

```
from google.colab import drive
drive.mount('/content/drive')
```

Mounted at /content/drive

```
!pip install bertopic[all] -q
!pip install sentence-transformers -q
```

WARNING: bertopic 0.17.0 does not provide the extra 'all'

```
150.6/150.6 kB 6.3 MB/s eta 0:00:00
363.4/363.4 MB 2.4 MB/s eta 0:00:00
13.8/13.8 MB 94.6 MB/s eta 0:00:00
24.6/24.6 MB 74.2 MB/s eta 0:00:00
883.7/883.7 kB 42.4 MB/s eta 0:00:00
664.8/664.8 MB 3.0 MB/s eta 0:00:00
211.5/211.5 MB 5.9 MB/s eta 0:00:00
56.3/56.3 MB 12.8 MB/s eta 0:00:00
127.9/127.9 MB 7.3 MB/s eta 0:00:00
207.5/207.5 MB 6.8 MB/s eta 0:00:00
21.1/21.1 MB 35.3 MB/s eta 0:00:00
```

BERTopic input

```
!pip install bertopic
!pip install plotly
!pip install umap-learn
```

```
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.1.5->bertopic) (2025.1)
Requirement already satisfied: tenacity>=6.2.0 in /usr/local/lib/python3.11/dist-packages (from plotly>=4.7.0->bertopic) (9.1.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-packages (from plotly>=4.7.0->bertopic) (24.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn>=1.0->bertopic) (3.3.0)
Requirement already satisfied: transformers<5.0.0, >=4.41.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transformers>=0.4.1->bertopic) (4.41.0)
Requirement already satisfied: torch>=1.11.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transformers>=0.4.1->bertopic) (2.5.1)
Requirement already satisfied: huggingface-hub>=0.20.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transformers>=0.4.1->bertopic) (0.20.0)
Requirement already satisfied: Pillow in /usr/local/lib/python3.11/dist-packages (from sentence-transformers>=0.4.1->bertopic) (10.4.0)
Requirement already satisfied: numba>=0.51.2 in /usr/local/lib/python3.11/dist-packages (from umap-learn>=0.5.0->bertopic) (0.59.0)
Requirement already satisfied: pynndescent>=0.5 in /usr/local/lib/python3.11/dist-packages (from umap-learn>=0.5.0->bertopic) (0.5.13)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.20.0->sentence-transformers) (3.13.1)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.20.0->sentence-transformers) (2024.10.0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.20.0->sentence-transformers) (6.0.2)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.20.0->sentence-transformers) (2.32.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.20.0->sentence-transformers) (4.12.0)
Requirement already satisfied: llvmlite<0.44, >=0.43.0dev0 in /usr/local/lib/python3.11/dist-packages (from numba>=0.51.2->umap-learn) (0.43.0dev0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas>=1.1.5->bertopic) (1.16.0)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (3.3)
Requirement already satisfied: Jinja2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (3.1.4)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (12.4.127)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (12.4.127)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (12.4.127)
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (9.1.0.70)
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (12.4.5.8)
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (11.2.1.3)
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (10.3.5.147)
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (11.6.1.9)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (0.6.2)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (12.4.127)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (12.4.127)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (3.2.0)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (1.13.1)
Requirement already satisfied: mpmath<1.4, >=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch>=1.11.0->sentence-transformers) (1.3.0)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0, >=4.41.0->bertopic) (2024.10.1)
Requirement already satisfied: tokenizers<0.22, >=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0, >=4.41.0->bertopic) (0.21.0)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0, >=4.41.0->bertopic) (0.4.3)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from Jinja2->torch>=1.11.0->sentence-transformers) (2.1.5)
Requirement already satisfied: charset-normalizer<4, >=2 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface-hub>=0.20.0->bertopic) (3.3.0)
Requirement already satisfied: idna<4, >=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface-hub>=0.20.0->bertopic) (3.10)
Requirement already satisfied: urllib3<3, >=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface-hub>=0.20.0->bertopic) (2.2.3)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface-hub>=0.20.0->bertopic) (2025.1.1)
Requirement already satisfied: plotly in /usr/local/lib/python3.11/dist-packages (5.24.1)
Requirement already satisfied: tenacity>=6.2.0 in /usr/local/lib/python3.11/dist-packages (from plotly) (9.1.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-packages (from plotly) (24.2)
Requirement already satisfied: umap-learn in /usr/local/lib/python3.11/dist-packages (0.5.7)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from umap-learn) (2.0.2)
Requirement already satisfied: scipy>=1.3.1 in /usr/local/lib/python3.11/dist-packages (from umap-learn) (1.14.1)
Requirement already satisfied: scikit-learn>=0.22 in /usr/local/lib/python3.11/dist-packages (from umap-learn) (1.6.1)
Requirement already satisfied: numba>=0.51.2 in /usr/local/lib/python3.11/dist-packages (from umap-learn) (0.60.0)
Requirement already satisfied: pynndescent>=0.5 in /usr/local/lib/python3.11/dist-packages (from umap-learn) (0.5.13)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from umap-learn) (4.67.1)
Requirement already satisfied: llvmlite<0.44, >=0.43.0dev0 in /usr/local/lib/python3.11/dist-packages (from numba>=0.51.2->umap-learn) (0.43.0dev0)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.11/dist-packages (from pynndescent>=0.5->umap-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn>=0.22->umap-learn) (3.3.0)
```

