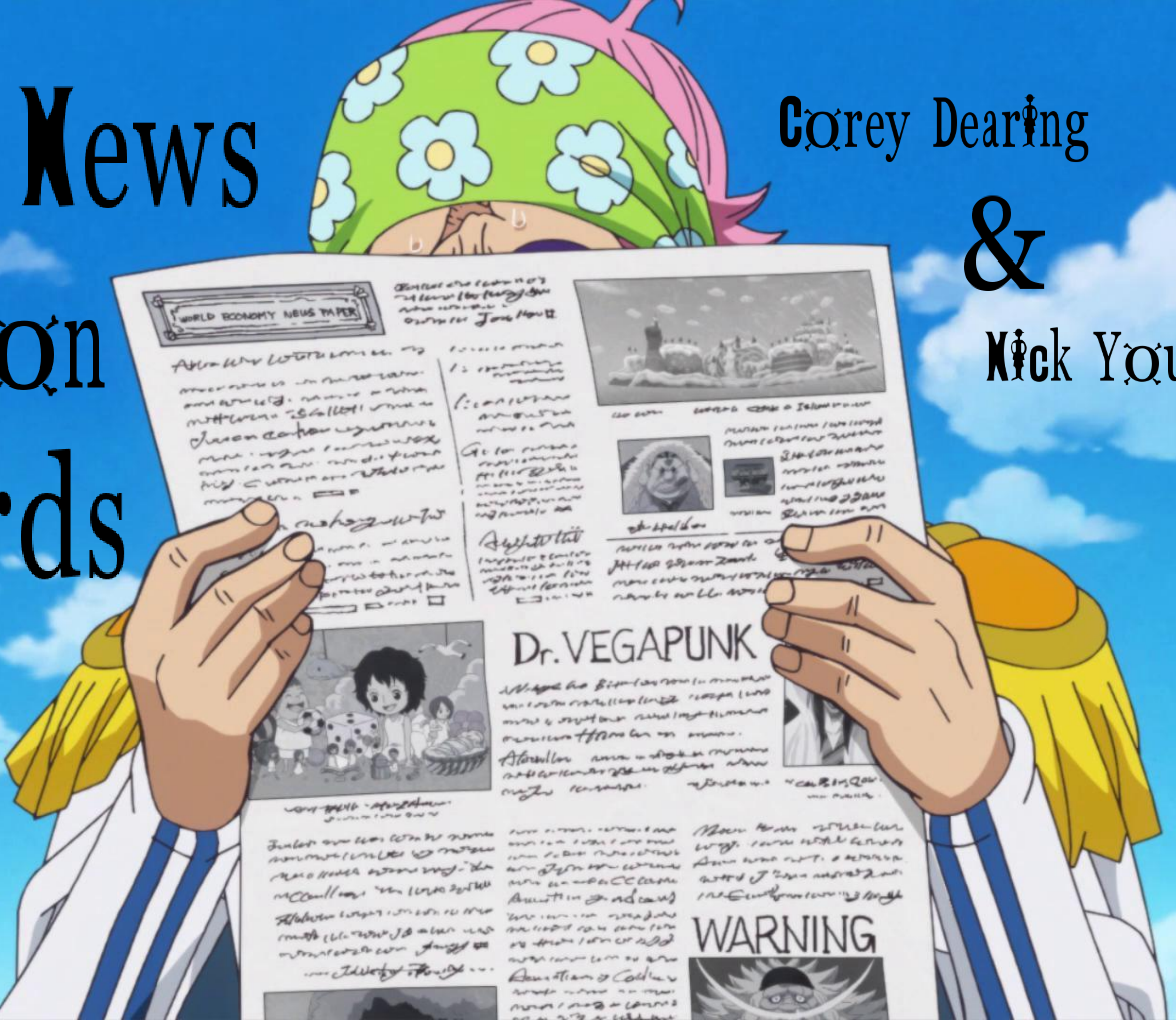
Building News Summarization Dashboards

Corey Dearing

&

Nick Young



DSC 592 Dr. Brown
Fall 2024

Project Outline

- GDELT APIv2
  - Currently stores 3 primary data table endpoints: events, gkg (global knowledge graph), and mentions.
- GKG Tools
  - Query, Parse, Tokenize, Vectorize, and Web Scraping Tools with Beautifulsoup.
- Zeroshot Classification
  - Labeling News Documents with Facebooks BART Multi-Genre Natural Language Inference (MNLI)
- TSNE
  - Representing Vectorized Article Clusters using t-Stochastic Neighbor Embeddings.
- Article Summarization
  - Article Summarization using BART CNN
- PyTesseract OCR
  - Extracting Textual Data from Images
- Plotly Dash News Dashboard

GDELT Project

- Open-source data pipeline that monitors news media worldwide in real-time to track global events, language, people, locations, organizations, and the tone of events.
- Updated every 15 minutes with various sources of data spanning back to January 1st, 1971.
- 15 min. updates typically contain 1000s of new records.

Watching The Entire World

GDELT monitors the world's news media from nearly every corner of every country in print, broadcast, and web formats, in over 100 languages, every moment of every day.

A Global Database of Society

Supported by Google Jigsaw, the GDELT Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day, creating a free open platform for computing on the entire world.

