**Analysis Report: TF-IDF Cosine Similarity Analysis of the Al Jazeera Gaza Corpus**

For Mini-Project 3 in our Digital Humanities course, I focused on analyzing the TF-IDF cosine similarity scores between article pairs in the Al Jazeera Gaza corpus. The goal was to measure how similar different articles were using Term Frequency–Inverse Document Frequency (TF-IDF) vectors and cosine similarity calculations. The dataset had been pre-processed and came in three filtered files:

tfidf-over-0.3.csv: All article pairs with similarity scores above 0.3

tfidf-over-0.3-len100.csv: Pairs with scores > 0.3 and more than 100 tokens

tfidf-over-0.3-len200.csv: Pairs with scores > 0.3 and more than 200 tokens

**1. Data Preparation and Cleaning**

I began by importing the datasets into Python using pandas. To improve readability and ensure consistency, I renamed the columns across all files. I also created two datetime columns (date1 and date2) based on article metadata (year, month, day). A new column pair\_date captured the earlier of the two dates in each article pair, which helped in creating a timeline for trend analysis. A source\_file column was also added to retain the origin of each pair.Once the three datasets were merged into a single DataFrame, I removed rows with missing dates or similarity scores. To allow for monthly aggregation, I extracted a month column using pandas Period objects.

**2. Statistical Overview and Top Similar Pairs**

I generated basic statistical summaries to understand the distribution and average of similarity scores. Then, to identify the most thematically aligned content, I extracted the top 10 article pairs (excluding self-comparisons) with the highest cosine similarity. These revealed which articles had the strongest conceptual and linguistic overlap.

**3. Visualization of Similarity Distributions and Patterns**

Using Plotly Express, I created two main visualizations:

a. Histogram of Similarity Scores  
This histogram compared score distributions across the three datasets. Saved as tfidf\_all\_similarity\_distribution.html, the visualization showed that most article pairs had scores just above the 0.3 threshold, while a few had scores exceeding 0.8. Overlaying histograms by source helped illustrate how filtering by article length influenced the results.

b. Monthly Similarity Trend Line  
To analyze thematic evolution over time, I calculated the monthly average similarity scores for each dataset. This trend line, saved as monthly\_similarity\_trend.html, revealed fluctuations likely linked to key events in Gaza. The line graph also highlighted variations depending on article length and time period.

4. Export for Gephi Network Analysis

To map article relationships as a network, I created edge and node lists compatible with Gephi:

Edges: These captured article pairs as connections, weighted by cosine similarity. I removed self-loops and filtered out weak connections (scores < 0.4) to simplify the visualization. The list was saved as edges.csv. I collected unique article IDs from both columns to form a list of nodes. The resulting nodes.csv contained basic identifiers, which can be enriched later with titles or dates. These exports allowed me to explore clusters, key themes, and influential articles using network analysis tools.

**Conclusion**

This project used TF-IDF cosine similarity to quantify how closely related articles in the Al Jazeera Gaza corpus were. The visualizations illustrated score distribution, temporal trends, and the impact of article length. Exporting to Gephi enabled further network-based interpretation, helping us trace thematic links and narrative structures during critical moments in Gaza. This analytical approach contributes to a deeper understanding of media framing and reporting patterns during major events.