Text Summarization

A project report submitted to Goa University

for the project of third year

Machine Learning

By

Yogeshwar Manerikar – 1928

Dhanraj Pai Raiturkar – 1940

Salria Pereira -1945

Sneha Valvaikar – 1958

Rashmi Jaiswar - 1962

under the guidance of

Prof. BASKAR SUNDARRAJAN
Department of Computer Science and Technology

GOA UNIVERSITY
Taleigao Plate

Certificate

This project report is the record of the work done by the candidates themselves during the period of study undermy guidance and that it has not previously formed on the basis for the award of any degree in the Goa University or elsewhere.

Mr. S Baskar

(Project Guide)

ABSTRACT

In this new era, where tremendous information is available on the internet, it is most important to provide the improved mechanism to extract the information quickly and most efficiently. It is very difficult for human beings to manually extract the summary of a large documents of text. There are plenty of text material available on the internet. So there is a problem of searching for relevant documents from the number of documents available, and absorbing relevant information from it. In order to solve the above two problems, the text summarization is very much necessary. Text summarization is the process of identifying the most important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meanings.

.

Very difficult for human beings to manually extract the summary of a large documents of text. Text summarization is the process of identifying the most important information in a document or set of related documents, and compressing them into a shorter version preserving its overall meaning.

Text Summarization is the process of creating a condensed form of text document which maintains significant information and general meaning of source text. Automatic text summarization becomes an important way of finding relevant information precisely in large text in a short time with little efforts.

ACKNOWLEDGMENTS

We would like to express our gratitude and appreciation to all those who gave us the possibilities to complete this report. Special thanks to ***professor Baskar Sundarrajan*** whose guidance and experience helped us in difficult times in our project, his stimulating suggestions and encouragement helped us in all time developing this project and writing this report. We also sincerely thank for the time spent proofreading and correcting our many mistakes.

Text summarization wit best accuracy with the cricket domain is challenge . project journey we get help from the ***Mrs. Siddhi bhandankar*** . She gives the explanation for every steps and motivation . Specially thankful for ***Mr. Ankit Mishra*** he is our junior and taking the interest in our project and giving the ideas .

We would also like to thank department of computer science, Goa University for giving us the opportunity to work with a project that will surely add to our achievements.

TABLE OF CONTENTS

INTRODUCTION

1) Text Summarization

The amount of information that is being generated in today‟s world is increasing tremendously with each day. There has literally been an explosion in the amount of information being generated in electronic form and exposed through the World Wide Web. This is especially true in the field of research and development where the frequency of the number of papers and articles being published is augmenting everyday. This presents the need for users to be able to skim through several different documents and quickly find the information that they are searching for. Summaries can help enormously in achieving this objective.

A summary refers to an abridged or a condensed version of a document. It is a concise and brief description of the original document outlining the most important points contained within it thereby removing the need to have to read the full text.

Some of the situations in which summaries are of enormous help are described below. Summaries are useful while reading news articles. Summaries of news articles help readers browse through the most important aspects of the article instead of having to read the full-length article. Research papers, we often have to go through many research papers and the most intuitive way to sort them would be by reading the abstract and the conclusion both of which represent the summary. Hence, summaries of documents and papers help tremendously while conducting research.

Hence, summarization systems prove to be very useful tools in various situations.

In this project we will be using the concept of **LSTM** and build a modelof text summarization which involves the concept of Natural Language Process to summary the articles.

2) Objectives

The objective of our project is to understand the concepts of a LSTM model and build a working model for text summarization by implementing LSTM.

In this project, we will be implementing the text summarization using LSTM (Long short term memory). We have collected the news content from inshorts from the sports category which is cricket related., generated summaries of articles using pretrained pipeline models , preprocessed, and then we feed the articles and summaries into the LSTM model which will be responsible for generating the summary.

3) Motivation

We must first understand how important this problem is to real world scenarios.
Some of the reasons for motivation of text summarization are as follows:

- To keep up with the world affairs by listening to news.
- People base investment decisions on stock market updates.
- People even go to movies largely on the basis of reviews they've seen.
- With summaries, People can make effective decisions in less time.

