NEWS HEADLINE GENERATION

"Exploring Creative Approaches for Accurate Headline Generation"

MODEL DESCRIPTION

OUR model starts by utilizing Named Entity Recognition (NER) and Latent Dirichlet Allocation (LDA) to extract entities and identify topics from news articles obtained from preprocessed data done by us. These techniques help in understanding the content and themes of the articles.

Next, Word2Vec embeddings are applied to capture semantic relationships between words in the text. This embedding allows the model to grasp the context and meaning of words, enhancing the headline generation process.

The core component of our model is an LSTM neural network, which is trained on the preprocessed news articles. The LSTM learns patterns and dependencies in the sequential data of the articles to generate headlines that accurately summarize the content.

To evaluate the quality of the generated headlines, our model utilizes ROUGE BLEU metrics. These metrics compare the generated headlines with reference headlines to assess their similarity and informativeness. Additionally, cosine similarity is employed to measure the similarity between vectors, aiding in comparing the generated headlines with reference headlines or different parts of the news articles.

DATASET

Data Loading and Preprocessing:

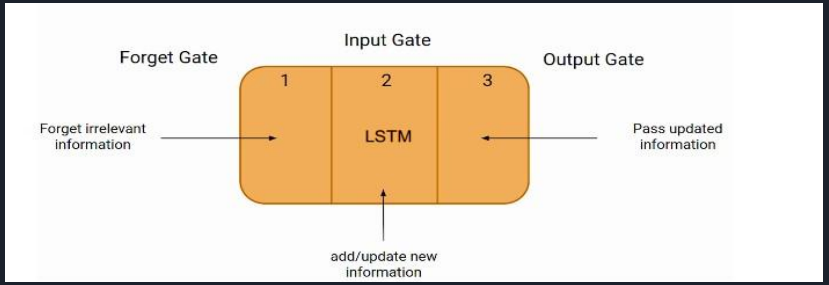
Loads the CNN/DailyMail dataset consisting of news articles and headlines in .story format

Preprocessed the text data by tokenizing and cleaning it.

```
1 df
```

	article	entities	tokens	topic
0	It's official: U.S. President Barack Obama wan...	[(U.S., GPE), (Barack Obama, PERSON), (Syria, ...	[it, 's, official, :, u.s., president, barack,...	3
1	(CNN) -- Usain Bolt rounded off the world cham...	[(CNN, ORG), (Usain Bolt, ORG), (Sunday, DATE)...	[(, cnn,), --, usain, bolt, rounded, off, the...	3
2	Kansas City, Missouri (CNN) -- The General Ser...	[(Kansas City, GPE), (Missouri, GPE), (CNN, OR...	[kansas, city, ,, missouri, (, cnn,), --, the...	3
3	Los Angeles (CNN) -- A medical doctor in Vanco...	[(Los Angeles, GPE), (CNN, ORG), (Vancouver, G...	[los, angeles, (, cnn,), --, a, medical, doct...	3
4	(CNN) -- Police arrested another teen Thursday...	[(CNN, ORG), (Thursday, DATE), (sixth, ORDINAL...	[(, cnn,), --, police, arrested, another, tee...	9
...	...	...	...	...
971	(CNN) -- Lake Urmia in Iran used to be a site ...	[(CNN, ORG), (Lake Urmia, FAC), (Iran, GPE), (...	[(, cnn,), --, lake, urmia, in, iran, used, t...	5
972	New York (CNN) -- New York Mayor Michael Bloom...	[(New York, GPE), (CNN, ORG), (New York, GPE),...	[new, york, (, cnn,), --, new, york, mayor, m...	3
973	(CNN) -- Within minutes of the news of his dea...	[(CNN, ORG), (minutes, TIME), (Twitter, PERSON...	[(, cnn,), --, within, minutes, of, the, news...	4
974	The national controversy over a surge of Centr...	[(Central American, NORP), (U.S., GPE), (this ...	[the, national, controversy, over, a, surge, o...	3
975	(CNN) -- Two U.S. spacecraft are set to crash ...	[(CNN, ORG), (Two, CARDINAL), (U.S., GPE), (Fr...	[(, cnn,), --, two, u.s., spacecraft, are, se...	8

