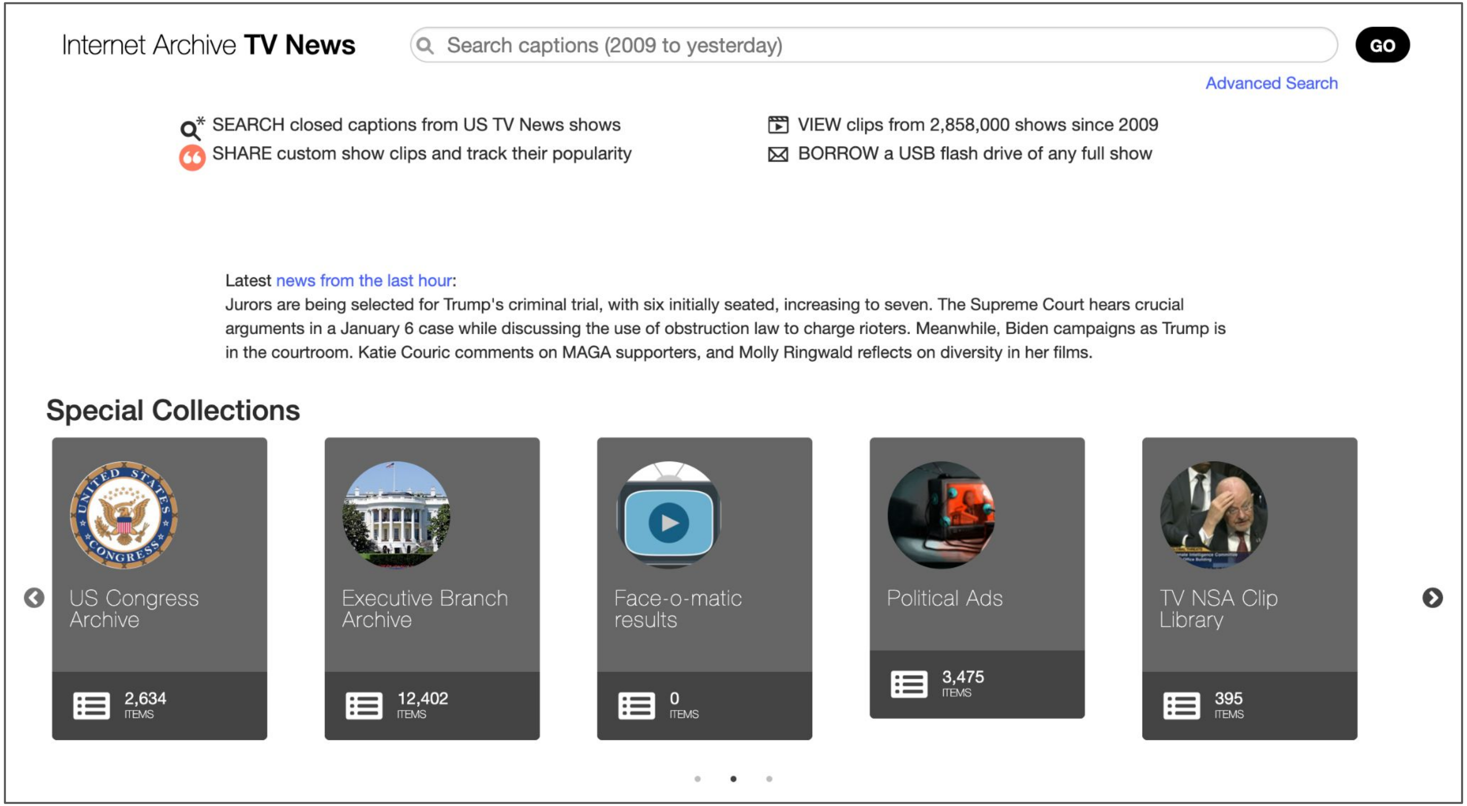
Utilizing Large Language Models for Semantic Search and Summarization of International Television News Archives

Sawood Alam *Internet Archive*
Mark Graham *Internet Archive*
Roger Macdonald *Internet Archive*
Kalev Leetaru *GDELT Project*

@ibnesayeed @waybackmachine @internetarchive

iPRES, September 17, 2024, Bijloke, Ghent, Belgium

Internet Archive: TV News Archive



SEARCH closed captions from US TV News shows

SHARE custom show clips and track their popularity

VIEW clips from 2,858,000 shows since 2009

BORROW a USB flash drive of any full show

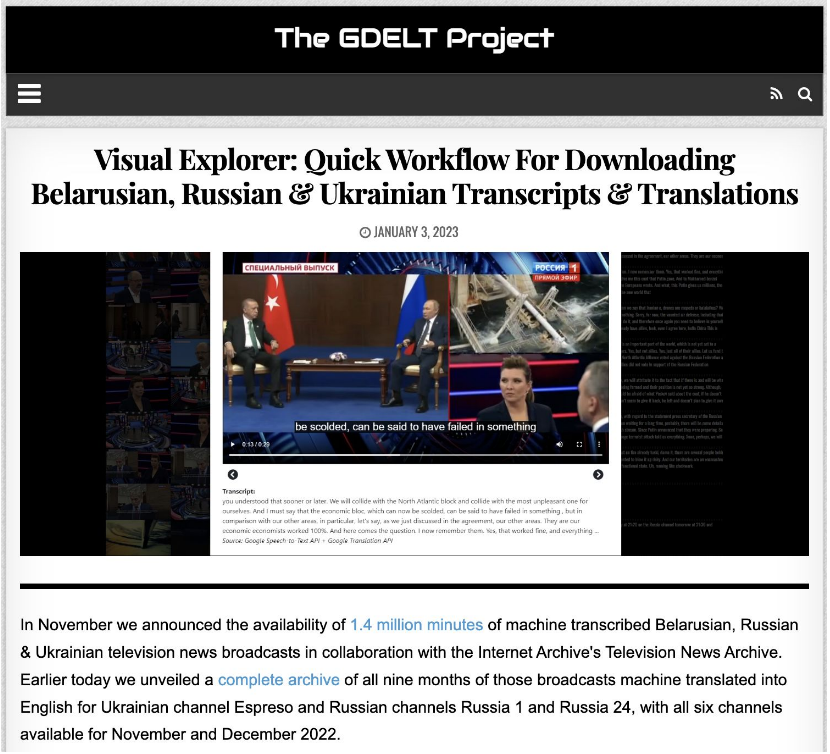
Latest news from the last hour:
Jurors are being selected for Trump's criminal trial, with six initially seated, increasing to seven. The Supreme Court hears crucial arguments in a January 6 case while discussing the use of obstruction law to charge rioters. Meanwhile, Biden campaigns as Trump is in the courtroom. Katie Couric comments on MAGA supporters, and Molly Ringwald reflects on diversity in her films.

