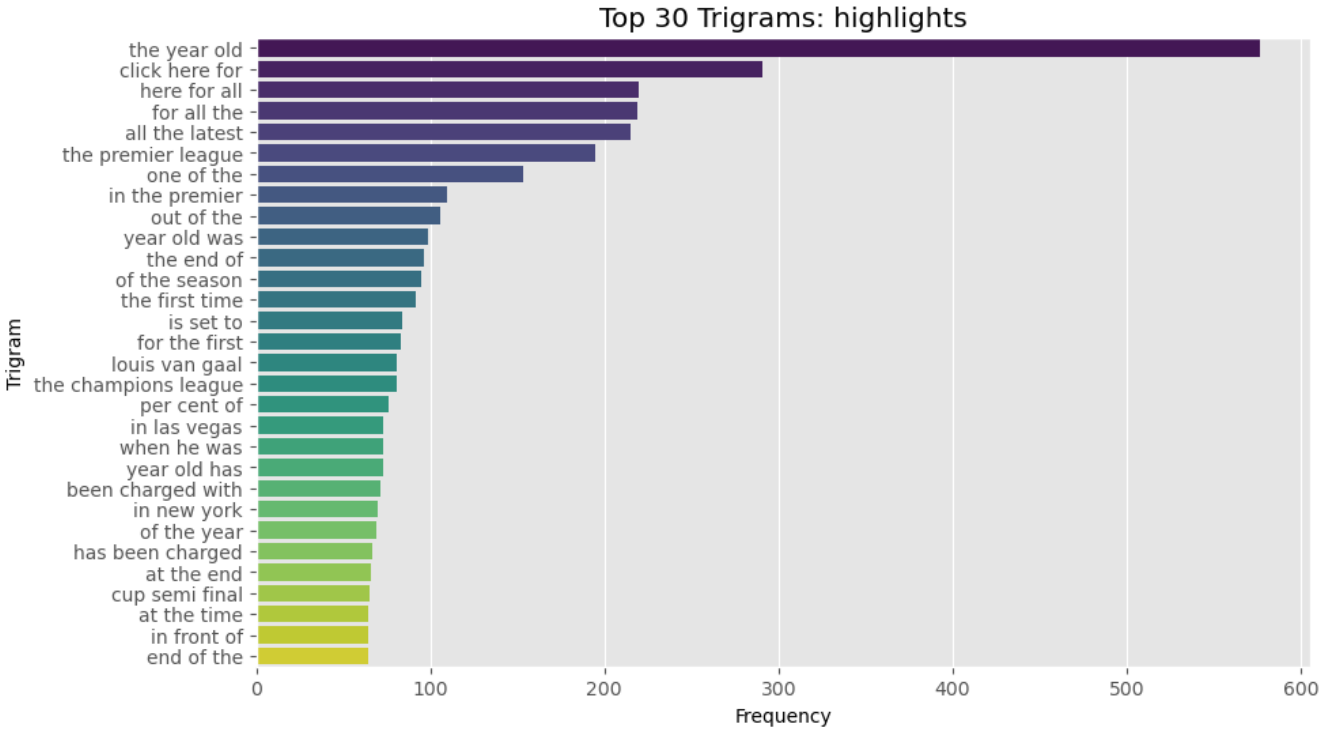
In [36]:
```
cv = CountVectorizer(ngram_range = (3,3))
X_count = cv.fit_transform(X)
X_count  = pd.DataFrame.sparse.from_spmatrix(X_count)
X_count.columns = sorted(cv.vocabulary_)
X_count.set_index(y.index, inplace=True)

all_tri_labels = X_count.sum().sort_values(ascending = False)[0:30]
```

In [37]:
```
# Create a bar plot
plt.figure(figsize=(10, 6))
sns.barplot(x=all_tri_labels.values, y=all_tri_labels.index, palette='viridis')

# Set plot labels and title
plt.xlabel('Frequency')
plt.ylabel('Trigram')
plt.title('Top 30 Trigrams: highlights')

# Display the plot
plt.show()
```



IV. Model building

Baseline - TFIDF Vectorizer

[TFIDF (https://medium.com/@ashins1997/text-summarization-f2542bc6a167#:~:text=Sentence%20scoring%20using%20tf%2Didf,word%20occurs%20in%20the%20document.)](https://medium.com/@ashins1997/text-summarization-f2542bc6a167#:~:text=Sentence%20scoring%20using%20tf%2Didf,word%20occurs%20in%20the%20document.)

Sentence scoring using tf-idf is one of the extractive approaches for text summarization.

`In TF-IDF (Term Frequency-Inverse Document Frequency)` text summarization, sentence scoring involves evaluating the importance of each sentence in a document. It does this by considering the frequency of words in the sentence (Term Frequency) and the uniqueness of those words across the entire document collection (Inverse Document Frequency). Sentences containing rare and important terms receive higher scores, making them more likely to be included in the summary. This aids in text summarization by selecting sentences that capture the key concepts and information, resulting in a concise and informative summary.

- **TF-IDF stands for Term Frequency** — Inverse Document Frequency. It is the product of two statistics.
- **Term Frequency (TF)** : It is the number of times the word occurs in the document.
- **Inverse Document Frequency (IDF)** : It is the measure of how much information the word provides, i.e., if it's common or rare across all documents.

Functions for steps in model building

1. Covert text to sentences : Converting a single text to list of sentences.

2. Pre-process text : Clean the sentences by removing unnecessary words, stopwords, punctuations, etc.
3. Create term frequency (tf) matrix : It shows the frequency of words in each sentence. We will calculate relative frequency to represent the tf instead of using actual frequency.

> It is calculated as t / T where,
>
> - t = Number of times the term appears in the document,
> - T = Total number of terms in the document

4. Create idf matrix : It shows the importance of words in each sentence with respect to the whole document.

> It is calculated as log_e(D/d) where,
>
> - D = Total number of documents,
> - d = Number of documents with term t in it

5. Calculate sentence tf-idf : It is the product of tf and idf for each word in the sentence and shows the importance of each word in the sentence.
6. Calculate sentence scores : Here score of the sentences are calculated as the average of the tf-idf value of words in the sentence. It is calculated as

> T / n where,
>
> - T = Total tf-idf of words in the sentence,
> - n = Number of distinct words in the sentence

7. Determine threshold : Threshold is the average value of the scores of the sentences.

> It is calculated as S / s where,
>
> - S = Total sum of scores of sentences,
> - s = Number of sentences

