Summary

The first task was to gather data from prominent Jewish newspapers in the US, download their historical data(last 3 years) and analyze them according to main subjects.

<u>Data Gathering:</u>

• <u>archive.org</u> proved to be a very useful resource as it stores snapshots of all web sites
  and can be queried using a query with website url and timestamp.

• A list of Jewish prominent newspaper names and their corresponding web sites was found
   using a chat got prompt.

Once the us newspaper list was generated, a subsequent calls were done to the archive using archive API With timestamp parameters to download sites to the local computer.

<u>Data Cleaning and preprocessing:</u>

• Python Beautiful soup was used to parse the website page file and h2 and h3 html headlines
   text were extracted.
• The extracted text was further cleaned from special characters and stop words

<u>Data Analysis:</u>

• Since subjects we're not known, clustering method was used specifically Latent Dirichle
   Allocation(LDA) to extract topics based upon word frequency in a topic.

• Second method utilized the summary of chat got api calls to summaries main subjects.

Topics found for instance were Religious topic and Foreign News.

<u>Data Visualization:</u>
• Word Cloud was extracted from the text
• Topic visualization using inter topic LDA tool.

<u>The Second task,</u>
Searched the internet for "Yom Kippur" segments of jewish leaders.
Data was scattered with no common pattern nor format(some were pdf, some text some video)
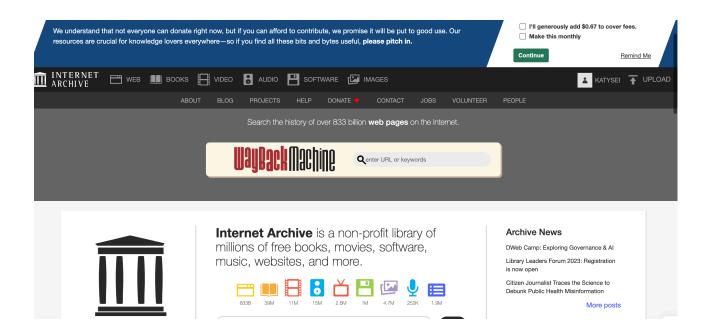A further exploration needed to be done.

Below some pics of the stages.
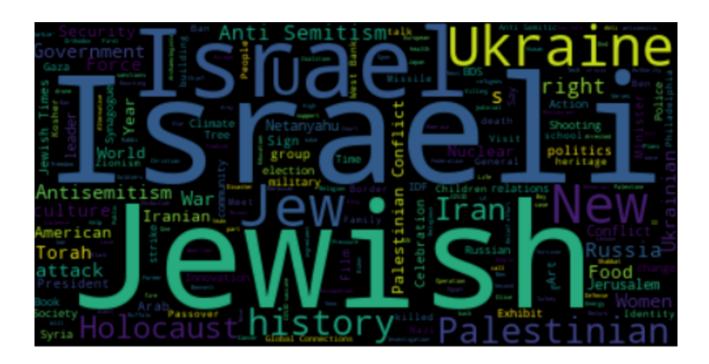Also a Jupyter notebook with code.

Jewish Newspaper List

	newspaper name	newspaper_website
0	The Forward	https://forward.com/
1	The Jewish Daily Forward	https://forward.com/
2	The Jewish Week	https://jewishweek.timesofisrael.com/
3	The New York Jewish Week	https://jewishweek.timesofisrael.com/
4	The Times of Israel	https://www.timesofisrael.com/
5	Jewish Telegraphic Agency (JTA)	https://www.jta.org/
6	The Jewish Journal of Greater Los Angeles	https://jewishjournal.com/
7	Jewish Exponent	https://jewishexponent.com/
8	The Algemeiner Journal	https://www.algemeiner.com/

Archive website.

Search the history of over 833 billion **web pages** on the Internet.

WayBackMachine   🔍 enter URL or keywords

Internet Archive is a non-profit library of millions of free books, movies, software, music, websites, and more.

833B   39M   11M   15M   2.6M   1M   4.7M   252K   1.9M

Archive News

DWeb Camp: Exploring Governance & AI

Library Leaders Forum 2023: Registration is now open

Citizen Journalist Traces the Science to Debunk Public Health Misinformation

More posts

Word Cloud

Topics

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 4 (9.1% of toker

	0	5,000	10,000	15,000

and
Jewish
Are
For
into
are
Human
South
Writing
Two
Greater
Society
Jersey
Israel,
Archival
Genealogical
Action
Saudi
Most
University
Bend
Arc:
monthly
Shows
Orthodox
People