✓ 2023 Topic modelling

```

!pip install bertopic umap-learn plotly nltk sentence-transformers -q

import re
import pandas as pd
import matplotlib.pyplot as plt
import nltk
from nltk.corpus import stopwords
from sentence_transformers import SentenceTransformer
from sklearn.feature_extraction.text import CountVectorizer
from bertopic import BERTopic
import plotly.io as pio

pio.renderers.default = "colab"

nltk.download("stopwords")
STOP_WORDS = stopwords.words("english")

###Load & preprocess data
DATA_PATH = "/content/drive/MyDrive/heat_posts2023.jsonl"
posts_df = pd.read_json(DATA_PATH, lines=True)

def clean_text(text: str) -> str:
    """Lowercase, remove URLs and non-alpha characters."""
    text = text.lower()
    text = re.sub(r"https?://\S+", "", text)
    text = re.sub(r"www\.\S+", "", text)
    text = re.sub(r"[^a-z\s]", "", text)
    return text

# Apply cleaning and parse dates
posts_df["clean_text"] = posts_df["text"].apply(clean_text)
posts_df["date"] = pd.to_datetime(posts_df["created_at"]).dt.date

texts = posts_df["clean_text"].tolist()
dates = posts_df["date"].tolist()

###Fit BERTopic model
# Using a smaller min_topic_size so even tiny clusters survive
topic_model = BERTopic(
    embedding_model = SentenceTransformer("all-MiniLM-L6-v2"),
    vectorizer_model = CountVectorizer(stop_words=STOP_WORDS, ngram_range=(1,2)),
    language = "english",
    calculate_probabilities = True,
    verbose = True,
    min_topic_size = 2,
)

topics, probs = topic_model.fit_transform(texts)
posts_df["topic"] = topics

###Auto-generate summaries for each topic
summary_map = {}
for t in topic_model.get_topic_info().Topic:
    if t == -1:
        continue
    top2 = [w for w,_ in topic_model.get_topic(t)[:2]]
    summary_map[t] = "_".join(top2)

posts_df["topic_summary"] = posts_df["topic"].map(lambda t: summary_map.get(t, "Other"))

print("Topic summaries:")
for t, label in summary_map.items():
    print(f" {t:2d} -> {label}")

###"Health / Energy / Policy / Climate" mapping
label_dict = {
    "Health": ["heatstroke", "hospital", "dehydration", "emergency", "risk", "dizzy", "overheat"],
    "Energy": ["power", "outage", "electricity", "grid", "air conditioning", "cooling", "blackout"],
    "Policy": ["government", "response", "support", "aid", "fund", "relief", "governor"],
    "Climate": ["climate", "crisis", "warming", "carbon", "emission", "anxiety"]
}

def assign_label(keywords, mapping):
    keys = set(w for w,_ in keywords)
    scores = {lbl: len(keys & set(words)) for lbl, words in mapping.items()}
    return max(scores, key=scores.get) if max(scores.values())>0 else "Other"

```

```
topic_label_map = {
    t: assign_label(topic_model.get_topic(t), label_dict)
    for t in summary_map.keys()
}
posts_df["semantic_label"] = posts_df["topic"].map(lambda t: topic_label_map.get(t, "Other"))

###Visualize topic-summary trends over time
time_df = posts_df.groupby(["date", "topic_summary"]).size().unstack(fill_value=0)
# pick top 10 summaries by overall counts
top10 = time_df.sum().sort_values(ascending=False).head(50).index

plt.figure(figsize=(14,6))
for lbl in top10:
    plt.plot(time_df.index, time_df[lbl], marker=".", label=lbl)

plt.title("Top-10 BERTopic Summaries Over Time")
plt.xlabel("Date")
plt.ylabel("Post Count")
plt.xticks(rotation=45)
plt.legend(title="Topic Summary", bbox_to_anchor=(1.02,1), loc="upper left")
plt.tight_layout()
plt.show()

###results
posts_df.to_csv("heatwave_posts_with_topics.csv", index=False)
pd.DataFrame(topic_model.topics_over_time(texts, dates)).to_csv("topics_over_time.csv", index=False)
```

```

[!nltk_data] Downloading package stopwords to /root/nltk_data...
[!nltk_data] Package stopwords is already up-to-date!
2025-04-26 16:13:19,432 - BERTopic - Embedding - Transforming documents to embeddings.
Batches: 100% 62/62 [00:36<00:00, 5.01it/s]

2025-04-26 16:13:55,877 - BERTopic - Embedding - Completed ✓
2025-04-26 16:13:55,879 - BERTopic - Dimensionality - Fitting the dimensionality reduction algorithm
2025-04-26 16:14:05,194 - BERTopic - Dimensionality - Completed ✓
2025-04-26 16:14:05,196 - BERTopic - Cluster - Start clustering the reduced embeddings
2025-04-26 16:14:11,837 - BERTopic - Cluster - Completed ✓
2025-04-26 16:14:11,846 - BERTopic - Representation - Fine-tuning topics using representation models.
2025-04-26 16:14:12,486 - BERTopic - Representation - Completed ✓

```

Topic summaries:

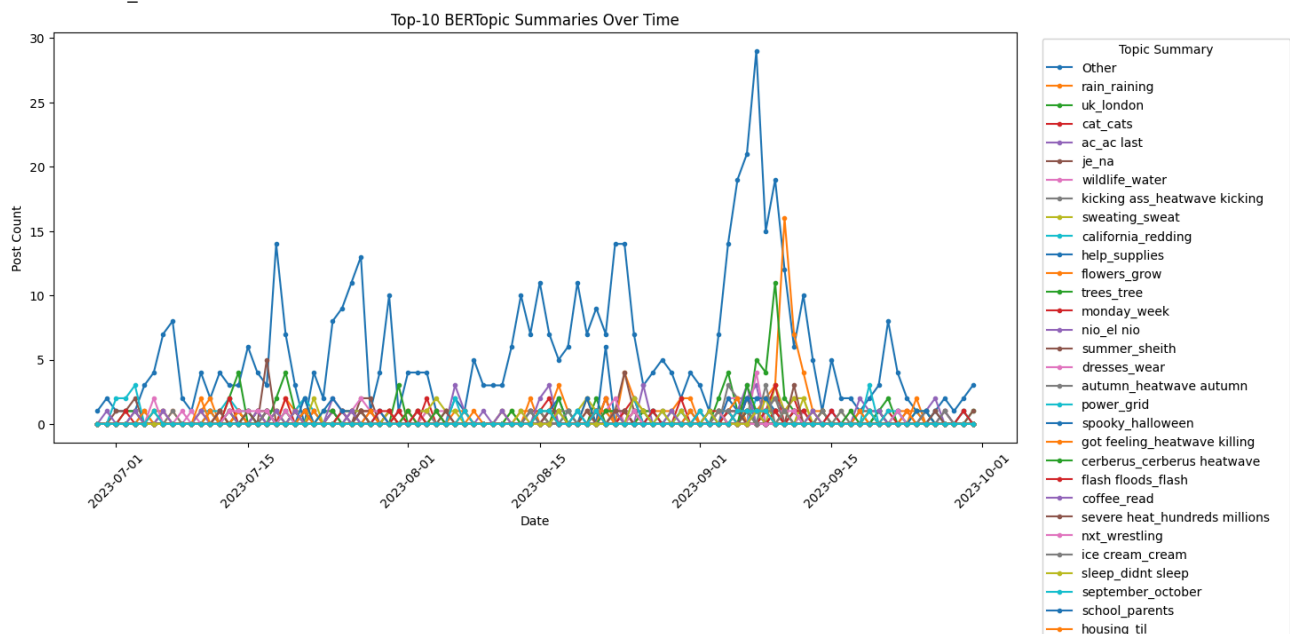
```

0 → rain_raining
1 → uk_london
2 → cat_cats
3 → ac_ac last
4 → je_na
5 → wildlife_water
6 → sweating_sweat
7 → kicking ass_heatwave kicking
8 → california_redding
9 → help_supplies
10 → monday_week
11 → trees_tree
12 → flowers_grow
13 → nio_el nio
14 → summer_sheith
15 → dresses_wear
16 → autumn_heatwave autumn
17 → power_grid
18 → got feeling_heatwave killing
19 → spooky_halloween
20 → cerberus_cerberus heatwave
21 → coffee_read
22 → flash_floods_flash
23 → nxt_wrestling
24 → severe heat_hundreds millions
25 → sleep_didnt sleep
26 → ice_cream_cream
27 → housing_til
28 → climate_change_climate
29 → art_ink_waiting
30 → overnight_killer
31 → school_parents
32 → september_october
33 → cases_conditioner
34 → chronic_chronic pain
35 → usually_check
36 → barbie_packing
37 → people_europe_killed people
38 → cosplay_graphic
39 → blob_ecosystems
40 → border_migrants
41 → italy_italy struggles
42 → paste_fried
43 → aloe_sunscreen
44 → candles_awww
45 → writing_actual
46 → melting_melting heatwave
47 → dinner_chicken
48 → corn_corn corn
49 → library_ac installed
50 → games_morning
51 → fur_applejack
52 → soul_music
53 → tomatoes_plants
54 → treat_allowed look
55 → add_lighting_totally get
56 → pool_enjoying
57 → current_uk_tabloid
58 → clothes_autumn
59 → _
60 → mail_email conference
61 → red_alerts_alerts
62 → water_remind
63 → pass_anticipating
64 → blinds_blackout
65 → portland_holiday
66 → acropolis_athens
67 → france_records
68 → gloom_backdrop thoughts
69 → bottom_row_rid
70 → phoenix_death_valley
71 → abbot_workers
72 → draw_ink
73 → looked_also involved
74 → vibe_telling
75 → heatwave_heatwave_chevronheatwave chevron
76 → anything_hot attend_mitchester

```

76 → anything not_green machine
 77 → cadaboo_billionaire volcano
 78 → spring heatwave_fires
 79 → arrived waiting_heatwave things
 80 → prompt_gingerrose
 81 → chilli_chilli heatwave
 82 → sf_times
 83 → people died_died
 84 → set_set record
 85 → netherlands_train
 86 → body_fun
 87 → smell_manure smell
 88 → underwear_adult stuff
 89 → cf_acclimated
 90 → fatsia_watering
 91 → novella_youll
 92 → abba_abba time
 93 → losing_losing dad
 94 → heatwave next_next week
 95 → ankle_coast
 96 → random september_hopefully ill
 97 → labor_day_labor
 98 → cookout_cheese
 99 → girlsweat_market
 100 → existential_dread
 101 → terrible_calmed
 102 → wake_sleep
 103 → hes_dog
 104 → therapy_ks
 105 → doesnt ac_air conditioning
 106 → living near_risk
 107 → ac say_bars
 108 → get back_back us
 109 → grain_china
 110 → jamboree_scout
 111 → comet_marine heatwave
 112 → city_ancestral city
 113 → contrasts granite_boulder light
 114 → universal_trees
 115 → loving_sun
 116 → burns_arizona
 117 → buy_avoiding strategy
 118 → september heatwave_september
 119 → celebrate happy_cooler temps
 120 → arizona_across rest
 121 → fan_eye
 122 → ahahaha_ahahaha always
 123 → september heatwave_september
 124 → ithot_jogging
 125 → ireland_heatwave ireland
 126 → anything impossible_dad hes
 127 → antibiotics_infection
 128 → banfossilfuels_banfossilfuels carbon
 129 → cum_abate
 130 → wisconsin_article
 131 → bike_musk
 132 → septembair_heatwave much
 133 → whether_absolutely wont
 134 → started_across south
 135 → latest_gothic
 136 → every_day_spent
 137 → wood_catching rays
 138 → brazil_rn xp
 139 → heatwave summer_assassin
 140 → media_continues
 141 → _
 142 → arriving_arriving long
 143 → deadly heatwave_spain surpass
 144 → learn live_failed
 145 → defence_blue blobs
 146 → sure run_fuck heatwave
 147 → ac unit_unit
 148 → breathe_back old
 149 → ecosystems_crisis
 150 → pillow_brandoninsertcreditcom
 151 → utilities_bad infrastructure
 152 → advice_actionable
 153 → allison_lives heart
 154 → christmas_chocolates
 155 → snow_advantage ticino
 156 → laundry_advantage laundry
 157 → spotify_anxiety resentment
 158 → sydney_average
 159 → good morning_cool newyork
 160 → fire_bom ok
 161 → coast_marshallblog_continuing heatwave
 162 → stream_alrighty
 163 → aced_aced room
 164 → added heatwave_bearable really
 165 → fires_intense
 166 → town_hall