Special Collections

Collection	Items
US Congress Archive	2,634 ITEMS
Executive Branch Archive	12,402 ITEMS
Face-o-matic results	0 ITEMS
Political Ads	3,475 ITEMS
TV NSA Clip Library	395 ITEMS

https://archive.org/details/tv

GDELT Project: Transcripts and Translations



Visual Explorer: Quick Workflow For Downloading Belarusian, Russian & Ukrainian Transcripts & Translations

JANUARY 3, 2023

be scolded, can be said to have failed in something

Transcript:
you understood that sooner or later. We will collide with the North Atlantic block and collide with the most unpleasant one for ourselves. And I must say that the economic bloc, which can now be scolded, can be said to have failed in something , but in comparison with our other areas, in particular, let's say, as we just discussed in the agreement, our other areas. They are our economic economists worked 100%. And here comes the question. I now remember them. Yes, that worked fine, and everything ...
Source: Google Speech-to-Text API + Google Translation API

In November we announced the availability of 1.4 million minutes of machine transcribed Belarusian, Russian & Ukrainian television news broadcasts in collaboration with the Internet Archive's Television News Archive. Earlier today we unveiled a complete archive of all nine months of those broadcasts machine translated into English for Ukrainian channel Espreso and Russian channels Russia 1 and Russia 24, with all six channels available for November and December 2022.

Program Inventory

https://storage.googleapis.com/data.gdeltproject.org/gdeltv3/iatv/visualexplorer/{CHANNEL}.{YYYYMMDD}.inventory.json

Transcript

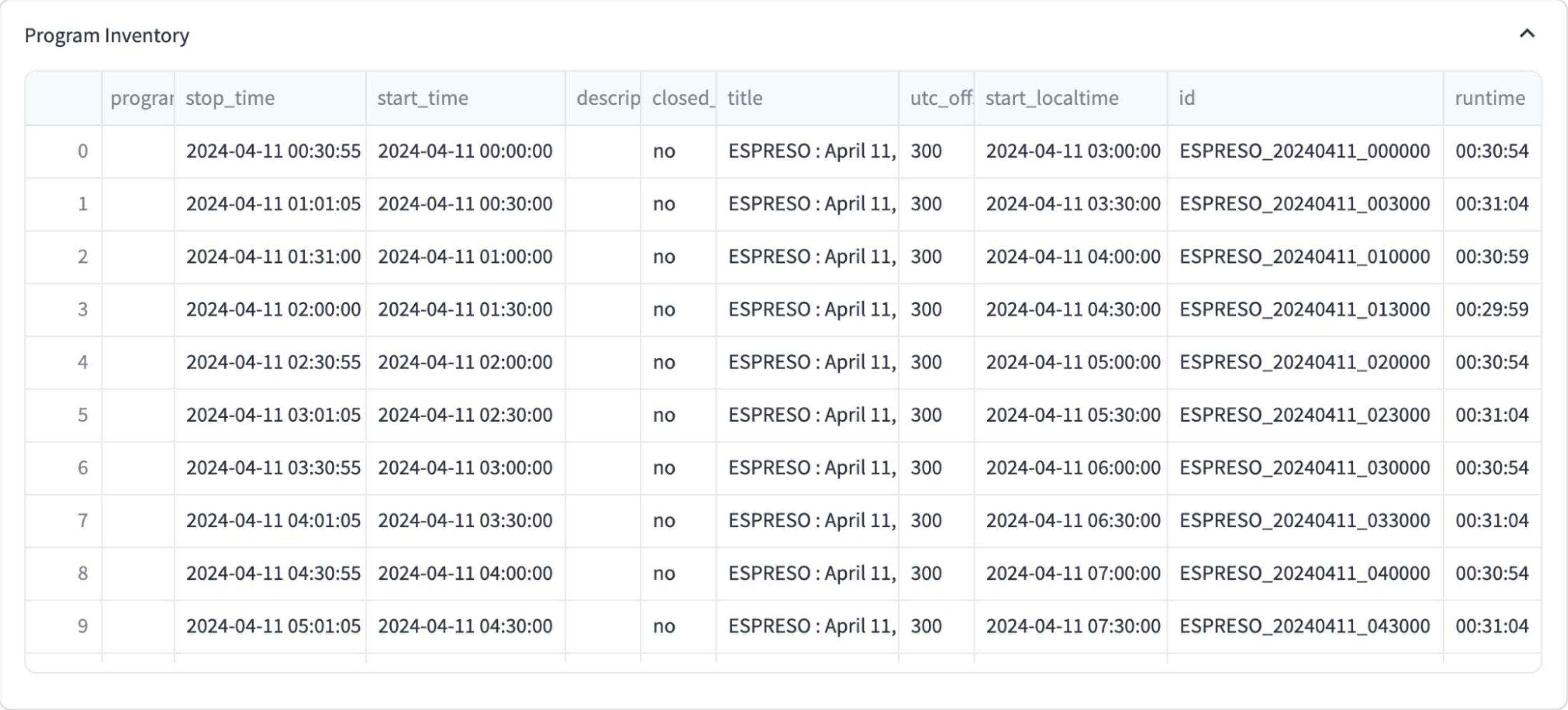
https://storage.googleapis.com/data.gdeltproject.org/gdeltv3/iatv/visualexplorer/{ID}.transcript.srt

Translation

https://storage.googleapis.com/data.gdeltproject.org/gdeltv3/iatv/visualexplorer/{ID}.transcript.en.srt

https://blog.gdeltproject.org/visual-explorer-quick-workflow-for-downloading-belarusian-russian-ukrainian-transcripts-translations/