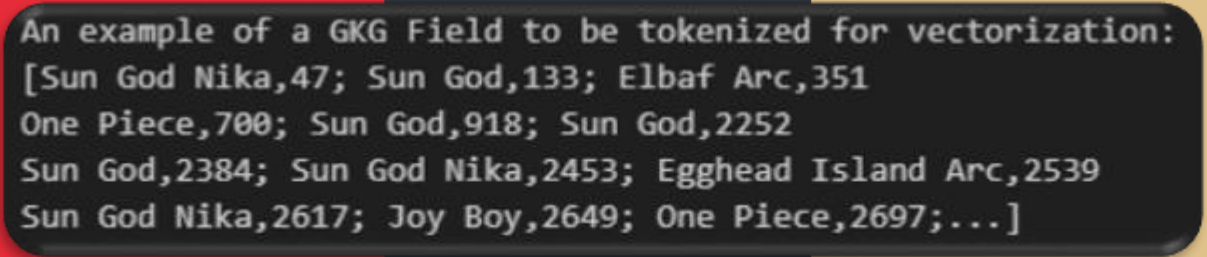
GDELT API Version 2.0

- Contains three main datasets recorded from GDELT – GKG, Events, and Mentions – We will be working with the GKG Database.
- GKG Database - Identifies persons/entities and tracks their appearances across different articles.
  - Records tone and themes underlying events.
  - Useful for identifying characters or an author from a manga.
- GKG expands GDELT's ability to quantify global human society beyond cataloging physical occurrences towards actually representing all of the latent dimensions, geography, and network structure of the global news.
- It applies an array of highly sophisticated natural language processing algorithms to each document to compute a range of codified metadata encoding key latent and contextual dimensions of the document.
- To sum up the GKG in a single sentence, it connects every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what's happening around the world, what its context is and who's involved, and how the world is feeling about it, every single day.

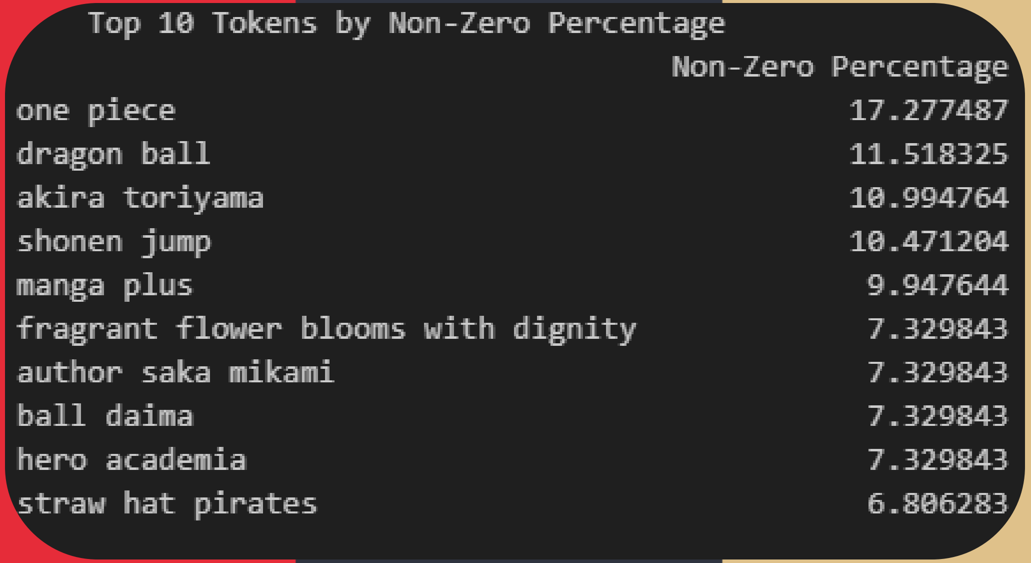
GKG Tools

- GKG Tools is a class designed for working with the GKG Global Knowledge Graph Table.
- GKG Tools can be used for custom queries date range and field filtering queries.
- GKG Tools can tokenize fields and vectorize records with respect to a set of common field tokens. Basic statistics of the vectorized space are available.
- GKG Tools vectorizer is a custom version of SKLearn TFIDF Vectorizer
- GKG Tools can be used to request document titles, headers, paragraphs, images.

```
Fields = ['amounts', 'persons', 'v2persons',
'allnames', 'themes', 'v2tone', 'v2location',
'locations', 'v2organizations', 'organizations',
'v2counts', 'counts']
```

