Mini-Project 3 involves analyzing and visualizing different aspects of Al-Jazeera corpus through different computational methods. The aim was to extract meaningful insights from the data by using different techniques such as n-gram frequency analysis, TF-IDF similarity network and topic modeling. To make the group work more collaborative, we use common platforms and divide the work among ourselves.

Distribution of work:

Following shows the workflow of each team member (Zeeshan, Sahrayar, and Ulya), it will help to ensure further reproducibility with updated data.

Repository Setup:

* Zeeshan forked the repository and added us to the fork on the GitHub and added other team members to work on the fork.
* Each member then cloned the forked repository of our local laptops for individual work.

Task Management:

* Zeeshan also created a Trello board to manage the tasks, update the work done and in progress, and to track each other’s progress. It was also helpful as each member can comment on the group members’ progress on task cards.

**Data Analysis Using Different Computational Methods:**

1. **N-gam Frequency Analysis:**

Responsibility: Ulya

Data set used: “1-gram-year-month.csv”

**Purpose:**

This part analyzes the frequency of unigrams in the Al-Jazeera corpus group by month and year. The analysis includes both absolute and relative frequency values. Focus was given to conflict-related and peace-related words. This part of the project helps to identify how the discourse around conflict and peace evolves over time in media reporting. The use of two types of frequency provides a more holistic view. These findings highlight

1. **Topic Modelling and Topic Trend Visualization:**

Responsibility: Zeeshan Karim

Dataset used: topic-model.csv, length.csv

Purpose:

In this part of analysis, the results of article length and topic modeling results are merges using data fields of year, month and day. This dual analysis reveals shifts in media attention and highlights imbalances in coverage on mainstream media like Al Jazeera. As appears in the output of n-gram, this method also reveals the casualty related topics dominate while topics related to humanitarian aid and diplomacy received limited attention.

1. **TF-IDF Similarity Network**

Responsibility: Sahrayar Morani

Data set used: tfidf\_analysis.py

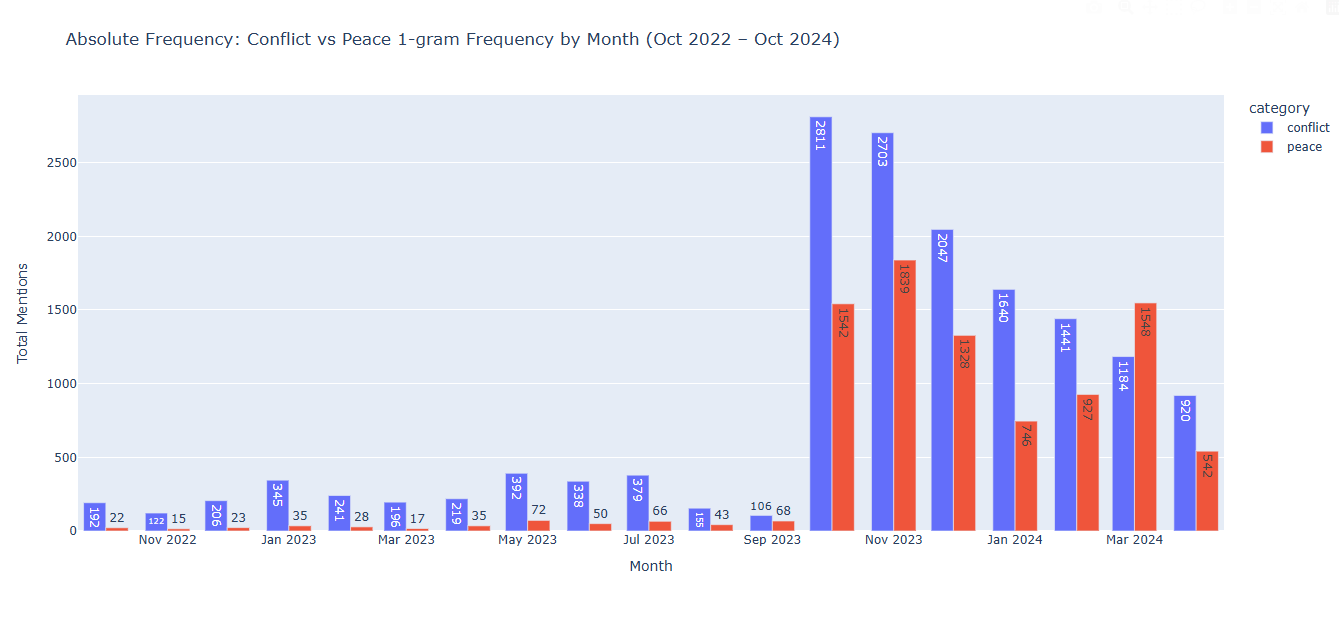
Purpose:

