```
# ----------------------------
# Install necessary Python packages
# ----------------------------
!pip install pandas matplotlib seaborn streamlit wordcloud pyngrok
```

```
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: streamlit in /usr/local/lib/python3.12/dist-packages (1.51.0)
Requirement already satisfied: wordcloud in /usr/local/lib/python3.12/dist-packages (1.9.4)
Requirement already satisfied: pyngrok in /usr/local/lib/python3.12/dist-packages (7.5.0)
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) (4.60.1)
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) (3.2.5)
Requirement already satisfied: altair!=5.4.0,!=5.4.1,<6,>=4.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (5.5.0
Requirement already satisfied: blinker<2,>=1.5.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (1.9.0)
Requirement already satisfied: cachetools<7,>=4.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (6.2.2)
Requirement already satisfied: click<9,>=7.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (8.3.1)
Requirement already satisfied: protobuf<7,>=3.20 in /usr/local/lib/python3.12/dist-packages (from streamlit) (5.29.5)
Requirement already satisfied: pyarrow<22,>=7.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (18.1.0)
Requirement already satisfied: requests<3,>=2.27 in /usr/local/lib/python3.12/dist-packages (from streamlit) (2.32.4)
Requirement already satisfied: tenacity<10,>=8.1.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (9.1.2)
Requirement already satisfied: toml<2,>=0.10.1 in /usr/local/lib/python3.12/dist-packages (from streamlit) (0.10.2)
Requirement already satisfied: typing-extensions<5,>=4.4.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (4.15.0)
Requirement already satisfied: watchdog<7,>=2.1.5 in /usr/local/lib/python3.12/dist-packages (from streamlit) (6.0.0)
Requirement already satisfied: gitpython!=3.1.19,<4,>=3.0.7 in /usr/local/lib/python3.12/dist-packages (from streamlit) (3.1.45)
Requirement already satisfied: pydeck<1,>=0.8.0b4 in /usr/local/lib/python3.12/dist-packages (from streamlit) (0.9.1)
Requirement already satisfied: tornado!=6.5.0,<7,>=6.0.3 in /usr/local/lib/python3.12/dist-packages (from streamlit) (6.5.1)
Requirement already satisfied: PyYAML>=5.1 in /usr/local/lib/python3.12/dist-packages (from pyngrok) (6.0.3)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.12/dist-packages (from altair!=5.4.0,!=5.4.1,<6,>=4.0->streamlit
Requirement already satisfied: jsonschema>=3.0 in /usr/local/lib/python3.12/dist-packages (from altair!=5.4.0,!=5.4.1,<6,>=4.0->
Requirement already satisfied: narwhals>=1.14.2 in /usr/local/lib/python3.12/dist-packages (from altair!=5.4.0,!=5.4.1,<6,>=4.0-
Requirement already satisfied: gitdb<5,>=4.0.1 in /usr/local/lib/python3.12/dist-packages (from gitpython!=3.1.19,<4,>=3.0.7->st
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: charset_normalizer<4,>=2 in /usr/local/lib/python3.12/dist-packages (from requests<3,>=2.27->stre
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.12/dist-packages (from requests<3,>=2.27->streamlit) (3.11
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.12/dist-packages (from requests<3,>=2.27->streamlit)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.12/dist-packages (from requests<3,>=2.27->streamlit)
Requirement already satisfied: smmap<6,>=3.0.1 in /usr/local/lib/python3.12/dist-packages (from gitdb<5,>=4.0.1->gitpython!=3.1.
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.12/dist-packages (from jinja2->altair!=5.4.0,!=5.4.1,<6
Requirement already satisfied: attrs>=22.2.0 in /usr/local/lib/python3.12/dist-packages (from jsonschema>=3.0->altair!=5.4.0,!=5
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /usr/local/lib/python3.12/dist-packages (from jsonschema>
Requirement already satisfied: referencing>=0.28.4 in /usr/local/lib/python3.12/dist-packages (from jsonschema>=3.0->altair!=5.4
Requirement already satisfied: rpds-py>=0.7.1 in /usr/local/lib/python3.12/dist-packages (from jsonschema>=3.0->altair!=5.4.0,!=
```

```
# ----------------------------
# Import required libraries
# ----------------------------
import pandas as pd     # For data manipulation
import numpy as np      # For numerical operations and random choice

# ----------------------------
# Define number of sample papers
# ----------------------------
n = 10000  # We will generate 10,000 sample research papers