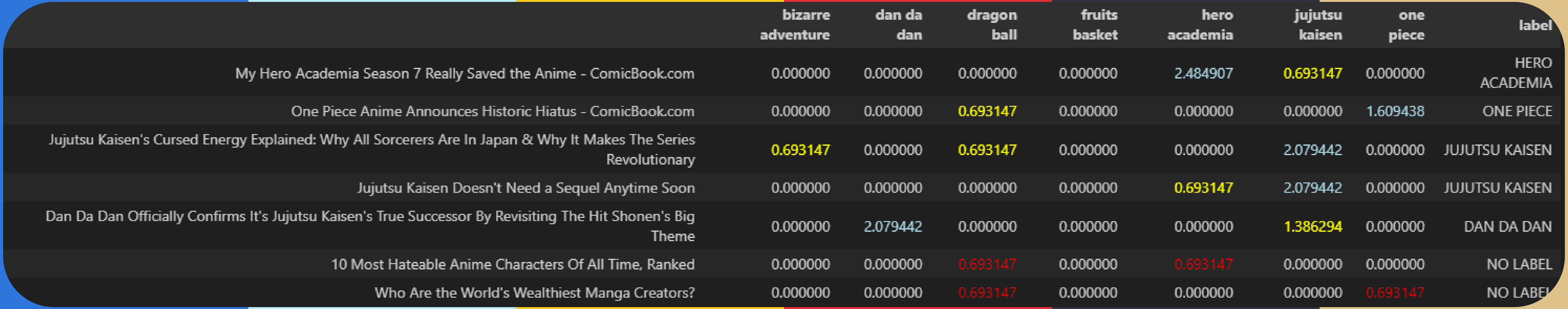


```
An example of a GKG Field to be tokenized for vectorization:
[Sun God Nika,47; Sun God,133; Elbaf Arc,351
One Piece,700; Sun God,918; Sun God,2252
Sun God,2384; Sun God Nika,2453; Egghead Island Arc,2539
Sun God Nika,2617; Joy Boy,2649; One Piece,2697;...]
```



```
    Top 10 Tokens by Non-Zero Percentage
                                      Non-Zero Percentage
one piece                                       17.277487
dragon ball                                     11.518325
akira toriyama                                  10.994764
shonen jump                                     10.471204
manga plus                                       9.947644
fragrant flower blooms with dignity              7.329843
author saka mikami                               7.329843
ball daima                                       7.329843
hero academia                                    7.329843
straw hat pirates                                6.806283
```

GKG Field Vectorizer



	bizarre adventure	dan da dan	dragon ball	fruits basket	hero academia	jujutsu kaisen	one piece	label
My Hero Academia Season 7 Really Saved the Anime - ComicBook.com	0.000000	0.000000	0.000000	0.000000	2.484907	0.693147	0.000000	HERO ACADEMIA
One Piece Anime Announces Historic Hiatus - ComicBook.com	0.000000	0.000000	0.693147	0.000000	0.000000	0.000000	1.609438	ONE PIECE
Jujutsu Kaisen's Cursed Energy Explained: Why All Sorcerers Are In Japan & Why It Makes The Series Revolutionary	0.693147	0.000000	0.693147	0.000000	0.000000	2.079442	0.000000	JUJUTSU KAISEN
Jujutsu Kaisen Doesn't Need a Sequel Anytime Soon	0.000000	0.000000	0.000000	0.000000	0.693147	2.079442	0.000000	JUJUTSU KAISEN
Dan Da Dan Officially Confirms It's Jujutsu Kaisen's True Successor By Revisiting The Hit Shonen's Big Theme	0.000000	2.079442	0.000000	0.000000	0.000000	1.386294	0.000000	DAN DA DAN
10 Most Hateable Anime Characters Of All Time, Ranked	0.000000	0.000000	0.693147	0.000000	0.693147	0.000000	0.000000	NO LABEL
Who Are the World's Wealthiest Manga Creators?	0.000000	0.000000	0.693147	0.000000	0.000000	0.000000	0.693147	NO LABEL

- GKG field vectorizer maps each articles field items into a space where the components are the field tokens.
- Different methods are available for weighting vectors by their non-zero token components.
- Method: Summation, Boolean Count, Log1p.
- The resulting vector representations can be used to final article clusters within their fields token space as well as for labeling, as seen above.

