DS250: Data Analytics and Vizualization Project

Analysis of News Repetition in Daily News Bulletin

Prepared By -
Tanmay Kumar Shrivastava
Uday Bhardwaj
Shivam

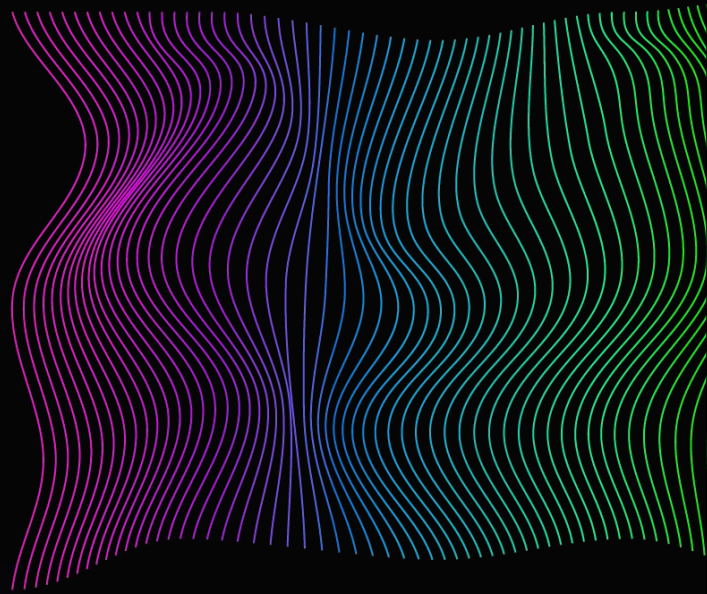
Submitted To-
Dr. Rajesh Kumar Mundotiya

Semester-
2023-24 M

Problem Statement

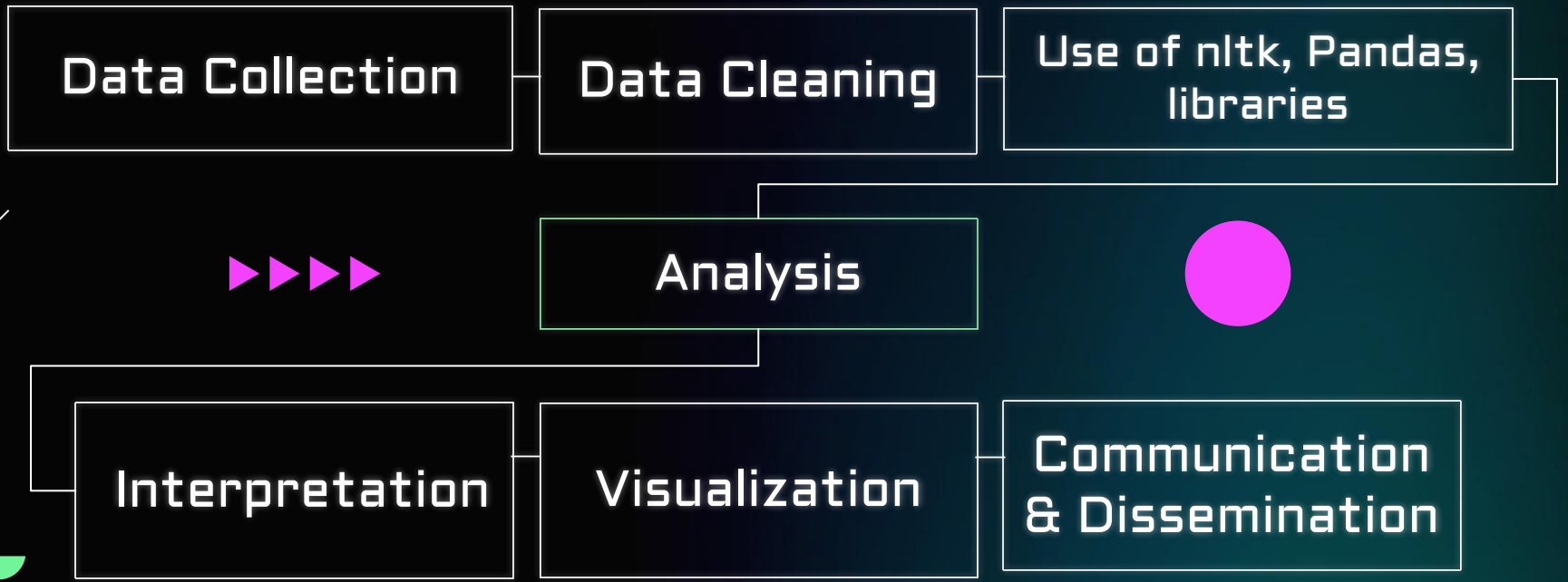
The modern news landscape is dominated by a continuous stream of information, often referred to as non-stop news. This relentless barrage and reptition of headlines can be overwhelming and difficult to make sense of.