# ----------------------------
# Generate sample dataset
# ----------------------------
df = pd.DataFrame({
    # Generate titles: "COVID Research Paper 0", "COVID Research Paper 1", ...
    "title": ["COVID Research Paper " + str(i) for i in range(n)],

    # Generate abstracts: "This study discusses COVID-19 impacts and findings 0", ...
    "abstract": ["This study discusses COVID-19 impacts and findings " + str(i) for i in range(n)],

    # Generate publication dates starting from 2018-01-01, one paper per day
    "publish_time": pd.date_range(start="2018-01-01", periods=n, freq="D"),
```

```python
    # Randomly assign a journal from a list of top journals
    "journal": np.random.choice(
        ["Nature", "Science", "The Lancet", "BMJ", "JAMA"], n
    ),

    # Randomly assign a source from a list of repositories
    "source_x": np.random.choice(
        ["PMC", "bioRxiv", "medRxiv", "WHO"], n
    )
})

# ----------------------------
# Save the sample dataset to a CSV file
# ----------------------------
df.to_csv("metadata_sample_10k.csv", index=False)  # index=False avoids adding row numbers

# ----------------------------
# Preview the first 5 rows of the dataset
# ----------------------------
df.head()
```

|   | title | abstract | publish_time | journal | source_x |
|---|-------|----------|--------------|---------|----------|
| 0 | COVID Research Paper 0 | This study discusses COVID-19 impacts and find... | 2018-01-01 | BMJ | PMC |
| 1 | COVID Research Paper 1 | This study discusses COVID-19 impacts and find... | 2018-01-02 | The Lancet | PMC |
| 2 | COVID Research Paper 2 | This study discusses COVID-19 impacts and find... | 2018-01-03 | Nature | medRxiv |
| 3 | COVID Research Paper 3 | This study discusses COVID-19 impacts and find... | 2018-01-04 | BMJ | PMC |
| 4 | COVID Research Paper 4 | This study discusses COVID-19 impacts and find... | 2018-01-05 | The Lancet | PMC |

Next steps:  [ Generate code with df ]  [ New interactive sheet ]

```python
# ----------------------------
# Basic info about the dataset
# ----------------------------
print("Dataset shape:", df.shape)  # rows and columns
print("\nData types:\n", df.dtypes)  # column types
print("\nMissing values:\n", df.isnull().sum())  # missing data per column

# Show summary statistics for numeric columns
df.describe()
```

```
Dataset shape: (10000, 5)
```

```python
# ----------------------------
# Convert 'publish_time' to datetime
# ----------------------------
df['publish_time'] = pd.to_datetime(df['publish_time'], errors='coerce')

# Extract year for analysis
df['year'] = df['publish_time'].dt.year

# Handle missing titles or abstracts
df['title'] = df['title'].fillna("Unknown Title")
df['abstract'] = df['abstract'].fillna("No Abstract")

# Optional: create word count for abstracts
df['abstract_word_count'] = df['abstract'].apply(lambda x: len(str(x).split()))

# Preview cleaned data
df.head()
```

| | title | abstract | publish_time | journal | source_x | year | abstract_word_count |
|---|---|---|---|---|---|---|---|
| min | 2018-01-01 00:00:00 | | | | | | |
| 0 | COVID Research Paper 0 | This study discusses COVID-19 impacts and find... | 2018-01-01 | BMJ | PMC | 2018 | 8 |
| 50% | 2031-09-09 12:00:00 | | | | | | |
| 1 | COVID Research Paper 1 | This study discusses COVID-19 impacts and find... | 2018-01-02 | The Lancet | PMC | 2018 | 8 |
| 75% | 2038-07-14 06:00:00 | | | | | | |
| 2 | COVID Research Paper 2 | This study discusses COVID-19 impacts and find... | 2018-01-03 | Nature | medRxiv | 2018 | 8 |
| | COVID Research | This study discusses COVID-19 impacts | 2018-01-04 | BMJ | PMC | 2018 | 8 |

Next steps:  [ Generate code with df ]  [ New interactive sheet ]

```python
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud

# ----------------------------
# 1. Publications per year
# ----------------------------
year_counts = df['year'].value_counts().sort_index()

plt.figure(figsize=(10,5))
sns.barplot(x=year_counts.index, y=year_counts.values, palette="viridis")
plt.xlabel("Year")
plt.ylabel("Number of Papers")
plt.title("Publications per Year")
plt.show()