The motivation here is to build such tool which is computationally efficient and creates summaries automatically.

Domain

1) Sports

We have chosen “Sports” as our project domain. Sport pertains to any form of competitive physical activity or game that aims to use, maintain or improve physical ability and skills while providing enjoyment to participants and, in some cases, entertainment to spectators. Sports can, through casual or organized participation, improve one's physical health. Hundreds of sports exist, from those between single contestants, through to those with hundreds of simultaneous participants, either in teams or competing as individuals. In certain sports such as racing, many contestants may compete, simultaneously or consecutively, with one winner; in others, the contest (a *match*) is between two sides, each attempting to exceed the other. Some sports allow a "tie" or "draw", in which there is no single winner; others provide tie-breaking records to ensure one winner and one loser. A number of contests may be arranged in a tournament producing a champion. Many sports leagues make an annual champion by arranging games in a regular sports season, followed in some cases by playoffs.

2) Cricket

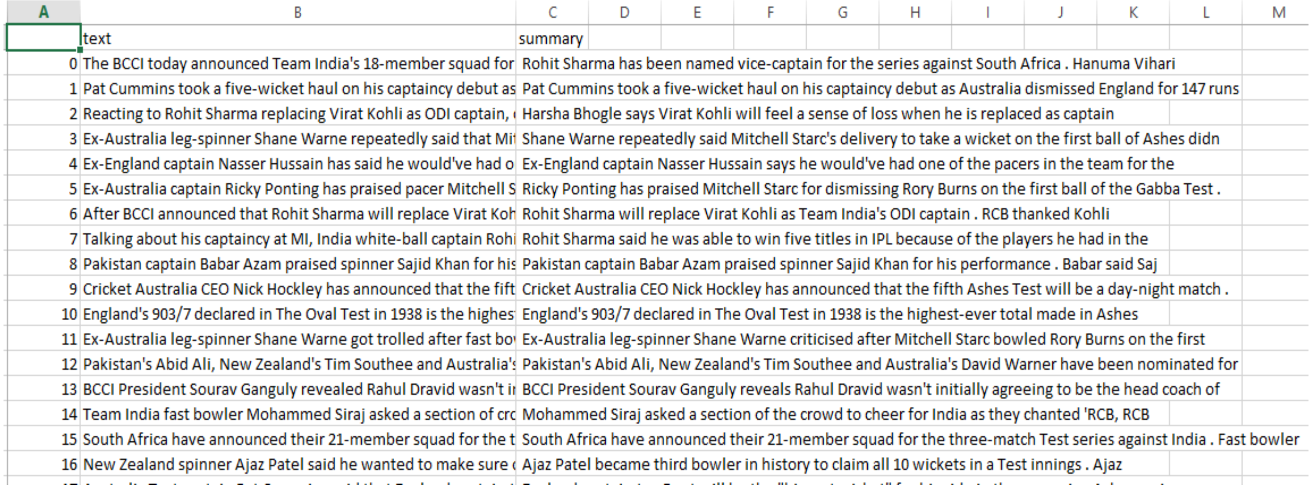
In Sports domain we have chosen Cricket Game News Articles as our data. Cricket is bat-and-ball game played between two teams of eleven players on a field at the Centre of which is a 22-yard (20-metre) pitch with a wicket at each end, each comprising two bails balanced on three stumps. The game proceeds when a player on the fielding team, called the bowler, "bowls" (propels) the ball from one end of the pitch towards the wicket at the other end. The batting side’s players score runs by striking the bowled ball with a bat and running between the wickets, while the fielding side tries to prevent this by keeping the ball within the field and getting it to either wicket, and also tries to dismiss each batter (so they are "out"). Means of dismissal include being bowled, when the ball hits the stumps and dislodges the bails, and by the fielding side either catching a hit ball before it touches the ground, or hitting a wicket with the ball before a batter can cross the crease line in front of the wicket to complete a run. When ten batters have been dismissed, the innings ends and the teams swap roles. The game is adjudicated by two umpires, aided by a third umpire and match referee in international matches.