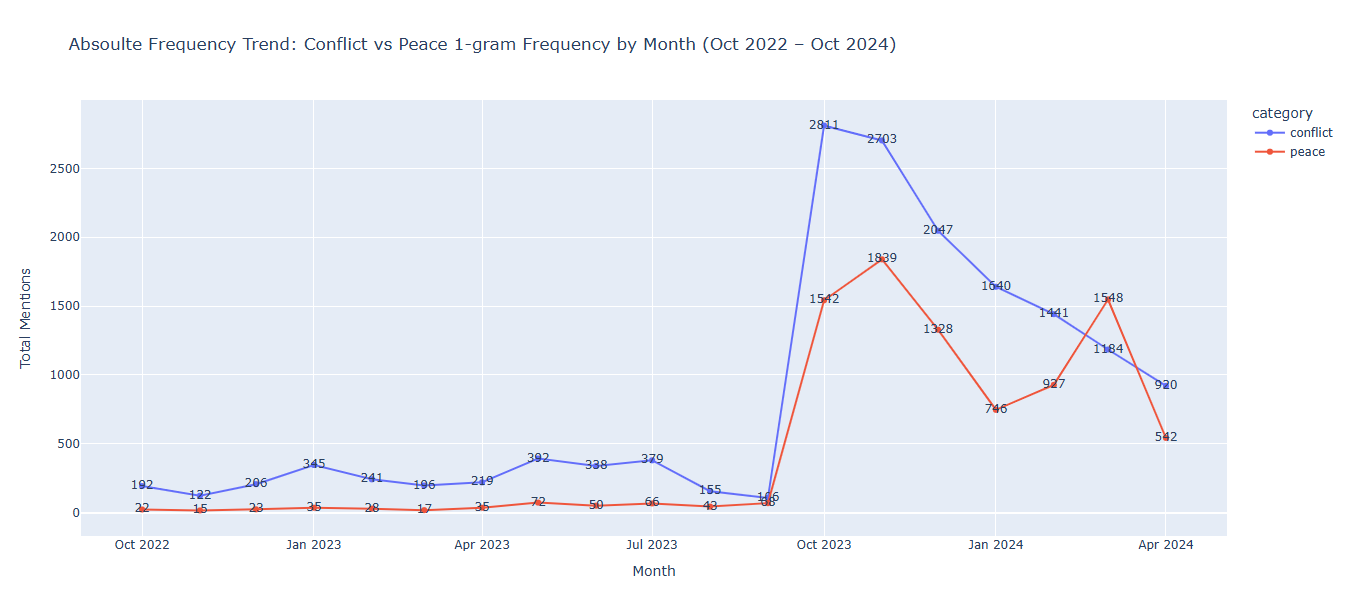
This process is done to identify similarity between articles based on TF-IDF values and to visualize a cluster of articles discussing similar topics from the corpus. The clusters of articles are visualized with tools like Gephi map to see how articles are interconnected. By focusing on articles from the corpus with similarity threshold above 0.3 the main goal was to identify the news coverage that discuss similar themes.