166 → town_var
 167 → better start_comin right
 168 → across every_attention craves
 169 → messages_along major
 170 → hat_crocheting smol
 171 → bees_absolutely grinding
 172 → fanwood_woman
 173 → boxes_amount boxes
 174 → thank_heatwave thanks
 175 → sure though_seems much
 176 → greece_battles wildfires
 177 → away puddle_date heatwave
 178 → public_black neighborhood
 179 → albuquerque_heatwave update
 180 → _
 181 → hospital_jazz
 182 → firewood_sashi
 183 → soup_cross stitching
 184 → allergies_pics
 185 → goodbye_ever experienced
 186 → additional_max_better yesterday
 187 → marine_disease
 188 → antonio_antonio incredibly
 189 → bring_autumn sound
 190 → keyboard_migraine
 191 → print_attempt screening
 192 → first_pumpkin_pumpkin spice
 193 → absurd_read_answers reckon
 194 → coolness_coolness finds
 195 → horny_pass_pass ya
 196 → crap_mini_cta
 197 → soon_heatwave_pls soon
 198 → back_little_burny
 199 → seal_heatwave_seal
 200 → calippo_another calippo
 201 → also_feeling_broke halfway
 202 → enters_amediting
 203 → reassuring_thats reassuring
 204 → parkin_got_montreal friends
 205 → bath_frozen_af im
 206 → dubai_earlier degree
 207 → anyone_got_bladeless
 208 → away_swampland_butterflies
 209 → chapter_analysis
 210 → bad_shade_blanket call
 211 → aging_aging like
 212 → delightful_also pnw
 213 → heatwave_year_fuck gonna
 214 → play_computer case
 215 → fix_outages
 216 → bad_wasfelt_heatwave rah
 217 → wet_address
 218 → add_stomach_animals heatwave
 219 → bad_week_heatwave looks
 220 → back_home_home
 221 → affirmed_affirmed badass
 222 → also_say_ay
 223 → red_heatwave_code red
 224 → heatwave_july_archives things
 225 → amid_chaotic_chaotic
 226 → office_hours_pt
 227 → save_planet_save
 228 → cute_blame hate



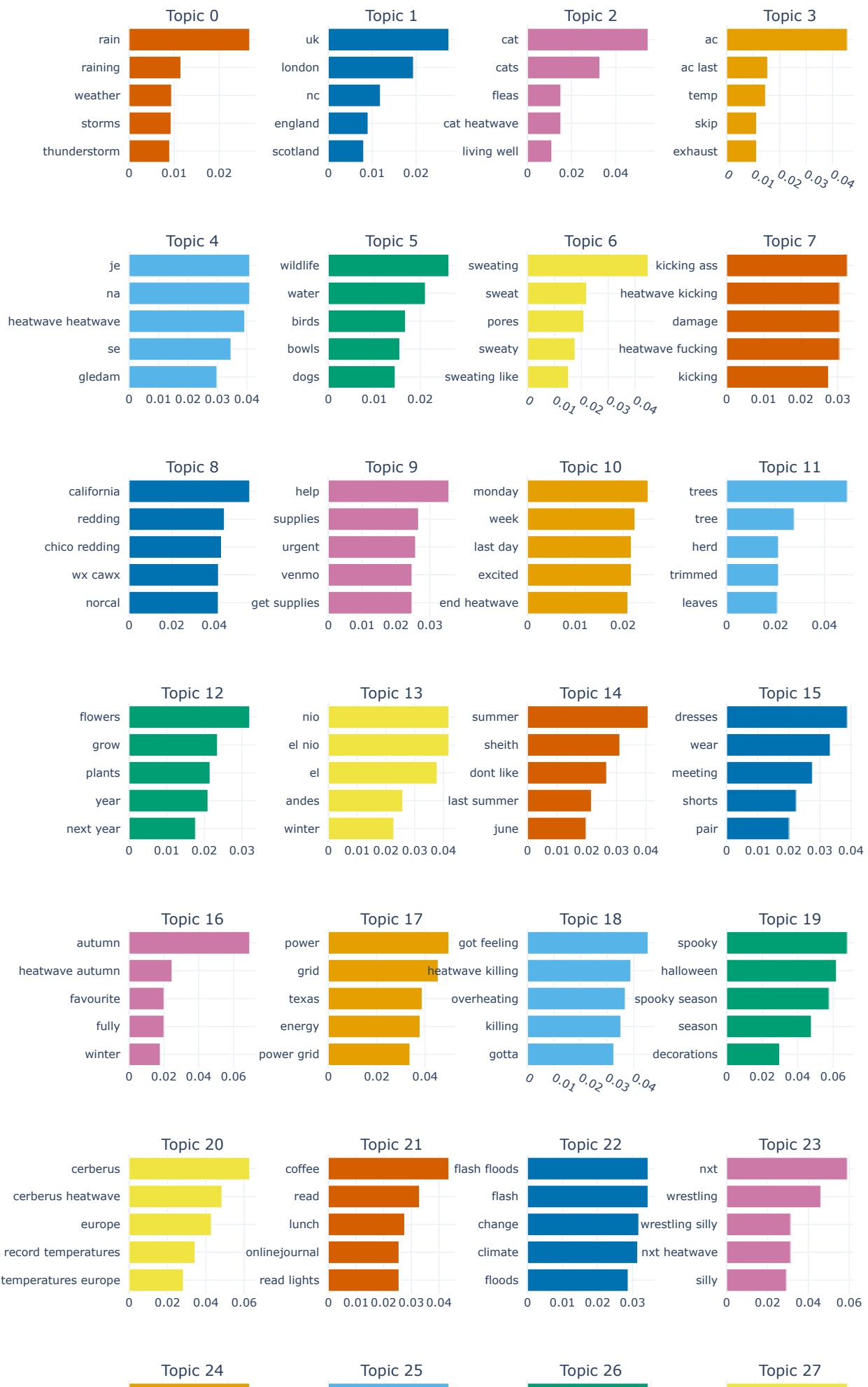
- overnight_killer
- art ink_waiting
- climate change_climate
- cosplay_graphic
- chronic_chronic pain
- cases_conditioner
- barbie_packing
- blob_ecosystems
- border_migrants
- people europe_killed people
- usually_check
- writing_actual
- library_ac installed
- melting_melting heatwave
- italy_italy struggles
- september heatwave_september
- paste_fried
- aloe_sunscreens