Data Set

The data being used for this project is specifically belonging to cricket domain. Data used have been gathered from a news website called "inshorts". We have collected the news content from inshorts from the sports category which is cricket related.

We have made a csv file containing rows of cricket news as our text column. In order to obtain a summary for this data we have used the pertained pipeline model. We have passed the csv file containing text column and created an array of summary for the corresponding text.

We have further combined the text column and summary array into a new pandas data frame, which we have used as our csv dataset to train and validate our model.



A	B	C
	text	summary
0	The BCCI today announced Team India's 18-member squad for	Rohit Sharma has been named vice-captain for the series against South Africa . Hanuma Vihari
1	Pat Cummins took a five-wicket haul on his captaincy debut as	Pat Cummins took a five-wicket haul on his captaincy debut as Australia dismissed England for 147 runs
2	Reacting to Rohit Sharma replacing Virat Kohli as ODI captain,	Harsha Bhogle says Virat Kohli will feel a sense of loss when he is replaced as captain
3	Ex-Australia leg-spinner Shane Warne repeatedly said that Mi	Shane Warne repeatedly said Mitchell Starc's delivery to take a wicket on the first ball of Ashes didn
4	Ex-England captain Nasser Hussain has said he would've had o	Ex-England captain Nasser Hussain says he would've had one of the pacers in the team for the
5	Ex-Australia captain Ricky Ponting has praised pacer Mitchell S	Ricky Ponting has praised Mitchell Starc for dismissing Rory Burns on the first ball of the Gabba Test .
6	After BCCI announced that Rohit Sharma will replace Virat Koh	Rohit Sharma will replace Virat Kohli as Team India's ODI captain . RCB thanked Kohli
7	Talking about his captaincy at MI, India white-ball captain Rohi	Rohit Sharma said he was able to win five titles in IPL because of the players he had in the
8	Pakistan captain Babar Azam praised spinner Sajid Khan for his	Pakistan captain Babar Azam praised spinner Sajid Khan for his performance . Babar said Saj
9	Cricket Australia CEO Nick Hockley has announced that the fift	Cricket Australia CEO Nick Hockley has announced that the fifth Ashes Test will be a day-night match .
10	England's 903/7 declared in The Oval Test in 1938 is the highes	England's 903/7 declared in The Oval Test in 1938 is the highest-ever total made in Ashes
11	Ex-Australia leg-spinner Shane Warne got trolled after fast bo	Ex-Australia leg-spinner Shane Warne criticised after Mitchell Starc bowled Rory Burns on the first
12	Pakistan's Abid Ali, New Zealand's Tim Southee and Australia's	Pakistan's Abid Ali, New Zealand's Tim Southee and Australia's David Warner have been nominated for
13	BCCI President Sourav Ganguly revealed Rahul Dravid wasn't i	BCCI President Sourav Ganguly reveals Rahul Dravid wasn't initially agreeing to be the head coach of
14	Team India fast bowler Mohammed Siraj asked a section of cro	Mohammed Siraj asked a section of the crowd to cheer for India as they chanted 'RCB, RCB
15	South Africa have announced their 21-member squad for the t	South Africa have announced their 21-member squad for the three-match Test series against India . Fast bowler
16	New Zealand spinner Ajaz Patel said he wanted to make sure	Ajaz Patel became third bowler in history to claim all 10 wickets in a Test innings . Ajaz

Scrapping data from:-

Crickbuize https://www.cricbuzz.com/cricket-news/latest-news
Times of India
Inshorts
ND TV https://sports.ndtv.com/cricket/news

Tools & Libraries

1) Pandas:

pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software.

2) Numpy:

NumPy is a Python library used for working with arrays.
It also has functions for working in domain of linear algebra, fourier transform, and matrices.
In Python we have lists that serve the purpose of arrays, but they are slow to process.
NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.
The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy.
Arrays are very frequently used in data science, where speed and resources are very important.

3) Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.