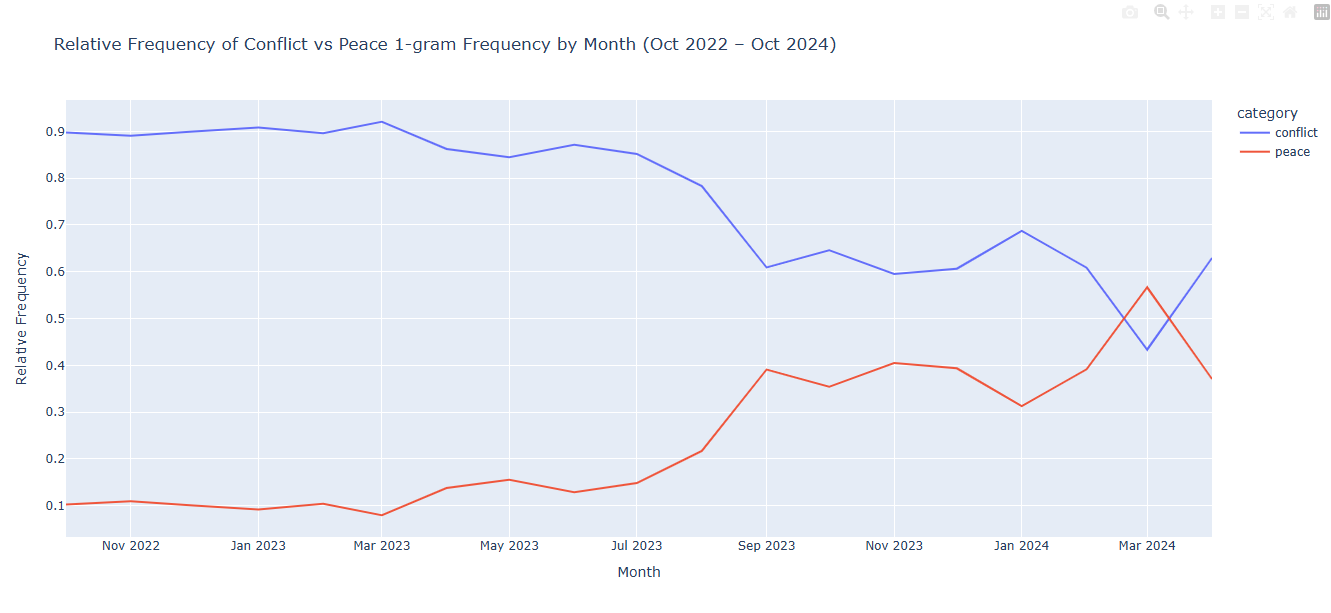
**1-Gram Frequency Analysis: Conflict and Peace Terms in the Al-Jazeera Corpus**

In n-grams I analyze n-gram-1-gram frequencies on a time-series dataset from Al-Jazeera news corpus. This was a large data set, so I focus only on the monthly usage patterns of unigrams between October 2022, and October 2024 I first only include months after the war started i.e after October 2023 till October 2024 and try to visualized how the terms related to peace and conflict have change over the time, but then thought it would better to know how these terms used before the war and how it has evolve over the time during war. The analysis consists of two python scripts: an exploration script and a visualization script. In the first script I extract the most frequent 1-grams and in the second script I explore how selected conflict and peace related terms vary over time. I used both absolute and relative frequencies to get a holistic idea.

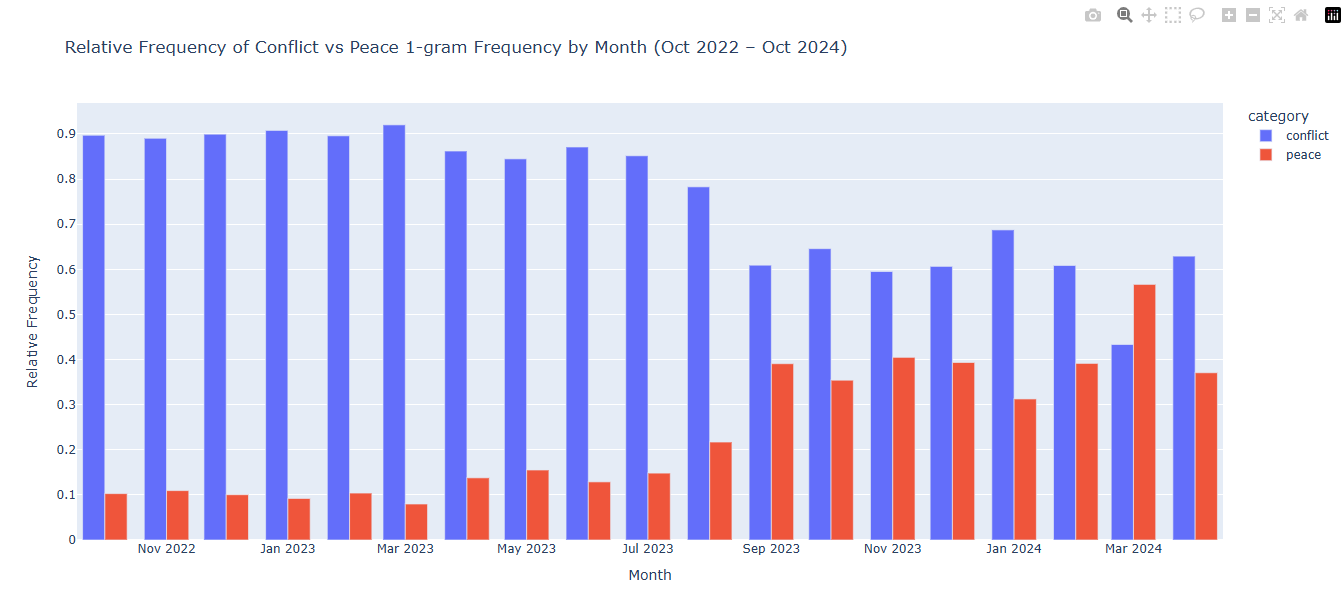
I used an n-gram frequency counting methodology specifically using 1-gram. The dataset “1-gram-year-month.csv” is opened and read using pandas. To isolate meaningful words from English most common stop words I used NLTK’s stop words. After calculating the absolute frequencies, the script identified the top 20 most frequent unigrams in the entire data set. This gave a big picture that which words were most common and dominant in the dataset.