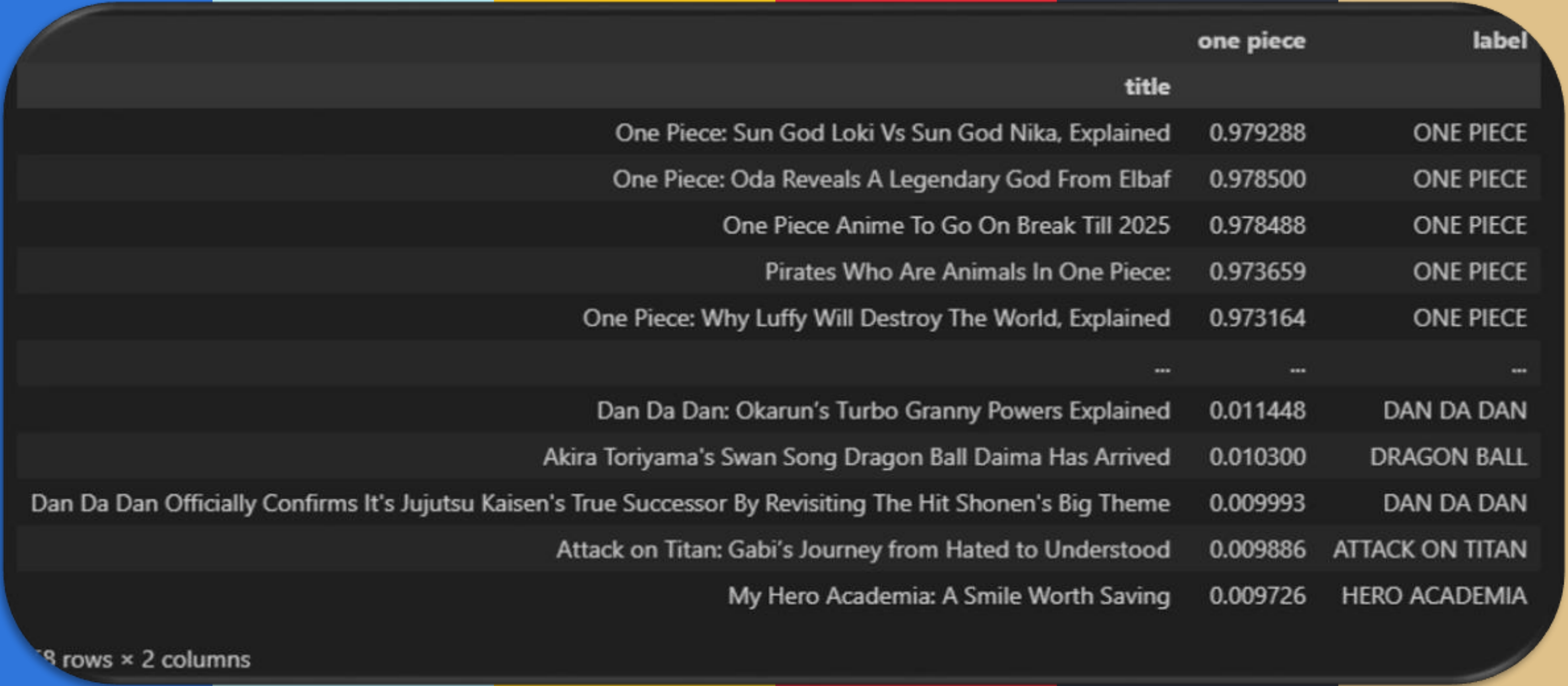
Beautifulsoup

- Parsing text from article titles.
  - Extracting headers, titles, and paragraphs.
- Retrieve articles only from Screenrant, Anime News Network, Gamerant, and Comicbook.com
  - Roughly 100 articles every Sunday.
  - Cut article into chunks to help with structuring and identifying the core paragraph in an article.
  - Article Labeling & Summarizing with Facebook's BART MNLI & CNN.

Zero-shot Classification with BART (MNLI)

- **Easy Integration:** Leverages Hugging Face's zero-shot classification pipeline for flexible, out-of-the-box text categorization without needing training data.
- **Model Choice:** Uses **facebook/bart-large-mnli**, a robust model pre-trained on the Multi-Genre Natural Language Inference (MNLI) dataset, making it effective for zero-shot tasks.
- **No Fine-Tuning Required:** Automatically classifies text into pre-defined labels, such as anime titles, without model retraining.
- **Efficient Processing:** Uses a batched approach for classifying multiple titles at once, improving processing time.
- **Scalable:** Easily applies to large datasets, making it ideal for classifying extensive lists like article titles in one go.