8. Generate summary : Generate a summary by extracting the sentences having scores greater than the threshold value.

In [38]:

```
# 1.Convert text to sentences
text = df['article'][1000]
sentences = sent_tokenize(text)

# Preprocess Text
ps = PorterStemmer()
def text_preprocessing(sentences):
    """
    Pre processing text to removeunnecessary words.
    """
#     print('Preprocessing text')
    stop_words = set(stopwords.words('english'))
    clean_words = []
    for sent in sentences:
        words = word_tokenize(sent)
        words = [ps.stem(word.lower()) for word in words if word.isalnum()]
        clean_words += [word for word in words if word not in stop_words]
    return clean_words

# 3.Create term frequency (tf) matrix

def create_tf_matrix(sentences: list) -> dict:
    """
    Here document refers to a sentence.
    TF(t) = (Number of times the term t appears in a document) /         (Total number of terms in the
    """
#     print('Creating tf matrix.')
    tf_matrix = {}
    for sentence in sentences:
        tf_table = {}
        words_count = len(word_tokenize(sentence))
        clean_words = text_preprocessing([sentence])
        # Determining frequency of words in the sentence
        word_freq = {}
        for word in clean_words:
            word_freq[word] = (word_freq[word] + 1) if word in  word_freq else 1
        # Calculating tf of the words in the sentence
        for word, count in word_freq.items():
            tf_table[word] = count / words_count
        tf_matrix[sentence[:15]] = tf_table
    return tf_matrix

# 4.Create idf matrix

def create_idf_matrix(sentences: list) -> dict:
    """
    IDF(t) = log_e(Total number of documents / Number of documents     with term t in it)
    """
#     print('Creating idf matrix.')

    idf_matrix = {}

    documents_count = len(sentences)
    sentence_word_table = {}

    # Getting words in the sentence
    for sentence in sentences:
        clean_words = text_preprocessing([sentence])
        sentence_word_table[sentence[:15]] = clean_words

    # Determining word count table with the count of sentences which contains the word.
    word_in_docs = {}
    for sent, words in sentence_word_table.items():
        for word in words:
            word_in_docs[word] = (word_in_docs[word] + 1) if word in word_in_docs else 1

    # Determining idf of the words in the sentence.
    for sent, words in sentence_word_table.items():
        idf_table = {}
        for word in words:
```

```
            idf_table[word] = math.log10(documents_count / float(word_in_docs[word]))

        idf_matrix[sent] = idf_table

    return idf_matrix

# 5.Calculate sentence tf-idf

def create_tf_idf_matrix(tf_matrix, idf_matrix) -> dict:
    """
    Create a tf-idf matrix which is multiplication of tf * idf individual words
    """
#     print('Calculating tf-idf of sentences.')

    tf_idf_matrix = {}

    for (sent1, f_table1), (sent2, f_table2) in zip(tf_matrix.items(), idf_matrix.items()):
        tf_idf_table = {}

        for (word1, value1), (word2, value2) in zip(f_table1.items(), f_table2.items()):
            tf_idf_table[word1] = float(value1 * value2)

        tf_idf_matrix[sent1] = tf_idf_table

    return tf_idf_matrix

# 6.Calculate sentence scores

def create_sentence_score_table(tf_idf_matrix) -> dict:
    """
    Determining average score of words of the sentence with its words tf-idf value.
    """
#     print('Creating sentence score table.')

    sentence_value = {}

    for sent, f_table in tf_idf_matrix.items():
        total_score_per_sentence = 0

        count_words_in_sentence = len(f_table)

        # Check if count_words_in_sentence is not zero
        if count_words_in_sentence != 0:
            for word, score in f_table.items():
                total_score_per_sentence += score

            sentence_value[sent] = total_score_per_sentence / count_words_in_sentence
        else:
            sentence_value[sent] = 0  # Set the score to 0 for empty sentences

    return sentence_value

# 7.Determine threshold
def find_average_score(sentence_value):
    """
    Calculate average value of a sentence form the sentence score table.
    """
#     print('Finding average score')

    sum = 0
    for val in sentence_value:
        sum += sentence_value[val]

    average = sum / len(sentence_value)

    return average


# 8.Generate summary
```

```
def generate_summary(sentences, sentence_value, threshold):
    """
    Generate a sentence for sentence score greater than average.
    """
#     print('Generating summary')

    sentence_count = 0
    summary = ''

    for sentence in sentences:
        if sentence[:15] in sentence_value and sentence_value[sentence[:15]] >= threshold:
            summary += sentence + " "
            sentence_count += 1

    return summary

# rouge score
def get_rouge_score(summary, abstract):
    scores = rouge.get_scores(summary, abstract)
    return scores
```

Sample Summary

Here we will test our tfidf summarizer as proof of concept. This cell will complete all the steps listed above and display the article as well as the reference summary and the newly generated one.

In [39]:

```
tf_matrix = create_tf_matrix(sentences)
# print('TF matrix', tf_matrix)

tf = tf_matrix
idf_matrix = create_idf_matrix(sentences)
# print('IDF matrix',idf_matrix)

idf = idf_matrix
tf_idf_matrix = create_tf_idf_matrix(tf_matrix, idf_matrix)
# print('TF-IDF matrix', tf_idf_matrix)
# print('First document tfidf',tf_idf_matrix[list(tf_idf_matrix.keys())[0]])

tf_idf = tf_idf_matrix
sentence_value = create_sentence_score_table(tf_idf_matrix)
# print('Sentence Scores', sentence_value)

threshold = find_average_score(sentence_value)
# print('Threshold', threshold)

summary = generate_summary(sentences, sentence_value, threshold)

print('\n\n')
print('Original Article: ')
print(df['article'][1000])

print('\n\n')
print('Predicted Summary: ')
print(summary)

print('\n\n')
print('Original Summary: ')
print(df['highlights'][1000])
```

```
Original Article:
Cristiano Ronaldo and Lionel Messi will go head-to-head once more in the race to be this season's top
scorer in the Champions League – although Luiz Adriano threatens to spoil the party. Both Barcelona an
d Real Madrid booked their spots in the semi-finals this week with victories over Paris Saint-Germain
and Atletico Madrid respectively. The planet's best footballers have scored eight times in Europe this
season. But Shakhtar Donetsk's Adriano, courted by Arsenal and Liverpool, has netted on nine occasions
this term. Cristiano Ronaldo, in action against Atletico Madrid on Wednesday evening, has scored eight
goals in Europe . Lionel Messi also has eight goals in the Champions League this term; one fewer than
Luiz Adriano . Ronaldo and Messi will both play at least two more times after Real Madrid and Barcelon
a reached the last four . Adriano, who moved to Donetsk in 2007, scored five against BATE Borsiov in t
he group stages. His performance that night made history, with the 27-year-old becoming only the secon
d player to score five times in a Champions League game. The other was Messi for Barcelona against Bay
er Leverkusen in 2012. He also scored the third quickest hat-trick in the competition's history (12 mi
nutes) as the Ukrainian side, knocked out by Bayern Munich in the round of 16, racked up the biggest-e
ver half-time lead (6-0) in Europe's premier tournament. 'I am in a good moment of my career and we'll
do what will be best for me and for the club,' said Adriano last month when quizzed over his future. A
driano, who netted five times against BATE Borisov in the group, has scored more goals than any other
player in the Champions League... he is out of contract in December and could move to the Premier Leag
ue . 'With my contract set to expire and many good performances, it'll be difficult to stay in Ukrain
e.' Arsenal have sent scouts to watch Adriano in recent months, while Liverpool are also keen on the B
razilian. His contract with Shakhtar Donetsk runs out at the end of the year. Ronaldo and Messi howeve
r, remain in pole-position to top the scoring charts with Barcelona and Real Madrid both in the hat fo
r the two-legged semi-finals to be played next month. Of the teams still in the pot, Neymar and Luis S
uarez of Barcelona, Real Madrid's Karim Benzema and former Manchester United and City striker Carlos T
evez, now plying his trade for Juventus, each have six goals. The draw for the last four will take pla
ce on Friday.


Predicted Summary:
Cristiano Ronaldo and Lionel Messi will go head-to-head once more in the race to be this season's top
scorer in the Champions League – although Luiz Adriano threatens to spoil the party. The planet's best
footballers have scored eight times in Europe this season. Cristiano Ronaldo, in action against Atleti
co Madrid on Wednesday evening, has scored eight goals in Europe . Adriano, who moved to Donetsk in 20
07, scored five against BATE Borsiov in the group stages. The other was Messi for Barcelona against Ba
yer Leverkusen in 2012. His contract with Shakhtar Donetsk runs out at the end of the year. The draw f
or the last four will take place on Friday.


Original Summary:
Luiz Adriano scored nine times for Shakhtar Donetsk in Europe this season .
The Brazilian is out of contract at the end of the year... both Arsenal and Liverpool are interested i
n signing the 27-year-old .
Cristiano Ronaldo and Lionel Messi have netted eight goals this season .
Real Madrid and Barcelona both in the Champions League semi-finals .
READ: Our reporters have their say on who will win the Champions League .
CLICK HERE for Sportsmail's guide to the Champions League final four .
```