# ----------------------------
# 2. Top 10 Journals
# ----------------------------
top_journals = df['journal'].value_counts().head(10)
plt.figure(figsize=(10,5))
sns.barplot(x=top_journals.values, y=top_journals.index, palette="magma")
plt.xlabel("Number of Papers")
plt.ylabel("Journal")
plt.title("Top 10 Journals Publishing COVID-19 Research")
plt.show()

# ----------------------------
# 3. Word Cloud for Titles
# ----------------------------
text = " ".join(df['title'].astype(str))
wc = WordCloud(width=800, height=400, background_color='white').generate(text)

plt.figure(figsize=(15,7))
plt.imshow(wc, interpolation='bilinear')
plt.axis('off')
plt.title("Word Cloud of Paper Titles")
plt.show()
```

```
/tmp/ipython-input-2708547185.py:11: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and s

    sns.barplot(x=year_counts.index, y=year_counts.values, palette="viridis")
```

Publications per Year

350 –

```python
# ----------------------------
# Create Streamlit app (app.py)
# ----------------------------
%%writefile app.py
import streamlit as st
import pandas as pd
import matplotlib.pyplot as plt
from wordcloud import WordCloud

st.title("CORD-19 Metadata Explorer")
st.write("Interactive dashboard for COVID-19 research papers")

# Load data
df = pd.read_csv("metadata.csv")
df['publish_time'] = pd.to_datetime(df['publish_time'], errors='coerce')
df['year'] = df['publish_time'].dt.year
df['title'] = df['title'].fillna("Unknown Title")

# Sidebar filter
st.sidebar.header("Filters")
year_range = st.sidebar.slider(
    "Select publication year range:",
    int(df['year'].min()),
    int(df['year'].max()),
    (int(df['year'].min()), int(df['year'].max()))
)

filtered_df = df[(df['year'] >= year_range[0]) & (df['year'] <= year_range[1])]

# Display sample data
st.subheader("Filtered Papers")
st.write(filtered_df.head())

# Publications per year
st.subheader("Publications Over Time")
year_counts = filtered_df['year'].value_counts().sort_index()
fig, ax = plt.subplots()
year_counts.plot(kind='bar')
plt.xlabel("Year")
plt.ylabel("Number of Papers")
st.pyplot(fig)

# Top Journals
st.subheader("Top Journals")
top_journals = filtered_df['journal'].value_counts().head(10)
st.bar_chart(top_journals)

# Word Cloud
st.subheader("Word Cloud of Titles")
text = " ".join(filtered_df['title'].astype(str))
wc = WordCloud(width=800, height=400).generate(text)
fig, ax = plt.subplots(figsize=(10,5))
ax.imshow(wc, interpolation='bilinear')
ax.axis('off')
st.pyplot(fig)
```

```
Writing app.py
```

```python
!pip install pyngrok streamlit
```

```
Requirement already satisfied: pyngrok in /usr/local/lib/python3.12/dist-packages (7.5.0)
Requirement already satisfied: streamlit in /usr/local/lib/python3.12/dist-packages (1.51.0)
Requirement already satisfied: PyYAML>=5.1 in /usr/local/lib/python3.12/dist-packages (from pyngrok) (6.0.3)
Requirement already satisfied: altair!=5.4.0,!=5.4.1,<6,>=4.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (5.5.0
```

```
Requirement already satisfied: blinker<2,>=1.5.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (1.9.0)
Requirement already satisfied: cachetools<7,>=4.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (6.2.2)
Requirement already satisfied: click<9,>=7.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (8.3.1)
Requirement already satisfied: numpy<3,>=1.23 in /usr/local/lib/python3.12/dist-packages (from streamlit) (2.0.2)
Requirement already satisfied: packaging<26,>=20 in /usr/local/lib/python3.12/dist-packages (from streamlit) (25.0)
Requirement already satisfied: pandas<3,>=1.4.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (2.2.2)
Requirement already satisfied: pillow<13,>=7.1.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (11.3.0)
Requirement already satisfied: protobuf<7,>=3.20 in /usr/local/lib/python3.12/dist-packages (from streamlit) (5.29.5)
Requirement already satisfied: pyarrow<22,>=7.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (18.1.0)
Requirement already satisfied: requests<3,>=2.27 in /usr/local/lib/python3.12/dist-packages (from streamlit) (2.32.4)
Requirement already satisfied: tenacity<10,>=8.1.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (9.1.2)