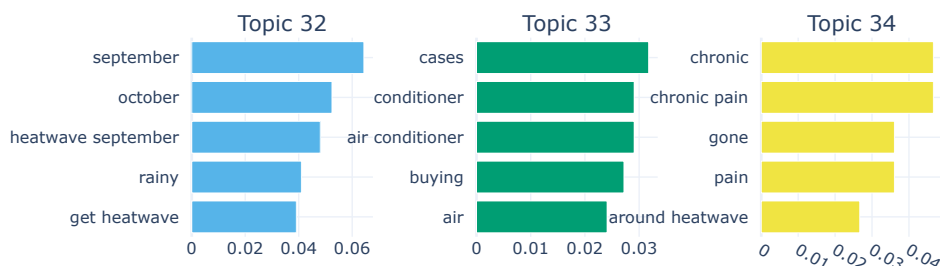
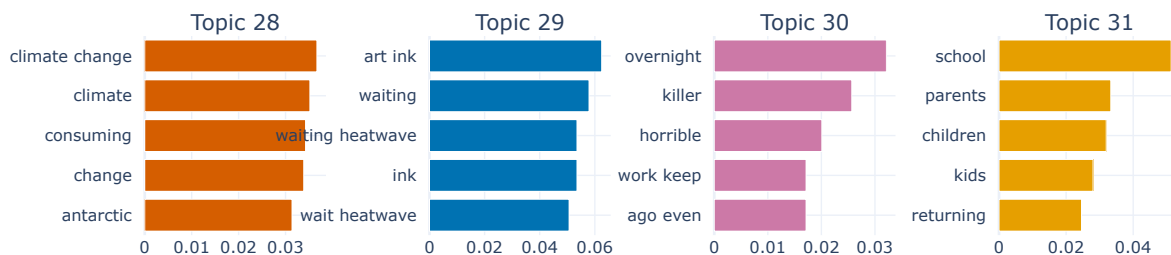
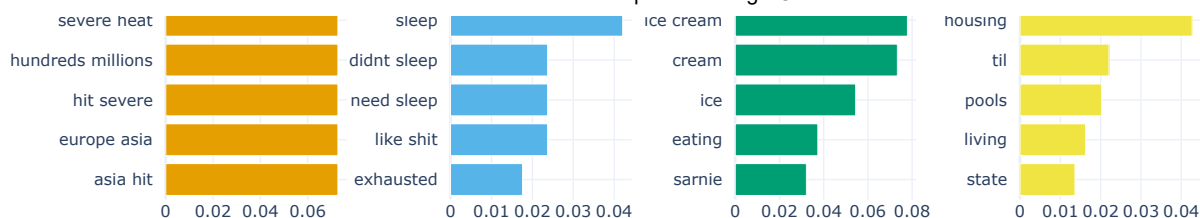
94it [00:05, 16.44it/s]

```
topic_model.visualize_barchart(top_n_topics=35).show()
```




