```
!pip install requests pandas matplotlib seaborn google-api-python-
client youtube-transcript-api sentence-transformers
```

```
Requirement already satisfied: requests in
/usr/local/lib/python3.12/dist-packages (2.32.4)
Requirement already satisfied: pandas in
/usr/local/lib/python3.12/dist-packages (2.2.2)
Requirement already satisfied: matplotlib in
/usr/local/lib/python3.12/dist-packages (3.10.0)
Requirement already satisfied: seaborn in
/usr/local/lib/python3.12/dist-packages (0.13.2)
Requirement already satisfied: google-api-python-client in
/usr/local/lib/python3.12/dist-packages (2.181.0)
Collecting youtube-transcript-api
  Downloading youtube_transcript_api-1.2.2-py3-none-any.whl.metadata
(24 kB)
Requirement already satisfied: sentence-transformers in
/usr/local/lib/python3.12/dist-packages (5.1.0)
Requirement already satisfied: charset_normalizer<4,>=2 in
/usr/local/lib/python3.12/dist-packages (from requests) (3.4.3)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.12/dist-packages (from requests) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.12/dist-packages (from requests) (2.5.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.12/dist-packages (from requests) (2025.8.3)
Requirement already satisfied: numpy>=1.26.0 in
/usr/local/lib/python3.12/dist-packages (from pandas) (2.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.12/dist-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in
/usr/local/lib/python3.12/dist-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in
/usr/local/lib/python3.12/dist-packages (from pandas) (2025.2)
Requirement already satisfied: contourpy>=1.0.1 in
/usr/local/lib/python3.12/dist-packages (from matplotlib) (1.3.3)
Requirement already satisfied: cycler>=0.10 in
/usr/local/lib/python3.12/dist-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
/usr/local/lib/python3.12/dist-packages (from matplotlib) (4.59.2)
Requirement already satisfied: kiwisolver>=1.3.1 in
/usr/local/lib/python3.12/dist-packages (from matplotlib) (1.4.9)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.12/dist-packages (from matplotlib) (25.0)
Requirement already satisfied: pillow>=8 in
/usr/local/lib/python3.12/dist-packages (from matplotlib) (11.3.0)
Requirement already satisfied: pyparsing>=2.3.1 in
/usr/local/lib/python3.12/dist-packages (from matplotlib) (3.2.3)
Requirement already satisfied: httplib2<1.0.0,>=0.19.0 in
/usr/local/lib/python3.12/dist-packages (from google-api-python-
client) (0.30.0)
Requirement already satisfied: google-auth!=2.24.0,!
=2.25.0,<3.0.0,>=1.32.0 in /usr/local/lib/python3.12/dist-packages
(from google-api-python-client) (2.38.0)
Requirement already satisfied: google-auth-httplib2<1.0.0,>=0.2.0
in /usr/local/lib/python3.12/dist-packages (from google-api-python-
client) (0.2.0)
Requirement already satisfied: google-api-core!=2.0.*,!=2.1.*,!
=2.2.*,!=2.3.0,<3.0.0,>=1.31.5 in /usr/local/lib/python3.12/dist-
packages (from google-api-python-client) (2.25.1)
Requirement already satisfied: uritemplate<5,>=3.0.1 in
/usr/local/lib/python3.12/dist-packages (from google-api-python-
client) (4.2.0)
Requirement already satisfied: defusedxml<0.8.0,>=0.7.1 in
/usr/local/lib/python3.12/dist-packages (from youtube-transcript-api)
(0.7.1)
Requirement already satisfied: transformers<5.0.0,>=4.41.0 in
/usr/local/lib/python3.12/dist-packages (from sentence-transformers)
(4.56.1)
Requirement already satisfied: tqdm in /usr/local/lib/python3.12/dist-
packages (from sentence-transformers) (4.67.1)
Requirement already satisfied: torch>=1.11.0 in
/usr/local/lib/python3.12/dist-packages (from sentence-transformers)
(2.8.0+cu126)
Requirement already satisfied: scikit-learn in
/usr/local/lib/python3.12/dist-packages (from sentence-transformers)
(1.6.1)
Requirement already satisfied: scipy in
/usr/local/lib/python3.12/dist-packages (from sentence-transformers)
(1.16.1)
Requirement already satisfied: huggingface-hub>=0.20.0 in
/usr/local/lib/python3.12/dist-packages (from sentence-transformers)
(0.34.4)
Requirement already satisfied: typing_extensions>=4.5.0 in
/usr/local/lib/python3.12/dist-packages (from sentence-transformers)
(4.15.0)
Requirement already satisfied: googleapis-common-protos<2.0.0,>=1.56.2
in /usr/local/lib/python3.12/dist-packages (from google-api-core!