4) Pyplot:

pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

5) Sklearn:

Scikit-learn (formerly scikits.learn andalsoknownas sklearn)isafree software machinelearning library for the Python programming language.[3] It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Scikit- learn is a NumFOCUS fiscally sponsored project.

6) Train_test_split:

Split arrays or matrices into random train and test subsets.
In machine learning, Train Test split activity is done to measure the performance of the machine learning algorithm when they are used to predict the new data which is not used to train the model.

7) Keras:

Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library.

8) Tokenizer:

This class allows to vectorize a text corpus, by turning each text into either a sequence of integers (each integer being the index of a token in a dictionary) or into a vector where the coefficient for each token could be binary, based on word count, based on tf-idf

9) Pad_sequences:

pad_sequences is used to ensure that all sequences in a list have the same length. By default this is done by padding 0 in the beginning of each sequence until each sequence has the same length as the longest sequence.

10) Layers:

Layers are the basic building blocks of neural networks in Keras. A layer consists of a tensor-in tensor-out computation function (the layer's call method) and some state, held in TensorFlow variables (the layer's weights). A Layer instance is callable, much like a function.

11) Model

Model groups layers into an object with training and inference features.

Arguments

- inputs: The input(s) of the model: a keras.Input object or list of keras.Input objects.
- outputs: The output(s) of the model. See Functional API example below.
- name: String, the name of the model.

12) Earlystopping:

Assuming the goal of a training is to minimize the loss. With this, the metric to be monitored would be 'loss', and mode would be 'min'. A model.fit() training loop will check at end of every epoch whether the loss is no longer decreasing, considering the min_delta and patience if applicable. Once it's found no longer decreasing, model.stop_training is marked True and the training terminates.

IMPLIMENTATION

Collection of data

Cleaning of text

- This is all for cleaning the text.
- We want the stop words for summery so we are not removing

Lambda x

- In the summary column adding the two words
- Reduction is a lambda calculus strategy to compute the value of the expression. In the current example, it consists of adding the bounds

Open word:- sostok

End of the word :- esotok

Counting the total words

- # Model to summarize the text between 0-100 words for Summary and 0-300 words for Text

Splitting the data

Imparting the train test split rom sklearn

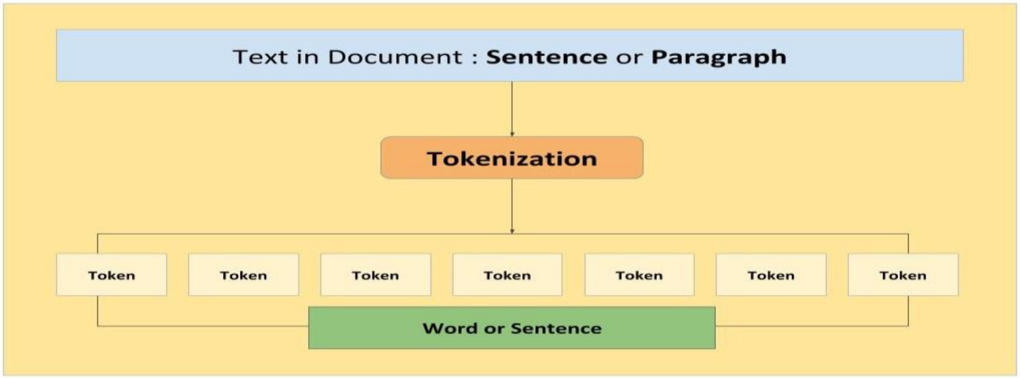
- Test size 0.1 because we have less data to train model

  So 90 train and 10 test
- The random state that you provide is used as a seed to the random number generator. This ensures that the random numbers are generated in the same order

Passing the numpy array to split

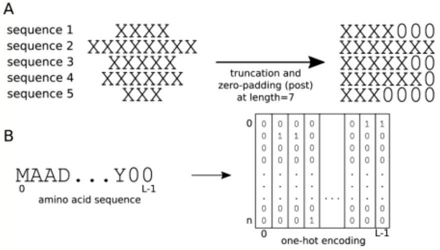
- **Pad_sequnce** is used to ensure that all sequences in a list have the same length. By default this is done by padding 0 in the beginning of each sequence until each sequence has the same length as the longest sequence.

