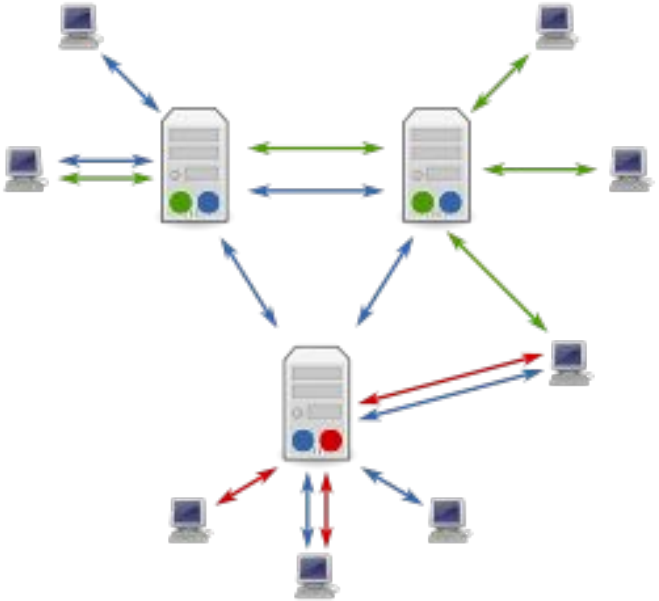
Improving Usenet Archives: How to Process Metadataless Materials

Kurt Lemai and Brian Mmari

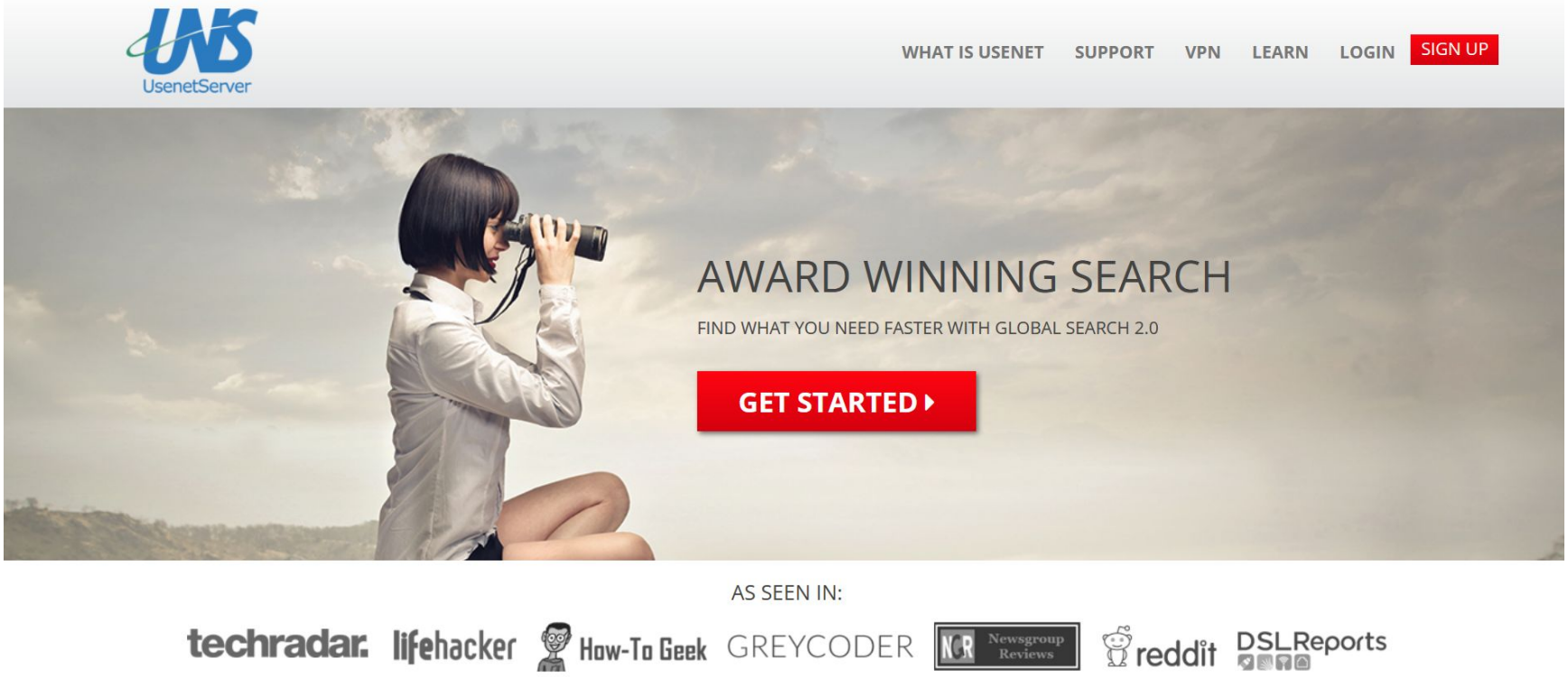
What is Usenet?