In [40]:
```
print(f'Original {len(sent_tokenize(text))} sentences, Summarized {len(sent_tokenize(summary))} sentenc
```

```
Original 18 sentences, Summarized 7 sentences
```

Quick Analysis

Here the tfidf model was able to create a small summary based on the sentence scores, summarizing the article from 18 sentences down to 7.

Applying to all articles - removal of outliers

Here we will run our tfidf model on all the articles in the dataframe, in an attempt to get an overall evaluation of the summaries generated by the tfidf model. In order to obtain consistent model results, we will limit the length of the articles to be 750 words since

In [41]:
```
df = df[df['articleWordCount']<=750]
```

In [42]:
```
len(df)
```

Out[42]: 7434

Processing DataFrame of articles and their highlights

Here for every article in the dataframe, this code tokenizes the articles into sentences, calculates the importance of each sentence using TF-IDF, and generates summaries by selecting the most important sentences based on a threshold. The generated summaries are stored in one list, while the original highlights are stored in another list for further analysis or comparison.

In [43]:
```
abstract_tfidf = []
introduction_tfidf = []


for i, row in df.iterrows():

    text = row['article']
    sentences = sent_tokenize(text)

    # Step 3: Create term frequency (tf) matrix for the current article
    tf_matrix = create_tf_matrix(sentences)

    # Step 4: Create idf matrix for the current article
    idf_matrix = create_idf_matrix(sentences)

    # Step 5: Calculate sentence tf-idf for the current article
    tf_idf_matrix = create_tf_idf_matrix(tf_matrix, idf_matrix)

    # Step 6: Calculate sentence scores for the current article
    sentence_value = create_sentence_score_table(tf_idf_matrix)

    # Step 7: Determine threshold for the current article
    threshold = find_average_score(sentence_value)

    # Step 8: Generate summary for the current article
    summary = generate_summary(sentences, sentence_value, threshold)

    abstract_tfidf.append(summary)
    introduction_tfidf.append(row['highlights'])
```

Printing the first 6 articles and summaries

In [44]:
```
abstract_tfidf[:5]
```

Out[44]:
```
["Ever noticed how plane seats appear to be getting smaller and smaller? With increasing numbers of pe
ople taking to the skies, some experts are questioning if having such packed out planes is putting pas
sengers at risk. They say that the shrinking space on aeroplanes is not only uncomfortable - it's putt
ing our health and safety in danger. More than squabbling over the arm rest, shrinking space on planes
putting our health and safety in danger? 'It is time that the DOT and FAA take a stand for humane trea
tment of passengers.' But could crowding on planes lead to more serious issues than fighting for space
in the overhead lockers, crashing elbows and seat back kicking? The distance between two seats from on
e point on a seat to the same point on the seat behind it is known as the pitch. While most airlines s
tick to a pitch of 31 inches or above, some fall below this. ",
 "'We then handed him over to the police.' 'I was drunk and thought I'd stand a good chance.' A police
spokesman said: 'He has been cautioned and will be sent for psychiatric evaluation. ",
 'Dougie Freedman is on the verge of agreeing a new two-year deal to remain at Nottingham Forest. Doug
ie Freedman is set to sign a new deal at Nottingham Forest . ',
 "Real Madrid? We'll see. ",
 "Scroll down for video . She filed for divorce in September 2014, citing 'irreconcilable difference
s'. Reports also emerged over the past week that he has received a breast enhancement. 'Bruce had sili
cone breast implants put in a few weeks ago,' a source told RadarOnline. 'He went with a smaller impla
nt because he didn't want to look ridiculous.' 'I will say that I think Bruce should tell his story hi
s way. She died at the scene . Kris filed for divorce from him last year . The Lexus was carrying 69-y
ear-old Kim Howe, who died from chest trauma at the scene. "]
```

In [45]:
```
introduction_tfidf[:5]
```

Out[45]:
```
['Experts question if  packed out planes are putting passengers at risk .\nU.S consumer advisory group
says minimum space must be stipulated .\nSafety tests conducted on planes with more leg room than airl
ines offer .',
 "Drunk teenage boy climbed into lion enclosure at zoo in west India .\nRahul Kumar, 17, ran towards a
nimals shouting 'Today I kill a lion!'\nFortunately he fell into a moat before reaching lions and was
rescued .",
 "Nottingham Forest are close to extending Dougie Freedman's contract .\nThe Forest boss took over fro
m former manager Stuart Pearce in February .\nFreedman has since lead the club to ninth in the Champio
nship .",
 'Fiorentina goalkeeper Neto has been linked with Liverpool and Arsenal .\nNeto joined Firoentina\xa0f
rom Brazilian outfit Atletico Paranaense in 2011 .\nHe is also wanted by PSG and Spanish clubs, accord
ing to his agent .\nCLICK HERE for the latest Liverpool news .',
 "Tell-all interview with the reality TV star, 69, will air on Friday April 24 .\nIt comes amid contin
uing speculation about his transition to a woman and following his involvement in a deadly car crash i
n February .\nThe interview will also be one of Diane Sawyer's first appearances on television followi
ng the sudden death of her husband last year ."]
```

Model Evaluation - Rouge Score

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used for evaluating the quality of text summarization and machine translation. These metrics assess the similarity between the generated summary and one or more reference summaries (human-written) to measure the performance of automatic summarization systems. The primary focus of ROUGE metrics is on content overlap and the ability of a system to capture the essential information from the source text.

In [46]:
```
# Initializing rouge metric
rouge = Rouge()
```

This code evaluates the quality of generated text summaries in a DataFrame (df_tfidf) by calculating ROUGE scores, specifically ROUGE-1 and ROUGE-L. ROUGE is a set of metrics for text summarization assessment.

It iterates through each row, comparing the generated summary ('summaries') with the original summary ('original_summaries'). The code calculates precision, recall, and F1 score for both ROUGE-1 and ROUGE-L.

The computed ROUGE scores are then added as new columns in the DataFrame for further analysis and comparison. This allows for a quantitative assessment of how well the generated summaries match the original highlights.

