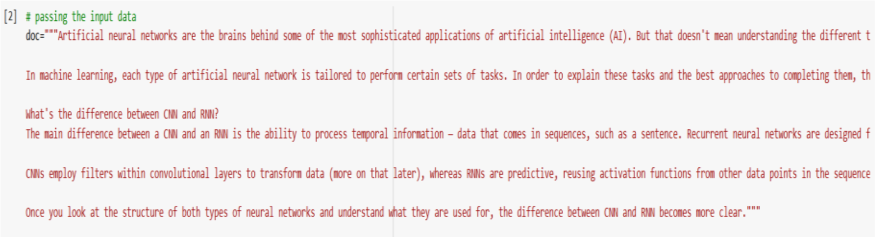
Data Collection and Preprocessing Phase

Date	11 March 2025
Team ID	740052
Project Title	AI-Based Intelligent Insight Extractor
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template:

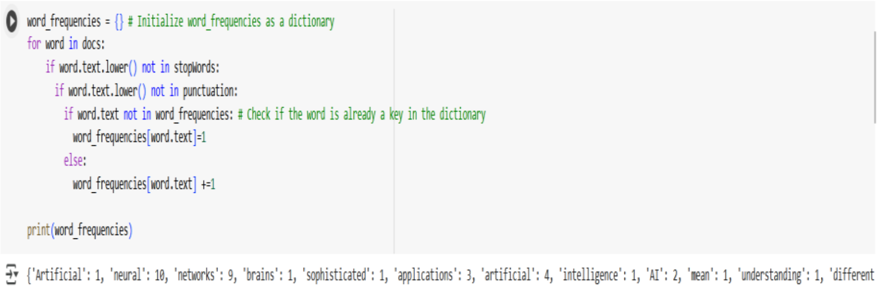
Data exploration and preprocessing for a text summarization project begin with thorough exploratory data analysis (EDA) to understand the structure, distribution, and relationships within the textual data. Handling missing or incomplete text entries is crucial, often addressed through techniques like data imputation or removal. Text data is cleaned and normalized through processes such as lowercasing, removing stop words, and stemming or lemmatization. Tokenization and vectorization (e.g., TF-IDF or word embeddings) convert text into numerical formats suitable for machine learning models. These steps ensure the dataset is well-prepared for training accurate summarization models.

<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>



```
[2] # passing the input data
    doc="""Artificial neural networks are the brains behind some of the most sophisticated applications of artificial intelligence (AI). But that doesn't mean understanding the different t

    In machine learning, each type of artificial neural network is tailored to perform certain sets of tasks. In order to explain these tasks and the best approaches to completing them, th

    What's the difference between CNN and RNN?
    The main difference between a CNN and an RNN is the ability to process temporal information – data that comes in sequences, such as a sentence. Recurrent neural networks are designed f

    CNNs employ filters within convolutional layers to transform data (more on that later), whereas RNNs are predictive, reusing activation functions from other data points in the sequence

    Once you look at the structure of both types of neural networks and understand what they are used for, the difference between CNN and RNN becomes more clear."""
```

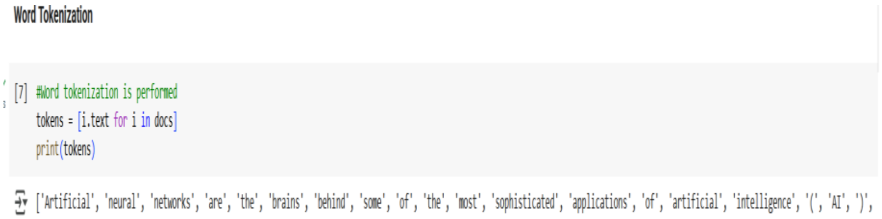
</td></tr>
<tr><td>Word Frequencies</td><td>