- **Tokenizer** Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms



Using padding='post'

Adding zeros at last to be the same size



Remove empty Summaries

- .i.e, which only have 'START' and 'END' tokens

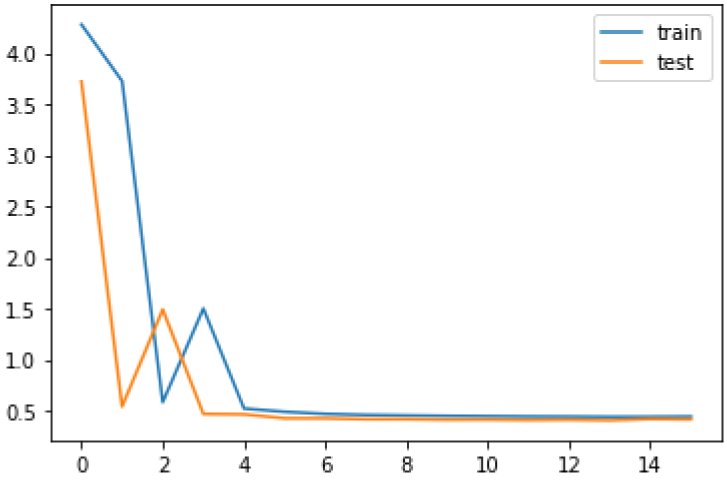
Importing the model

- LSTM
- Embedding
- Dense
- Input

Train the model

- We are using three encoders to get good output
- latent_dim = 300
- embedding_dim = 200

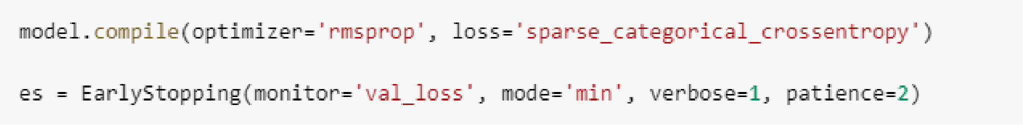
result for lstm



Decoder:-

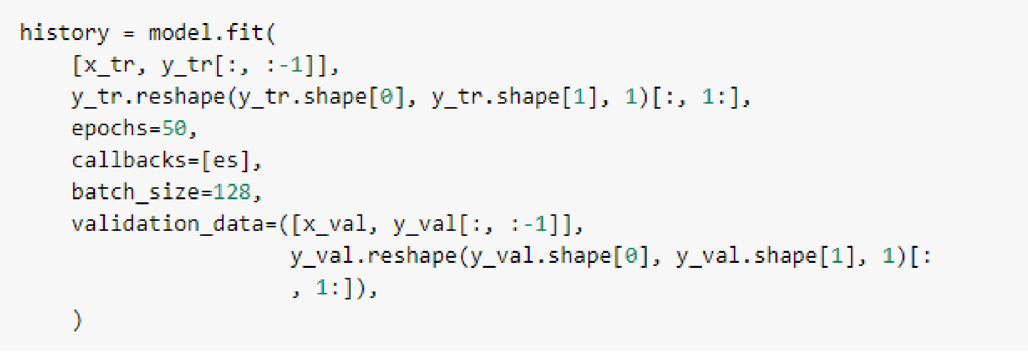
- Encode the input as state vectors.
- Generate empty target sequence of length 1
- Populate the first word of target sequence with the start word. “sostok”
- Exit condition: either hit max length or find the stop word.” Eostok”