To better understand the nature of non-stop news, this project aims to conduct a comparative analysis of daily non-stop news videos using a combination of **data collection, cleaning, Data Analysis and visualization, and natural language processing techniques.**



```
if ("PROJECT" == "SUCCESSFUL");
    cout << "FULL MARKS";
else
    cout << "DECISION UPTO SIR";
```

PROJECT ARCHITECTURE



INITIAL APPROACH

Our Initial approach was to find a previously available dataset for the task. During our search, we encountered few datasets related to the topic but there were some limitations which were either in the dataset or in the method to obtain them. We will just briefly discuss one of them as follow:

We found a dataset on Kaggle which contained news articles published by a News Channel during a time-span of 10 years. However, the dataset contained the news articles which were already distinct and had very less similarity in them.

Moving on, we tried to obtain the data from the Youtube videos of Indian News Channels directly. For this task, We required a model which could either scrap the subtitles of the videos or could convert the voice to text from the video.

Challenges Addressed During Data Collection

Problems / Errors faced in making of the project during collection of data

CHALLENGES FACED DURING THE DATA COLLECTION OF THE PROJECT

1

All Daily News weren't having CC (captions) for transcribing the video

2

Lost almost half a month in data collection and cleaning

3

Logic Applied for different languages but the errors and complexity increased

4

Tried many NLP methdology but it did not gave the desired result as per requirements

euro news.

FINALIZATION OF THE PROJECT IDEA

After such long and rigorous efforts of our team, we finally found out that EURO NEWS can help to make our project possible by accepting all requirements. We analyzed for 512 days (Morning, Mid-Day, Evening)

ABOUT THE PROJECT

Technical Stack

> Python programming language
> Web scraping libraries (e.g., youtube-transcribing-api etc.)
> Natural language processing libraries (e.g., nltk, re)
> Hugging Face Transformers (e.g., all-MiniLM-L6-v2, Sentence-Transformers)
> Data visualization libraries (e.g., Matplotlib, Seaborn etc.)

DATA COLLECTION: YouTube-Transcribing API

The Youtube Transcript API can get the youtube video subtitles/transcripts by giving the video id.

- The YouTube Transcript API is a great tool that allows developers to programmatically **extract transcripts or captions from YouTube videos.** A transcript is essentially a written version of the spoken content in a video. Captions, on the other hand, can include both spoken words and descriptions of relevant sounds, making content even more accessible for individuals with hearing impairments.

- This API provides a bridge between the dynamic world of video content and the easily searchable world of text. It opens up numerous opportunities in various domains, including accessibility, content analysis, content localization, search engine optimisation and beyond. .

DATA ANALYSIS:
DATASET (upto 1536 rows)

INDEX	NAME	DETAILS
1.	Channel	Channel from which news has been transcribed
2.	Date	The date on which the Headlines broadcasted
3.	Time	The timing of the Broadcast- Morning, Midday, Evening
4.	Location	Based country of news channel: India/ Outside India
5.	Headlines	Raw headlines collected

DATA PRE-PROCESSING:
Text Preprocessing and Cleaning

- **Text Normalization and Tokenization:** Apply text normalization techniques to normalize the dataset and extraneous whitespace. Tokenize the text into individual words or meaningful units.