Social Media before social Media!

```
slrn 0.9.8.0 ** Press '?' for help, 'q' to quit. ** Server: localhost
 1! -    5   53:[Peter Flynn ]  2 Re: Confused
 2  D        6:[Christian Ga]    cool part
->  D  100   15:[David Kastru]     └>
 4  -        20:[Christian Ga]       ├>
[692/698 unread] Group: comp.text.tex                        — 9/222 (4%)
From: David Kastrup <dak@gnu.org>
Newsgroups: comp.text.tex
Subject: Re: cool part
Date: 24 May 2004 22:06:31 +0200

Christian Gammelgaard <cgammelXXX@stud.auc.dk> writes:

> Hello there
> Does anyone have a smart way to make a cool \part{} page?
> I have a boring one, where the number is reprecentet in roman..... and
> nothing else...

\usepackage{graphicx}
\renewcommand{\thepart}{\reflectbox{\Roman{part}}}

should be very cool.

--
David Kastrup, Kriemhildstr. 15, 44793 Bochum
UKTUG FAQ: <URL:http://www.tex.ac.uk/cgi-bin/texfaq2html>
~
1252 : Re: cool part                                         — 1/20 (All)
SPC:Pgdn  B:PgUp  u:Un-Mark-as-Read  f:Followup  n:Next  p:Prev  q:Quit
```



How can we access usenet today?



Usenet Archives

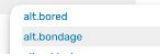
Alt Comp Microsoft Misc Net News Rec Sci Soc Talk

Posts Groups

search subjects & author names

ALTERNATIVE GROUPS

alt.bored
alt.bondage
alt.cable-ip
alt.bullshit
alt.business
alt.biz.misc
alt.bitterness



INTERNET ARCHIVE

WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES

ABOUT BLOG PROJECTS HELP DONATE CONTACT JOBS

Usenet

Usenet Archive

Usenet is a worldwide distributed Internet discussion system. It was developed from the general purpose UUCP dial-up netw
Truscott and Jim Ellis conceived the idea in 1979 and it was established in 1980. Users read and post messages (called artic
categories, known as newsgroups. Usenet resembles a bulletin board system (BBS) in many respects, and is the precursor t

More...

COLLECTION ABOUT

79,449 Results

Search

Search metadata

Search text contents (no results)