In [47]:

```
df_tfidf = pd.DataFrame()

df_tfidf['summaries'] = abstract_tfidf

df_tfidf['original_summaries'] = introduction_tfidf

# Initialize the ROUGE scorer
rouge = Rouge()

# Initialize lists to store results

rouge_1_p = []  # ROUGE-1 Precision
rouge_1_r = []  # ROUGE-1 Recall
rouge_1_f = []  # ROUGE-1 F1
rouge_l_p = []  # ROUGE-L Precision
rouge_l_r = []  # ROUGE-L Recall
rouge_l_f = []  # ROUGE-L F1

for i, row in df_tfidf.iterrows():
    scores = get_rouge_score(row['summaries'], row['original_summaries'])
    rouge_1_p.append(scores[0]['rouge-1']['p'])
    rouge_1_r.append(scores[0]['rouge-1']['r'])
    rouge_1_f.append(scores[0]['rouge-1']['f'])
    rouge_l_p.append(scores[0]['rouge-l']['p'])
    rouge_l_r.append(scores[0]['rouge-l']['r'])
    rouge_l_f.append(scores[0]['rouge-l']['f'])


df_tfidf['rouge_1_p'] = rouge_1_p
df_tfidf['rouge_1_r'] = rouge_1_r
df_tfidf['rouge_1_f'] = rouge_1_f
df_tfidf['rouge_L_p'] = rouge_l_p
df_tfidf['rouge_L_r'] = rouge_l_r
df_tfidf['rouge_L_f'] = rouge_l_f


df_tfidf.head()
```

Out[47]:

	summaries	original_summaries	rouge_1_p	rouge_1_r	rouge_1_f	rouge_L_p	rouge_L_r	rouge_L_f
0	Ever noticed how plane seats appear to be getting smaller and smaller? With increasing numbers of people taking to the skies, some experts are questioning if having such packed out planes is putti...	Experts question if packed out planes are putting passengers at risk .\nU.S consumer advisory group says minimum space must be stipulated .\nSafety tests conducted on planes with more leg room th...	0.136364	0.454545	0.209790	0.127273	0.424242	0.195804
1	'We then handed him over to the police.' 'I was drunk and thought I'd stand a good chance.' A police spokesman said: 'He has been cautioned and will be sent for psychiatric evaluation.	Drunk teenage boy climbed into lion enclosure at zoo in west India .\nRahul Kumar, 17, ran towards animals shouting 'Today I kill a lion!'\nFortunately he fell into a moat before reaching lions an...	0.090909	0.088235	0.089552	0.060606	0.058824	0.059701
2	Dougie Freedman is on the verge of agreeing a new two-year deal to remain at Nottingham Forest. Dougie Freedman is set to sign a new deal at Nottingham Forest .	Nottingham Forest are close to extending Dougie Freedman's contract .\nThe Forest boss took over from former manager Stuart Pearce in February .\nFreedman has since lead the club to ninth in the C...	0.300000	0.214286	0.250000	0.250000	0.178571	0.208333
3	Real Madrid? We'll see.	Fiorentina goalkeeper Neto has been linked with Liverpool and Arsenal .\nNeto joined Firoentina from Brazilian outfit Atletico Paranaense in 2011 .\nHe is also wanted by PSG and Spanish clubs, acc...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	Scroll down for video . She filed for divorce in September 2014, citing 'irreconcilable differences'. Reports also emerged over the past week that he has received a breast enhancement. 'Bruce had ...	Tell-all interview with the reality TV star, 69, will air on Friday April 24 .\nIt comes amid continuing speculation about his transition to a woman and following his involvement in a deadly car c...	0.137500	0.224490	0.170543	0.125000	0.204082	0.155039

Rouge-l Recall

ROUGE-L recall is a critical metric in text summarization evaluation because it places a strong emphasis on capturing the longest common subsequence between the generated summary and the reference summary. This is particularly valuable because it:

- Assesses Content Overlap: ROUGE-L recall evaluates how well the generated summary captures the most important content from the reference summary. It ensures that the key information in the original text is retained in the summary.
- Considers Sequencing: It takes into account the order in which words appear in the summaries. This means that not only should the same words be present, but they should also be arranged in a coherent and logical sequence.
- Emphasizes Quality: ROUGE-L recall prioritizes the quality of the summary over its length. It encourages the generation of concise summaries that faithfully represent the source text's essential information.
- Sensitive to Substantial Content: Longer common subsequences are more likely to represent significant content. ROUGE-L recall ensures that the most vital parts of the text are included in the summary, helping to maintain its informativeness.
- Performance Assessment: In the context of summarization, evaluating recall is crucial because it measures how well the system remembers and includes critical information. This is a fundamental aspect of evaluating the overall effectiveness of a summary.

In essence, ROUGE-L recall is a vital metric in text summarization as it assesses content fidelity, structure, and the ability to capture essential information, aligning well with the objectives of text summarization.

In [48]:
```
sorted_df = df_tfidf.sort_values(by=['rouge_1_r'], ascending=False)
sorted_df.reset_index(drop=True)

top_5_rows = sorted_df.head(5)

top_5_rows
```

Out[48]:

	summaries	original_summaries	rouge_1_p	rouge_1_r	rouge_1_f	rouge_L_p	rouge_L_r	rouge_L_f
6981	This incredible video shows the moment a world freediving champion jumped into the world's second deepest underwater sink hole. French free diver jumped into Dean's Blue Hole in the Bahamas. Guill...	French free diver jumped into Dean's Blue Hole in the Bahamas .\nGuillame Néry is seen at the edge before taking the plunge .\nThe hole is 660ft (200 metres) deep, although he doesn't go to the b...	0.360656	1.00000	0.530120	0.360656	1.00000	0.530120
2563	Bangladesh beat fellow World Cup quarter-finalists Pakistan by 79 runs in the first one-day international in Dhaka. Tamim Iqbal and Mushfiqur Rahim scored centuries as Bangladesh made 329 for six ...	Bangladesh beat fellow World Cup quarter-finalists Pakistan by 79 runs .\nTamim Iqbal and Mushfiqur Rahim scored centuries for Bangladesh .\nBangladesh made 329 for six and Pakistan could only mus...	0.389474	1.00000	0.560606	0.389474	1.00000	0.560606
6449	Spanish researchers say climate change impacted human migration. Until 1.4 million years ago it was too cold to inhabit southeast Spain. But then the climate warmed to 13°C (55°F) and became more ...	Spanish researchers say climate change impacted human migration .\nUntil 1.4 million years ago it was too cold to inhabit southeast Spain .\nBut then the climate warmed to 13°C (55°F) and became m...	0.352941	1.00000	0.521739	0.352941	1.00000	0.521739
5013	Genetically engineering plants and crops to change their DNA has been a cause of much controversy in recent years. Scientists in Belgium say all sweet potatoes (stock image shown) contain 'foreign...	Scientists in Belgium say all sweet potatoes contain 'foreign DNA'\nAgrobacterium bacteria in the crop exchanges genes between species .\nThis makes sweet potatoes a 'natural genetically modified ...	0.364583	1.00000	0.534351	0.364583	1.00000	0.534351
663	The common ancestor of humans may have had tentacles, a scientist has claimed. It seemingly puts to bed another theory that suggests our ancestors were much more simple, worm-like creatures. A Rus...	Russian scientist says distant ancestor of humans had tentacles .\nThey lived more than 540 million years ago and used them for food .\nIt's likely they also had a complex nervous system like we d...	0.282759	0.97619	0.438503	0.282759	0.97619	0.438503