To get a focused view, I selected two set two categories: conflict-related terms (such as attacks, strikes, and killed) and peace related terms (such as peace, ceasefire, relief etc). The data set was filtered so the output includes only these words, and then monthly frequency terms were calculated for both absolute count and relative frequencies (used this method to know the proportion of words used in the corpus relative for that month). These methods are selected to meet the objective of the project, as to explore how media language reflects changing discourse revolves around violence and humanitarian narrative in a region under conflict or war. The use of both absolute and relatives’ values gives a holistic overview and avoids mistakes and gives a fair comparison across time.



The first set of visualization presents absolute frequency trends of conflict and peace words over the 24-month period. I used both line and bar graph for clear picture. We can see that before war started both words were used less, however still conflict words come more often. This may be as the region is porn to conflict before the start of war. But after the war started in 2023 conflict-related words reached its spike, but it gradually diminishes even in March 2024 peace related terms showed more trends than conflict related words. There could be two possible reasons, one may be as war goes on the media sees how people are suffering and try to cover the humanitarian aspect. Or people just get tired of the war and stop talking about it. If we look at overall trend pf the terms related to conflict, it’s moving downward while peace related terms are not steady and changing over different periods.

The second set of visualization is a comparison of relative frequencies. It helps to understand how often a word appears compared to the total words, not just the raw number. Overall, the words related to conflict came more often than peace related words. But in some months peace related words showed in more proportion than conflict related words. This shows us how the media shifts its coverage over different periods.



Using unigram frequences is a good way to track themes over time in a big data set like news articles. However, it has limitations as it does not show the context of the words used, secondly it also cannot tell the meaning of the words like if a word has more than one meaning like strike, which can be either attack or protest. Despite these limitations, unigram works best for this project and the selection of two words of different categories like peace and conflict are clear. The pattern over time steady and elative frequency also helps to make sure that the results are not affected by more or fewer articles.

**Topic Modelling and Topic Trend Visualization**

This section of the project examines the distribution of media coverage on selected topics related to the Gaza conflict (2023–2024) in the Al-Jazeera news corpus. The analysis investigates how media narratives shifted before and during the conflict by observing the frequency of selected topics such as casualties, media/journalists, diplomatic resolutions, human rights, and humanitarian crises.

The main argument I depicted from this analysis is that while there is an expected surge in reporting on immediate conflict casualties when the war breaks out, other essential dimensions of the conflict, particularly media transparency, diplomatic negotiations, human rights, and humanitarian aid receive inconsistent, comparatively low, or unstained attention. The data visualisation clearly reveals this imbalance, offering insights into how conflict narratives are constructed and prioritized by media outlets in crisis settings.

The analysis consisted of two Python scripts:

Data Preparation and Merging Script:

The first script loaded two datasets — one containing topic modelling results (topic-model.csv) and the other containing article length information (length.csv). These were merged using an inner join on the common date columns ['year', 'month', 'day']. Rows with a Topic ID of -1 were filtered out, as they indicated unclassified or irrelevant content. A mapping of Topic IDs to descriptive names was created by concatenating the top three keywords associated with each topic, making them human-interpretable for visualization.

