Milestone 1: Data Understanding and Exploration

Noor Magdy 16005059 - Mohamed Wael 19011272

Introduction

Natural Language Processing (NLP) tasks heavily depend on the quality and characteristics of the input data. Before applying any summarization techniques, it is essential to understand the structure, content, and limitations of the dataset being used. This milestone focuses on exploring and analyzing the dataset in order to gain insights into its size, structure, and textual properties.

The goal of Milestone 1 is to ensure that the dataset is suitable for summarization tasks and to identify potential challenges that may affect later stages of the project. Through basic data inspection, cleaning, and exploratory data analysis, this milestone establishes a solid foundation for the subsequent implementation of extractive and abstractive summarization methods.

Data Overview

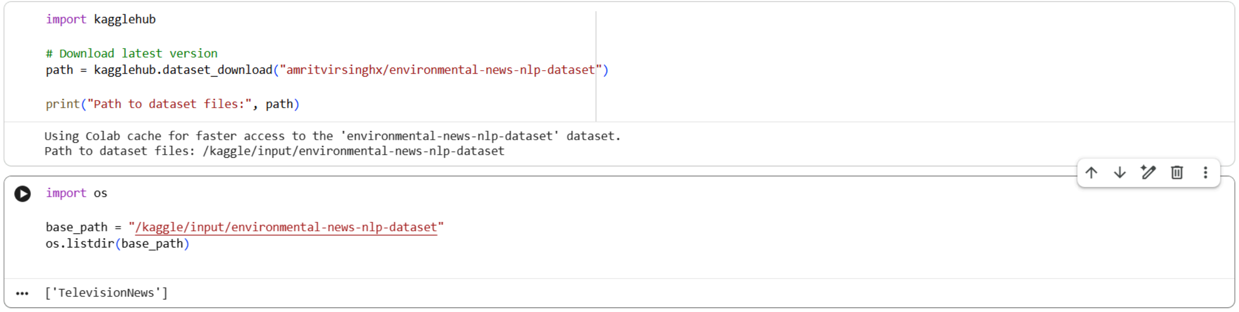
The dataset used in this project is the Environmental News NLP Dataset, obtained from Kaggle. It consists of short news snippets extracted from television news programs related to environmental and public policy topics.

Each record represents a single news snippet along with additional metadata such as broadcast date and source. The dataset contains approximately 95,000 records, making it sufficiently large for exploratory NLP analysis.

The primary column of interest for this project is the Snippet column, which contains the textual data used for summarization in later milestones.

Link of the Data: Environmental News NLP Dataset

Dataset Access and Directory Inspection



```
import kagglehub

# Download latest version
path = kagglehub.dataset_download("amritvirsinghx/environmental-news-nlp-dataset")

print("Path to dataset files:", path)
```

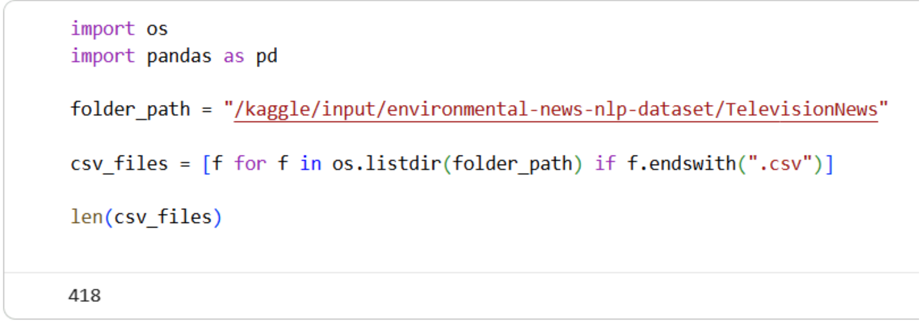
```
Using Colab cache for faster access to the 'environmental-news-nlp-dataset' dataset.
Path to dataset files: /kaggle/input/environmental-news-nlp-dataset
```

```
import os

base_path = "/kaggle/input/environmental-news-nlp-dataset"
os.listdir(base_path)
```

```
['TelevisionNews']
```

The dataset was downloaded directly into the Google Colab environment, allowing efficient access without manual file handling. After obtaining the dataset path, the directory structure was inspected to verify the contents of the dataset. This step confirmed that the dataset files were successfully loaded and organized within a single main folder **(TelevisionNews)**, which contains the CSV files used for analysis. Verifying the directory structure ensured that the dataset was correctly prepared for subsequent loading and processing steps.