Model Training



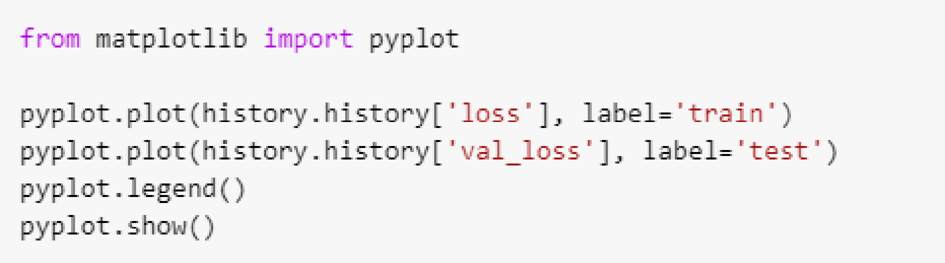
```
model.compile(optimizer='rmsprop', loss='sparse_categorical_crossentropy')

es = EarlyStopping(monitor='val_loss', mode='min', verbose=1, patience=2)
```

In this above step, we have compile the model and define EarlyStopping to stop training the model once the validation loss metric has stopped decreasing.



```
history = model.fit(
    [x_tr, y_tr[:, :-1]],
    y_tr.reshape(y_tr.shape[0], y_tr.shape[1], 1)[:, 1:],
    epochs=50,
    callbacks=[es],
    batch_size=128,
    validation_data=([x_val, y_val[:, :-1]],
                     y_val.reshape(y_val.shape[0], y_val.shape[1], 1)[:
                     , 1:]),
    )
```

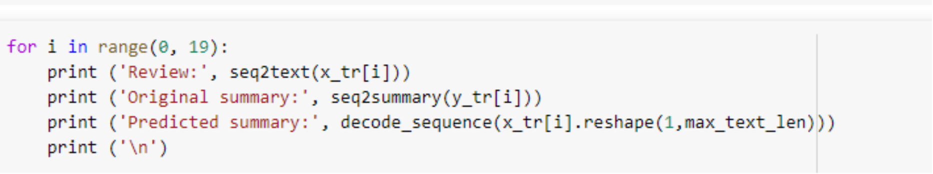
Next, we have use the model.fit() method to fit the training data where we can define the batch size to be 128. Send the text and summary (excluding the last word in summary) as the input, and a reshaped summary tensor comprising every word (starting from the second word) as the output. Besides, this to enable validation during the training phase, we have send the validation data as well.



```
from matplotlib import pyplot

pyplot.plot(history.history['loss'], label='train')
pyplot.plot(history.history['val_loss'], label='test')
pyplot.legend()
pyplot.show()
```

Here we have plot the training and validation loss metrics observed during the training.

Performance

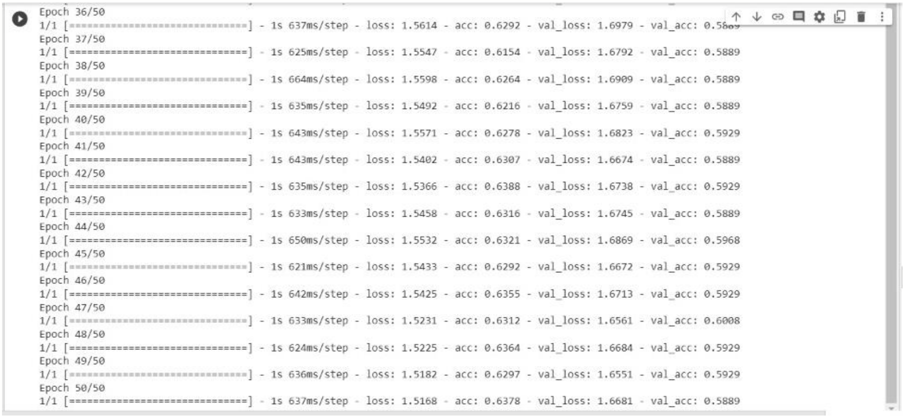


```
for i in range(0, 19):
    print ('Review:', seq2text(x_tr[i]))
    print ('Original summary:', seq2summary(y_tr[i]))
    print ('Predicted summary:', decode_sequence(x_tr[i].reshape(1,max_text_len)))
    print ('\n')
```