Zero-shot Classification

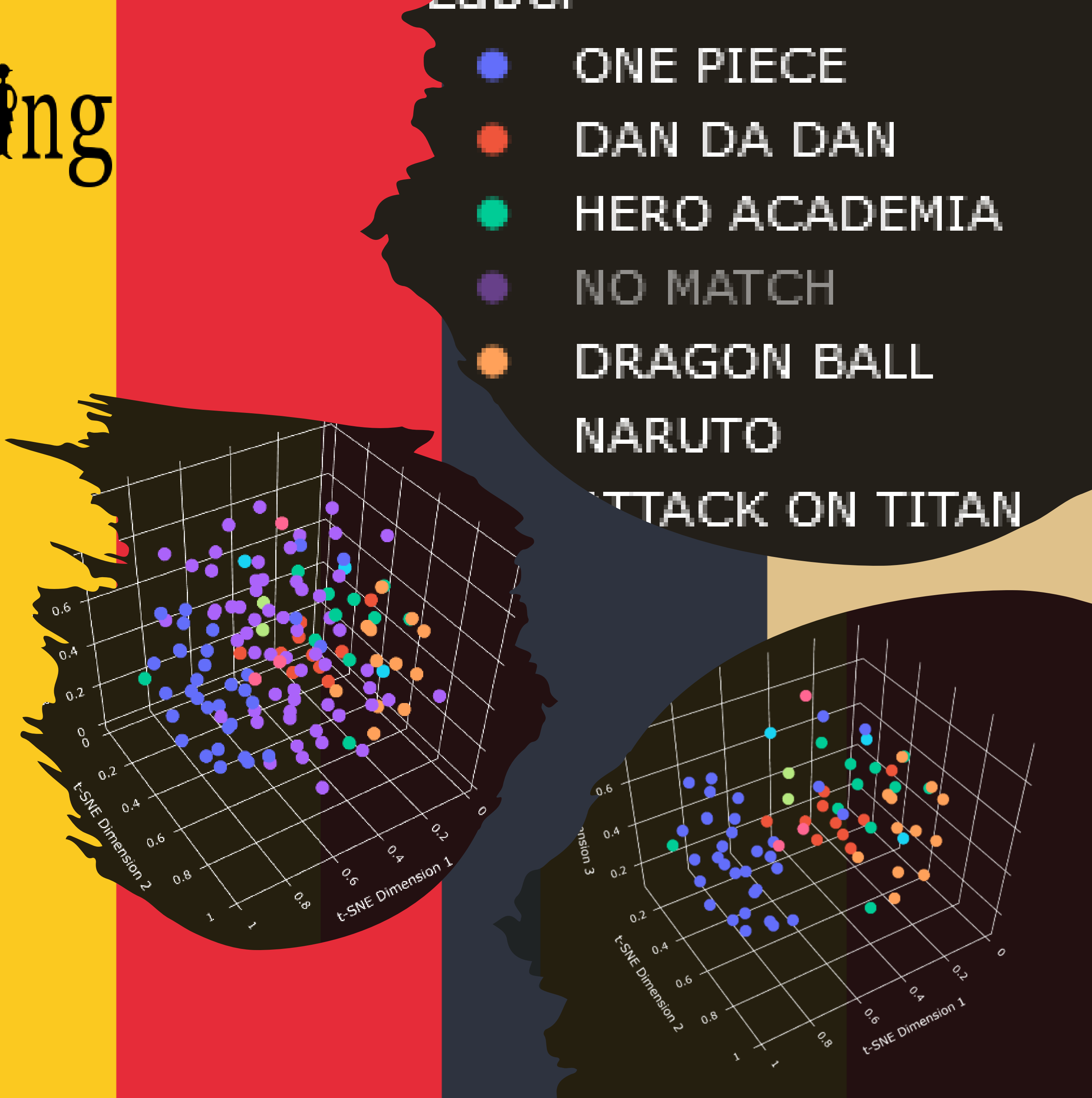


	one piece	label
title		
One Piece: Sun God Loki Vs Sun God Nika, Explained	0.979288	ONE PIECE
One Piece: Oda Reveals A Legendary God From Elbaf	0.978500	ONE PIECE
One Piece Anime To Go On Break Till 2025	0.978488	ONE PIECE
Pirates Who Are Animals In One Piece:	0.973659	ONE PIECE
One Piece: Why Luffy Will Destroy The World, Explained	0.973164	ONE PIECE
...	...	...
Dan Da Dan: Okarun’s Turbo Granny Powers Explained	0.011448	DAN DA DAN
Akira Toriyama's Swan Song Dragon Ball Daima Has Arrived	0.010300	DRAGON BALL
Dan Da Dan Officially Confirms It's Jujutsu Kaisen's True Successor By Revisiting The Hit Shonen's Big Theme	0.009993	DAN DA DAN
Attack on Titan: Gabi’s Journey from Hated to Understood	0.009886	ATTACK ON TITAN
My Hero Academia: A Smile Worth Saving	0.009726	HERO ACADEMIA

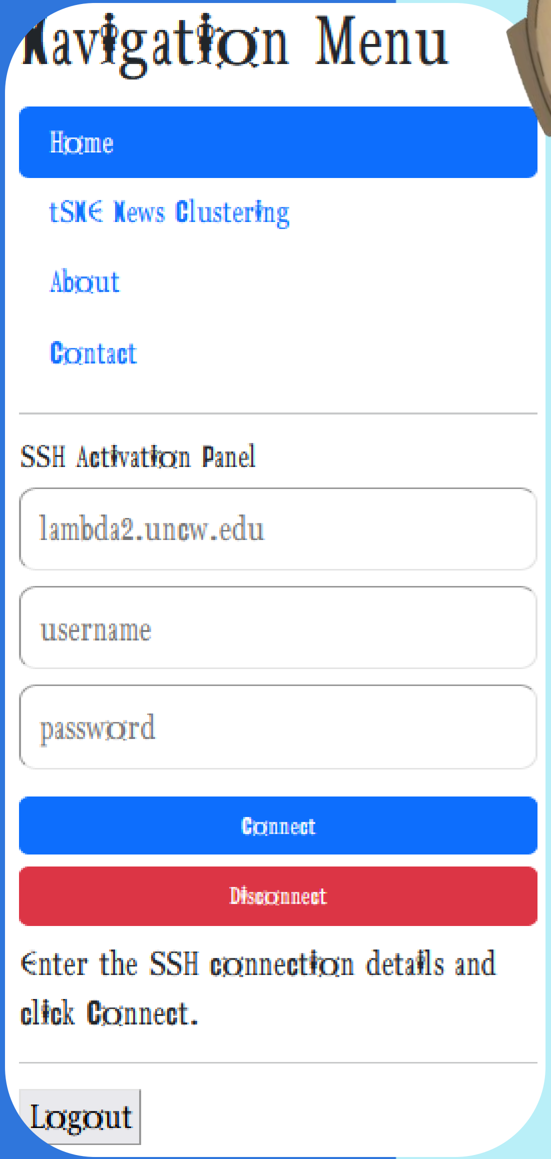
8 rows × 2 columns

t-SNE News Clustering

- **Dimensionality Reduction:** t-SNE maps high-dimensional article title vectors to a 2D/3D space, preserving relative distances and clusters.
- **Visual Clusters:** Helps visualize similarities between titles, revealing patterns and groupings.
- **Interpretability:** Converts complex text data into an intuitive, easy-to-understand 3D scatter plot.
- The GKG Tools Vectorizer creates a vector basis of the tokenized GDELT Field to cluster articles in a higher dimensional vector space (ex. 'persons').