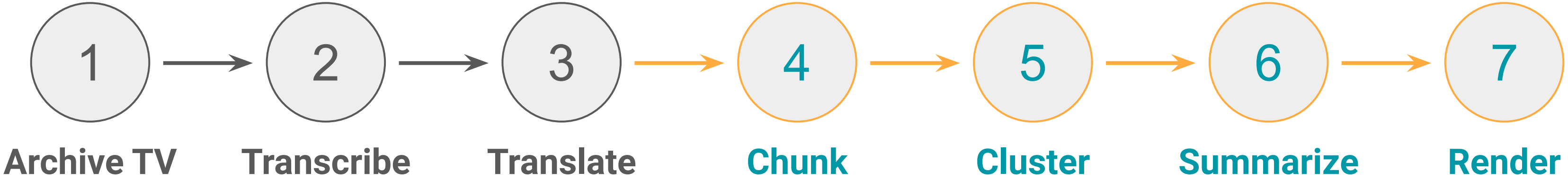
Program Inventory of a Channel



Program Inventory

	prograr	stop_time	start_time	descrip	closed_	title	utc_off	start_localtime	id	runtime
0		2024-04-11 00:30:55	2024-04-11 00:00:00		no	ESPRESO : April 11,	300	2024-04-11 03:00:00	ESPRESO_20240411_000000	00:30:54
1		2024-04-11 01:01:05	2024-04-11 00:30:00		no	ESPRESO : April 11,	300	2024-04-11 03:30:00	ESPRESO_20240411_003000	00:31:04
2		2024-04-11 01:31:00	2024-04-11 01:00:00		no	ESPRESO : April 11,	300	2024-04-11 04:00:00	ESPRESO_20240411_010000	00:30:59
3		2024-04-11 02:00:00	2024-04-11 01:30:00		no	ESPRESO : April 11,	300	2024-04-11 04:30:00	ESPRESO_20240411_013000	00:29:59
4		2024-04-11 02:30:55	2024-04-11 02:00:00		no	ESPRESO : April 11,	300	2024-04-11 05:00:00	ESPRESO_20240411_020000	00:30:54
5		2024-04-11 03:01:05	2024-04-11 02:30:00		no	ESPRESO : April 11,	300	2024-04-11 05:30:00	ESPRESO_20240411_023000	00:31:04
6		2024-04-11 03:30:55	2024-04-11 03:00:00		no	ESPRESO : April 11,	300	2024-04-11 06:00:00	ESPRESO_20240411_030000	00:30:54
7		2024-04-11 04:01:05	2024-04-11 03:30:00		no	ESPRESO : April 11,	300	2024-04-11 06:30:00	ESPRESO_20240411_033000	00:31:04
8		2024-04-11 04:30:55	2024-04-11 04:00:00		no	ESPRESO : April 11,	300	2024-04-11 07:00:00	ESPRESO_20240411_040000	00:30:54
9		2024-04-11 05:01:05	2024-04-11 04:30:00		no	ESPRESO : April 11,	300	2024-04-11 07:30:00	ESPRESO_20240411_043000	00:31:04