- **Stemming or Lemmatization:** Apply stemming or lemmatization to reduce words to their root forms.
  Note- Stemming was not performed as it was observed that it was mostly damaging the information which can be used to find similarity between the sentences.

- **Removal of Stop Words:** Eliminate common words that don't add significant meaning to the text. It was done with re module.

DATA ANALYSIS:
Models Used for Analysis

Hugging Face

Models Used	Utilize Hugging Face transformers (all-MiniLM-L6-v2 and clips/mfaq) or other NLP libraries to perform analysis on the transcribed text.
We have used two different models for the different kinds of similarity techniques Cosine Similarity: all-MiniLM-L6-v2 Semantic Similarity: clips/mfaq	

```
from sentence_transformers import SentenceTransformer,util
model = SentenceTransformer('all-MiniLM-L6-v2')
```

Use of `all-MiniLM-L6-v2`

This is a sentence-transformers model. It maps sentences and paragraphs to a 384 dimensional dense vector space and can be used for tasks like clustering or semantic search.

DATA ANALYSIS:
More about all-MiniLM-L6-v2

Hugging Face

```
for sentence, embedding in zip(sentences, embeddings):
    print("Sentence:", sentence)
    print("Embedding:", embedding)
    print("")
```

Our model is intented to be used as a sentence and short paragraph encoder.
During the use of the model, we passed the news headline as the input and it would create a set of low-dimensional vectors which could be further used for the comparison between such vectors to obtain the similarity score. **We applied this model for finding the Cosine Similarity score.**

One advantage is that, this model can be used to embed sentences which have large length. This was highly important as the set of News Headlines obtained still remained large.

DATA ANALYSIS:
clips/mfaq

It also provides embeddings. The model is trained from the available FAQ dataset. It produces a set of vectors which can be further used to compare between each other to find the similarity scores.

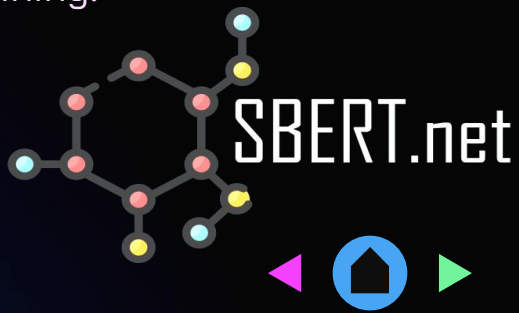
Similar to the previous model, we have used the model to provide us with the embeddings of a particular news headline which would be used to compare against each other to find the similarity. **We have used this model for the calculation of the values for Semantic similarity.**

```
for sentence, embedding in zip(sentences, embeddings):
    print("Sentence:", sentence)
    print("Embedding:", embedding)
    print("")
```

DATA ANALYSIS:
Sentence Embedding, Sentence Similarity Semantic Search

SentenceTransformers is a Python framework for state-of-the-art sentence, text and image embeddings.

We decided to use this framework to compute sentence / text embeddings for more than 100 languages. But, to avoid errors we have to focused on one language i.e English. These embeddings can then be compared e.g. with cosine-similarity to find sentences with a similar meaning. This can be useful for semantic textual similar, semantic search, or paraphrase mining.

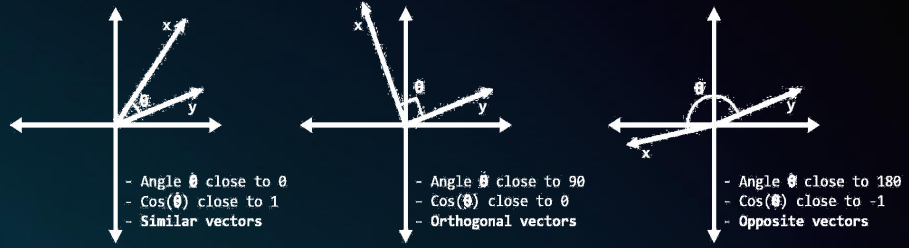


DATA ANALYSIS:
Sentence Embedding, Sentence Similarity Semantic Search

> Semantic Similarity

Semantic similarity measures the degree to which two words or phrases have the same meaning. It considers the contextual meaning of words and phrases, taking into account their relationships to other words and their usage in different contexts.