CSV File Enumeration:



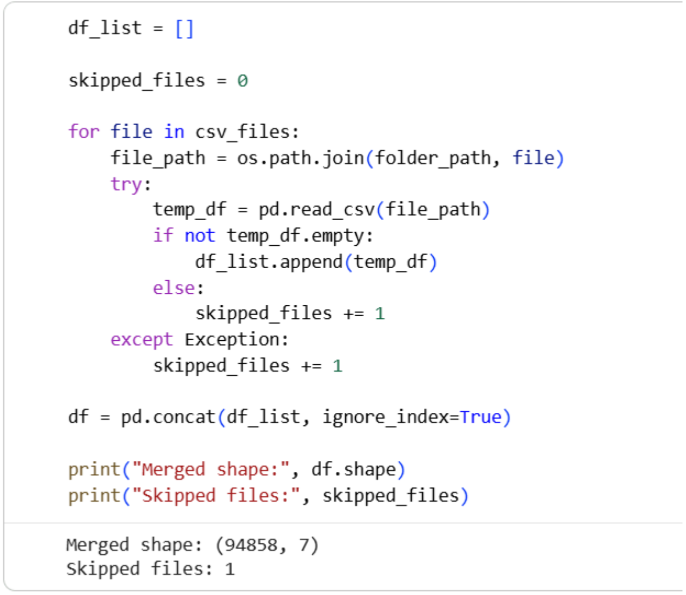
```
import os
import pandas as pd

folder_path = "/kaggle/input/environmental-news-nlp-dataset/TelevisionNews"

csv_files = [f for f in os.listdir(folder_path) if f.endswith(".csv")]

len(csv_files)
```

```
418
```

- The dataset folder was inspected to locate all CSV files.
- A total of 418 CSV files were found within the **TelevisionNews** directory.
- This confirmed that the dataset is split across multiple files.
- The files were later combined into a single DataFrame for analysis.

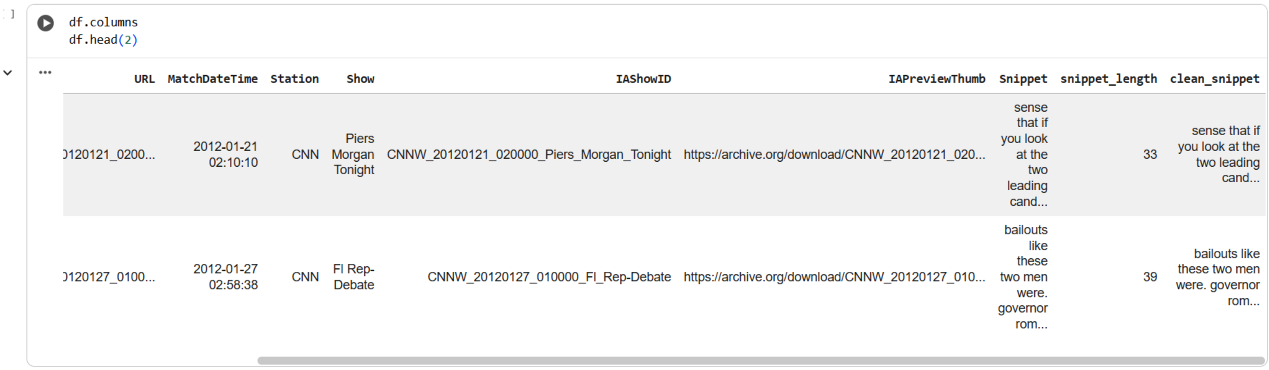
The CSV files were loaded and merged into a single dataset. Empty or invalid files were skipped, resulting in a final DataFrame of 94,858 records with 7 columns.

The dataset columns and a sample of records were inspected to understand the data structure and verify successful loading, with particular focus on the Snippet text field.



```
df_list = []

skipped_files = 0

for file in csv_files:
    file_path = os.path.join(folder_path, file)
    try:
        temp_df = pd.read_csv(file_path)
        if not temp_df.empty:
            df_list.append(temp_df)
        else:
            skipped_files += 1
    except Exception:
        skipped_files += 1

df = pd.concat(df_list, ignore_index=True)

print("Merged shape:", df.shape)
print("Skipped files:", skipped_files)
```