Methodology



1 Archive TV

2 Transcribe

3 Translate

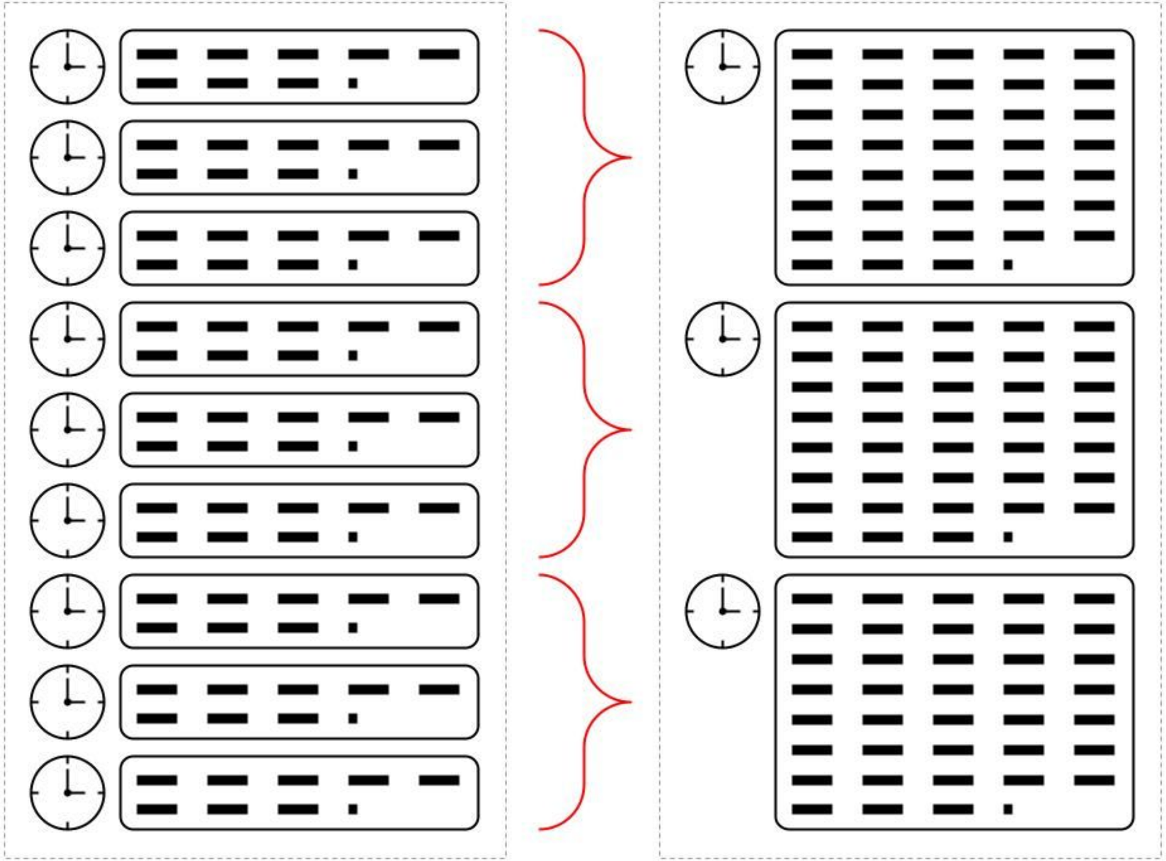
4 Chunk

5 Cluster

6 Summarize

7 Render

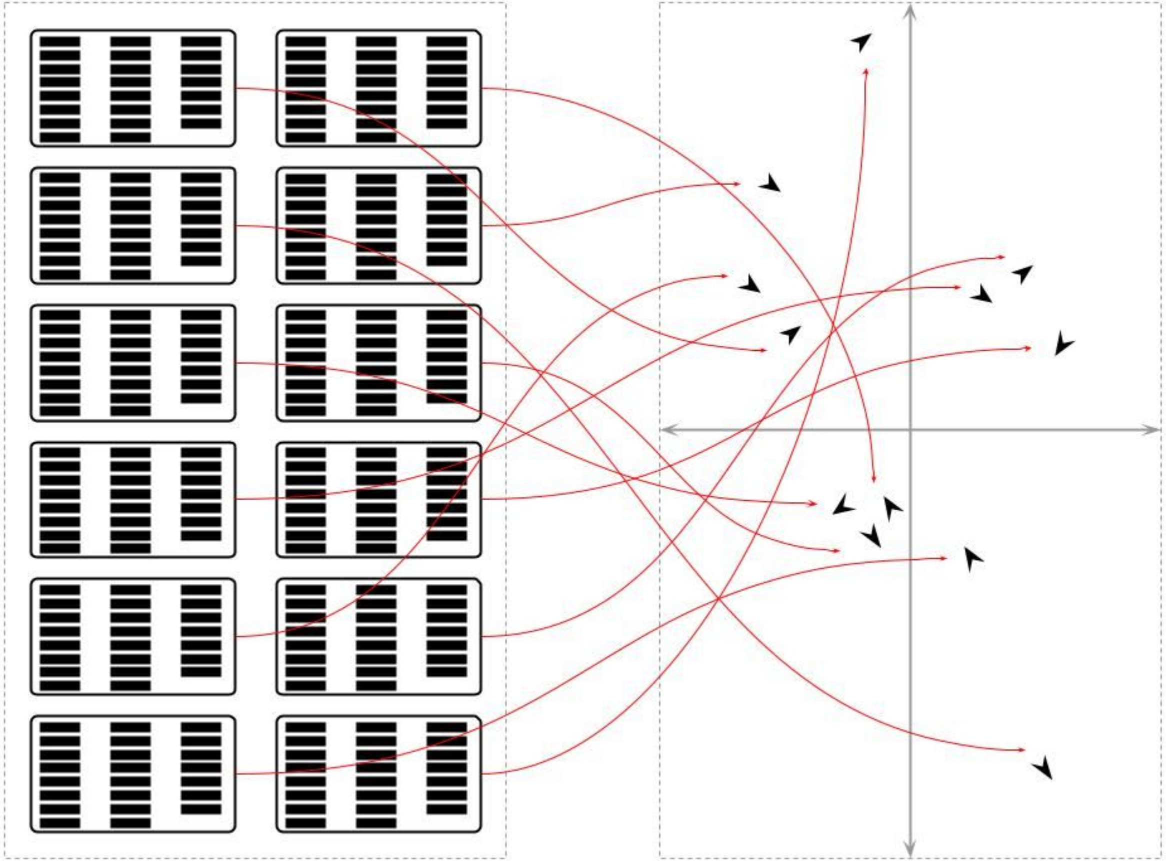
Transcript Re-Chunking (Composing Smaller Chunks)



Considerations

- Length
- Temporal alignment
- Sentence boundaries
- Overlaps
- Empty chunks

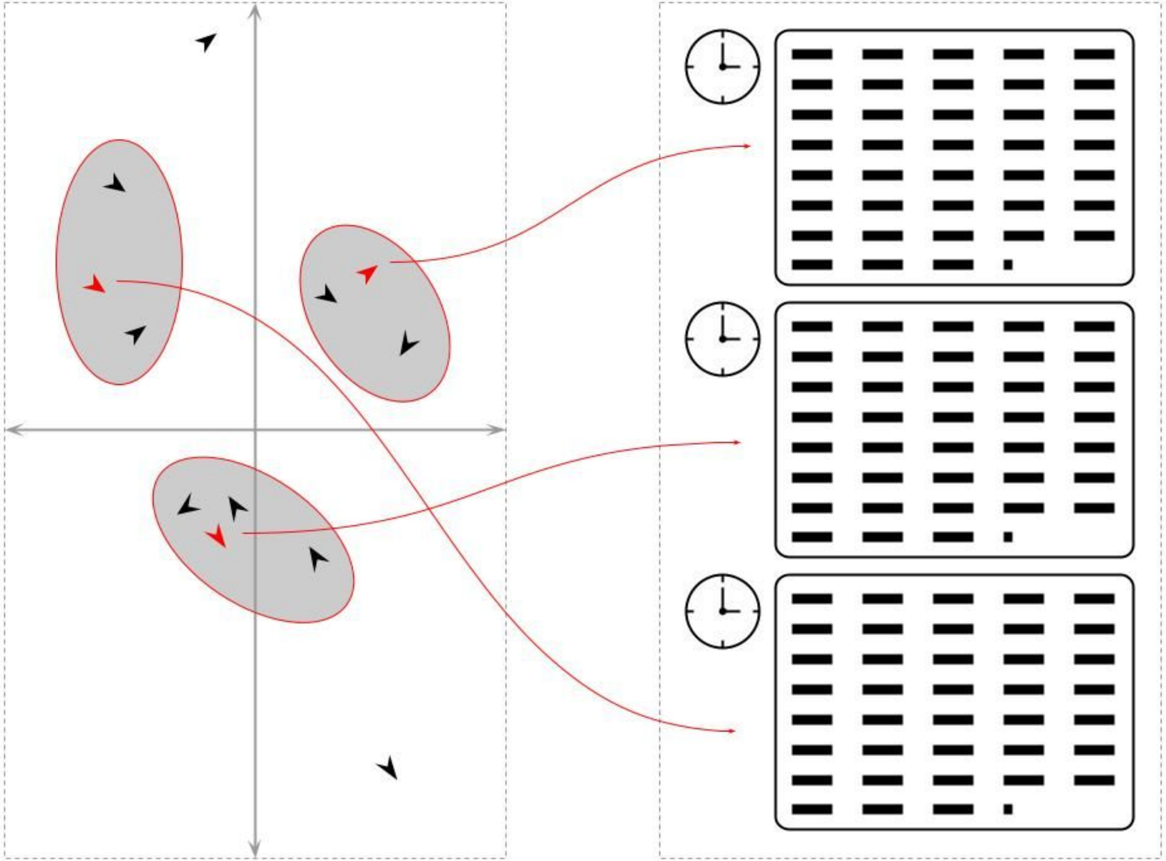
Transcript Chunk Vectorization (Doc2Vec)