Accessing BART-CNN Model

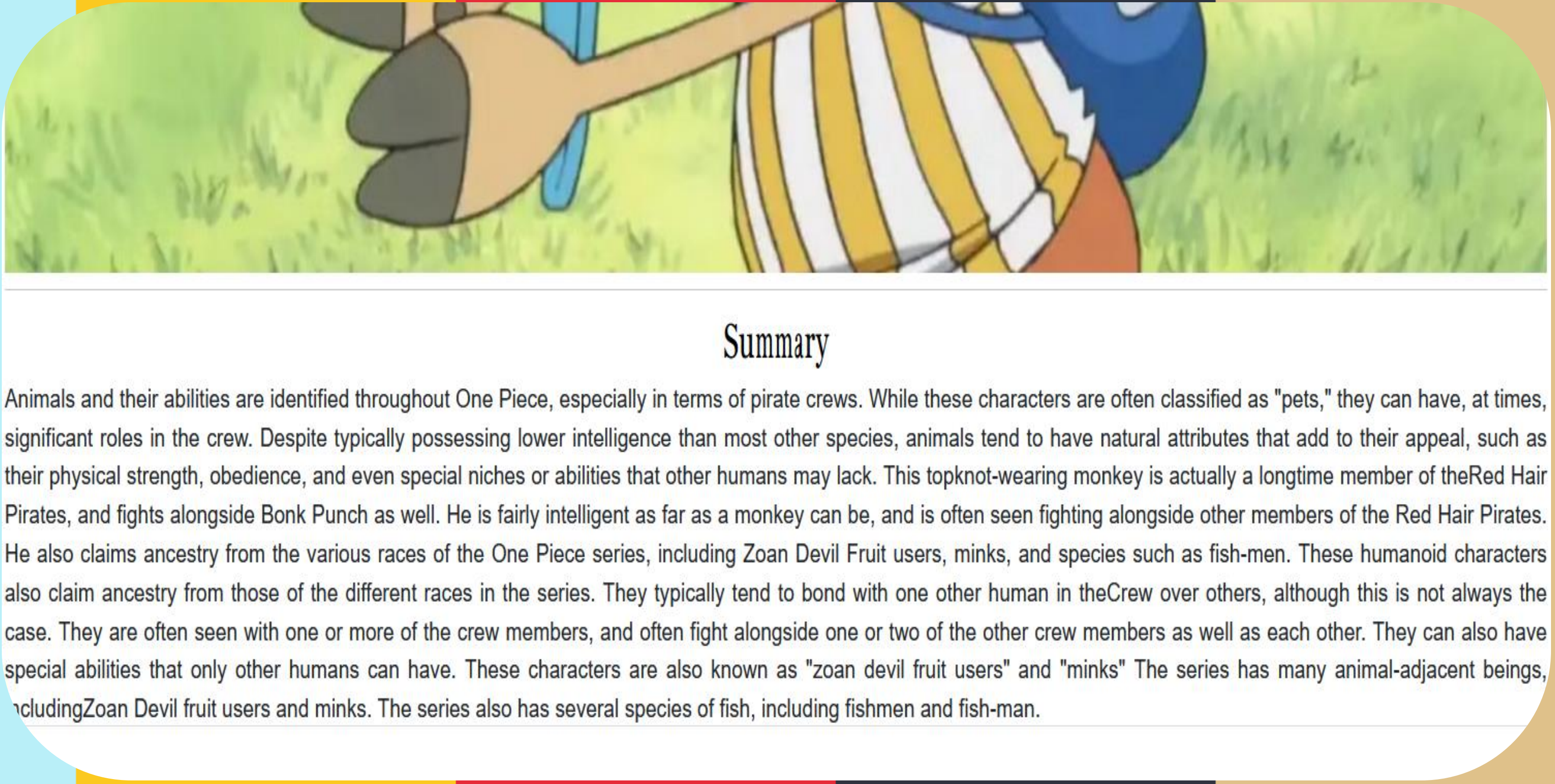


lambda2.uncw.edu

username

password

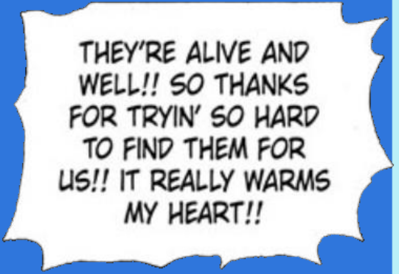
Enter the SSH connection details and click Connect.