> Filters

Sort by: Weekly views Title Date published Creator

Giganews Usenet Collection

25,328 items
3.7 terabytes

Usenet Historical Collection

1,019 items
646.1 gigabytes

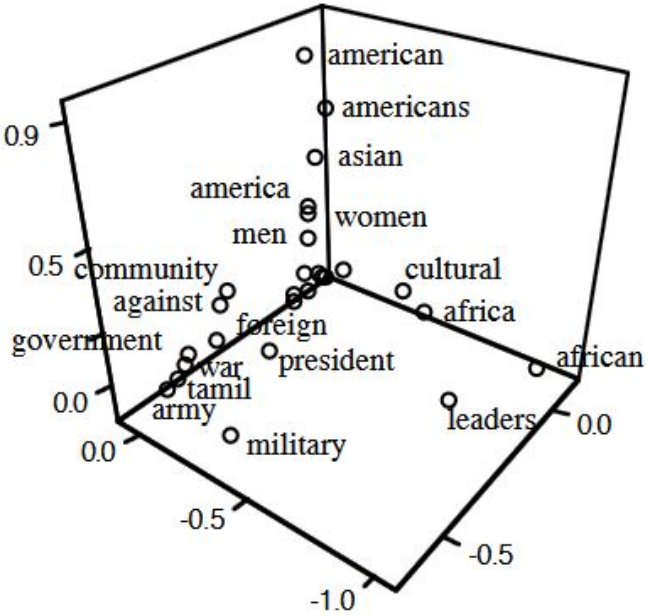
Usenet groups within alt.sex from giganews.com

107 10 0

Usenet grou
comp.lang f
giganews.c

97

What have other people done?



(a) Factor loadings of topic words

Figure 2.6: Example of conversation map interface

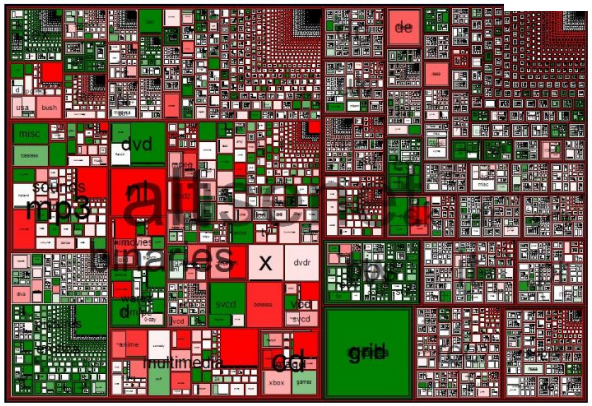


Figure 2.9: Sample tree map [63]

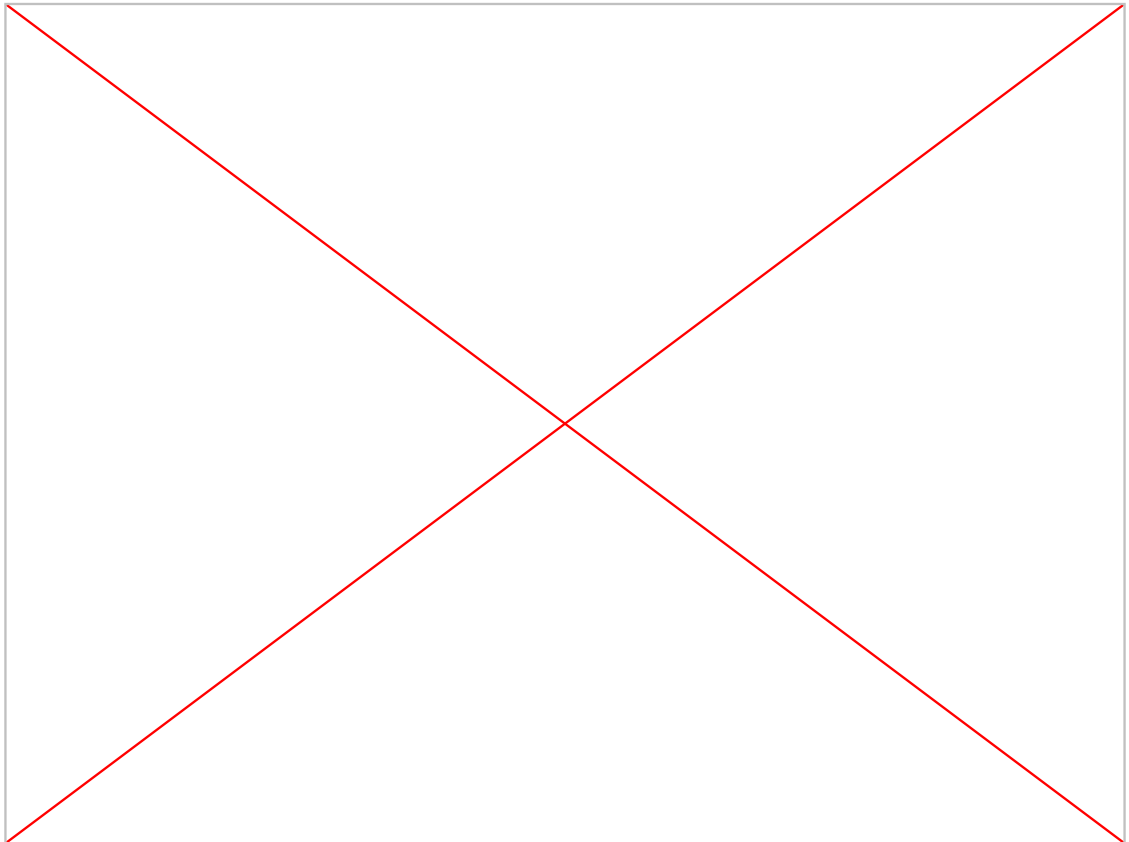
What are our goals?