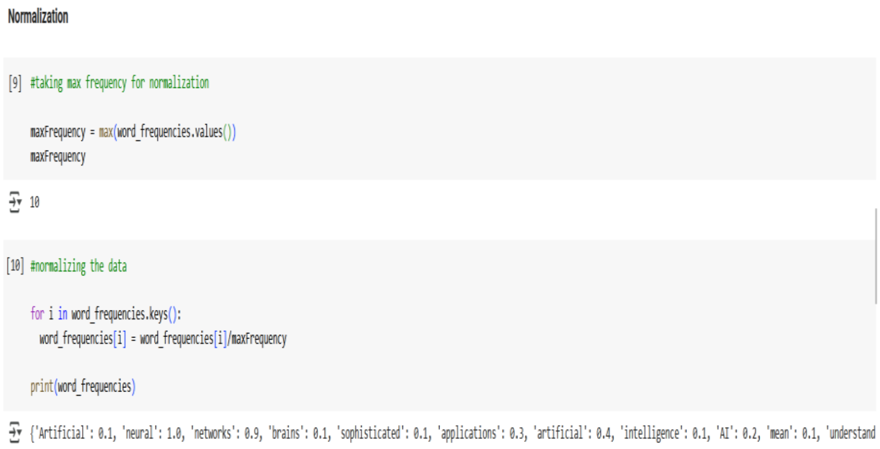


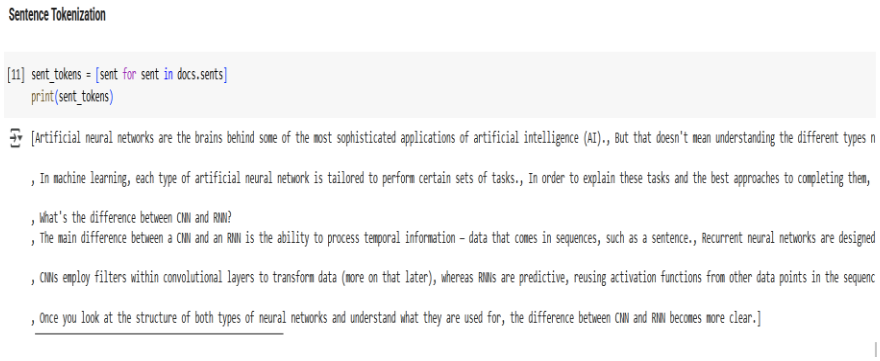
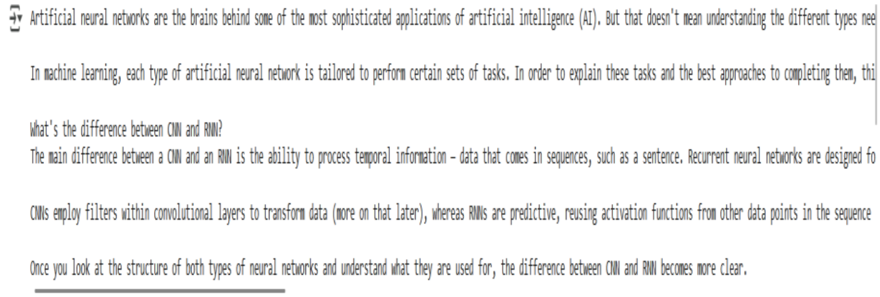
```
word_frequencies = {} # Initialize word_frequencies as a dictionary
for word in docs:
    if word.text.lower() not in stopWords:
      if word.text.lower() not in punctuation:
        if word.text not in word_frequencies: # Check if the word is already a key in the dictionary
          word_frequencies[word.text]=1
        else:
          word_frequencies[word.text] +=1

print(word_frequencies)
```

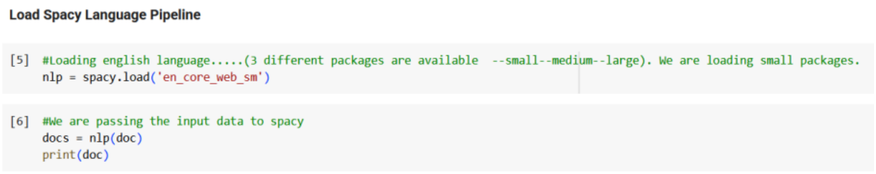
```
{'Artificial': 1, 'neural': 10, 'networks': 9, 'brains': 1, 'sophisticated': 1, 'applications': 3, 'artificial': 4, 'intelligence': 1, 'AI': 2, 'mean': 1, 'understanding': 1, 'different
```