MODEL DESCRIPTION

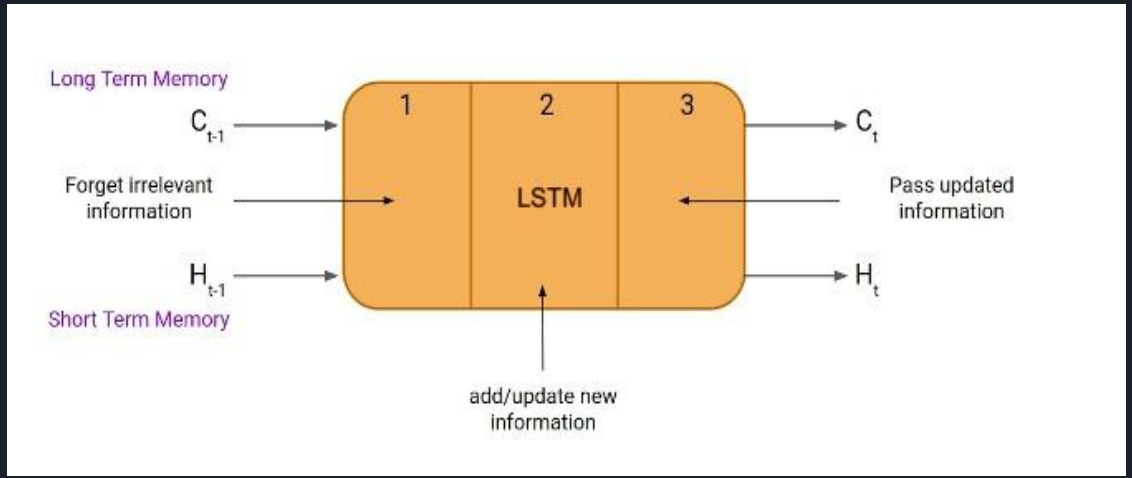


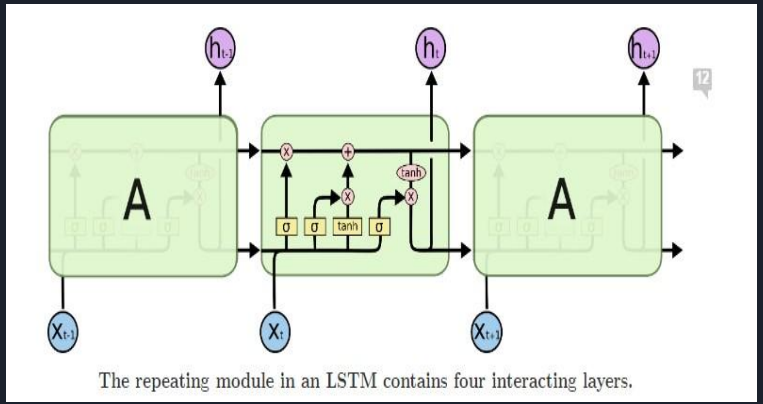
LSTM Model for Headline Generation:

Constructed an LSTM-based neural network model for headline generation.