DATA ANALYSIS:
Sentence Embedding, Sentence Similarity Semantic Search




```
A:
[[ 1  2  2]
 [ 3  2  2]
 [-2  1 -3]]

B:
[[ 4  2  4]
 [ 2 -2  5]
 [ 3  4 -4]]

Cosine Similarity:
[0.88888889 0.5066404
 0.41739194]
```

Cosine Similarity <

Cosine similarity measures the similarity between two vectors of numbers. It calculates the cosine of the angle between the two vectors, which represents the degree to which the vectors point in the same direction.

DATA ANALYSIS:
Sentence Embedding, Sentence Similarity Semantic Search

P.O.D	Semantic Similarity	Cosine Similarity
Applications	Natural language processing (NLP) tasks	Recommendation systems, information retrieval, anomaly detection
Input	Word embeddings	Vectors
Output	A number between 0 and 1, where 0 indicates that the vectors are completely dissimilar and 1 indicates that the vectors are identical	

DATA ANALYSIS:
FINAL DATASET (upto 1024 rows)

Data Frame

S.No	Date	Technique	Morning-Midday	Midday-Evening	Morning-Evening	Average Similarity
1	2023-01-01	Cosine Similarity	73.351413	44.613236	34.173316	50.712655
2	2023-01-01	Semantic Similarity	96.062630	97.445554	95.561874	96.356686
3	2023-01-02	Cosine Similarity	57.157195	59.162945	42.400160	52.906767
4	2023-01-02	Semantic Similarity	98.220742	98.225093	98.085129	98.176988

Question: Which of these methods is better way to accurately describe the similarity?

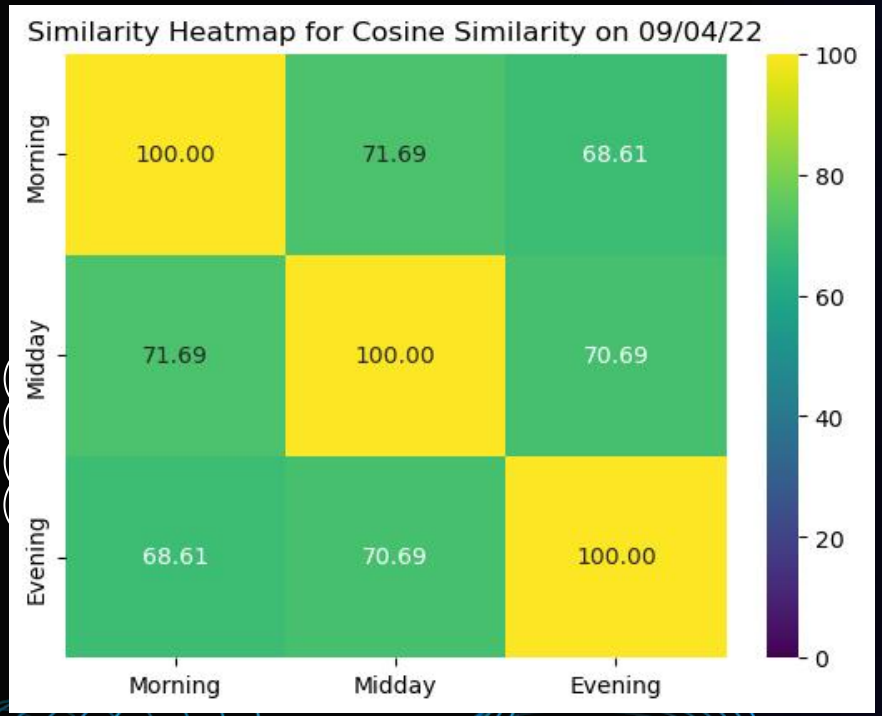
DATA ANALYSIS:
FINAL DATASET (upto 1536 rows)