Topic Word Scores





✓ 2024 Topic modelling

```
!pip install bertopic umap-learn plotly nltk sentence-transformers -q
```

```
import re
import pandas as pd
import matplotlib.pyplot as plt
import nltk
from nltk.corpus import stopwords
from sentence_transformers import SentenceTransformer
from sklearn.feature_extraction.text import CountVectorizer
from bertopic import BERTopic
import plotly.io as pio
```

```
pio.renderers.default = "colab"
```

```
nltk.download("stopwords")
STOP_WORDS = stopwords.words("english")
```

```
###Load data
DATA_PATH = "/content/drive/MyDrive/heat_posts2024_standardized.jsonl"
posts_df = pd.read_json(DATA_PATH, lines=True)
```

```
def clean_text(text: str) -> str:
    """Lowercase, remove URLs and non-alpha characters."""
    text = text.lower()
    text = re.sub(r"https?://\S+", "", text)
    text = re.sub(r"www\.\S+", "", text)
    text = re.sub(r"^[a-z\s]", "", text)
    return text
```

```
posts_df["clean_text"] = posts_df["text"].apply(clean_text)
posts_df["date"] = pd.to_datetime(posts_df["created_at"]).dt.date
```

```
texts = posts_df["clean_text"].tolist()
dates = posts_df["date"].tolist()
```

```
### Fit BERTopic model
# Using a smaller min_topic_size so even tiny clusters survive
```

```

topic_model = BERTopic(
    embedding_model = SentenceTransformer("all-MiniLM-L6-v2"),
    vectorizer_model = CountVectorizer(stop_words=STOP_WORDS, ngram_range=(1,2)),
    language = "english",
    calculate_probabilities = True,
    verbose = True,
    min_topic_size = 2,
)

topics, probs = topic_model.fit_transform(texts)
posts_df["topic"] = topics

# Auto-generate summaries for each topic
summary_map = {}
for t in topic_model.get_topic_info().Topic:
    if t == -1:
        continue
    top2 = [w for w,_ in topic_model.get_topic(t)[:2]]
    summary_map[t] = "_".join(top2)

posts_df["topic_summary"] = posts_df["topic"].map(lambda t: summary_map.get(t, "Other"))

print("Topic summaries:")
for t, label in summary_map.items():
    print(f" {t:2d} → {label}")

# "Health / Energy / Policy / Climate" mapping
label_dict = {
    "Health": ["heatstroke", "hospital", "dehydration", "emergency", "risk", "dizzy", "overheat"],
    "Energy": ["power", "outage", "electricity", "grid", "air conditioning", "cooling", "blackout"],
    "Policy": ["government", "response", "support", "aid", "fund", "relief", "governor"],
    "Climate": ["climate", "crisis", "warming", "carbon", "emission", "anxiety"]
}

def assign_label(keywords, mapping):
    keys = set(w for w,_ in keywords)
    scores = {lbl: len(keys & set(words)) for lbl, words in mapping.items()}
    return max(scores, key=scores.get) if max(scores.values())>0 else "Other"

topic_label_map = {
    t: assign_label(topic_model.get_topic(t), label_dict)
    for t in summary_map.keys()
}
posts_df["semantic_label"] = posts_df["topic"].map(lambda t: topic_label_map.get(t, "Other"))

### visualize topic-summary trends over time
# pivot
time_df = posts_df.groupby(["date", "topic_summary"]).size().unstack(fill_value=0)
# pick top 10 summaries by overall counts
top10 = time_df.sum().sort_values(ascending=False).head(50).index

plt.figure(figsize=(14,6))
for lbl in top10:
    plt.plot(time_df.index, time_df[lbl], marker=".", label=lbl)

plt.title("Top-10 BERTopic Summaries Over Time")
plt.xlabel("Date")
plt.ylabel("Post Count")
plt.xticks(rotation=45)
plt.legend(title="Topic Summary", bbox_to_anchor=(1.02,1), loc="upper left")
plt.tight_layout()
plt.show()

### results
posts_df.to_csv("heatwave_posts_with_topics.csv", index=False)
pd.DataFrame(topic_model.topics_over_time(texts, dates)).to_csv("topics_over_time.csv", index=False)

```

```

[!nltk_data] Downloading package stopwords to /root/nltk_data...
[!nltk_data] Package stopwords is already up-to-date!
2025-04-26 15:37:45,242 - BERTopic - Embedding - Transforming documents to embeddings.
Batches: 100% 127/127 [01:23<00:00, 4.49it/s]

2025-04-26 15:39:09,189 - BERTopic - Embedding - Completed ✓
2025-04-26 15:39:09,191 - BERTopic - Dimensionality - Fitting the dimensionality reduction algorithm
2025-04-26 15:39:37,989 - BERTopic - Dimensionality - Completed ✓
2025-04-26 15:39:37,992 - BERTopic - Cluster - Start clustering the reduced embeddings
2025-04-26 15:40:38,053 - BERTopic - Cluster - Completed ✓
2025-04-26 15:40:38,061 - BERTopic - Representation - Fine-tuning topics using representation models.
2025-04-26 15:40:39,044 - BERTopic - Representation - Completed ✓

Topic summaries:
0 → india_delhi
1 → music_boogie nights
2 → rain_thunderstorm
3 → september_october
4 → writing_story
5 → ac_ac broke
6 → antarctic_antarctica
7 → canada_environment
8 → sleep_bed
9 → week_next week
10 → sweat_sweating
11 → heatwave_ecosearch_ecosearch
12 → summer_finally_summer
13 → reaching_range
14 → covid_covid heatwave
15 → birds_bird
16 → plants_plant
17 → fuck_heatwave_fuck
18 → wrestling_nxt
19 → stream_streaming
20 → ssts_sat
21 → dog_dogs
22 → mexico_times likely
23 → really_like_hand
24 → heatwave_heatwave_aim think
25 → flowers_blooms
26 → power_generator
27 → pop_heatwave getting
28 → disabled_pls
29 → lucky_heatwave_cooler
30 → headache_terrible headache
31 → prison_womens
32 → county_wildfires
33 → hajj_hundreds
34 → third_supposed last
35 → marine_coral
36 → math_druid
37 → finland_norway
38 → tea_lemonade
39 → stay_hydrated_hydrated
40 → wear_dress
41 → climate_heatwave_ecosearch
42 → hot_outside_acerbicotterskysocial looking
43 → pacific_north pacific
44 → feel_youre sitting
45 → bike_ride
46 → genai_llms
47 → australia_breaking winter
48 → spent_magical
49 → noir_film
50 → recordbreaking_million
51 → heatwave_finally_broke
52 → angeles_times_angeles
53 → abc_world_david muir
54 → york_new york
55 → kids_children
56 → eastern_europe_mediterranean heatwave
57 → coffee_iced coffee
58 → portland_seattle
59 → heatwave_day_day heatwave
60 → japan_much japan
61 → ad_glassyeyed_allowed commute
62 → italy_italian
63 → saturday_august
64 → art_thermos
65 → fireworks_hot take
66 → engineers_blackouts
67 → perfect_weather_roll
68 → fur_cuddle
69 → boiler_radiator
70 → brain_castle
71 → continue_happen_happen get
72 → greece_bake seasons
73 → acropolis_hottest hours
74 → ahead_treat_aussie comedian
75 → furryart_furry
76 → scrolling_anymore ive