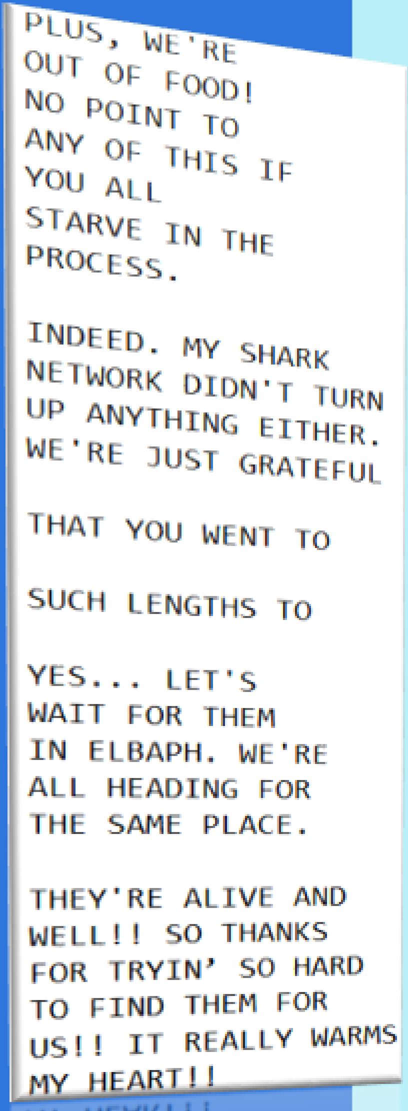
Summary

Animals and their abilities are identified throughout One Piece, especially in terms of pirate crews. While these characters are often classified as "pets," they can have, at times, significant roles in the crew. Despite typically possessing lower intelligence than most other species, animals tend to have natural attributes that add to their appeal, such as their physical strength, obedience, and even special niches or abilities that other humans may lack. This topknot-wearing monkey is actually a longtime member of theRed Hair Pirates, and fights alongside Bonk Punch as well. He is fairly intelligent as far as a monkey can be, and is often seen fighting alongside other members of the Red Hair Pirates. He also claims ancestry from the various races of the One Piece series, including Zoan Devil Fruit users, minks, and species such as fish-men. These humanoid characters also claim ancestry from those of the different races in the series. They typically tend to bond with one other human in theCrew over others, although this is not always the case. They are often seen with one or more of the crew members, and often fight alongside one or two of the other crew members as well as each other. They can also have special abilities that only other humans can have. These characters are also known as "zoan devil fruit users" and "minks" The series has many animal-adjacent beings, ncludingZoan Devil fruit users and minks. The series also has several species of fish, including fishmen and fish-man.

Image to Text Extraction — OCR



THEY'RE ALIVE AND
WELL!! SO THANKS
FOR TRYIN' SO HARD
TO FIND THEM FOR
US!! IT REALLY WARMS
MY HEART!!



PLUS, WE'RE
OUT OF FOOD!
NO POINT TO
ANY OF THIS IF
YOU ALL
STARVE IN THE
PROCESS.