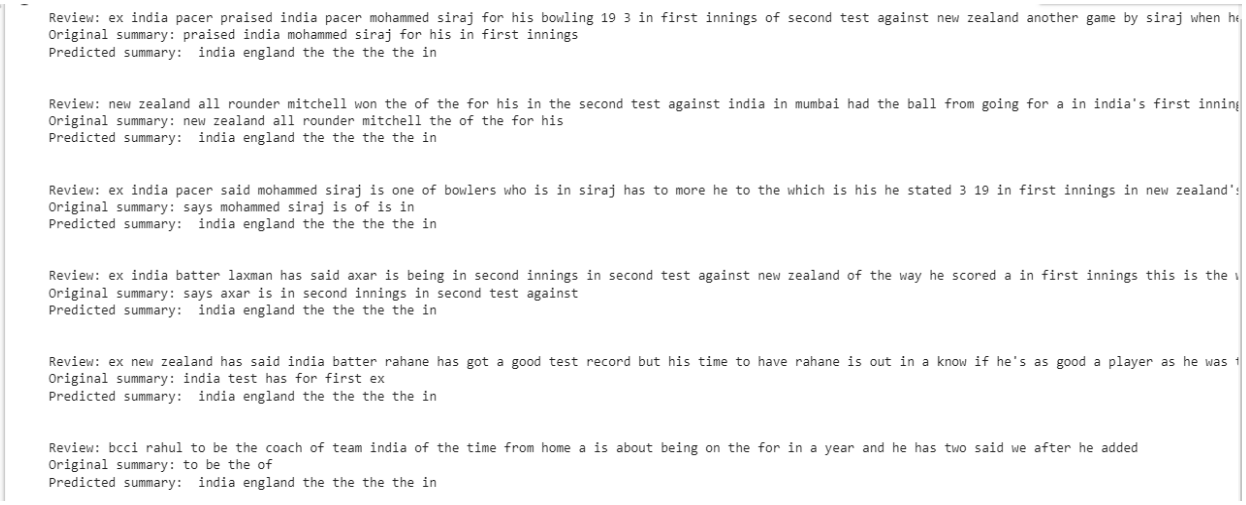
Generated the predictions by sending in the text