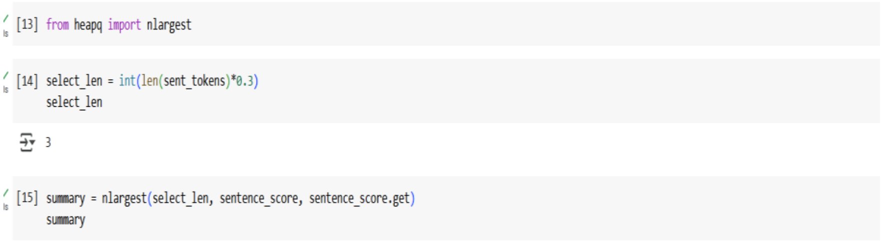
</td></tr>
</table>

Word Tokenization	
	Word Tokenization
	[7] #Word tokenization is performed
	tokens = [i.text for i in docs]
	print(tokens)
	['Artificial', 'neural', 'networks', 'are', 'the', 'brains', 'behind', 'some', 'of', 'the', 'most', 'sophisticated', 'applications', 'of', 'artificial', 'intelligence', '(', 'AI', ')',

Normalization	
	Normalization
	[9] #taking max frequency for normalization
	maxFrequency = max(word_frequencies.values())
	maxFrequency
	10
	[10] #normalizing the data
	for i in word_frequencies.keys():
	  word_frequencies[i] = word_frequencies[i]/maxFrequency
	print(word_frequencies)
	{'Artificial': 0.1, 'neural': 1.0, 'networks': 0.9, 'brains': 0.1, 'sophisticated': 0.1, 'applications': 0.3, 'artificial': 0.4, 'intelligence': 0.1, 'AI': 0.2, 'mean': 0.1, 'understand

Sentence Tokenization	 Sentence Tokenization [11] sent_tokens = [sent for sent in docs.sents] print(sent_tokens) [Artificial neural networks are the brains behind some of the most sophisticated applications of artificial intelligence (AI)., But that doesn't mean understanding the different types n , In machine learning, each type of artificial neural network is tailored to perform certain sets of tasks., In order to explain these tasks and the best approaches to completing them, , What's the difference between CNN and RNN? , The main difference between a CNN and an RNN is the ability to process temporal information – data that comes in sequences, such as a sentence., Recurrent neural networks are designed , CNNs employ filters within convolutional layers to transform data (more on that later), whereas RNNs are predictive, reusing activation functions from other data points in the sequenc , Once you look at the structure of both types of neural networks and understand what they are used for, the difference between CNN and RNN becomes more clear.]
Data Preprocessing Code Screenshots	
Loading Data	 Artificial neural networks are the brains behind some of the most sophisticated applications of artificial intelligence (AI). But that doesn't mean understanding the different types nee In machine learning, each type of artificial neural network is tailored to perform certain sets of tasks. In order to explain these tasks and the best approaches to completing them, thi What's the difference between CNN and RNN? The main difference between a CNN and an RNN is the ability to process temporal information – data that comes in sequences, such as a sentence. Recurrent neural networks are designed fo CNNs employ filters within convolutional layers to transform data (more on that later), whereas RNNs are predictive, reusing activation functions from other data points in the sequence Once you look at the structure of both types of neural networks and understand what they are used for, the difference between CNN and RNN becomes more clear.

Load Spacy Pipeline	 **Load Spacy Language Pipeline**
	`[5] #Loading english language.....(3 different packages are available  --small--medium--large). We are loading small packages.`
	`    nlp = spacy.load('en_core_web_sm')`
	`[6] #We are passing the input data to spacy`
	`    docs = nlp(doc)`
	`    print(doc)`

Text Transformation	 `[13] from heapq import nlargest`
	`[14] select_len = int(len(sent_tokens)*0.3)`
	`     select_len`
	`3`
	`[15] summary = nlargest(select_len, sentence_score, sentence_score.get)`
	`     summary`

Feature Engineering	Attached the codes in final submission.
Save Processed Data	-