Considerations

- Bag-of-Words vs. Transformer
- Language(s)
- New vocabulary
- Stop-words

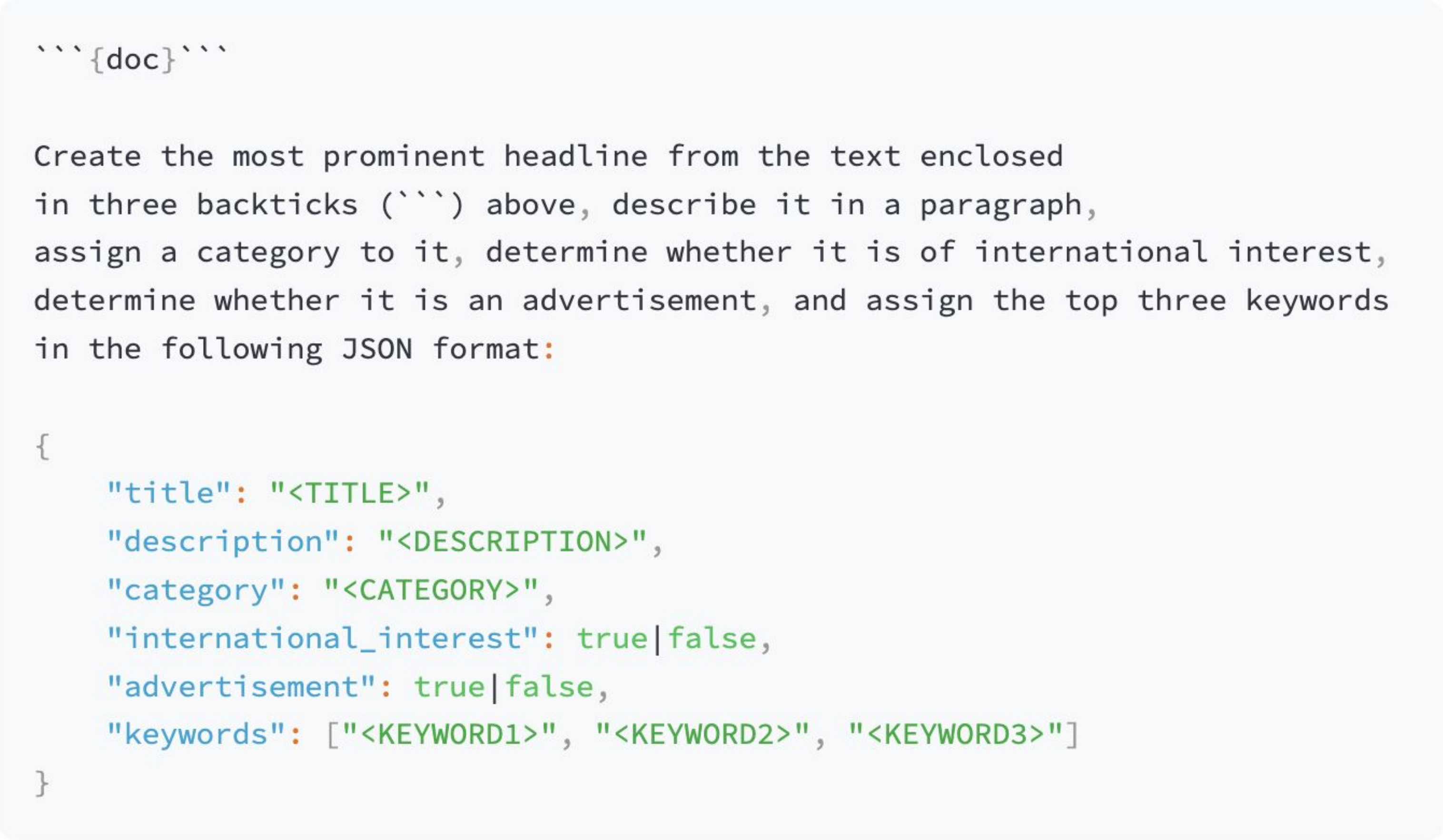
Transcript Chunk Clustering and Sampling



Considerations

- Clustering algorithm (KNN, DBSCAN, etc.)
- Sampling strategy (counts and positions)
- t-SNE projections

LLM Prompt for Transcript Chunk Summarization

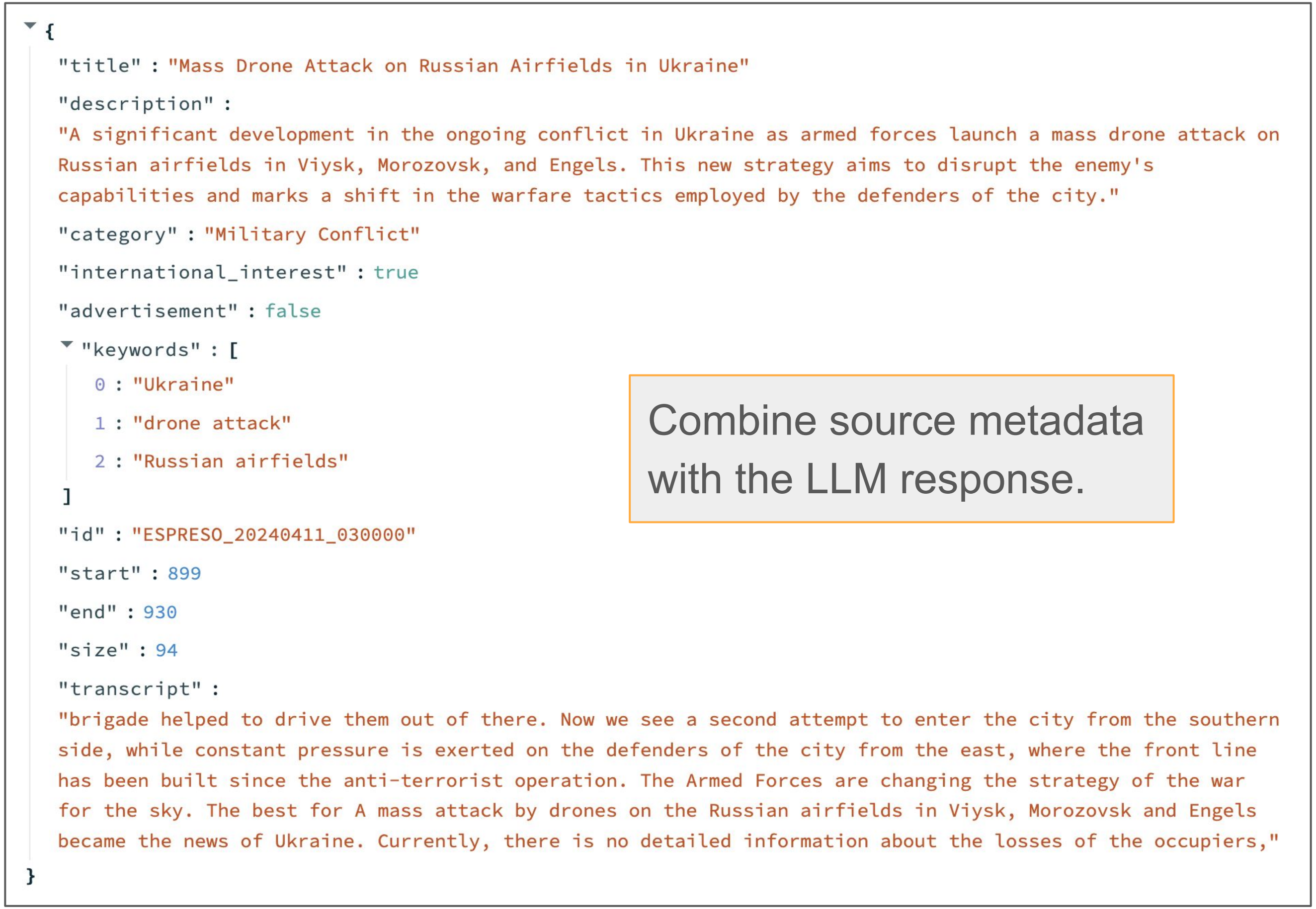


~~~
```{doc}```

Create the most prominent headline from the text enclosed
in three backticks (```) above, describe it in a paragraph,
assign a category to it, determine whether it is of international interest,
determine whether it is an advertisement, and assign the top three keywords
in the following JSON format:

{
    "title": "<TITLE>",
    "description": "<DESCRIPTION>",
    "category": "<CATEGORY>",
    "international_interest": true|false,
    "advertisement": true|false,
    "keywords": ["<KEYWORD1>", "<KEYWORD2>", "<KEYWORD3>"]
}
~~~

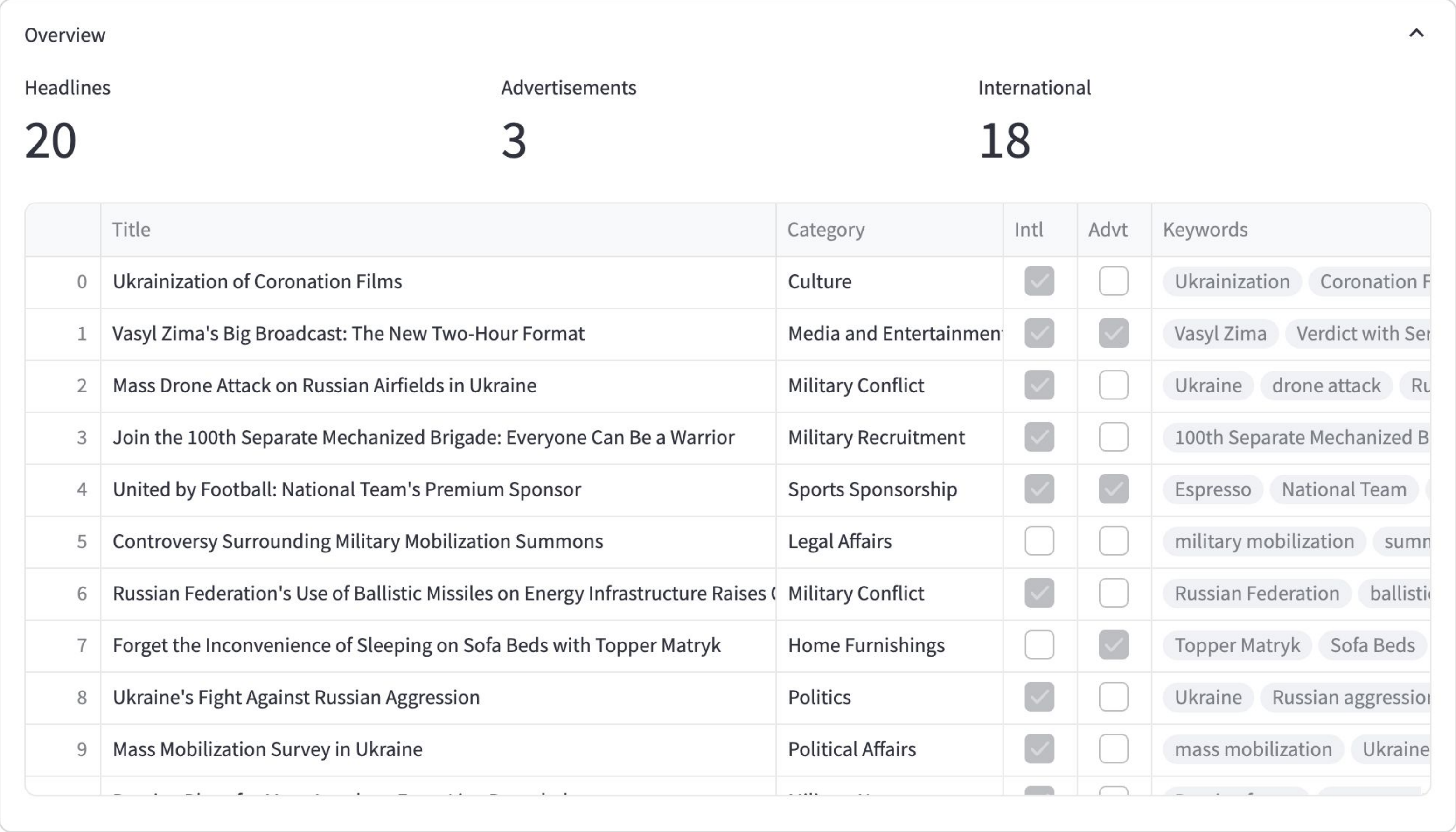
News Summary Object