```

```

76 → searching_anymore ave
77 → ffxiv_blues
78 → plenty_avoid
79 → peoples_hostile
80 → mendocino county_mendocino
81 → curtains_blinds
82 → ktla_citys
83 → bearpigxxbskysocial_bought plugins
84 → goals speak_goals
85 → article run_conspiracy
86 → houses_etc
87 → nice weekend_stuck
88 → outages_reply
89 → superman_livewire
90 → ecosearch summer_renters
91 → fruit_pluots
92 → morocco_least people
93 → autumn_fall
94 → fish_eaten
95 → batteries_grid
96 → phone_alerts
97 → arizona_hoa
98 → lights flickering_flickering
99 → pakistan_heatwave climatecrisis
100 → sale_ad
101 → midwest northeast_midwest
102 → even stay_laptop
103 → ihr_good morning
104 → ireland_back sunshine
105 → vote_braved ny
106 → one yeah_behind pittsburgh
107 → daily digest_digest
108 → monkeys_wildlife
109 → climate breakdown_breakdown
110 → edmonton_inarow
111 → artfight_art fight
112 → texas_without power
113 → colors_added heatwave
114 → warmer night_degrees warmer
115 → warmer_yeah
116 → pasta_dinner
117 → shower_gel
118 → oven_bake
119 → bumblebees_bees
120 → haiku_haiku heatwave
121 → ai_grandma
122 → humidity_although theres
123 → california_next days
124 → deal heatwave_deal
125 → bail_debilitating
126 → mediterranean_marine heatwave
127 → schools_friday heatwave
128 → th th_th
129 → house_unit
130 → garden_garden update
131 → eternal_agapetheism
132 → great day_ur
133 → energy heatwave_amreading
134 → fans_airconned
135 → birthweight_study finds
136 → cant face_cleaned refilled
137 → actually killing_augh bad
138 → ac cant_bump
139 → coast_coast heat
140 → go away_heatwave go
141 → spiders_room
142 → really need_outside
143 → dewpoint_maryland heatwave
144 → midwest_heatwave ecosearch
145 → paris_paris olympics
146 → hydrate_hydrate hydrate
147 → crne_shipping uk
148 → forecast_map
149 → painting_dotartblogkarinwander
150 → genai_air get
151 → connotation_negative connotation
152 → sinuses_aggravates
153 → draw_sketch
154 → absolutely correct_correct degrees
155 → hot sticky_sticky
156 → apparently weve_back said
157 → china_northern china
158 → caleb_caleb graves
159 → moment heatwave_window open
160 → lager_autumnal light
161 → socks heatwave_socks
162 → christmas_song
163 → algiers_egypt_algiers
164 → europe_ukraine
165 → boring_britfunk
166 → ...

```

166 → may_know
167 → body_period
168 → awaiting heatwave_coming way
169 → walks_two
170 → back haha_back everyone
171 → air purifier_purifier
172 → anyone help_done today
173 → cosplay_drama
174 → christmas_garfield
175 → heatstroke alerts_japan heatwave
176 → terrible_absolutely shrinkage
177 → las_las vegas
178 → heatwave extremeheat_extremeheat
179 → heatwave across_records highest
180 → change heating_heating weather
181 → begins_tomorrow
182 → sdgs_humanrights
183 → balkan_croatia
184 → hair_braid
185 → art taintedstarbskysocial_da fridge
186 → active moment_anyway heres
187 → come handy_experience mine
188 → neighbor_chin
189 → beach_beach heatwave
190 → wave_vs_technical
191 → another month_didnt pack
192 → comics_comic
193 → acres_ktvq
194 → depressed_affairs
195 → fucking_always coincides
196 → oh boy_boy
197 → brother_heatwave weekend
198 → admin_admin done
199 → stew_pieces
200 → british heatwave_complaining
201 → happen_accidental
202 → fulfillment_dawntrail
203 → angeles_los angeles
204 → already better_break weekend
205 → arts_street
206 → dubai_dubai experiencing
207 → festival coming_coming soon
208 → france_theyve
209 → going office_office
210 → im available_folk im
211 → ontario_another freak
212 → bots_rescue bots
213 → iowa_flooding
214 → england_carnage scotland
215 → journalists_images
216 → ready_ready heatwave
217 → room_airconditioned house
218 → ice cream_cream
219 → also known_belgian
220 → dangerous_apnewscomarticleuse cautious
221 → anatomically_anatomically impossible
222 → grain_corn
223 → melon_actually quite
224 → la_also thats
225 → ago massive_always kept
226 → spain_barcelona zaragoza
227 → deadly heat_centerpoint texas
228 → lounging_lounging around
229 → naming_ago people
230 → marine_heat_marine
231 → alternatively snuggling_bamboo crisp
232 → awaiting end_avocado toast
233 → pass_pass pass
234 → waves likely_colorado
235 → back worked_ban day
236 → cake_alternating chocolate
237 → dahlias_flowers
238 → always makes_around like
239 → belly hold_bit worrying
240 → ch_siberian
241 → sick_chronological
242 → landscape_asteriskwordpresscom
243 → added background_break day
244 → activity insects_aiding
245 → spc severeweather_weatherawareday weather
246 → video_day video
247 → ac low_alright gotten
248 → high pressure_pressure
249 → indian_pacific
250 → utena_ancient strange
251 → tube_tube london
252 → kindness_cards
253 → channel_afraid little
254 → around neck_neck
255 → bike_emergency