Getting average rouge score

In [49]:
```
avg_precision = df_tfidf['rouge_1_p'].mean()
avg_recall = df_tfidf['rouge_1_r'].mean()
avg_f1 = df_tfidf['rouge_1_f'].mean()
avg_l_precision = df_tfidf['rouge_L_p'].mean()
avg_l_recall = df_tfidf['rouge_L_r'].mean()
avg_l_f1 = df_tfidf['rouge_L_f'].mean()
```

In [50]:
```
# Print the results
print("Average ROUGE-1 Precision: ", avg_precision)
print("Average ROUGE-1 Recall: ", avg_recall)
print("Average ROUGE-1 F1: ", avg_f1)
print("Average ROUGE-L Precision: ", avg_l_precision)
print("Average ROUGE-L Recall: ", avg_l_recall)
print("Average ROUGE-L F1: ", avg_l_f1)
```

```
Average ROUGE-1 Precision:  0.1843339164898426
Average ROUGE-1 Recall:  0.3177265870616042
Average ROUGE-1 F1:  0.21824338563616905
Average ROUGE-L Precision:  0.1741664523212486
Average ROUGE-L Recall:  0.3004502518561328
Average ROUGE-L F1:  0.20621636978713345
```

Analysis - Recall

Recall measures the proportion of relevant information that our summarization model successfully captures from the source articles and presents in the generated summaries. In our business case, a higher recall score indicates that our model is effective at retrieving and including important content from news articles in the summaries. This is crucial as it ensures that readers and businesses are well-informed and that the generated summaries are a valuable source of information. Interpretation:

- ROUGE-L Recall Score (0.3002): The ROUGE-L Recall score of 0.3002 indicates that our summarization model captures approximately 30.02% of the content present in the reference "highlights." This signifies that TFIDF model is fairly effective at recalling important content from the reference summaries.
- Information Retrieval: A higher ROUGE-L Recall score suggests that our summarization model is successful at retrieving critical information and core concepts from the source CNN/DailyMail articles. It aligns with our goal of efficiently summarizing news content.
- Enhanced Understanding: The ROUGE-L Recall score suggests that the summaries generated by our model provide a substantial amount of information that is present in the reference "highlights." This is valuable for enhancing the understanding of complex news topics.
- Supporting Market Research: For our goal of enhancing market research, a higher ROUGE-L Recall score is beneficial. It implies that our model is adept at capturing market-relevant information from the source articles, which is essential for tracking news developments and gaining insights into the business environment.

Implications:

- Information Access and Time Efficiency: The high ROUGE-L Recall score indicates that our summarization model efficiently retrieves essential information. This can save time for individuals and businesses as they quickly grasp key points from news articles.
- Improved Understanding: The model's ability to capture content from the "highlights" enhances the accessibility of complex topics. This is beneficial for individuals looking to better understand important news and events.
- Critical Thinking and Broader Perspective: Readers can efficiently scan summaries from multiple sources to gain a well-rounded view of a topic. This promotes critical thinking and a broader perspective, aligning with our business case's goal.
- Market Intelligence and Adaptation: The high ROUGE-L Recall score indicates that businesses can effectively use the generated summaries for competitive intelligence and market research. Staying updated on industry developments and adapting to market changes becomes more feasible.
- Market Research and Content Curation: Summarization supports market research by summarizing customer sentiment, emerging trends, and competitor strategies. Additionally, it facilitates content curation for media companies and content aggregators, ultimately improving user engagement and retention.

Iteration 2 - Seq2Seq (Run in Google Colab)

The next iteration of modeling that was deployed was an RNN Seq2Seq Model using google colab. The use of google colab was necessary due to the high RAM needs of running the model itself. Here is a description of the modeling process that was used in the notebook:

Attention Layer

An attention layer is a critical component in modern neural network architectures, particularly in sequence-to-sequence models, and it plays a fundamental role in enhancing the model's ability to focus on specific parts of input data. Here's an explanation of the concept of an attention layer:

1. Purpose of Attention: The primary purpose of an attention layer is to enable a neural network to "pay attention" to different parts of the input sequence when making predictions. It allows the model to assign varying degrees of importance or relevance to different elements in the input sequence.
2. Sequences and Alignment: Attention layers are commonly used in tasks involving sequences, such as machine translation, text summarization, and speech recognition. When processing sequences, it's important to consider the alignment between the elements of the input and output sequences. The attention mechanism helps in aligning these sequences.
3. Mechanism: The attention mechanism works by computing a weighted sum of the elements in the input sequence, where the weights are dynamically determined based on the context of the current decoding step. In other words, it identifies which parts of the input sequence are most relevant for generating the current output.
4. Calculation of Attention Weights: Attention weights are calculated by comparing the current state of the model (decoder) to each element in the input sequence. These comparisons are often performed using a similarity function, such as dot product or cosine similarity. The results are then transformed into probabilities using a softmax function, making them sum to 1.

Benefits:

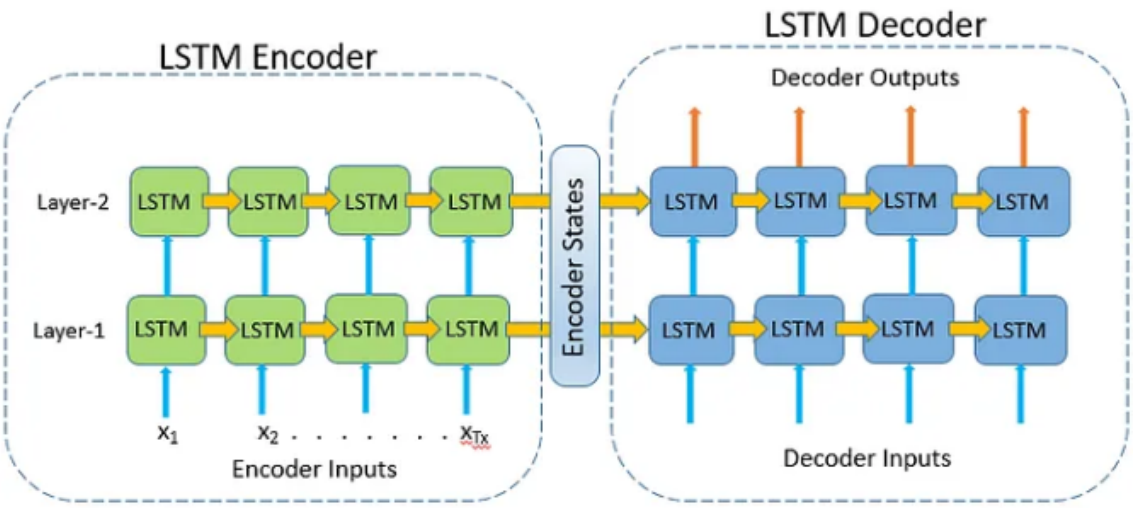
- Enhanced Performance: Attention mechanisms have significantly improved the performance of sequence-to-sequence models in various NLP tasks.
- Handling Variable-Length Sequences: Attention allows the model to handle input sequences of variable length and generate output sequences of varying length.
- Improved Interpretability: Attention mechanisms provide insights into which parts of the input data are influential in making specific predictions.