```
{
  "title" : "Mass Drone Attack on Russian Airfields in Ukraine"
  "description" :
  "A significant development in the ongoing conflict in Ukraine as armed forces launch a mass drone attack on
  Russian airfields in Viysk, Morozovsk, and Engels. This new strategy aims to disrupt the enemy's
  capabilities and marks a shift in the warfare tactics employed by the defenders of the city."
  "category" : "Military Conflict"
  "international_interest" : true
  "advertisement" : false
  "keywords" : [
    0 : "Ukraine"
    1 : "drone attack"
    2 : "Russian airfields"
  ]
  "id" : "ESPRESO_20240411_030000"
  "start" : 899
  "end" : 930
  "size" : 94
  "transcript" :
  "brigade helped to drive them out of there. Now we see a second attempt to enter the city from the southern
  side, while constant pressure is exerted on the defenders of the city from the east, where the front line
  has been built since the anti-terrorist operation. The Armed Forces are changing the strategy of the war
  for the sky. The best for A mass attack by drones on the Russian airfields in Viysk, Morozovsk and Engels
  became the news of Ukraine. Currently, there is no detailed information about the losses of the occupiers,"
}
```

Combine source metadata with the LLM response.

An Overview of a Daily News Summary of a Channel

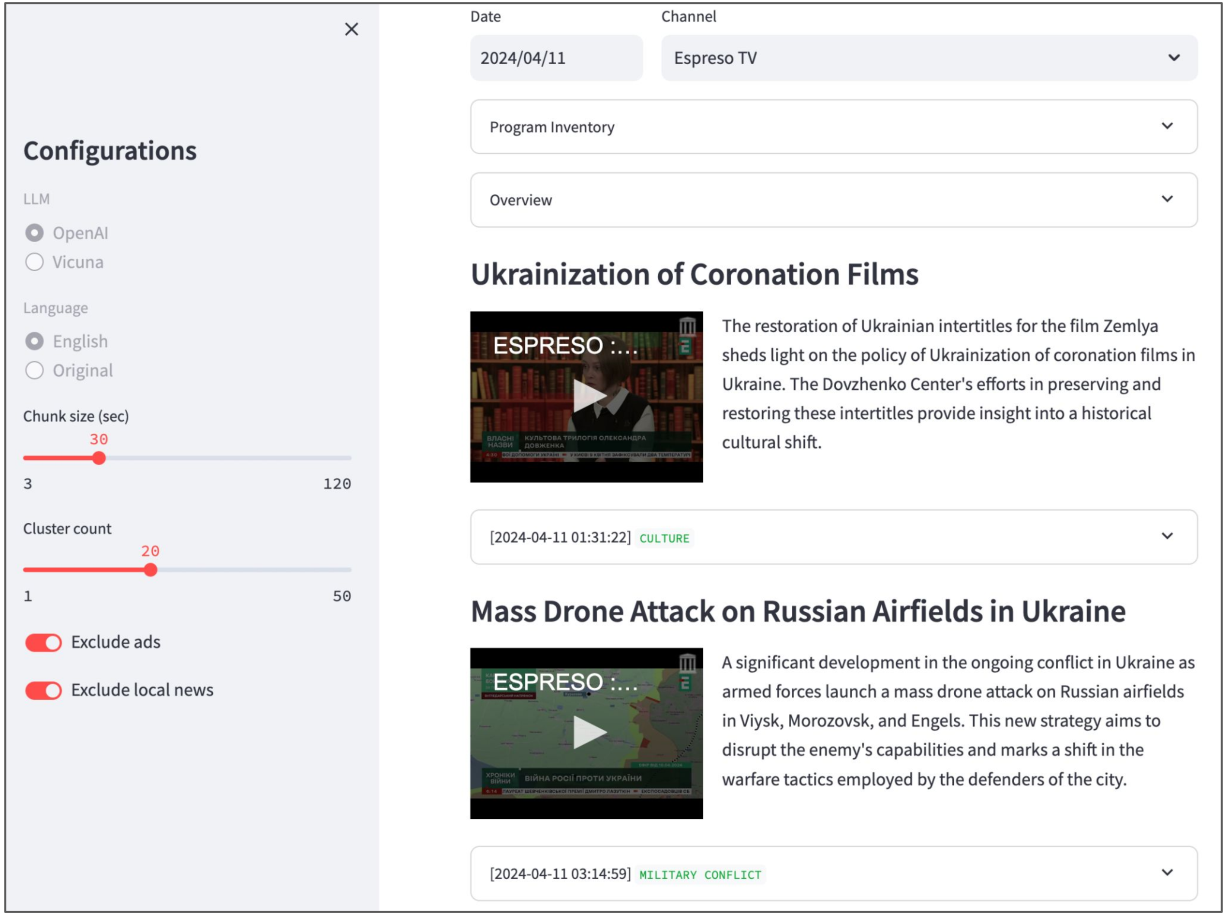


Overview

Headlines	Advertisements	International
20	3	18

	Title	Category	Intl	Advt	Keywords
0	Ukrainization of Coronation Films	Culture	☑	□	Ukrainization Coronation F
1	Vasyl Zima's Big Broadcast: The New Two-Hour Format	Media and Entertainmen	☑	☑	Vasyl Zima Verdict with Ser
2	Mass Drone Attack on Russian Airfields in Ukraine	Military Conflict	☑	□	Ukraine drone attack Ru
3	Join the 100th Separate Mechanized Brigade: Everyone Can Be a Warrior	Military Recruitment	☑	□	100th Separate Mechanized B
4	United by Football: National Team's Premium Sponsor	Sports Sponsorship	☑	☑	Espresso National Team
5	Controversy Surrounding Military Mobilization Summons	Legal Affairs	□	□	military mobilization sumn