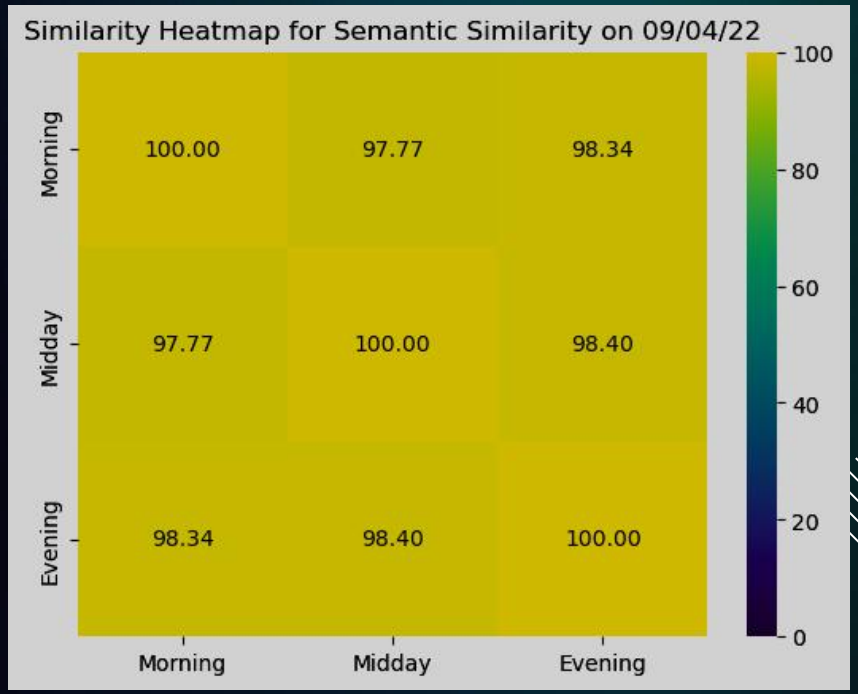
Ans.
The cosine similarity is quite sensitive to the order of the words and the way the text has been written, while semantic similarity is not. Semantic similarity is difficult to compute whereas cosine similarity is simple.

Overall, Semantic similarity is quite an accurate way to describe the similarity but we get more accurate results by applying cosine similarity. Another reason for the difference in the similarities can be attributed to the information addition in the latest news.

DATA VIZUALIZATION:

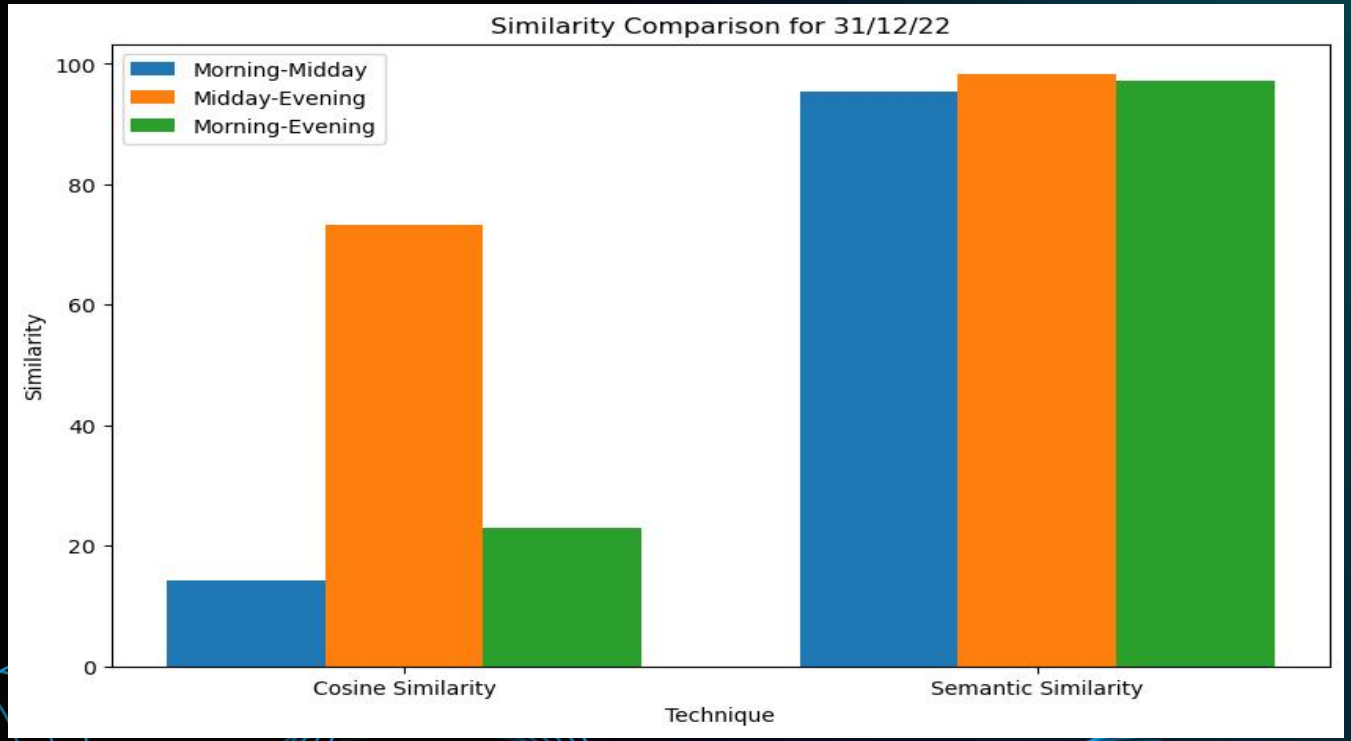
Similarity of a particular day (HeatMap)





DATA VIZUALIZATION:

Similarity of a particular day (Bar Chart)



DATA VIZUALIZATION:

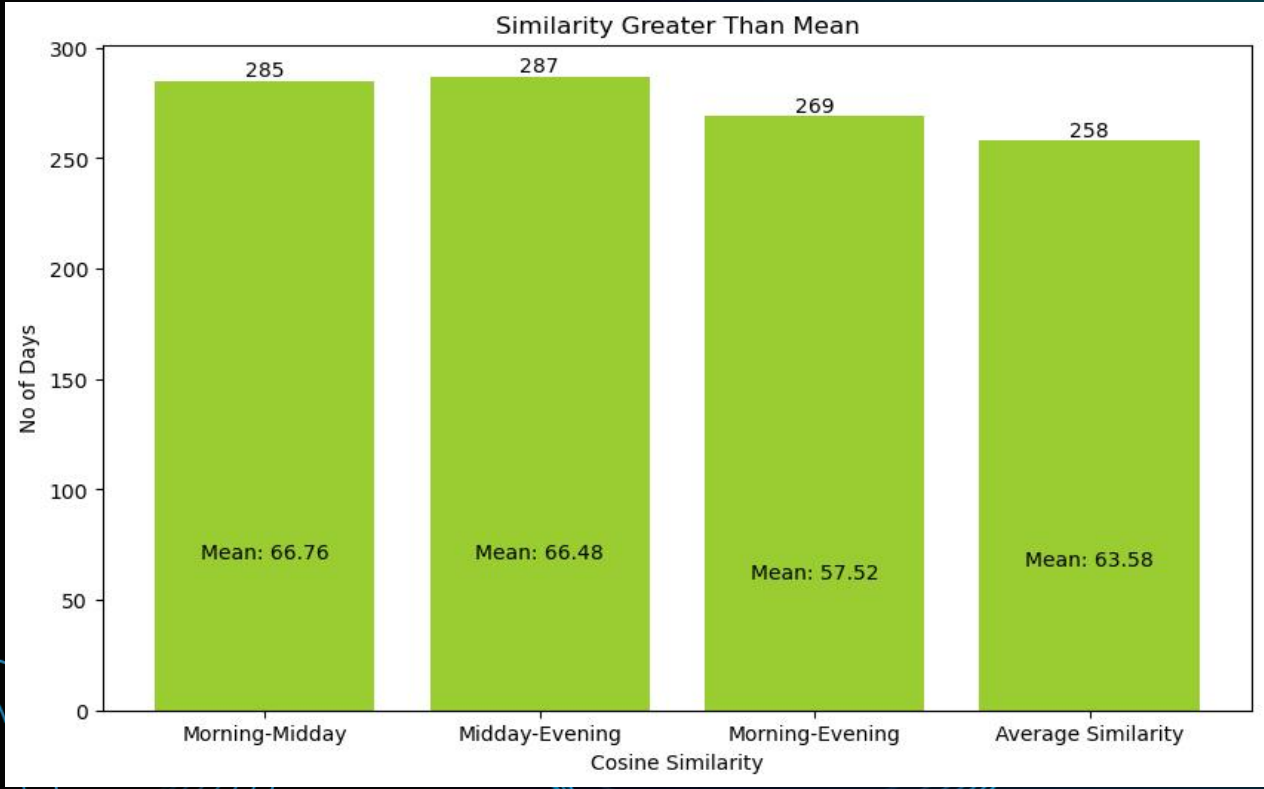
Similarity of a particular day (Word Cloud)