- Find better ways to search usenet
- Visualize usenet with contemporary methods
- Enable new ways of browsing usenet

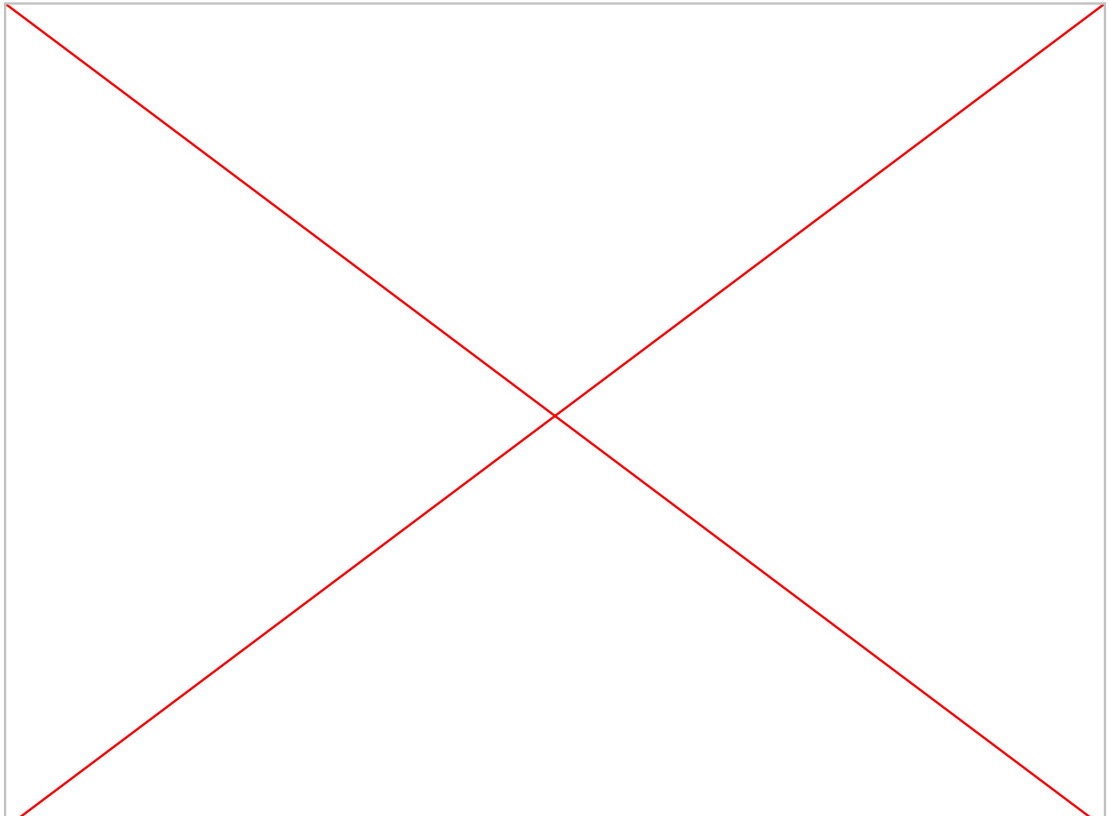
Methods

- 20 Usegroups Dataset cleaned from Huggingface
- Vectorize Messages (all-MiniLM-L6-v2)
- Store Vectors into a Vector Database (FAISS)
- 2D Projection (UMAP)
- Vector Search
- Retrieval Augment Generation (gpt-4o)
- Deploy on Dash App with ngrok as server