A Seq2Seq (Sequence-to-Sequence) model is a type of neural network architecture designed for various natural language processing (NLP) tasks, including text summarization. The primary purpose of a Seq2Seq model is to transform input sequences into output sequences of varying lengths. In the context of text summarization, it's used to convert long articles or documents into shorter, coherent summaries.

Here's an overview of the steps involved in preparing, building, and deploying a Seq2Seq model for encoder-decoder text summarization:

Building the Seq2Seq Model - 3 Stacked LSTM

Lets first have a look at a visual of how an LSTM model looks:



Some important terminologies

- `Return Sequences = True` : When the return sequences parameter is set to True, LSTM produces the hidden state and cell state for every timestep

- `Return State = True` : When return state = True, LSTM produces the hidden state and cell state of the last timestep only
- `Initial State` : This is used to initialize the internal states of the LSTM for the first timestep
- `latent_dim` : It denotes the number of hidden units.

The model took over 2 hours to run and was done in google colab due to the high RAM demand. Here is a snapshot of the model summary:

```
Model: "model"
__________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #   Connected to
==================================================================================================
 input_1 (InputLayer)        [(None, 750)]             0         []

 embedding (Embedding)       (None, 750, 100)          3656400   ['input_1[0][0]']

 lstm (LSTM)                 [(None, 750, 300),        481200    ['embedding[0][0]']
                              (None, 300),
                              (None, 300)]

 input_2 (InputLayer)        [(None, None)]            0         []

 lstm_1 (LSTM)               [(None, 750, 300),        721200    ['lstm[0][0]']
                              (None, 300),
                              (None, 300)]

 embedding_1 (Embedding)     (None, None, 100)         697600    ['input_2[0][0]']

 lstm_2 (LSTM)               [(None, 750, 300),        721200    ['lstm_1[0][0]']
                              (None, 300),
                              (None, 300)]

 lstm_3 (LSTM)               [(None, None, 300),       481200    ['embedding_1[0][0]',
                              (None, 300),                        'lstm_2[0][1]',
                              (None, 300)]                        'lstm_2[0][2]']

 attention_layer (Attention  ((None, None, 300),       180300    ['lstm_2[0][0]',
 Layer)                       (None, None, 750))                  'lstm_3[0][0]']

 concat_layer (Concatenate)  (None, None, 600)         0         ['lstm_3[0][0]',
                                                                  'attention_layer[0][0]']

 time_distributed (TimeDist  (None, None, 6976)        4192576   ['concat_layer[0][0]']
 ributed)

==================================================================================================
Total params: 11131676 (42.46 MB)
Trainable params: 11131676 (42.46 MB)
Non-trainable params: 0 (0.00 Byte)
```

And here is a snapshot of the model being compiled:

```
history=model.fit([x_tr,y_tr[:,:-1]], y_tr.reshape(y_tr.shape[0],y_tr.shape[1], 1)[:,1:]
```

```
Epoch 1/15
71/71 [==============================] - 506s 7s/step - loss: 4.5387 - val_loss: 3.9446
Epoch 2/15
71/71 [==============================] - 481s 7s/step - loss: 3.9580 - val_loss: 3.8895
Epoch 3/15
71/71 [==============================] - 481s 7s/step - loss: 3.9185 - val_loss: 3.8623
Epoch 4/15
71/71 [==============================] - 483s 7s/step - loss: 3.9044 - val_loss: 3.8554
Epoch 5/15
71/71 [==============================] - 480s 7s/step - loss: 3.8981 - val_loss: 3.8486
Epoch 6/15
71/71 [==============================] - 480s 7s/step - loss: 3.8929 - val_loss: 3.8466
Epoch 7/15
71/71 [==============================] - 480s 7s/step - loss: 3.8874 - val_loss: 3.8411
Epoch 8/15
71/71 [==============================] - 477s 7s/step - loss: 3.8809 - val_loss: 3.8322
Epoch 9/15
 1/71 [..............................] - ETA: 7:53 - loss: 3.6862
```

Epoch vs Val Loss

Accessing the Notebook

- Due to the absurdly high GPU demand, the RNN model was run in google colab with the help of GPU runtime.

Access the notebook here (https://colab.research.google.com/drive/11kJO7o52TKAnjS2fnDLYp4TWvwXJLRQ_?usp=sharing) and make a copy if you wish to use the notebook.

WARNING this notebook requires a high demand of RAM to run and likely will crash if you do not have RAM up to 40GB. Please just view the notebook if you do not have the processing power to do so.

This is the link (https://github.com/andrewkoji/Capstone_Text_Summarization_Model/blob/main/Seq2Seq_Model_Attn_Layer_text_summarizer.ipynb to the notebook in the github repository if you would like to just view the notebook and its results.

Results

Review: least migrants trying reach europe libya killed boat capsized emerged tonight also claimed body one migrant died another boat making perilous trip north africa italy tossed overboard trafficker circling sharks stories emerged migrants took advantage calm weather cross mediterranean past weekend scroll video rescued italian coast guard involved operation rescue migrants coast sicily april treatment refugee stretchered italian coast guard vessel palermo harbour italy april pregnant woman helped boat sicilian porto empedocle harbor monday italy coast guard helped save migrants monday capsized boat waters libya red cross volunteer carries baby wrapped blanket migrants disembarked sicilian porto empedocle harbor italy monday charity save children said capsized boat carrying people flipped around hours leaving libyan coast survivors mostly sub saharan africa rescued brought italian port monday weekend migrants already died crossing mediterranean far year number sharp rise period said geneva based international organisation migration migrants arrived italy last year packed boats people smugglers yesterday alleged trafficker one boat threw migrant overboard died asphyxiation hold vessel witnesses said body attacked sharks following boat suspected people smuggler guinea arrested sicily rescued along passengers thought also face manslaughter charges february people drowned attempting crossing cold weather rough seas saved rescued migrants onboard italian finance guard boat porto empedocle sicily april migrants arrived italy last year packed boats people smugglers calm weather migrants coast guard boat arrive sicilian porto empedocle harbor italy monday italian coast guard taking part rescue operation coast sicily april migrants took advantage calm weather cross mediterranean past weekend number migrant boats trying reach eu africa risen recent weeks fine weather makes route safer total

rescued since friday reached boats italian coast guard navy ships merchant vessels delivered migrants rescued boats italian ports throughout yesterday pregnant woman board one italian navy ship died journey shore another pregnant migrant eritrea gave birth meanwhile eu frontex border agency said traffickers trying recover boat fired shots air warn away coastguard sparking concerns safety rescue workers migrants save children humanitarian groups called eu bolster rescue operations number boats soars summer usually italian coast guard said monday recovered nine bodies boat carrying migrants sank coast libya number migrant boats trying reach eu africa risen recent weeks fine weather makes route safer total rescued since friday reached boats migrants arrive sicilian porto empedocle harbor italy monday total rescued since friday reached boats migrant wheeled hospital malta flown island woman believed somali italy handles largest number migrant arrivals eu become increasingly alarmed breakdown law order libya home two rival governments loosely aligned militia forces growing militant islamist movement giovanna di save children official assisted new arrivals told times year old nigerian whose brother killed muslim fundamentalists told tortured electric shocks traffickers money travel many talking fleeing fundamentalists dimitris eu commissioner migration said europe must adapt deal rising numbers migrants unprecedented influx migrants borders particular refugees unfortunately new norm said brussels people entered eu illegally last year many travelled syria eritrea somalia libya made perilous journey across sea italy

Original summary: around migrants trying to reach europe died when their boat capsized body of one migrant on another vessel was thrown to sharks stories emerged after more than crossed mediterranean at weekend

Predicted summary: the year old is the in the league in the league the year old is the the the league in the league the year old is the the the league in the league

Rouge Score-l recall: 0.0

As you can see above, the model performed poorly, producing incoherent text that failed to provide any sort of summary to the article. The Rouge score indicates that there is virtually no overlap to the article.