To get a clear picture, only include data from years 2023 and 2024, and filtered the rest.

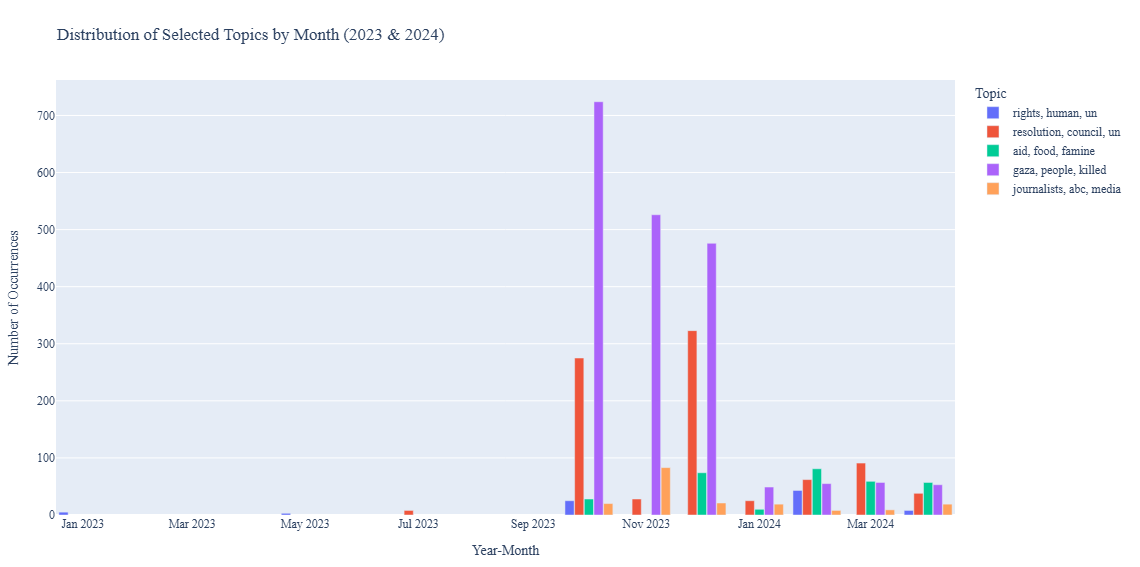
To narrow the focus, the data was filtered for the years 2023 and 2024, and five specific topics were selected based on their relevance to the Gaza conflict. These topics included conflict casualties, journalists and media, diplomatic resolutions, human rights, and humanitarian aid. A new column, Year Month, was created by combining the year and month columns for chronological ordering. Topic occurrences were then counted for each month.

Visualization Script:

The second script used Plotly Express to create interactive visualisations: a grouped bar chart and a line graph. The bar graph plotted the number of articles per topic by month, while the line graph showed topic trends over time. I used Plotly Express for this visualization because it offers interactive and clear representations. We can see that, as expected, the topic “gaza, people, killed” had a huge spike in October and November 2023 when the war broke out. This sharp increase made sense as the initial news coverage focused heavily on casualties and violence.

However, if we look at other topics like “journalists, abc, media” or “rights, human, un”, their coverage remained quite low and inconsistent throughout the period. Logically, after covering the immediate conflict, media should have focused more on reporting conditions, diplomatic resolutions, and human rights issues — but these topics did not receive comparable attention, which shows a significant imbalance in media focus. Even the topic related to “aid, food, famine”, which highlights the humanitarian crisis, was also underreported, especially when compared to the casualty reports.

These visualizations aimed to explore patterns in how media attention shifted between different topics during the Gaza conflict.



The second visualization was a line graph depicting the innovations of the selected topics over time. This graph facilitated the visualization of oscillations and persistent trends. It vividly illustrated that although conflict-related topics were predominant during the early phase of the conflict, there was no enduring peak in the coverage of humanitarian or diplomatic themes in the subsequent months. By the 3rd month of 2024, the “aid, food, famine” topic had a modest rise, but still, fell short of expectations based on the magnitude of the humanitarian disaster.A graph showing the growth of the year

AI-generated content may be incorrect.