=2.0.*,!=2.1.*,!=2.2.*,!=2.3.0,<3.0.0,>=1.31.5->google-api-python-
client) (1.70.0)
Requirement already satisfied: protobuf!=3.20.0,!=3.20.1,!=4.21.0,!
=4.21.1,!=4.21.2,!=4.21.3,!=4.21.4,!=4.21.5,<7.0.0,>=3.19.5 in
/usr/local/lib/python3.12/dist-packages (from google-api-core!=2.0.*,!
=2.1.*,!=2.2.*,!=2.3.0,<3.0.0,>=1.31.5->google-api-python-client)
(5.29.5)
Requirement already satisfied: proto-plus<2.0.0,>=1.22.3 in
/usr/local/lib/python3.12/dist-packages (from google-api-core!=2.0.*,!
=2.1.*,!=2.2.*,!=2.3.0,<3.0.0,>=1.31.5->google-api-python-client)
(1.26.1)
```

```
Requirement already satisfied: cachetools<6.0,>=2.0.0 in
/usr/local/lib/python3.12/dist-packages (from google-auth!=2.24.0,!
=2.25.0,<3.0.0,>=1.32.0->google-api-python-client) (5.5.2)
Requirement already satisfied: pyasn1-modules>=0.2.1 in
/usr/local/lib/python3.12/dist-packages (from google-auth!=2.24.0,!
=2.25.0,<3.0.0,>=1.32.0->google-api-python-client) (0.4.2)
Requirement already satisfied: rsa<5,>=3.1.4 in
/usr/local/lib/python3.12/dist-packages (from google-auth!=2.24.0,!
=2.25.0,<3.0.0,>=1.32.0->google-api-python-client) (4.9.1)
Requirement already satisfied: filelock in
/usr/local/lib/python3.12/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (3.19.1)
Requirement already satisfied: fsspec>=2023.5.0 in
/usr/local/lib/python3.12/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (2025.3.0)
Requirement already satisfied: pyyaml>=5.1 in
/usr/local/lib/python3.12/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (6.0.2)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in
/usr/local/lib/python3.12/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (1.1.9)
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.12/dist-packages (from python-dateutil>=2.8.2-
>pandas) (1.17.0)
Requirement already satisfied: setuptools in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (75.2.0)
Requirement already satisfied: sympy>=1.13.3 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (1.13.3)
Requirement already satisfied: networkx in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (3.5)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (3.1.6)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.6.77 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (12.6.77)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.6.77 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (12.6.77)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.6.80 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (12.6.80)
Requirement already satisfied: nvidia-cudnn-cu12==9.10.2.21 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (9.10.2.21)
Requirement already satisfied: nvidia-cublas-cu12==12.6.4.1 in
```

```
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (12.6.4.1)
Requirement already satisfied: nvidia-cufft-cu12==11.3.0.4 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (11.3.0.4)
Requirement already satisfied: nvidia-curand-cu12==10.3.7.77 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (10.3.7.77)
Requirement already satisfied: nvidia-cusolver-cu12==11.7.1.2 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (11.7.1.2)
Requirement already satisfied: nvidia-cusparse-cu12==12.5.4.2 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (12.5.4.2)
Requirement already satisfied: nvidia-cusparselt-cu12==0.7.1 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (0.7.1)
Requirement already satisfied: nvidia-nccl-cu12==2.27.3 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (2.27.3)
Requirement already satisfied: nvidia-nvtx-cu12==12.6.77 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (12.6.77)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.6.85 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (12.6.85)