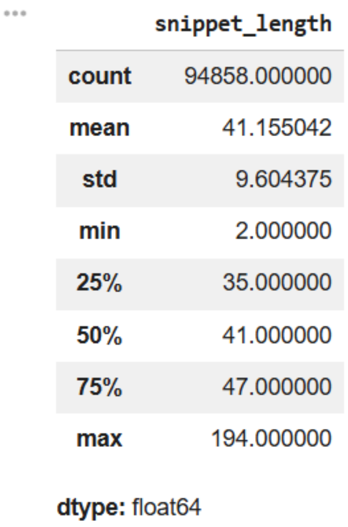
```
Merged shape: (94858, 7)
Skipped files: 1
```



```
df.columns
df.head(2)
```

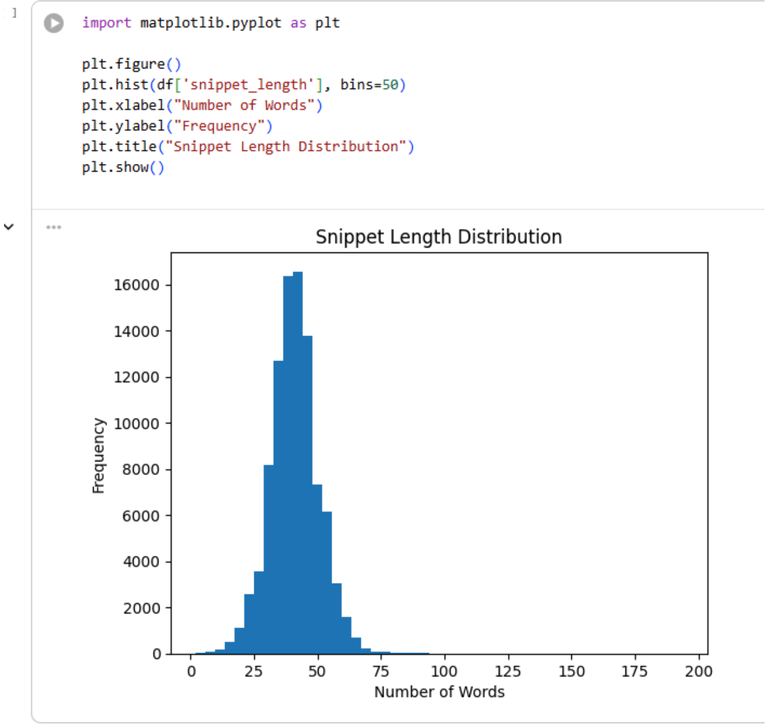
	URL	MatchDateTime	Station	Show	IAShowID	IAPreviewThumb	Snippet	snippet_length	clean_snippet
	0120121_0200...	2012-01-21 02:10:10	CNN	Piers Morgan Tonight	CNNW_20120121_020000_Piers_Morgan_Tonight	https://archive.org/download/CNNW_20120121_020...	sense that if you look at the two leading cand...	33	sense that if you look at the two leading cand...
	0120127_0100...	2012-01-27 02:58:38	CNN	Fl Rep-Debate	CNNW_20120127_010000_Fl_Rep-Debate	https://archive.org/download/CNNW_20120127_010...	bailouts like these two men were. governor rom...	39	bailouts like these two men were. governor rom...

The dataset contains 94,858 snippets with an average length of approximately 41 words. Most snippets fall within a relatively narrow range, with half of the snippets containing between 35 and 47 words. The shortest snippet contains 2 words, while the longest reaches 194 words. These results indicate that the dataset primarily consists of short, already condensed text, which may limit the extent of further summarization.



	snippet_length
count	94858.000000
mean	41.155042
std	9.604375
min	2.000000
25%	35.000000
50%	41.000000
75%	47.000000
max	194.000000