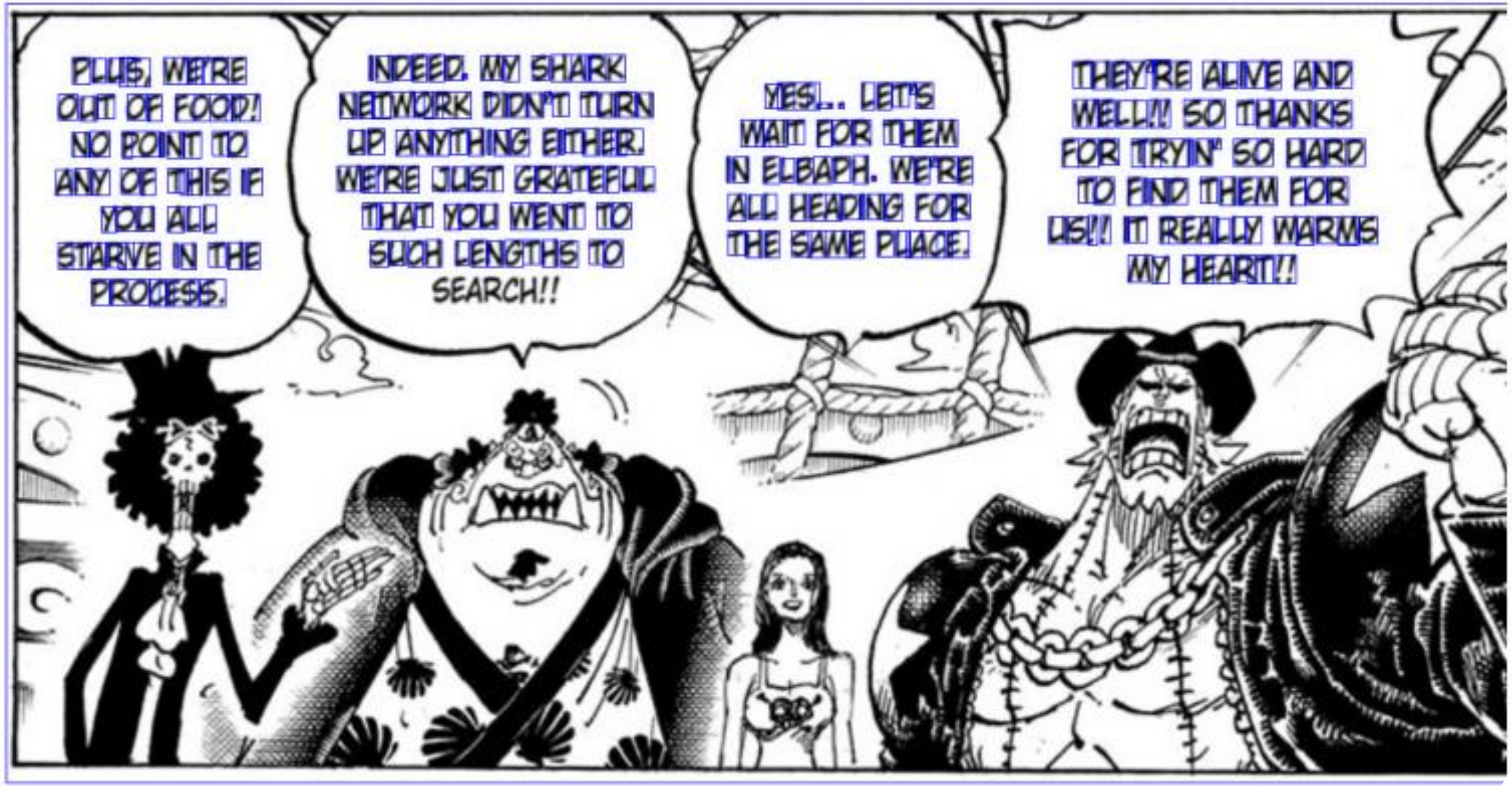
INDEED. MY SHARK
NETWORK DIDN'T TURN
UP ANYTHING EITHER.
WE'RE JUST GRATEFUL

THAT YOU WENT TO

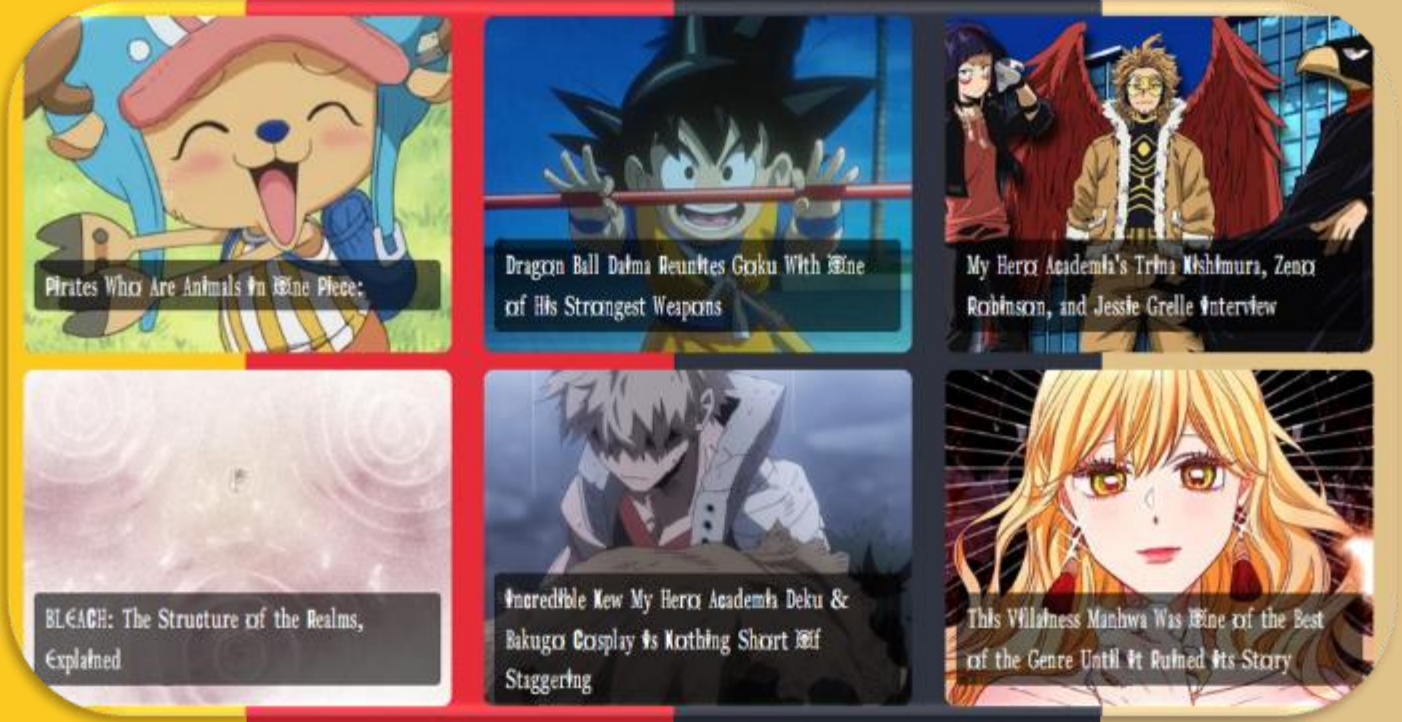
SUCH LENGTHS TO

YES... LET'S
WAIT FOR THEM
IN ELBAPH. WE'RE
ALL HEADING FOR
THE SAME PLACE.

THEY'RE ALIVE AND
WELL!! SO THANKS
FOR TRYIN' SO HARD
TO FIND THEM FOR
US!! IT REALLY WARMS
MY HEART!!


THEY'RE ALIVE AND WELL!! SO THANKS FOR TRYIN' SO HARD TO FIND THEM FOR US!! IT REALLY WARMS MY HEART!!
YES... LET'S WAIT FOR THEM IN ELBAPH. WE'RE ALL HEADING FOR THE SAME PLACE.
INDEED. MY SHARK NETWORK DIDN'T TURN UP ANYTHING EITHER. WE'RE JUST GRATEFUL THAT YOU WENT TO SUCH LENGTHS TO SEARCH!!
PLUS, WE'RE OUT OF FOOD! NO POINT TO ANY OF THIS IF YOU ALL STARVE IN THE PROCESS.

One Piece News Dashboard



Pirates Who Are Animals in One Piece:

Dragon Ball Daima Reunites Goku With One of His Strongest Weapons

My Hero Academia's Trina Nishimura, Zeno Robinson, and Jessie Grelle Interview

BLEACH: The Structure of the Realms, Explained

Incredible New My Hero Academia Deku & Bakugo Cosplay Is Nothing Short Of Staggering

This Villainess Manhwa Was One of the Best of the Genre Until It Ruined Its Story

Future Work

- Add GKG vectorized field token statistics for clusters and Articles to the article information modal.
- Add tSNE cluster panel tools:
  - Filter by multiple vector fields, locations, tags.
- Explore accuracy and performance measures between labeling and clustering methods.
  - GKG Vectorized Labels vs. BART MNLI Zero-shot
  - GKG Field Vectors vs. SKLearn TDIDF Title vectors.
- Automate GDELT GKG manga & anime query to check for new manga articles periodically.
- Apply classification and clustering models to incoming data.
- Dockerize app and deploy.

Thank You!