256 → alaska_ak
 257 → tornado_art gaming
 258 → balkans hit_hit blackouts
 259 → la county_emergency
 260 → three hours_run
 261 → moss_air sweet
 262 → af particularly_brand protestant
 263 → biden right_bunch grumpy
 264 → brigade_fire brigade
 265 → doors_windows
 266 → european union_union
 267 → abb_abb always
 268 → cat_abby castor
 269 → window_flat
 270 → dnp_wide
 271 → shaved_selfie
 272 → merlin_accompanied withering
 273 → alcohol based_arent thing
 274 → brazil_awkward
 275 → climate news_texas
 276 → yall_bluesky minutes
 277 → bbc_across half
 278 → absence_absence due
 279 → olympics_western europe
 280 → wax_melting heatwave
 281 → gearsgnomes_heatwave gearsgnomes
 282 → captain_captain cold
 283 → ago must_bit cooling
 284 → big_oil_homicide
 285 → hrm heatwave_uh hrm
 286 → power_back_power
 287 → ordered_airco
 288 → brakes_brakes one
 289 → scorcher_going scorcher
 290 → protect_address climate
 291 → consumption maximum_heatwave pushes
 292 → black crumb_clips looks
 293 → green blue_red green
 294 → adapted_fine
 295 → accidentally timed_afternoon evening
 296 → need shower_shower
 297 → action day_proalertweather
 298 → despite roasty_gone nice
 299 → mint_cat
 300 → bloom cool_cottage garden
 301 → cold front_front
 302 → southern vt_vt made
 303 → ableist society_article heatwave
 304 → valley swim_swim club
 305 → bskyappprofileweat_advertising havent
 306 → small talk_ableist nonsense
 307 → nowhere_acclimitization
 308 → groyper_cookout groyper
 309 → ac fun_ac landlords
 310 → ontario_tornadoes
 311 → captaincold_chen
 312 → advantage great_almost bad
 313 → morning bluesky_bluesky
 314 → ago lil_also entirely
 315 → according know_breaking incredible
 316 → skittles_rainbow
 317 → appear ill_greece appear
 318 → amazon_amazon workers
 319 → could crest_crest
 320 → japan sees_st day
 321 → wave washington_washington post
 322 → fashion_slutty
 323 → circuits_cool updates
 324 → absolutely agree_americans ubiquitous
 325 → swing state_state nevada
 326 → also realises_appropriate uncomfortable
 327 → fur_confirmed wearing
 328 → already predicting_befor year
 329 → actually know_anyway considered
 330 → ac building_access week
 331 → anything moved_attacked
 332 → blame_heatwave blame
 333 → care plants_hitting tomorrow
 334 → back settled_get could
 335 → dont ac_chicago
 336 → cruel_completely
 337 → broken could_could less
 338 → angelamariemortonblogspotcomseriesnippetshtmlm_angelamariemortonblogspotcomseriesnippetshtmlm currently
 339 → cat_already gets
 340 → ass boobs_beach onlyfans
 341 → meteorologist_accuweather
 342 → adds oppression_advertising times
 343 → allowed idea_also snowstorm
 344 → app_phone
 345 → neighbors_country see

```

346 → global warming_warming
347 → balkans first_countries sizzle
348 → permission_landlord
349 → brother_also local
350 → cold middle_bad everything
351 → expensive_aircon rocker
352 → dana_da maioralheatwave
353 → kings_talking heads
354 → south korea_korea
355 → heatwave duration_duration
356 → actually looking_adhesive
357 → amid wildfires_climate grief
358 → brunch_brunch social
359 → meds_interfere
360 → family_big brother
361 → digital_ac online
362 → earliest heatwave_earliest
363 → looking forward_backlash happens
364 → heatwave edition_tokyo people
365 → bavarian_bavarian chancellor
366 → flat_velux
367 → impacts heat_impacts
368 → air conditioning_conditioning
369 → forest_methane
370 → myadventureswithsuperman_livewire
371 → tan_ago glorious
372 → bergen_always seem
373 → awaits_us_awaits
374 → village_road
375 → netherlands_amsterdam caught
376 → busy_af stuff
377 → aggravated_aggravated mandatory
378 → laundry_drying
379 → deepers_closin
380 → cant last_already outside
381 → bricedailyphoto_heatwave bricedailyphoto
382 → accent_vocal
383 → stories_americans offended
384 → da officially_heat middle
385 → ethan_page_ethan
386 → also hide_bask
387 → iowans_could extremely
388 → debating_fan
389 → badly slept_crashing im
390 → alex apparently_alex
391 → gumbo_jambalaya
392 → season im_youre welcome
393 → go outside_area weekend
394 → also gotta_brain fog
395 → salty lemonade_pound
396 → oz_dehydration
397 → erdtree want_heatwave shall
398 → heatwave proofofcat_every wave
399 → tokyo_checklist
400 → buys_boycotting heatwave
401 → cornwall live_cornwall
402 → days feeling_envy still
403 → apartment facade_decompressing
404 → alaska athens_alpine heatwave
405 → cdc_change destruction
406 → psa uk_heatwave ugh
407 → goodbye heatwave_already like
408 → budsorgukheatwavetip thank_budsorgukheatwavetip
409 → consistently bad_deal much
410 → found online_day deconstructingkremlinpropagandasubstackcompmemesmond
411 → coming london_hows heatwave
412 → headcanon_related
413 → season_companion found
414 → beallyoucanbophellthreadcc_beallyoucanbophellthreadcc thewhothismonthbskysocial
415 → average temperature_actonclimate
416 → yn_accomplish
417 → teeth_brush teeth
418 → bikini top_corner elf
419 → bones proving_challenge half
420 → gorgeous heatwave_heatwave gorgeous
421 → field_around rain

```