6	Russian Federation's Use of Ballistic Missiles on Energy Infrastructure Raises (	Military Conflict	☑	□	Russian Federation ballisti
7	Forget the Inconvenience of Sleeping on Sofa Beds with Topper Matryk	Home Furnishings	□	☑	Topper Matryk Sofa Beds
8	Ukraine's Fight Against Russian Aggression	Politics	☑	□	Ukraine Russian aggressio
9	Mass Mobilization Survey in Ukraine	Political Affairs	☑	□	mass mobilization Ukraine

An Interactive Rendition of a Daily Summary



×

Configurations

LLM

● OpenAI

○ Vicuna

Language

● English

○ Original

Chunk size (sec)

30

3 120

Cluster count

20

1 50

☑ Exclude ads

☑ Exclude local news

Date: 2024/04/11

Channel: Espreso TV

Program Inventory

Overview

Ukrainization of Coronation Films

ESPRESO :...
ВЛАСНІ НАЗВИ
КУЛЬТОВА ТРИЛОГІЯ ОЛЕКСАНДРА ДОВЖЕНКА
4:30
ВОЇ ДОПОМОГИ УКРАЇНІ
У КИЄВІ 9 КВІТНЯ ЗАФІКСУВАЛИ ДВА ТЕМПЕРАТУРІ

The restoration of Ukrainian intertitles for the film Zemlya sheds light on the policy of Ukrainization of coronation films in Ukraine. The Dovzhenko Center's efforts in preserving and restoring these intertitles provide insight into a historical cultural shift.

[2024-04-11 01:31:22] CULTURE

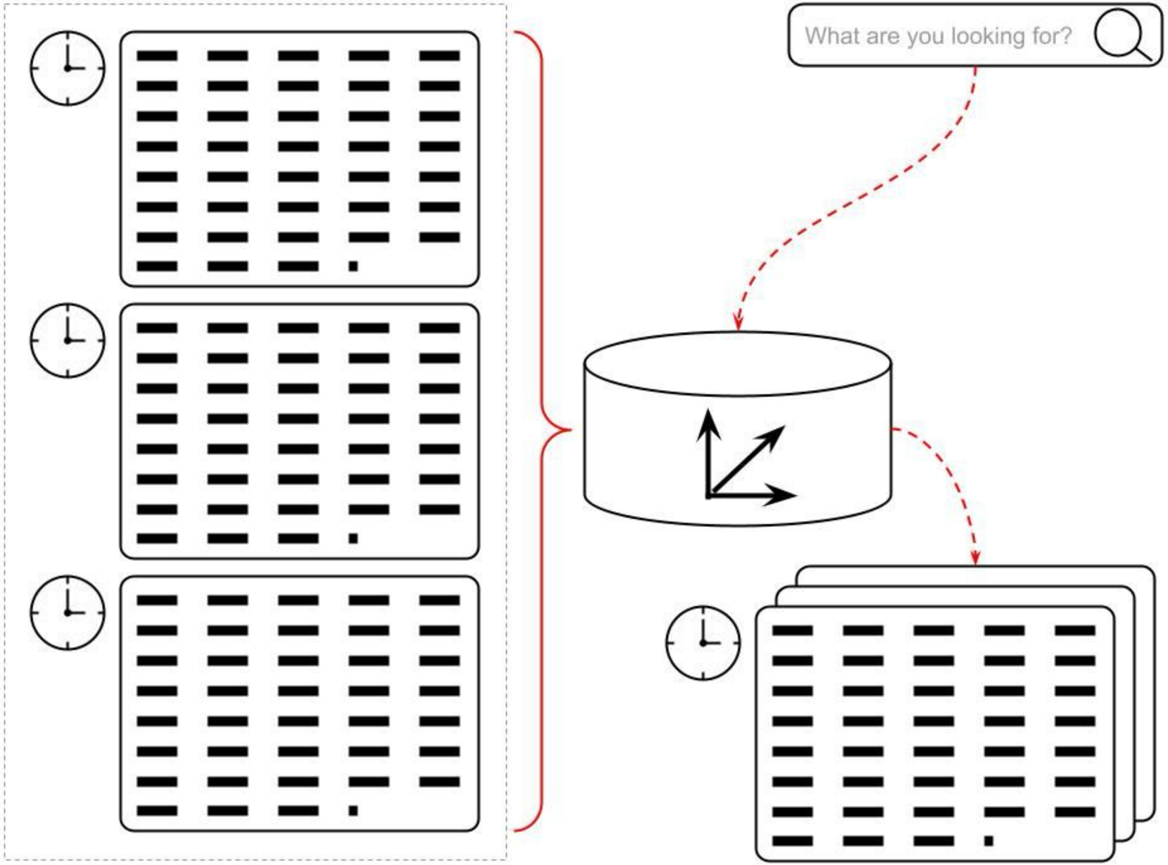
Mass Drone Attack on Russian Airfields in Ukraine

ESPRESO :...
ЕФІР ВІД 10.04.2024
ХРОНІКИ ВІЙНИ
ВІЙНА РОСІЇ ПРОТИ УКРАЇНИ
6:14
ЛАУРЕАТ ШЕВЧЕНКІВСЬКОЇ ПРЕМІЇ ДМИТРО ЛАЗУТКІН
ЕКСПОСАДОВЦІВ СБ

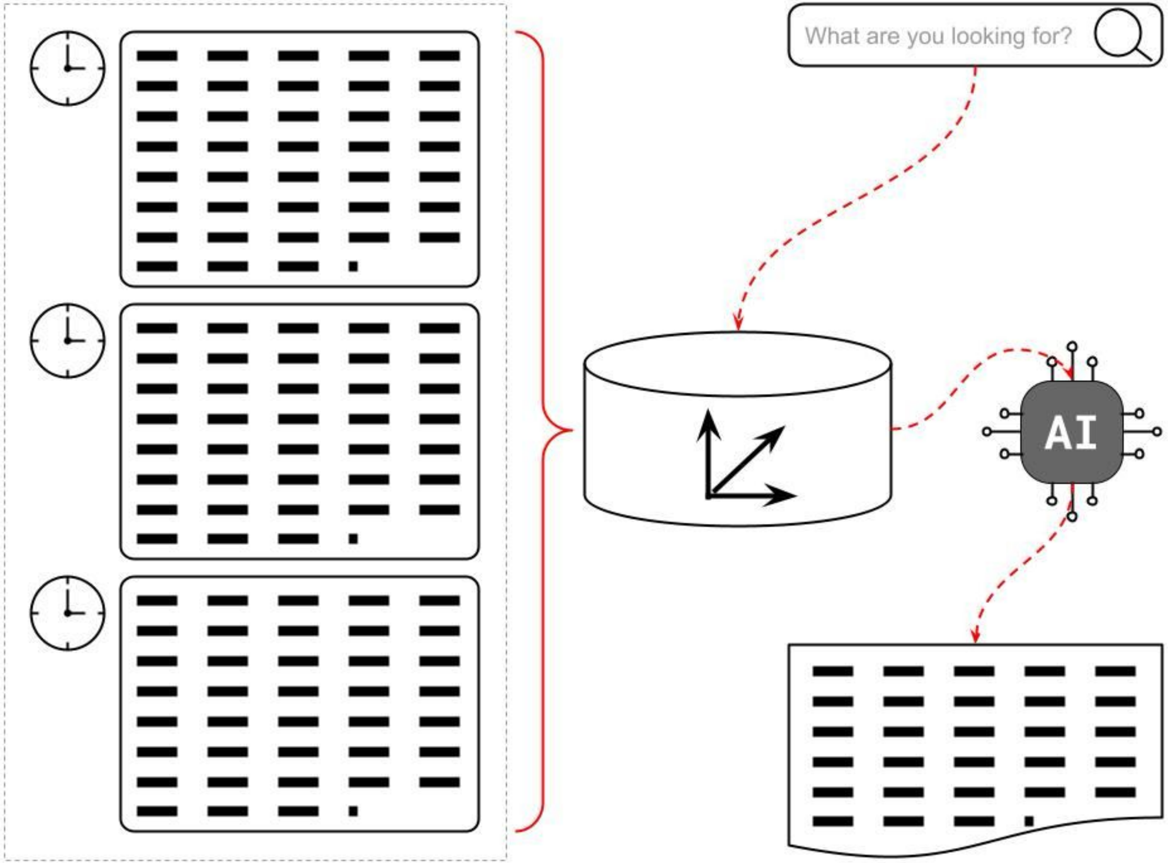
A significant development in the ongoing conflict in Ukraine as armed forces launch a mass drone attack on Russian airfields in Viysk, Morozovsk, and Engels. This new strategy aims to disrupt the enemy's capabilities and marks a shift in the warfare tactics employed by the defenders of the city.

[2024-04-11 03:14:59] MILITARY CONFLICT

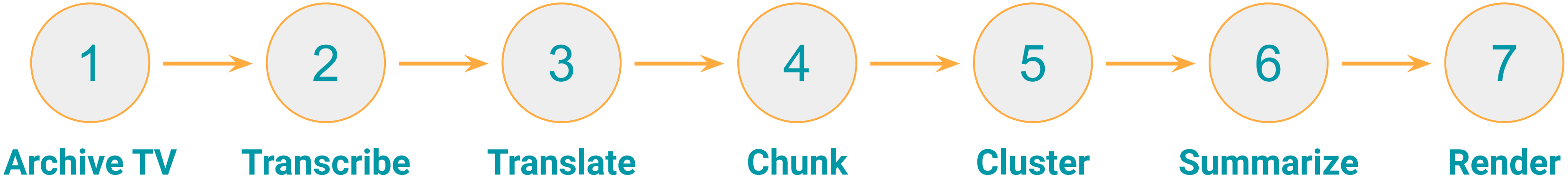
Semantic Search



Retrieval-Augmented Generation (RAG)



Conclusions



1 Archive TV → 2 Transcribe → 3 Translate → 4 Chunk → 5 Cluster → 6 Summarize → 7 Render

Code: https://github.com/internetarchive/newsum
Demo: https://newsum.sawood-dev.us.archive.org/

@ibnesayeed @waybackmachine @internetarchive