dtype: float64

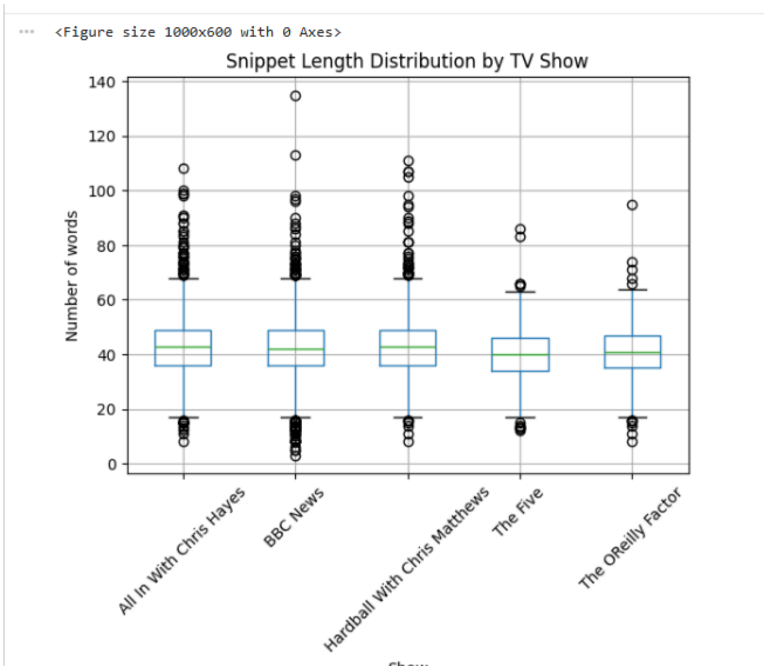


```
import matplotlib.pyplot as plt

plt.figure()
plt.hist(df['snippet_length'], bins=50)
plt.xlabel("Number of Words")
plt.ylabel("Frequency")
plt.title("Snippet Length Distribution")
plt.show()
```

The histogram shows that most snippets are concentrated between 30 and 50 words, indicating that the dataset primarily consists of short, already condensed text. Only a small number of snippets are substantially longer, limiting the potential for strong compression in most summarization cases.

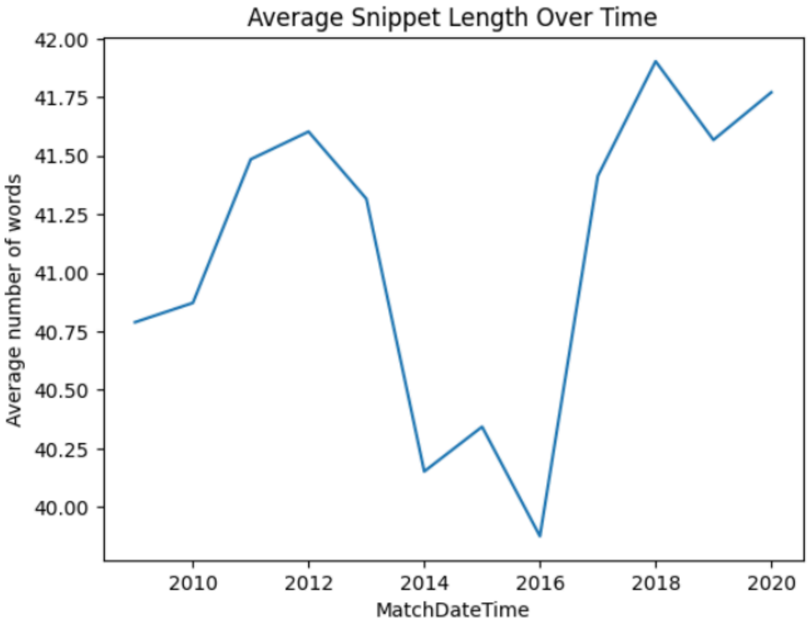
The boxplots indicate that, despite originating from different programs, the majority of snippets across all shows have **similar median lengths**, generally centered around **40 words**. This suggests a consistent editorial or broadcasting style in how short news segments are presented, regardless of the specific program.

While the central tendencies are similar, the presence of outliers varies across shows. Some programs, such as *BBC News* and *Hardball with Chris Matthews*, exhibit a wider spread and a larger number of longer snippets. These outliers indicate occasional extended commentary or discussion segments, which may offer greater opportunities for summarization compared to the typical short snippets.



```
<Figure size 1000x600 with 0 Axes>
```

The average snippet length remains stable over time, with only minor fluctuations. This indicates consistent formatting of news snippets across years and suggests that summarization difficulty is not significantly affected by temporal changes in snippet length.



Conclusion

This milestone presented an initial exploration and analysis of the Environmental News NLP dataset, focusing on understanding its structure, content, and statistical properties. The dataset consists of a large collection of televised news snippets with consistent formatting, covering multiple programs and spanning several years.

The analysis showed that snippet lengths are relatively stable across the dataset, with most snippets centered around a moderate word count and only a small number of extreme cases. This consistency was further observed across different television shows and over time, suggesting standardized editorial practices in broadcast news production.

These findings are important for downstream NLP tasks, particularly text summarization. The absence of strong structural variation indicates that differences in summarization performance are more likely to stem from linguistic and semantic complexity rather than variations in input length. Overall, this milestone confirms that the dataset is clean, well-structured, and suitable for implementing and evaluating both extractive and abstractive summarization methods in the following milestones.