Evaluation matrix:-

We get 0.2 loss function



```
Epoch 36/50
1/1 [==============================] - 1s 637ms/step - loss: 1.5614 - acc: 0.6292 - val_loss: 1.6979 - val_acc: 0.5889
Epoch 37/50
1/1 [==============================] - 1s 625ms/step - loss: 1.5547 - acc: 0.6154 - val_loss: 1.6792 - val_acc: 0.5889
Epoch 38/50
1/1 [==============================] - 1s 664ms/step - loss: 1.5598 - acc: 0.6264 - val_loss: 1.6909 - val_acc: 0.5889
Epoch 39/50
1/1 [==============================] - 1s 635ms/step - loss: 1.5492 - acc: 0.6216 - val_loss: 1.6759 - val_acc: 0.5889
Epoch 40/50
1/1 [==============================] - 1s 643ms/step - loss: 1.5571 - acc: 0.6278 - val_loss: 1.6823 - val_acc: 0.5929
Epoch 41/50
1/1 [==============================] - 1s 643ms/step - loss: 1.5402 - acc: 0.6307 - val_loss: 1.6674 - val_acc: 0.5889
Epoch 42/50
1/1 [==============================] - 1s 635ms/step - loss: 1.5366 - acc: 0.6388 - val_loss: 1.6738 - val_acc: 0.5929
Epoch 43/50
1/1 [==============================] - 1s 633ms/step - loss: 1.5458 - acc: 0.6316 - val_loss: 1.6745 - val_acc: 0.5889
Epoch 44/50
1/1 [==============================] - 1s 650ms/step - loss: 1.5532 - acc: 0.6321 - val_loss: 1.6869 - val_acc: 0.5968
Epoch 45/50
1/1 [==============================] - 1s 621ms/step - loss: 1.5433 - acc: 0.6292 - val_loss: 1.6672 - val_acc: 0.5929
Epoch 46/50
1/1 [==============================] - 1s 642ms/step - loss: 1.5425 - acc: 0.6355 - val_loss: 1.6713 - val_acc: 0.5929
Epoch 47/50
1/1 [==============================] - 1s 633ms/step - loss: 1.5231 - acc: 0.6312 - val_loss: 1.6561 - val_acc: 0.6008
Epoch 48/50
1/1 [==============================] - 1s 624ms/step - loss: 1.5225 - acc: 0.6364 - val_loss: 1.6684 - val_acc: 0.5929
Epoch 49/50
1/1 [==============================] - 1s 636ms/step - loss: 1.5182 - acc: 0.6297 - val_loss: 1.6551 - val_acc: 0.5929
Epoch 50/50
1/1 [==============================] - 1s 637ms/step - loss: 1.5168 - acc: 0.6378 - val_loss: 1.6681 - val_acc: 0.5889
```



```
Review: ex india pacer praised india pacer mohammed siraj for his bowling 19 3 in first innings of second test against new zealand another game by siraj when he
Original summary: praised india mohammed siraj for his in first innings
Predicted summary:  india england the the the the in


Review: new zealand all rounder mitchell won the of the for his in the second test against india in mumbai had the ball from going for a in india's first inning
Original summary: new zealand all rounder mitchell the of the for his
Predicted summary:  india england the the the the in


Review: ex india pacer said mohammed siraj is one of bowlers who is in siraj has to more he to the which is his he stated 3 19 in first innings in new zealand's
Original summary: says mohammed siraj is of is in
Predicted summary:  india england the the the the in


Review: ex india batter laxman has said axar is being in second innings in second test against new zealand of the way he scored a in first innings this is the
Original summary: says axar is in second innings in second test against
Predicted summary:  india england the the the the in


Review: ex new zealand has said india batter rahane has got a good test record but his time to have rahane is out in a know if he's as good a player as he was
Original summary: india test has for first ex
Predicted summary:  india england the the the the in


Review: bcci rahul to be the coach of team india of the time from home a is about being on the for in a year and he has two said we after he added
Original summary: to be the of
Predicted summary:  india england the the the the in
```

Here are a few notable summaries generated.

Limitation to existing system

1. Existing system does not contain large set of data.
2. Missing on attention file.
3. Existing system doesn’t give proper accuracy score.

Future Scope

1. Design a UI

   Where one can enter text in the textbox and on click of a button it will summarize the text.

2. Include the whole sports domain

   For the existing system we have just considered Cricket as a part of sport domain, later on other sports like football, volleyball, tennis etc. can be included

3. In future we can use GPT2 for better and fast progress.

Conclusion

As with time internet is growing at a very fast rate and with-it data and information is also increasing. It will be difficult for human to summarize large amount of data. Thus, there is a need of automatic text summarization because of this huge amount of data.

We have made a basic automation text summarization. We have learned all the basics of machine learning and tried to implement abstractive method using LSTM

REFERENCE

1. Text summarization in NLP using spacy | Python review
   https://youtu.be/5mY6a3QbIXM

2. Tokenization
   https://youtu.be/5mY6a3QbIXM

3. research paper
   http://irgu.unigoa.ac.in/drs/handle/unigoa/4463

   http://irgu.unigoa.ac.in/drs/handle/unigoa/5965
4. Aproch towords topic
   https://youtu.be/-wcrzwNmiAU
5. kaggle
   https://www.kaggle.com/sandeepbhogaraju/text-summarization-with-seq2seq-model
6. Key words for text
   https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/

7. Row data to clean data NLTK
   https://machinelearningmastery.com/clean-text-machine-learning-python/

8. Steps
   https://youtu.be/XO97Uon83Os

9. models
   https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/

   https://github.com/DeepsMoseli/Bidirectiona-LSTM-for-text-summarization-/blob/master/lstm_Attention.py

   https://medium.com/analytics-vidhya/seq2seq-abstractive-summarization-using-lstm-and-attention-mechanism-code-da2e9c439711

   https://www.kaggle.com/singhabhiiitkgp/text-summarization-using-lstm