This part of the project highlighted a clear imbalance in how Al-Jazeera covered the ongoing conflict. Main focus was given to casualties when the war started and this make sense, but other important topics like humanitarian aid, human rights, and journalism did not get much attention. Using topic modelling with article length data helped me track these patterns over time. While this method doesn’t show the exact context, it made it easy to see which topics were talked about most. It showed how media coverage can shape public understanding by focusing on some issues more than others.

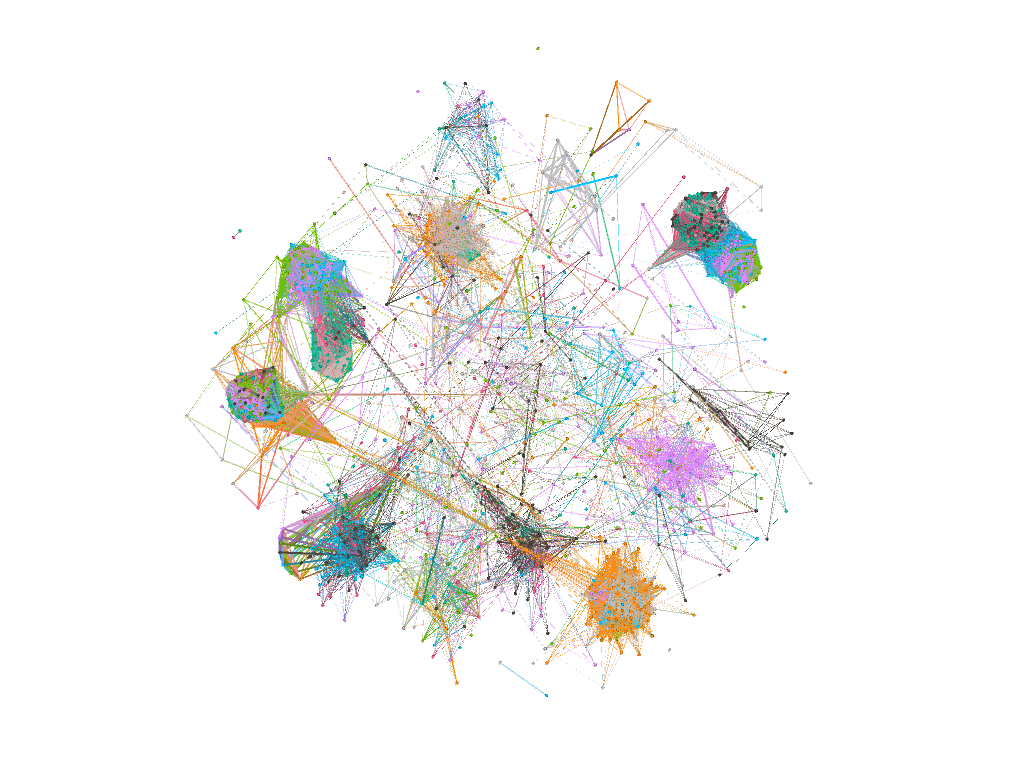
3- TF-IDF and Similarity Visualisation

In this analysis, I explore the textual patterns and thematic structures within the Al-Jazeera news corpus related to the Gaza Conflict (2023–2024), using Term Frequency-Inverse Document Frequency (TF-IDF) as a core method for identifying significant keywords and article similarities. To achieve this, I used tools such as Gephi for network visualization and Plotly for temporal data plotting, drawing insights from a CSV file generated through TF-IDF scoring. This approach enabled the mapping of connections between articles, revealing the frequency and thematic overlap across the dataset.

The TF-IDF which includes the use of Gephi and Plotly to visualize the data in the csv constructed from the Gaza Conflict (2023-2024) in the Al-Jazeera news corpus. The purpose of TF-IDF was to see frequency and similarities between articles and find bunches of articles which talked about similar topics based on text frequency.

Exploration:

The main purpose of my exploration was initially to map the articles in Gephi to see the bunches of articles which formed clusters linking to each other with similar topics, this was accomplished by first creating nodes and edges from the csv file provided under tfidf folder titled ‘tfidf-over-0.3’. Ploting the Nodes and Edges results in this visual:



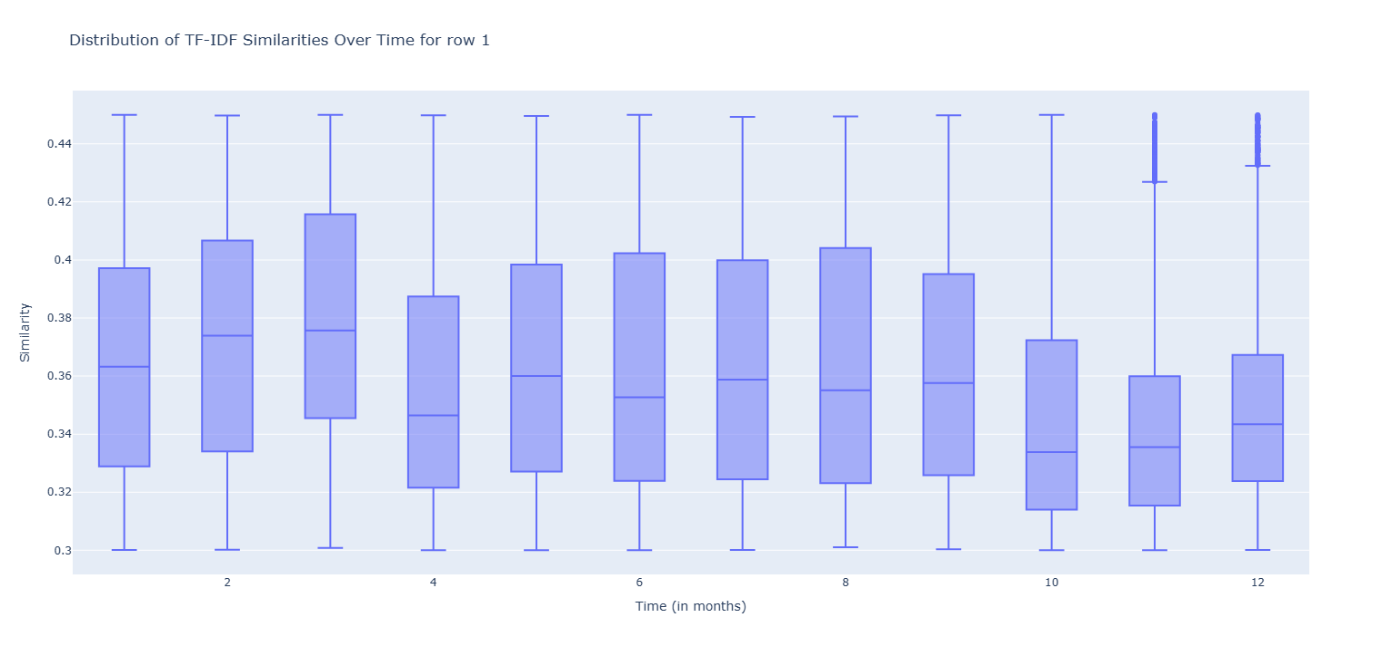
We can see the plots connected by lines, the closer the plots the more similar topics they have and the lines connecting them link them to the next related articles which fall under a different topic within the context of the Gaza conflict.

Visualization and plots:

From the exploration I thought it best to produce a visualization of the most number of similarities which lay between the 0.3 and 0.45 mark. It would show the average similarities and with outliers over a period of months from both the datasets.

It would be more beneficial to link this with topic modelling to get an greater understanding of the data and to visualize it in an easier manner to see the main topics which are linked to each other, mainly UN, resolution, Gaza and Human Rights.

Plot from the first visualization:



The plot was pretty widespread and ranged at a varying mean from 0.34-0.38. Hovering on the box plots gives more information about the plots. We can see that there are some outliers in the months 11 and 12.

Plot from the second visualization:

A graph with a row of rectangular objects

AI-generated content may be incorrect.

This plot had outliers in the first and second month with an increasing mean and the spread as we move through the months towards the end. Comparing these two graphs can help is understand the majority of the linking topics being talked about during the Gaza conflict (2023-2024) from the Al-Jazeera corpus.

The TF-IDF analysis, supported by Gephi and Plotly visualisations, provided valuable insights into the thematic clustering of news articles covering the Gaza Conflict. The visualisations highlighted meaningful patterns across time, including outliers and shifts in article similarity, pointing to evolving narratives in the news cycle. Future work could integrate topic modelling to deepen these insights and offer a more refined understanding of discourse trends throughout the conflict period.