Unfortunately, the model performed as poorly on all the articles and will require some serious fine tuning for improvement. Due to time constraint, and the knowledge of so many state of the art premade models out there, I decided to pivot to what ends up being my final model.

Moving forward

Since we do not forsee ourselves building a model that will match the "state of the art" models that are already built for this task, we will switch to a fine-tuned model that we can pull from huggingface.

Iteration 3 - BART out of the box model

Click here (https://huggingface.co/facebook/bart-large-cnn) for the link to the BART model

BART (large-sized model), fine-tuned on CNN Daily Mail

BART model pre-trained on English language, and fine-tuned on CNN Daily Mail. It was introduced in the paper BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension by Lewis et al. and first released in this repository (https://github.com/pytorch/fairseq/tree/master/examples/bart).

Model description

- BART is a transformer encoder-encoder (seq2seq) model with a bidirectional (BERT-like) encoder and an autoregressive (GPT-like) decoder. BART is pre-trained by
- (1) corrupting text with an arbitrary noising function
- (2) learning a model to reconstruct the original text.
- BART is particularly effective when fine-tuned for text generation (e.g. summarization, translation) but also works well for comprehension tasks (e.g. text classification, question answering). This particular checkpoint has been fine-tuned on CNN Daily Mail, a large collection of text-summary pairs.

In [51]:
```
from transformers import pipeline
```

In [52]:
```
summarizer = pipeline("summarization", model="facebook/bart-large-cnn")

ARTICLE = """Cristiano Ronaldo and Lionel Messi will go head-to-head once more in the race to be this s
             Both Barcelona and Real Madrid booked their spots in the semi-finals this week with victori
             The planet's best footballers have scored eight times in Europe this season. But Shakhtar D
             Cristiano Ronaldo, in action against Atletico Madrid on Wednesday evening, has scored eight
             Lionel Messi also has eight goals in the Champions League this term; one fewer than Luiz Ad
             Ronaldo and Messi will both play at least two more times after Real Madrid and Barcelona re
             Adriano, who moved to Donetsk in 2007, scored five against BATE Borsiov in the group stages
             His performance that night made history, with the 27-year-old becoming only the second play
             The other was Messi for Barcelona against Bayer Leverkusen in 2012. He also scored the thir
             'I am in a good moment of my career and we'll do what will be best for me and for the club,
             Adriano, who netted five times against BATE Borisov in the group, has scored more goals tha
             'With my contract set to expire and many good performances, it'll be difficult to stay in U
             His contract with Shakhtar Donetsk runs out at the end of the year. Ronaldo and Messi howev
             Of the teams still in the pot, Neymar and Luis Suarez of Barcelona, Real Madrid's Karim Ben
             The draw for the last four will take place on Friday."""
print(summarizer(ARTICLE, max_length=130, min_length=75, do_sample=False))
```

```
[{'summary_text': "Cristiano Ronaldo and Lionel Messi have both scored eight Champions League goals th
is season. But Shakhtar Donetsk's Luiz Adriano has netted on nine occasions. Arsenal and Liverpool are
interested in signing the Brazilian. Adriano is out of contract in December and could move to the Prem
ier League. Barcelona and Real Madrid will play in the semi-finals next month. The draw for the last f
our will take place on Friday."}]
```

In [53]:
```
df['highlights'][1000]
```

Out[53]:
```
"Luiz Adriano scored nine times for Shakhtar Donetsk in Europe this season .\nThe Brazilian is out of
contract at the end of the year... both Arsenal and Liverpool are interested in signing the 27-year-ol
d .\nCristiano Ronaldo and Lionel Messi have netted eight goals this season .\nReal Madrid and Barcelo
na both in the Champions League semi-finals .\nREAD: Our reporters have their say on who will win the
Champions League .\nCLICK HERE for Sportsmail's guide to the Champions League final four ."
```

In [54]:
```
df_bart = df[df['articleWordCount'] <= 750]
```

In [55]:
```
import textwrap
```

In [56]:
```
def wrap(x):
    return textwrap.fill(x,replace_whitespace=False,fix_sentence_endings=True)
```

In [57]:
```
def summary_trf(num):
    print('Original Article\n')
    print(wrap(df_bart['article'][num]))
    print('\nSummary')
    result = summarizer(df_bart['article'][num])
    return result[0]['summary_text']
```

In [58]:
```
summary_trf(1)
```

```
Original Article

A drunk teenage boy had to be rescued by security after jumping into a
lions' enclosure at a zoo in western India.  Rahul Kumar, 17,
clambered over the enclosure fence at the Kamla Nehru Zoological Park
in Ahmedabad, and began running towards the animals, shouting he would
'kill them'. Mr Kumar explained afterwards that he was drunk and
'thought I'd stand a good chance' against the predators.  Next level
drunk: Intoxicated Rahul Kumar, 17, climbed into the lions' enclosure
at a zoo in Ahmedabad and began running towards the animals shouting
'Today I kill a lion!'  Mr Kumar had been sitting near the enclosure
when he suddenly made a dash for the lions, surprising zoo security.
The intoxicated teenager ran towards the lions, shouting: 'Today I
kill a lion or a lion kills me!'  A zoo spokesman said: 'Guards had
earlier spotted him close to the enclosure but had no idea he was
planing to enter it.  'Fortunately, there are eight moats to cross
before getting to where the lions usually are and he fell into the
second one, allowing guards to catch up with him and take him out.
'We then handed him over to the police.'  Brave fool: Fortunately, Mr
Kumar  fell into a moat as he ran towards the lions and could be
rescued by zoo security staff before reaching the animals (stock
image) Kumar later explained: 'I don't really know why I did it.  'I
was drunk and thought I'd stand a good chance.'  A police spokesman
said: 'He has been cautioned and will be sent for psychiatric
evaluation.  'Fortunately for him, the lions were asleep and the zoo
guards acted quickly enough to prevent a tragedy similar to that in
Delhi.'  Last year a 20-year-old man was mauled to death by a tiger in
the Indian capital after climbing into its enclosure at the city zoo.

Summary
```

Out[58]:
```
"Rahul Kumar, 17, clambered over enclosure fence at\xa0Kamla Nehru Zoological Park. He ran towards the
animals shouting 'Today I kill a lion or a lion kills me!' Fortunately, Mr Kumar fell into a moat and
was rescued by zoo security. He has been cautioned and will be sent for psychiatric evaluation."
```