Morning (9/4/22)

Midday (9/4/22)

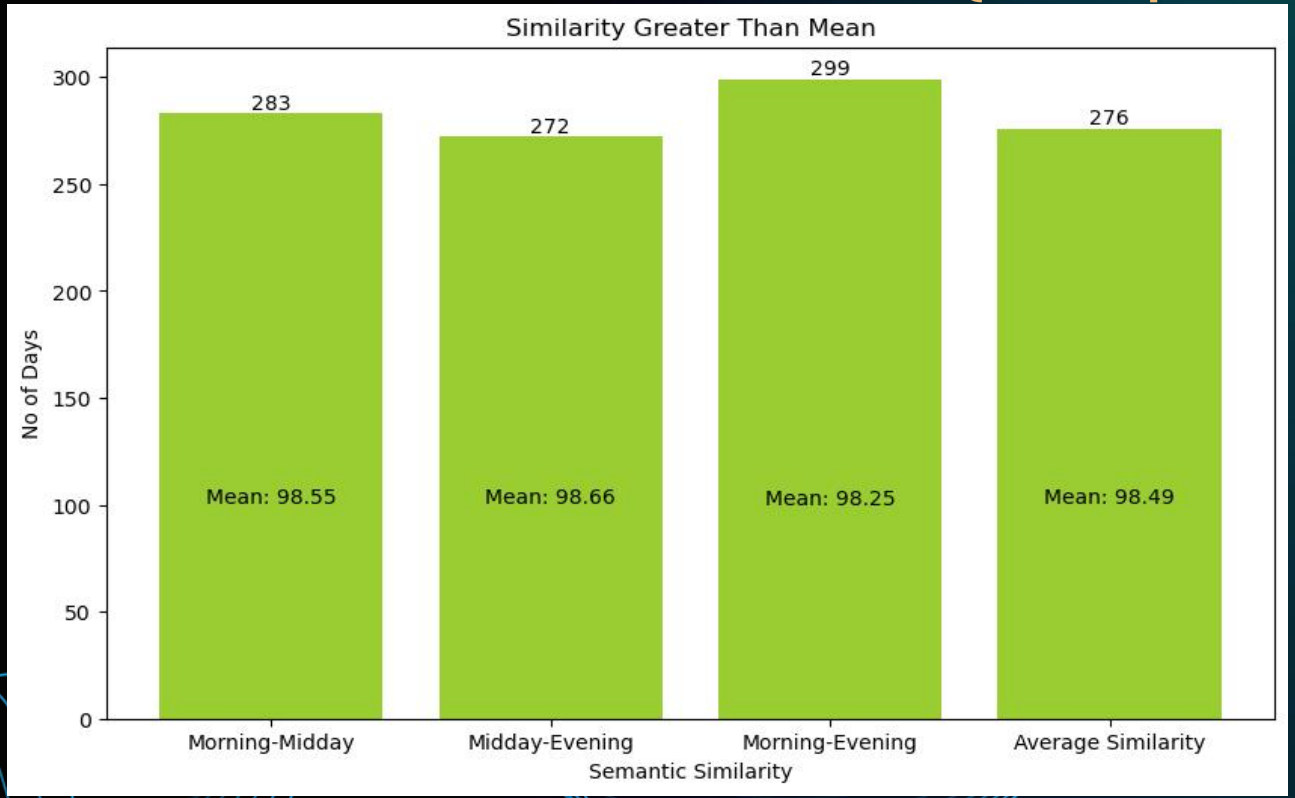
Evening (9/4/22)