Requirement already satisfied: nvidia-cufile-cu12==1.11.1.6 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (1.11.1.6)
Requirement already satisfied: triton==3.4.0 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (3.4.0)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.12/dist-packages (from
transformers<5.0.0,>=4.41.0->sentence-transformers) (2024.11.6)
Requirement already satisfied: tokenizers<=0.23.0,>=0.22.0 in
/usr/local/lib/python3.12/dist-packages (from
transformers<5.0.0,>=4.41.0->sentence-transformers) (0.22.0)
Requirement already satisfied: safetensors>=0.4.3 in
/usr/local/lib/python3.12/dist-packages (from
transformers<5.0.0,>=4.41.0->sentence-transformers) (0.6.2)
Requirement already satisfied: joblib>=1.2.0 in
/usr/local/lib/python3.12/dist-packages (from scikit-learn->sentence-
transformers) (1.5.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in
/usr/local/lib/python3.12/dist-packages (from scikit-learn->sentence-
transformers) (3.6.0)
Requirement already satisfied: pyasn1<0.7.0,>=0.6.1 in
/usr/local/lib/python3.12/dist-packages (from pyasn1-modules>=0.2.1-
>google-auth!=2.24.0,!=2.25.0,<3.0.0,>=1.32.0->google-api-python-client) (0.6.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.12/dist-packages (from sympy>=1.13.3->torch>=1.11.0->sentence-transformers) (1.3.0)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.12/dist-packages (from jinja2->torch>=1.11.0->sentence-transformers) (3.0.2)
Downloading youtube_transcript_api-1.2.2-py3-none-any.whl (485 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 485.0/485.0 kB 12.1 MB/s eta 0:00:00
```

```python
#Module 1 & 2
import requests
import pandas as pd
import re
from googleapiclient.discovery import build

# --- API Key ---
API_KEY = "AIzaSyCh0Up2u318a7bVrIB2x-GbibLuZ4Ee58I"

# ========================================
# □ Step 1: Get Channel ID from Username/Handle
# ========================================
def get_channel_id(username_or_handle, api_key=API_KEY):
    url = f"https://www.googleapis.com/youtube/v3/channels?part=id&forUsername={username_or_handle}&key={api_key}"
    response = requests.get(url).json()

    if "items" not in response or len(response["items"]) == 0:
        search_url = f"https://www.googleapis.com/youtube/v3/search?part=snippet&type=channel&q={username_or_handle}&key={api_key}"
        response = requests.get(search_url).json()
        if "items" in response and len(response["items"]) > 0:
            return response["items"][0]["snippet"]["channelId"]
        else:
            return None
    else:
        return response["items"][0]["id"]

# ========================================
# □ Step 2: Fetch 50 Videos from Channel
# ========================================
def fetch_50_videos(channel_id, api_key=API_KEY):
    url = "https://www.googleapis.com/youtube/v3/search"
    params = {
        "part": "snippet",
        "channelId": channel_id,
        "maxResults": 50,
        "order": "date",
        "type": "video",
        "key": api_key
    }
    response = requests.get(url, params=params).json()
    videos = [
        {
            "videoId": item["id"]["videoId"],
            "title": item["snippet"]["title"],
            "publishedAt": item["snippet"]["publishedAt"]
        }
        for item in response.get("items", [])
    ]
    return pd.DataFrame(videos)

# ========================================
# □ Step 3: Fetch Multiple Channels
# ========================================
channels = ["MrBeast", "indiatoday", "lofi2307"]  # □ add more here
all_data = []

for ch in channels:
    ch_id = get_channel_id(ch)
    if ch_id:
        df = fetch_50_videos(ch_id)