Requirement already satisfied: toml<2,>=0.10.1 in /usr/local/lib/python3.12/dist-packages (from streamlit) (0.10.2)
Requirement already satisfied: typing-extensions<5,>=4.4.0 in /usr/local/lib/python3.12/dist-packages (from streamlit) (4.15.0)
Requirement already satisfied: watchdog<7,>=2.1.5 in /usr/local/lib/python3.12/dist-packages (from streamlit) (6.0.0)
Requirement already satisfied: gitpython!=3.1.19,<4,>=3.0.7 in /usr/local/lib/python3.12/dist-packages (from streamlit) (3.1.45)
Requirement already satisfied: pydeck<1,>=0.8.0b4 in /usr/local/lib/python3.12/dist-packages (from streamlit) (0.9.1)
Requirement already satisfied: tornado!=6.5.0,<7,>=6.0.3 in /usr/local/lib/python3.12/dist-packages (from streamlit) (6.5.1)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.12/dist-packages (from altair!=5.4.0,!=5.4.1,<6,>=4.0->streamlit
Requirement already satisfied: jsonschema>=3.0 in /usr/local/lib/python3.12/dist-packages (from altair!=5.4.0,!=5.4.1,<6,>=4.0->
Requirement already satisfied: narwhals>=1.14.2 in /usr/local/lib/python3.12/dist-packages (from altair!=5.4.0,!=5.4.1,<6,>=4.0-
Requirement already satisfied: gitdb<5,>=4.0.1 in /usr/local/lib/python3.12/dist-packages (from gitpython!=3.1.19,<4,>=3.0.7->st
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.12/dist-packages (from pandas<3,>=1.4.0->streaml
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.12/dist-packages (from pandas<3,>=1.4.0->streamlit) (2025.
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.12/dist-packages (from pandas<3,>=1.4.0->streamlit) (202
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.12/dist-packages (from requests<3,>=2.27->stre
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.12/dist-packages (from requests<3,>=2.27->streamlit) (3.11
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.12/dist-packages (from requests<3,>=2.27->streamlit)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.12/dist-packages (from requests<3,>=2.27->streamlit)
Requirement already satisfied: smmap<6,>=3.0.1 in /usr/local/lib/python3.12/dist-packages (from gitdb<5,>=4.0.1->gitpython!=3.1.
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.12/dist-packages (from jinja2->altair!=5.4.0,!=5.4.1,<6
Requirement already satisfied: attrs>=22.2.0 in /usr/local/lib/python3.12/dist-packages (from jsonschema>=3.0->altair!=5.4.0,!=5
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /usr/local/lib/python3.12/dist-packages (from jsonschema>
Requirement already satisfied: referencing>=0.28.4 in /usr/local/lib/python3.12/dist-packages (from jsonschema>=3.0->altair!=5.4
Requirement already satisfied: rpds-py>=0.7.1 in /usr/local/lib/python3.12/dist-packages (from jsonschema>=3.0->altair!=5.4.0,!=
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-packages (from python-dateutil>=2.8.2->pandas<3,>=1.4.
```

```python
from pyngrok import ngrok
```

```python
# ==========================================
# Step 2: Register your ngrok authtoken
# ------------------------------------------
# Replace the authtoken below with your own.
# This authenticates your ngrok account so you
# can create tunnels without errors.
# ==========================================

!ngrok config add-authtoken 36FIoPclz9NMa5eoPGXFQsI8JEU_7Kxobgj6nj5mhed8KAff3
```

```
Authtoken saved to configuration file: /root/.config/ngrok/ngrok.yml
```

```python
# ==========================================
# Step 3: Start the Streamlit app
# ------------------------------------------
# --server.port sets Streamlit to run on port 8501
# --server.address 0.0.0.0 allows external access
# '&' at the end runs Streamlit in the background.
# Make sure your Streamlit script is named: app.py
# ==========================================

!streamlit run app.py --server.port 8501 --server.address 0.0.0.0 &
```

```
Collecting usage statistics. To deactivate, set browser.gatherUsageStats to false.


  You can now view your Streamlit app in your browser.

  URL: http://0.0.0.0:8501

  Stopping...
```

```python
from pyngrok import ngrok

# Open a public tunnel to the Streamlit port (8501)
public_url = ngrok.connect(8501)
```