DATA VIZUALIZATION:
Values Greater than Mean (Bar plot)



DATA VIZUALIZATION:

Values Greater than Mean (Bar plot)



RESULT

Here are some examples of how the project demonstrates how news from around the world can increase similarities:

- The word cloud reveals that the terms "ukraine" and "russia" are the most prominent in the word cloud, implying that these are the two most important topics to people worldwide like Russia-Ukraine War.
- Also, in the other word cloud, the terms "Israel" and "Gaza" are prominent in the word cloud,

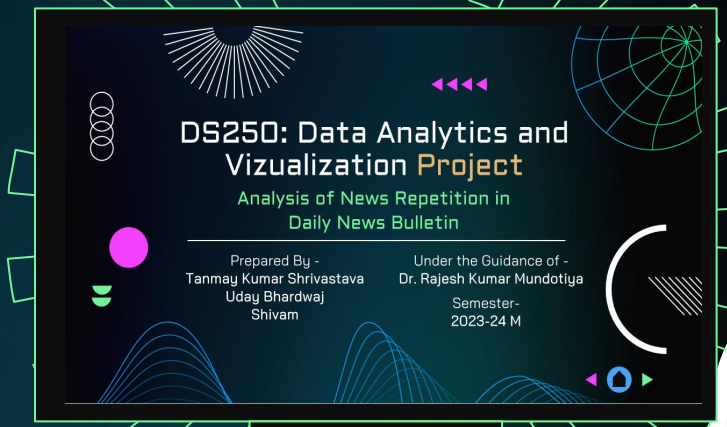
The same thing can be seen in more creative way-

If news is a global village, then similarities are the town square. When news from anyplace in the world breaks, it spreads to every corner of the globe, bringing people together to share their tales and viewpoints. Similarly, this project targets this global conversation, and it demonstrates how similar our lives may be, despite our differences in culture and background.

REFERENCES

- https://medium.com/@jawadkr111/youtube-transcript-api-how-to-deal-with-video-content-017ab88e8180
- https://www.exchange4media.com/media-tv-news/nclt-approves-proposal-for-acquisition-of-television-home-shopping-network-report-131018.html
- https://github.com/jdepoix/youtube-transcript-api
- https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing
- https://www.geeksforgeeks.org/python-lemmatization-with-nltk/
- https://www.learndatasci.com/glossary/cosine-similarity/
- https://www.turing.com/kb/guide-on-word-embeddings-in-nlp
- https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2#background
- https://huggingface.co/clips/mfaq

THANK YOU!


DS250: Data Analytics and Vizualization Project
Analysis of News Repetition in Daily News Bulletin
Prepared By -
Tanmay Kumar Shrivastava
Uday Bhardwaj
Shivam
Under the Guidance of -
Dr. Rajesh Kumar Mundotiya
Semester-
2023-24 M