        df["channel"] = ch
        all_data.append(df)
        print(f"□ {len(df)} videos fetched from {ch}")
    else:
        print(f"□ Channel not found: {ch}")

# Combine into one DataFrame
final_df = pd.concat(all_data, ignore_index=True)

# ========================================
# □ Step 4: Save to CSV
# ========================================
youtube = build("youtube", "v3", developerKey=API_KEY)
safe_name = "_".join([re.sub(r'[^A-Za-z0-9]+', '_', c) for c in
channels])
output_file = f"{safe_name}_50videos.csv"
final_df.to_csv(output_file, index=False, encoding="utf-8-sig")

print(f"\n□ Final DataFrame shape: {final_df.shape}")
print(f"□ Data saved as {output_file}")
final_df.head()
```

```
□ 3 videos fetched from MrBeast
□ 26 videos fetched from indiatoday
□ 50 videos fetched from lofi2307
□ Final DataFrame shape: (79, 4)
□ Data saved as MrBeast_indiatoday_lofi2307_50videos.csv
```

```
{"summary":"{\n  \"name\": \"final_df\",\n  \"rows\": 79,\n
\"fields\": [\n    {\n      \"column\": \"videoId\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 79,\n        \"samples\": [\n
\"xAAfs5D8I2o\",\n          \"Chfvwn2W6z4\",\n
\"WgAqA7VLeNc\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"title\",\n      \"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 77,\n        \"samples\": [\n          \"India
fashion week- Kavita Bhartia collection\",\n          \"NON STOP
HEART-BROKEN\\ud83d\\udc94 SAD MASHUP PART - 7 | Best Broken/Sad
Playlist By @lofi2307  | #instatrending\",\n          \"Milind Soman
Behind The Scene\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"publishedAt\",\n      \"properties\": {\n        \"dtype\":
\"object\",\n        \"num_unique_values\": 71,\n        \"samples\":
[\n          \"2025-09-03T16:04:28Z\",\n          \"2007-04-
09T18:32:22Z\",\n          \"2025-04-19T18:00:08Z\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"channel\",\n      \"properties\":
{\n        \"dtype\": \"category\",\n        \"num_unique_values\":
3,\n        \"samples\": [\n          \"MrBeast\",\n
\"indiatoday\",\n          \"lofi2307\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"final_df"}
```

```python
#Module 3
import pandas as pd
from youtube_transcript_api import YouTubeTranscriptApi
from youtube_transcript_api._errors import TranscriptsDisabled,
NoTranscriptFound

# === Step 1: Load video data from Module 2 ===
input_file = "MrBeast_indiatoday_lofi2307_50videos.csv"  # change to
your actual file
df = pd.read_csv(input_file)

# === Step 2: Function to fetch transcript ===
def get_transcript(video_id):
    try:
        transcript = YouTubeTranscriptApi().fetch(video_id,
languages=['en'])
        # join all snippet texts into one string
        text = " ".join([snip.text for snip in transcript])
        return text
    except (TranscriptsDisabled, NoTranscriptFound):
        return "Transcript not available"
    except Exception as e:
        return f"Error: {str(e)}"

# === Step 3: Extract transcripts for all videos ===
df["transcript"] = df["videoId"].apply(get_transcript)

# === Step 4: Save final CSV ===
output_file = input_file.replace(".csv", "_with_transcripts.csv")
df.to_csv(output_file, index=False, encoding="utf-8-sig")

print(f"□ Final CSV saved as {output_file}")
print(df.head(3))
```

```
□ Final CSV saved as
MrBeast_indiatoday_lofi2307_50videos_with_transcripts.csv
       videoId                  title           publishedAt  channel  \
0  Chfvwn2W6z4   S.T.A.L.K.E.R. Clip  2007-04-09T18:32:22Z  MrBeast
1  A1GTlvv1mcs  Taru Christmas dance  2006-12-23T23:52:19Z  MrBeast
2  rE3X0NlQnpE    Taru Mortal Combat  2006-08-24T18:14:58Z  MrBeast

                 transcript
0  Transcript not available