Vector Mapping With UMAP 2d Projection

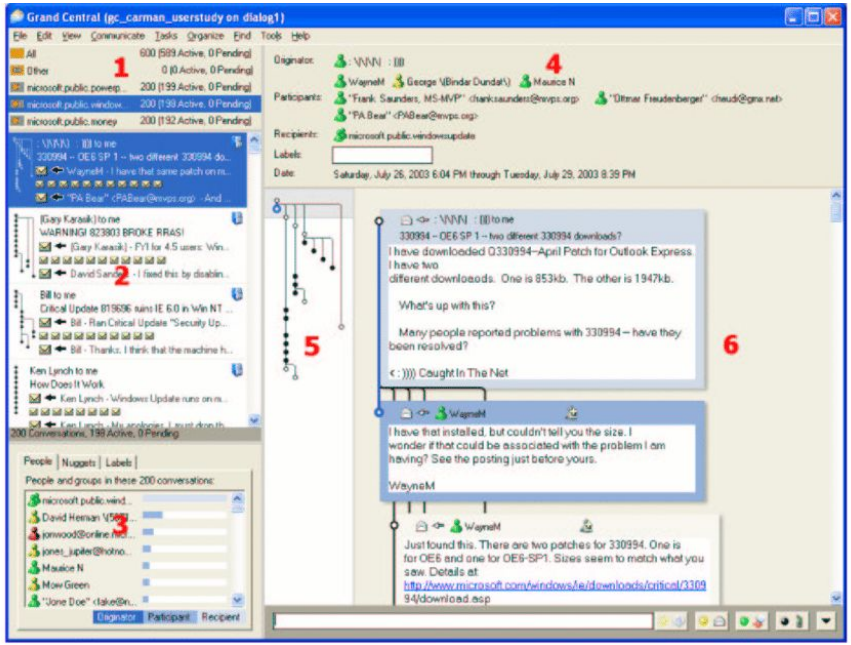


Vector Search and RAG



Limitations and Future Improvements

- This is not a system meant to be comprehensive – it is a heuristic
- Bias in vectorization
- Expand to a larger dataset
- Consider other forms of mapping
- Linking RAG Chat to texts



References

Ibrahim, Rowaida Khalil, Subhi R. M. Zeebaree, Karwan Jacksi, Mohammed A. M. Sadeeq, Hanan M. Shukur, and Ahmed Alkhayyat. "Clustering Document Based Semantic Similarity System Using TFIDF and K-Mean." In *2021 International Conference on Advanced Computer Applications (ACA)*, 28–33. Maysan, Iraq: IEEE, 2021. https://doi.org/10.1109/ACA52198.2021.9626822.

"Internet Archive: Digital Library of Free & Borrowable Texts, Movies, Music & Wayback Machine." Accessed April 25, 2025. https://archive.org/details/usenet.

Nguyen, Ha Dung, Thi-Hoang Anh Nguyen, and Thanh Binh Nguyen. "A Proposed Large Language Model-Based Smart Search for Archive System." arXiv, January 13, 2025. https://doi.org/10.48550/arXiv.2501.07024.

Paolillo, J.C. "Visualizing Usenet: A Factor-Analytic Approach." In *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, 10 pp.-, 2000. https://doi.org/10.1109/HICSS.2000.926716.

Sack, Warren. "Conversation Map: A Content-Based Usenet Newsgroup Browser." In *Proceedings of the 5th International Conference on Intelligent User Interfaces*, 233–40. New Orleans Louisiana USA: ACM, 2000. https://doi.org/10.1145/325737.325856.

"UsenetArchives.Com." Accessed April 25, 2025. https://usenetarchives.com/.

Wang, Mari. "NewsView: A Recommender System for Usenet Based on FAST Data Search," 2004. https://www.duo.uio.no/handle/10852/9103.