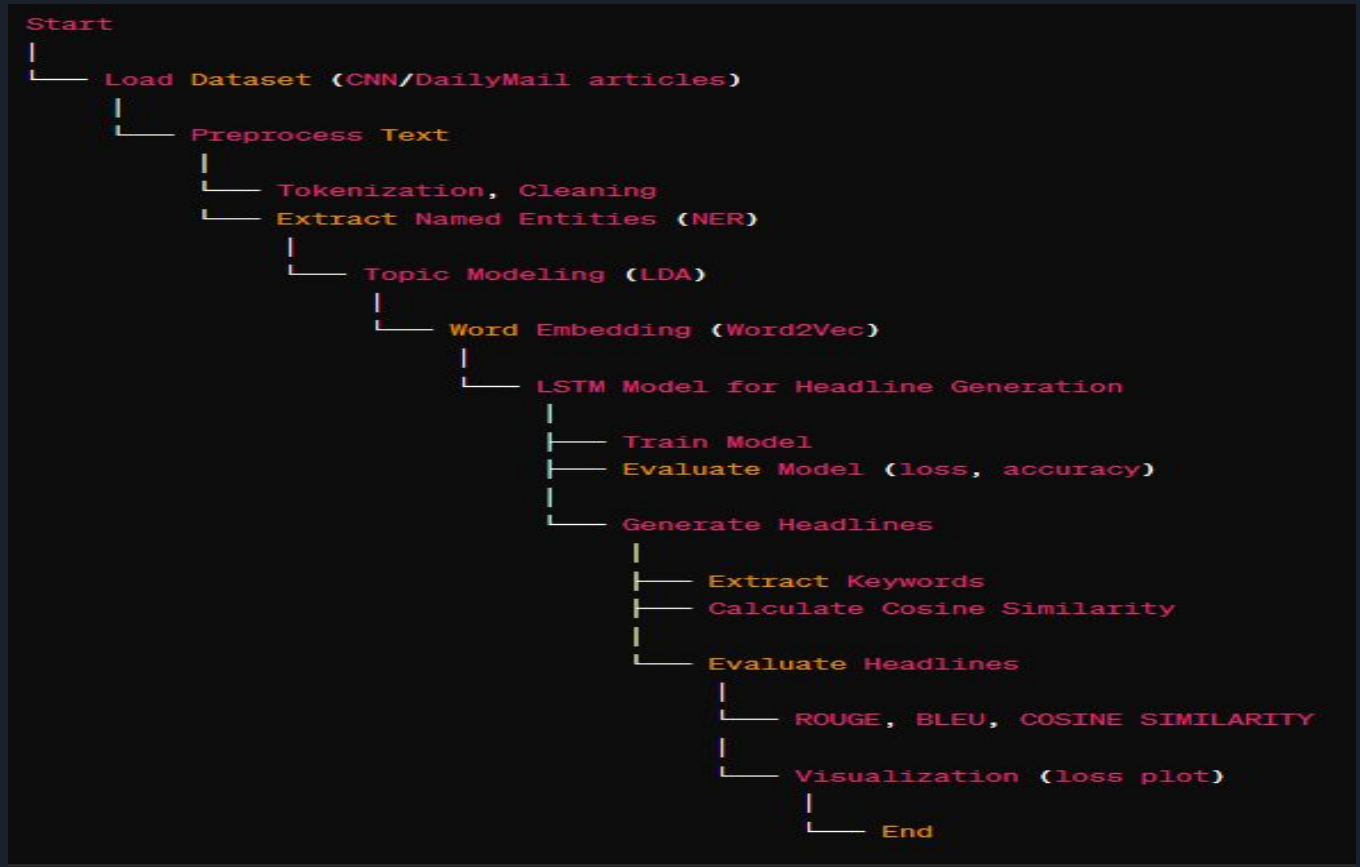
Trains the model using the preprocessed dataset.

Implements cosine similarity for keyword-based similarity between texts.





The repeating module in an LSTM contains four interacting layers.


Start
Load Dataset (CNN/DailyMail articles)
Preprocess Text
Tokenization, Cleaning
Extract Named Entities (NER)
Topic Modeling (LDA)
Word Embedding (Word2Vec)
LSTM Model for Headline Generation
Train Model
Evaluate Model (loss, accuracy)
Generate Headlines
Extract Keywords
Calculate Cosine Similarity
Evaluate Headlines
ROUGE, BLEU, COSINE SIMILARITY
Visualization (loss plot)
End

WHY ‘S ?

LDA: Discovers abstract topics in documents, informing relevant headline generation.

NER: Identifies entities in news articles such as place ,person, organization enriching headlines with informative content.

Word2Vec: Embeds words, aiding LSTM in understanding context for headline generation as it captures semantic relationships.

LSTM: Captures long-term dependencies for headline generation from news articles.

ROUGE: Evaluates headline quality by comparing generated and reference headlines.measures the overlap between the generated headlines and the reference headlines in terms of n-gram overlap, providing insights into the informativeness and fluency of the generated headlines.

BLEU: Measures headline similarity to reference headlines based on n-gram precision rewarding the generation of n-grams present in the reference headlines.

Cosine Similarity: Measures similarity between vectors, useful for comparing headlines or text parts.