1  Transcript not available
2  Transcript not available
```

```
#trying
!pip install youtube-transcript-api --upgrade
!pip install requests pandas

import pandas as pd
from youtube_transcript_api import YouTubeTranscriptApi
from youtube_transcript_api._errors import TranscriptsDisabled,
NoTranscriptFound

# === Step 1: Load video data ===
input_file = "MrBeast_indiatoday_lofi2307_50videos.csv"  # replace
with your file
df = pd.read_csv(input_file)

# === Step 2: Function to fetch transcript ===
def fetch_full_transcript(video_id):
    try:
        transcript = YouTubeTranscriptApi().fetch(video_id,
languages=['en'])
        # Join full transcript text
        text = " ".join([snip.text for snip in transcript])
        return text if text.strip() else "Transcript empty"
    except (TranscriptsDisabled, NoTranscriptFound):
        return "Transcript not available"
    except Exception as e:
        return f"Error: {str(e)}"

# === Step 3: Apply to all videos ===
df["transcript"] = df["videoId"].apply(fetch_full_transcript)

# === Step 4: Save final output ===
output_file = input_file.replace(".csv", "_with_transcripts.csv")
df.to_csv(output_file, index=False, encoding="utf-8-sig")

print(f"□ Final transcripts saved to {output_file}")
print(df.head(3))
```

```
Requirement already satisfied: youtube-transcript-api in
/usr/local/lib/python3.12/dist-packages (1.2.2)
Requirement already satisfied: defusedxml<0.8.0,>=0.7.1 in
/usr/local/lib/python3.12/dist-packages (from youtube-transcript-api)
(0.7.1)
Requirement already satisfied: requests in
/usr/local/lib/python3.12/dist-packages (from youtube-transcript-api)
(2.32.4)
Requirement already satisfied: charset_normalizer<4,>=2 in
/usr/local/lib/python3.12/dist-packages (from requests->youtube-
transcript-api) (3.4.3)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.12/dist-packages (from requests->youtube-
transcript-api) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.12/dist-packages (from requests->youtube-
transcript-api) (2.5.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.12/dist-packages (from requests->youtube-
transcript-api) (2025.8.3)
Requirement already satisfied: requests in
/usr/local/lib/python3.12/dist-packages (2.32.4)
Requirement already satisfied: pandas in
/usr/local/lib/python3.12/dist-packages (2.2.2)
Requirement already satisfied: charset_normalizer<4,>=2 in
/usr/local/lib/python3.12/dist-packages (from requests) (3.4.3)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.12/dist-packages (from requests) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.12/dist-packages (from requests) (2.5.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.12/dist-packages (from requests) (2025.8.3)
Requirement already satisfied: numpy>=1.26.0 in
/usr/local/lib/python3.12/dist-packages (from pandas) (2.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.12/dist-packages (from pandas) (2.9.0.post0)
```

```
Requirement already satisfied: pytz>=2020.1 in
/usr/local/lib/python3.12/dist-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in
/usr/local/lib/python3.12/dist-packages (from pandas) (2025.2)
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.12/dist-packages (from python-dateutil>=2.8.2-
>pandas) (1.17.0)
□ Final transcripts saved to
MrBeast_indiatoday_lofi2307_50videos_with_transcripts.csv
       videoId                  title           publishedAt  channel  \
0  Chfvwn2W6z4   S.T.A.L.K.E.R. Clip  2007-04-09T18:32:22Z  MrBeast
1  A1GTlvv1mcs  Taru Christmas dance  2006-12-23T23:52:19Z  MrBeast
2  rE3X0NlQnpE    Taru Mortal Combat  2006-08-24T18:14:58Z  MrBeast

                 transcript
0  Transcript not available
1  Transcript not available
2  Transcript not available
```

```
#Module 4
# □ Module 4: Clean transcripts in combined CSV file

import re
import pandas as pd
import nltk
from nltk.corpus import stopwords

# Download stopwords (only first time)
nltk.download("stopwords")
stop_words = set(stopwords.words("english"))

# Input combined CSV file
input_file =
"MrBeast_indiatoday_lofi2307_50videos_with_transcripts.csv"
# Output cleaned file
output_file =
"MrBeast_indiatoday_lofi2307_50videos_with_clean_transcripts.csv"

def clean_text(text):
    if not isinstance(text, str):
        return ""
    text = text.lower()  # lowercase
    text = re.sub(r"[^a-z\s]", " ", text)  # remove punctuation &
numbers
    words = [w for w in text.split() if w not in stop_words]  # remove
stopwords
    return " ".join(words)

# Load dataset
df = pd.read_csv(input_file)
```

```
# Clean transcript column
df = df.rename(columns={"transcript": "clean_transcript"})
df["clean_transcript"] = df["clean_transcript"].apply(clean_text)

# Save cleaned dataset
df.to_csv(output_file, index=False, encoding="utf-8")
print(f"□ Cleaned file saved as {output_file}")
```

```
□ Cleaned file saved as
MrBeast_indiatoday_lofi2307_50videos_with_clean_transcripts.csv
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
#Module 5
# □ Module 5: Basic Text Analysis on Cleaned Transcripts

import pandas as pd
import matplotlib.pyplot as plt
from collections import Counter

# Load cleaned dataset
input_file =
"MrBeast_indiatoday_lofi2307_50videos_with_clean_transcripts.csv"
df = pd.read_csv(input_file)

# □ Step 1: Add transcript length stats
df["word_count"] = df["clean_transcript"].apply(lambda x:
len(str(x).split()))
df["char_count"] = df["clean_transcript"].apply(lambda x: len(str(x)))

# □ Step 2: Get most common words (per channel)
def get_top_words(texts, n=10):
    all_words = " ".join(texts).split()
    return Counter(all_words).most_common(n)

channels = df["channel"].unique()
for ch in channels:
    top_words = get_top_words(df[df["channel"] == ch]
["clean_transcript"], 10)
    print(f"\n□ Top words for {ch}:")
    for word, freq in top_words:
        print(f"{word}: {freq}")

# □ Step 3: Visualization → Distribution of transcript lengths
plt.figure(figsize=(10,6))
for ch in channels:
    df[df["channel"] == ch]["word_count"].hist(alpha=0.5, bins=20,
label=ch)
```

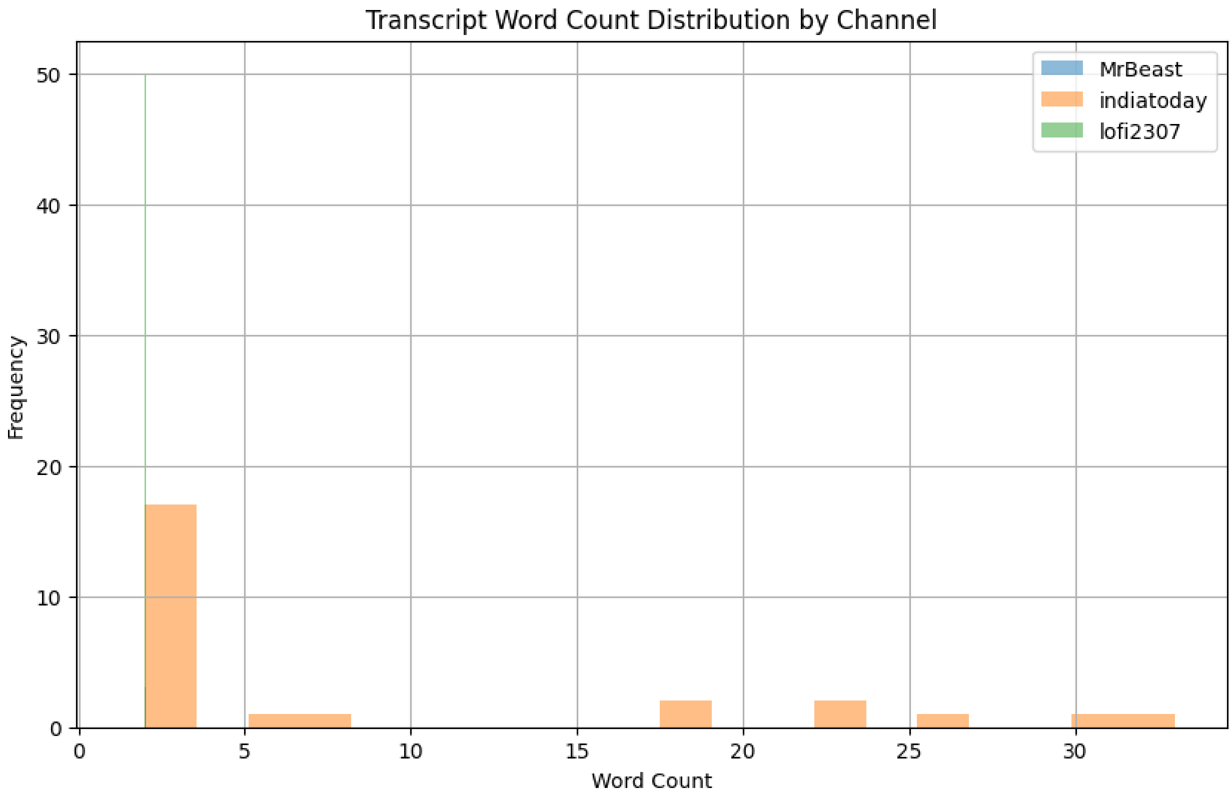
```
plt.title("Transcript Word Count Distribution by Channel")
plt.xlabel("Word Count")
plt.ylabel("Frequency")
plt.legend()
plt.show()

# Save updated file with stats
output_file =
"MrBeast_indiatoday_lofi2307_50videos_with_clean_transcripts_stats.csv
"
df.to_csv(output_file, index=False, encoding="utf-8")
print(f"\n□ Updated file saved with stats: {output_file}")
```

```
□ Top words for MrBeast:
transcript: 3
available: 3

□ Top words for indiatoday:
transcript: 17
available: 17
one: 6
like: 6
mind: 5
india: 4
today: 4
rocks: 4
always: 4
tip: 4

□ Top words for lofi2307:
transcript: 50
available: 50
```



```
□ Updated file saved with stats:
MrBeast_indiatoday_lofi2307_50videos_with_clean_transcripts_stats.csv

# □ Module 6: Video Index with SentenceTransformer Embeddings

!pip install sentence-transformers --upgrade
!pip install polars --upgrade

import pandas as pd
from sentence_transformers import SentenceTransformer
import numpy as np

# □ Step 1: Load cleaned dataset with stats
input_file =
"MrBeast_indiatoday_lofi2307_50videos_with_clean_transcripts_stats.csv
"
df = pd.read_csv(input_file)

# □ Step 2: Choose best model (from previous evaluation, e.g., all-
MiniLM-L6-v2)
model_name = "all-MiniLM-L6-v2"
model = SentenceTransformer(model_name)

# □ Step 3: Embed titles and clean transcripts
print("Embedding video titles...")
```

```
title_embeddings = model.encode(df["title"].tolist(),
show_progress_bar=True)

print("Embedding clean transcripts...")
transcript_embeddings = model.encode(df["clean_transcript"].tolist(),
show_progress_bar=True)

# □ Step 4: Concatenate embeddings (title + transcript)
combined_embeddings = np.hstack([title_embeddings,
transcript_embeddings])
print("Combined embedding shape:", combined_embeddings.shape)

# □ Step 5: Append embeddings to original dataframe
embedding_cols = [f"emb_{i}" for i in
range(combined_embeddings.shape[1])]
embeddings_df = pd.DataFrame(combined_embeddings,
columns=embedding_cols)
df_final = pd.concat([df.reset_index(drop=True), embeddings_df],
axis=1)

# □ Step 6: Save final video index
output_file_csv = "Video_Index_MrBeast_indiatoday_lofi2307.csv"
output_file_parquet =
"Video_Index_MrBeast_indiatoday_lofi2307.parquet"

df_final.to_csv(output_file_csv, index=False, encoding="utf-8")
df_final.to_parquet(output_file_parquet, index=False)

print(f"\n□ Final video index saved:\nCSV → {output_file_csv}\nParquet
→ {output_file_parquet}")
```

```
Requirement already satisfied: sentence-transformers in
/usr/local/lib/python3.12/dist-packages (5.1.0)
Requirement already satisfied: transformers<5.0.0,>=4.41.0 in
/usr/local/lib/python3.12/dist-packages (from sentence-transformers)
(4.56.1)
Requirement already satisfied: tqdm in /usr/local/lib/python3.12/dist-
packages (from sentence-transformers) (4.67.1)
Requirement already satisfied: torch>=1.11.0 in
/usr/local/lib/python3.12/dist-packages (from sentence-transformers)
(2.8.0+cu126)
Requirement already satisfied: scikit-learn in
/usr/local/lib/python3.12/dist-packages (from sentence-transformers)
(1.6.1)
Requirement already satisfied: scipy in
/usr/local/lib/python3.12/dist-packages (from sentence-transformers)
(1.16.1)
Requirement already satisfied: huggingface-hub>=0.20.0 in
/usr/local/lib/python3.12/dist-packages (from sentence-transformers)
(0.34.4)
```

```
Requirement already satisfied: Pillow in
/usr/local/lib/python3.12/dist-packages (from sentence-transformers)
(11.3.0)
Requirement already satisfied: typing_extensions>=4.5.0 in
/usr/local/lib/python3.12/dist-packages (from sentence-transformers)
(4.15.0)
Requirement already satisfied: filelock in
/usr/local/lib/python3.12/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (3.19.1)
Requirement already satisfied: fsspec>=2023.5.0 in
/usr/local/lib/python3.12/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (2025.3.0)
Requirement already satisfied: packaging>=20.9 in
/usr/local/lib/python3.12/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (25.0)
Requirement already satisfied: pyyaml>=5.1 in
/usr/local/lib/python3.12/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (6.0.2)
Requirement already satisfied: requests in
/usr/local/lib/python3.12/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (2.32.4)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in
/usr/local/lib/python3.12/dist-packages (from huggingface-hub>=0.20.0-
>sentence-transformers) (1.1.9)
Requirement already satisfied: setuptools in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (75.2.0)
Requirement already satisfied: sympy>=1.13.3 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (1.13.3)
Requirement already satisfied: networkx in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (3.5)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (3.1.6)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.6.77 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (12.6.77)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.6.77 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (12.6.77)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.6.80 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (12.6.80)
Requirement already satisfied: nvidia-cudnn-cu12==9.10.2.21 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (9.10.2.21)
Requirement already satisfied: nvidia-cublas-cu12==12.6.4.1 in
```

```
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (12.6.4.1)
Requirement already satisfied: nvidia-cufft-cu12==11.3.0.4 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (11.3.0.4)
Requirement already satisfied: nvidia-curand-cu12==10.3.7.77 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (10.3.7.77)
Requirement already satisfied: nvidia-cusolver-cu12==11.7.1.2 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (11.7.1.2)
Requirement already satisfied: nvidia-cusparse-cu12==12.5.4.2 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (12.5.4.2)
Requirement already satisfied: nvidia-cusparselt-cu12==0.7.1 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (0.7.1)
Requirement already satisfied: nvidia-nccl-cu12==2.27.3 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (2.27.3)