In [59]:
```
articles_org = []
highlights_org = []
for i, row in df_bart.iterrows():
    articles_org.append(row['article'])
    highlights_org.append(row['highlights'])
    if i ==5:
        break
```

In [60]:

```
bart_summaries = []
for highlight, article in zip(highlights_org, articles_org):
    print("Predicted Summary: ")
    summary = summarizer(article,max_length=130, min_length=75, do_sample=False)[0]['summary_text']
    print(summarizer(article,max_length=130, min_length=75, do_sample=False)[0]['summary_text'])
    bart_summaries.append(summary)
    print('\n')
    print("Reference Summary: ")
    print(highlight)
    print('\n')
    result = get_rouge_score(summary, highlight)
    print("Rouge-1 Score Recall: ")
    print(result[0]['rouge-1']['r'])
    print('\n')
```

Predicted Summary:
U.S consumer advisory group set up by the Department of Transportation said that while the government is happy to set standards for animals flying on planes, it doesn't stipulate a minimum amount of space for humans. Tests conducted by the FAA use planes with a 31 inch pitch, a standard which on some airlines has decreased. Many economy seats on United Airlines have 30 inches of room, while some airlines offer as little as 28 inches.

Reference Summary:
Experts question if packed out planes are putting passengers at risk .
U.S consumer advisory group says minimum space must be stipulated .
Safety tests conducted on planes with more leg room than airlines offer .

Rouge-l Score Recall:
0.3939393939393939

Predicted Summary:
Rahul Kumar, 17, clambered over enclosure fence at Kamla Nehru Zoological Park. He ran towards the animals shouting 'Today I kill a lion or a lion kills me!' Fortunately, Mr Kumar fell into a moat and was rescued by zoo security. He has been cautioned and will be sent for psychiatric evaluation. Last year a 20-year-old man was mauled to death by a tiger in Delhi.

Reference Summary:
Drunk teenage boy climbed into lion enclosure at zoo in west India .
Rahul Kumar, 17, ran towards animals shouting 'Today I kill a lion!'
Fortunately he fell into a moat before reaching lions and was rescued .

Rouge-l Score Recall:
0.6470588235294118

Predicted Summary:
Dougie Freedman is on the verge of agreeing a new two-year deal. Freedman has stabilised Forest since he replaced Stuart Pearce. Forest made an audacious attempt on the play-off places when Freedman replaced Pearce but have tailed off in recent weeks. That has not prevented Forest's ownership making moves to secure Freedman on a contract for the next two seasons.

Reference Summary:
Nottingham Forest are close to extending Dougie Freedman's contract .
The Forest boss took over from former manager Stuart Pearce in February .
Freedman has since lead the club to ninth in the Championship .

Rouge-l Score Recall:
0.39285714285714285

Predicted Summary:
Fiorentina goalkeeper Neto is wanted by a number of top European clubs, according to his agent Stefano Castagna. Liverpool were linked with a move for the 25-year-old earlier in the season when Simon Mignolet was dropped from the side. A January move for Neto never materialised but the former Atletico Paranaense keeper looks certain to leave the Florence-based club in the summer.

Reference Summary:
Fiorentina goalkeeper Neto has been linked with Liverpool and Arsenal .
Neto joined Firoentina from Brazilian outfit Atletico Paranaense in 2011 .
He is also wanted by PSG and Spanish clubs, according to his agent .
CLICK HERE for the latest Liverpool news .

Rouge-l Score Recall:
0.4864864864864865

```
Predicted Summary:
The former Olympian and reality TV star, 65, will speak in a 'far-ranging' interview with Sawyer for a special edition of '20/20' on Friday April 24. The interview comes amid growing speculation about the father-of-six's transition to a woman, and follows closely behind his involvement in a deadly car crash in California in February. Rumors started swirling around Jenner's gender identity last year, when he emerged from a Beverly Hills clinic with his Adam's apple shaved down. His behavior over the past year also fueled speculation as he began embracing an increasingly female appearance.


Reference Summary:
Tell-all interview with the reality TV star, 69, will air on Friday April 24 .
It comes amid continuing speculation about his transition to a woman and following his involvement in a deadly car crash in February .
The interview will also be one of Diane Sawyer's first appearances on television following the sudden death of her husband last year .


Rouge-l Score Recall:
0.6122448979591837


Predicted Summary:
The prize porker, known as Pigwig, had fallen into the pool in Ringwood, Hampshire. His owners had been taking him for a walk around the garden when the animal plunged into the water and was unable to get out. Two fire crews and a specialist animal rescue team had to use slide boards and strops to haul the huge black pig from the small pool. Firefighters were also called out to rescue a horse which fell into a pool in West Sussex.


Reference Summary:
Giant pig fell into the swimming pool at his home in Ringwood, Hampshire .
It took the efforts of a team of firefighters to winch him out of the water .
A wayward horse also had to be rescued from a swimming pool in Sussex .


Rouge-l Score Recall:
0.5294117647058824
```

Analysis

Predicted Summary 1:

- ROUGE-L Score: 0.3939 This summary does not closely match the reference summary in terms of content and structure. It provides some information about a U.S. consumer advisory group and issues with seat space on planes, but it lacks key details present in the reference summary. The ROUGE-L score is relatively low, indicating poor recall and overlap with the reference summary.

Predicted Summary 2:

- ROUGE-L Score: 0.6471 This summary is a better match to the reference summary, capturing the main incident involving a teenager at a zoo. The ROUGE-L score is relatively high, indicating good recall and overlap with the reference summary.

Predicted Summary 3:

- ROUGE-L Score: 0.3929 This summary provides information about Dougie Freedman's contract extension but lacks some of the key details from the reference summary. The ROUGE-L score is relatively low, indicating poor recall and overlap with the reference summary.

Predicted Summary 4:

- ROUGE-L Score: 0.4865 This summary captures the main idea of Fiorentina goalkeeper Neto being wanted by European clubs but is slightly different from the reference summary. The ROUGE-L score suggests reasonable recall and overlap with the

reference summary. In summary, the model's performance varies across different examples. It performs reasonably well in some cases, closely matching the reference summary, but less accurately in others, leading to lower ROUGE-L scores.

Predicted Summary 5:

- ROUGE-L Score: 0.6122 This summary discusses an upcoming interview with a reality TV star amid speculation about their transition to a woman. It also mentions their involvement in a deadly car crash. While the summary captures the main points, there are some differences from the reference summary, such as the age of the TV star. The ROUGE-L score suggests a reasonably good recall and overlap with the reference summary.

Predicted Summary 6:

- ROUGE-L Score: 0.5294 This summary describes the rescue of a giant pig named Pigwig that fell into a swimming pool in Ringwood, Hampshire. It mentions the involvement of firefighters in rescuing both the pig and a horse that fell into a pool in Sussex. The summary provides a general overview of the events but lacks some details present in the reference summary. The ROUGE-L score suggests moderate recall and overlap with the reference summary

Moving Forward

The model may benefit from further training or fine-tuning to improve its summarization quality, especially for complex or nuanced content. Additionally, it's important to note that ROUGE scores are just one way to evaluate summarization quality, and human evaluation and context-specific criteria are often necessary for a comprehensive assessment of the model's performance

In [ ]: