drooid

Rishiraj Sutar

Question 2) . Create a timeline summarization plot or list, that shows the major events of the IsraelHamas war that had occurred during the time-period mentioned in the Json file.

Data Description

The dataset provided is a JSON file containing a collection of 37,000 news articles sourced from various media outlets. The articles span a timeframe from October 2023 to March 2024 and primarily cover topics related to the Israel-Hamas conflict. However, it is noted that not all articles within this dataset focus on the conflict; some cover unrelated topics.

Objective

Our purpose is solely to extract events. To streamline the process, we're retaining only the titles as they succinctly encapsulate the content of the article body.

Data Pre-processing

1. **Data Cleaning**

   a) **Noise Removal:**

   - **Remove Punctuation and Special Characters**: Strip out punctuation marks (e.g., `!`, `?`, `#`) and special characters (e.g., `@`, `$`, `%`).
   - **Eliminate HTML Tags**: Remove HTML tags that may be present due to web scraping.
   - **Filter Irrelevant Symbols**: Identify and remove irrelevant symbols or encoding errors that do not contribute to the text analysis.

   b) **Lowercasing:**

   Convert all text to lowercase to ensure uniformity. This step helps in reducing the number of unique tokens, thus simplifying the analysis.

2. **Tokenization:**

   Tokenization involves splitting the text into individual words or tokens.

3. **Stop Words Removal:**

   Stop words are common words (e.g., "the", "is", "in") that often do not carry significant meaning and can be removed to focus on the more informative parts of the text.

   - **Identify Stop Words**: Use predefined lists of stop words or create a custom list tailored to the dataset.
   - **Remove Stop Words**: Filter out stop words from the tokenized text.

4. **Lemmatization**

   Lemmatization reduces words to their base or root form, known as the lemma. This helps in normalizing the text and reducing the number of unique words without losing meaningful information.

Text-to-Numbers:

Machine learning algorithms do not comprehend letters. So, we must encode them in a format understandable by machines. To imbue sentences with meaning, we will utilize Spacy's pre-trained Word Embedding model, en_core_web_lg.

Each article is converted into 300th dimensional array. Thus, we obtained a matrix of dimensions 37000×300.

Clustering algorithm:

After obtaining the feature matrix, we group the titles using the DBSCAN algorithm. Unlike K-Means, DBSCAN automatically determines the number and sizes of clusters.

Cluster Index	Size
25	34
198	36
183	37
100	45
22	51
6	63
7	205
185	206
0	1033
-1	15806

To ensure similar sentences are clustered together, we aim for a higher number of classes. The -1 class represents unclustered sentences, while other numbers are cluster indexes. We choose cluster 0 because it has the most topics.

Selecting one event for a day:

- To organize the sentences chronologically and filter them by relevance, we'll display one article per day to maintain a clear and consistent timeline.
- Since multiple sentences may pertain to the same event each day, we need a method to select the most representative one.
- We can cluster the daily titles, then choose the sentence that is closest to the center of each cluster, effectively capturing the essence of the event.
- For each day, we can calculate the central vector(group mean) and choose the title which is closest to it.

Selecting the most important events:

From the 66 events obtained, we have to select the most 20 important Israel-Hamas War events to be put in the timeline. The detailed process is given below:

- **Embedding Sentences:** We used a pre-trained BERT model to convert each event's title into a numerical embedding, capturing the semantic meaning of the text.
- **Calculating Distances:** Pairwise cosine distances between these embeddings were calculated to measure how different each event is from the others.
- **Selecting Distinct Events:** An iterative algorithm identified 20 events with the greatest semantic differences, ensuring a diverse and representative set of key events.
- **Output List:** The final list of 20 distinct events was compiled for inclusion in the timeline, providing a broad and comprehensive overview of the conflict.

DATE	EVENT
2023, 10, 11	INTERVIEW: UNICEF has 'every hope' for more Gaza convoys
2023, 10, 13	Gaza's disappearing internet, visualized
2023, 10, 18	IDF claims Israel 'almost at full operational control' of northern Gaza, warns Hamas chief: 'Will meet the barrels of our guns'
2023, 10, 21	Hamas 'admits it's BROKEN' truce to carry out deadly Jerusalem attack as Israel issues chilling response
2023, 10, 30	'There's a threshold to Muslims' tolerance,' Iran's defense chief says of Gaza tragedy
2023, 11, 30	Brazil Supreme Court Rejects Bolsonaro's Request To Travel To Israel
2023, 12, 3	Russia Foreign Ministry: US veto means continued 'bloodshed' in Palestine's Gaza
2023, 12, 5	'Operation Al-Aqsa Flood' Day 78: UNSC resolution criticized as 'meaningless,' hundreds of thousands evacuate central Gaza
2023, 12, 12	There's no such thing as a 'Ramadan truce'
2023, 12, 15	'I've lost contact': Surviving Israeli bombs amid communications blackout
2023, 12, 19	Netanyahu calls Hamas's demands 'ludicrous' and proceeds with plans for a ground invasion in Rafah.
2023, 12, 22	Early Milky Way's 'Shiva & Shakti' Building Blocks Found
2023, 12, 23	Mohamed Salah 'shares pain' of grieving families amid Israel-Hamas war
2023, 12, 25	Qatar slams int'l inaction on Gaza conflict, calls for probe into Israeli 'crimes'
2024, 3, 8	'Wake the F@#K Up Shri': Palestine Supporters Storm Indian American Congressman's Michigan Home
2024, 3, 12	Turkey's Erdogan: UN Security Council has 'not fulfilled responsibility' in Gaza
2024, 3, 15	Turkish MP dies from heart attack immediately after saying 'Israel will suffer Allah's wrath' (VIDEO)
2024, 3, 22	UN chief appeals for end to Gaza's 'nightmare'
2024, 3, 24	Israeli forces 'destroy 500 Gaza tunnel entrances' to entomb Hamas as they close in on terror group's tyrant
2024, 3, 30	Gaza ceasefire by Ramadan 'looking tough': US president Joe Biden

timeline summarization plot of Israel-Hamas war