Requirement already satisfied: nvidia-nvtx-cu12==12.6.77 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (12.6.77)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.6.85 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (12.6.85)
Requirement already satisfied: nvidia-cufile-cu12==1.11.1.6 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (1.11.1.6)
Requirement already satisfied: triton==3.4.0 in
/usr/local/lib/python3.12/dist-packages (from torch>=1.11.0->sentence-
transformers) (3.4.0)
Requirement already satisfied: numpy>=1.17 in
/usr/local/lib/python3.12/dist-packages (from
transformers<5.0.0,>=4.41.0->sentence-transformers) (2.0.2)
Requirement already satisfied: regex!=2019.12.17 in
/usr/local/lib/python3.12/dist-packages (from
transformers<5.0.0,>=4.41.0->sentence-transformers) (2024.11.6)
Requirement already satisfied: tokenizers<=0.23.0,>=0.22.0 in
/usr/local/lib/python3.12/dist-packages (from
transformers<5.0.0,>=4.41.0->sentence-transformers) (0.22.0)
Requirement already satisfied: safetensors>=0.4.3 in
/usr/local/lib/python3.12/dist-packages (from
transformers<5.0.0,>=4.41.0->sentence-transformers) (0.6.2)
Requirement already satisfied: joblib>=1.2.0 in
/usr/local/lib/python3.12/dist-packages (from scikit-learn->sentence-
transformers) (1.5.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in
/usr/local/lib/python3.12/dist-packages (from scikit-learn->sentence-
transformers) (3.6.0)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/usr/local/lib/python3.12/dist-packages (from sympy>=1.13.3-
>torch>=1.11.0->sentence-transformers) (1.3.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.12/dist-packages (from jinja2->torch>=1.11.0-
>sentence-transformers) (3.0.2)
Requirement already satisfied: charset_normalizer<4,>=2 in
/usr/local/lib/python3.12/dist-packages (from requests->huggingface-
hub>=0.20.0->sentence-transformers) (3.4.3)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.12/dist-packages (from requests->huggingface-
hub>=0.20.0->sentence-transformers) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.12/dist-packages (from requests->huggingface-
hub>=0.20.0->sentence-transformers) (2.5.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.12/dist-packages (from requests->huggingface-
hub>=0.20.0->sentence-transformers) (2025.8.3)
Requirement already satisfied: polars in
/usr/local/lib/python3.12/dist-packages (1.25.2)
Collecting polars
  Downloading polars-1.33.1-cp39-abi3-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (15 kB)
Downloading polars-1.33.1-cp39-abi3-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (39.7 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 39.7/39.7 MB 26.5 MB/s eta
0:00:00
pting uninstall: polars
    Found existing installation: polars 1.25.2
    Uninstalling polars-1.25.2:
      Successfully uninstalled polars-1.25.2
ERROR: pip's dependency resolver does not currently take into account
all the packages that are installed. This behaviour is the source of
the following dependency conflicts.
cudf-polars-cu12 25.6.0 requires polars<1.29,>=1.25, but you have
polars 1.33.1 which is incompatible.
Successfully installed polars-1.33.1

/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/
_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your
settings tab (https://huggingface.co/settings/tokens), set it as
secret in your Google Colab and restart your session.
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to
access public models or datasets.
  warnings.warn(
{"model_id":"e70e328a03524fd88b2e6e0010beb93e","version_major":2,"version_minor":0}

{"model_id":"ac190ee21cb544c3ab88a300e167b1c9","version_major":2,"version_minor":0}

{"model_id":"68db3c787d52483bb483c59295761a87","version_major":2,"version_minor":0}

{"model_id":"9a23360cb6ce43f9933452ab134a0d4f","version_major":2,"version_minor":0}

{"model_id":"4ca14ad464724b9783d821604dd395a7","version_major":2,"version_minor":0}

{"model_id":"cad4bf7ee7334f3bbd25a53dffb9ea56","version_major":2,"version_minor":0}

{"model_id":"04d6960b403a4533a925884841d1ff14","version_major":2,"version_minor":0}

{"model_id":"a99a9ba3608a46c491481bd5cb2c7e95","version_major":2,"version_minor":0}

{"model_id":"d1e0a48654904037ac39f4292d72bef9","version_major":2,"version_minor":0}

{"model_id":"c07c1d594c444e9d8ffbed4934fdf706","version_major":2,"version_minor":0}

{"model_id":"4403c0393dff4691805160370a914dca","version_major":2,"version_minor":0}

Embedding video titles...

{"model_id":"6afd9380bc704c4dbe8cac0b7c5dba21","version_major":2,"version_minor":0}

Embedding clean transcripts...

{"model_id":"b84699f758064e58a7b004ed60aa3c17","version_major":2,"version_minor":0}

Combined embedding shape: (79, 768)

□ Final video index saved:
CSV → Video_Index_MrBeast_indiatoday_lofi2307.csv
Parquet → Video_Index_MrBeast_indiatoday_lofi2307.parquet
```