Lab 1 Report - Corpus Construction

1. Objective

To construct a structured news corpus from a real-world dataset for use in subsequent NLP tasks.

---

2. Dataset

Fake and Real News Dataset (Kaggle)
Files used:

- `Fake.csv`
- `True.csv`
  Each file contains news articles with title and text fields.

---

3. Methodology

1. Loaded both CSV files using Pandas.
2. Assigned binary labels:
   - Fake → 0
   - Real → 1
3. Merged datasets into a single DataFrame.
4. Removed short or empty articles.
5. Randomly sampled 5000 articles.
6. Saved the structured corpus as:
   - `news_corpus.jsonl`
   - `news_corpus.csv`

---

4. Final Dataset Structure

Column	Description
title	News headline

Column	Description
text	Article content
label	0 (Fake), 1 (Real)

5. Conclusion

A clean, labeled, and structured corpus was successfully created.
This dataset will be used for preprocessing, sentiment analysis, and machine learning tasks in subsequent labs.

6. Git Link:-

(Lab 1 - Git)[https://github.com/MYTH-il/NLP/blob/main/lab1_corpus.ipynb]