

Modern NLP Approach - Extractive Summarization

▼ Load The Dataset

```
# Mount Google Drive (if using Colab)
from google.colab import drive
drive.mount('/content/drive')

import pandas as pd

# Load cleaned dataset
input_path = "/content/drive/My Drive/NLP/Assignment_3/fomc_transcripts_spacy_cleaned.csv"
df = pd.read_csv(input_path)
df['Date'] = pd.to_datetime(df['Date'])
```

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount() again.

df

		URL	Date	Year	Month	Day	Content	date
0	www.federalreserve.gov/monetarypolicy/...	https://www.federalreserve.gov/monetarypolicy/...	2025-03-19	2025	2025	03-01	19	Home Monetary Policy Federal Open Market Committ...
1	www.federalreserve.gov/monetarypolicy/...	https://www.federalreserve.gov/monetarypolicy/...	2025-01-29	2025	2025	01-01	29	Home Monetary Policy Federal Open Market Committ...
2	www.federalreserve.gov/monetarypolicy/...	https://www.federalreserve.gov/monetarypolicy/...	2024-12-18	2024	2024	12-01	18	Home Monetary Policy Federal Open Market Committ...
3	www.federalreserve.gov/monetarypolicy/...	https://www.federalreserve.gov/monetarypolicy/...	2024-11-07	2024	2024	11-01	7	Home Monetary Policy Federal Open Market Committ...
4	www.federalreserve.gov/monetarypolicy/...	https://www.federalreserve.gov/monetarypolicy/...	2024-09-18	2024	2024	09-01	18	Home Monetary Policy Federal Open Market Committ...
...

			https://																
61	www.federalreserve.gov/	2015-07-29	2015	2015	2015-07-01	29	Home	Monetary	Policy	Federal	Open Market	Commit...							2015-07-31
	monetarypolicy/...																		
			https://																
62	www.federalreserve.gov/	2015-06-17	2015	2015	2015-06-01	17	Home	Monetary	Policy	Federal	Open Market	Commit...							2015-06-30
	monetarypolicy/...																		
			https://																
63	www.federalreserve.gov/	2015-04-29	2015	2015	2015-04-01	29	Home	Monetary	Policy	Federal	Open Market	Commit...							2015-04-30
	monetarypolicy/...																		
			https://																
64	www.federalreserve.gov/	2015-03-18	2015	2015	2015-03-01	18	Home	Monetary	Policy	Federal	Open Market	Commit...							2015-03-31
	monetarypolicy/...																		
			https://																
65	www.federalreserve.gov/	2015-01-28	2015	2015	2015-01-01	28	Home	Monetary	Policy	Federal	Open Market	Commit...							2015-01-31
	monetarypolicy/...																		

66 rows × 20 columns

Next steps: [Generate code with df](#) [View recommended plots](#) [New interactive sheet](#)

```
# Show summary info, including dtypes and memory usage
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 66 entries, 0 to 65
Data columns (total 20 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   URL              66 non-null    object  
 1   Date             66 non-null    datetime64[ns]
 2   Year             66 non-null    int64  
 3   Month            66 non-null    object  
 4   Day              66 non-null    int64  
 5   Content          66 non-null    object  
 6   date             66 non-null    object  
 7   GDP              66 non-null    float64 
 8   CPI              66 non-null    float64 
 9   Unemployment    66 non-null    float64 
 10  FedFundsRate    66 non-null    float64 
 11  SNR_Ratio       66 non-null    float64 
 12  StopWord_Ratio  66 non-null    float64 
 13  Redundancy_Ratio 66 non-null    float64 
 14  SpecialChar_Density 66 non-null    float64 
 15  Semantic_Density 66 non-null    float64
```

```
15  Semantic_Diversity      66 non-null    float64
16  Digit_Ratio            66 non-null    float64
17  Noise_Score             66 non-null    float64
18  Quality_Score           66 non-null    float64
19  cleaned_text            66 non-null    object
dtypes: datetime64[ns](1), float64(12), int64(2), object(5)
memory usage: 10.4+ KB
```

❖ Sorting the Date and Index

```
import pandas as pd

# 1) Convert the 'Date' column from object (string) to datetime
df['Date'] = pd.to_datetime(df['Date'], errors='coerce')

# 2) (Optional) Drop or inspect rows where conversion failed
# print(df[df['Date'].isna()])

# 3) Sort the DataFrame by the new datetime 'Date' column in descending order
df = df.sort_values(by='Date', ascending=False)

# 4) (Optional) Reset index if you want a clean integer index
df = df.reset_index(drop=True)

# Now df['Date'] is a datetime dtype and the rows are sorted newest → oldest
#print(df[['doc_id', 'Date']].head())
```

```
# 1) Make sure your index is reset 0...N-1
df = df.reset_index(drop=True)

# 2) Rebuild doc_id as "doc1", "doc2", ... up to "docN"
df['doc_id'] = 'doc' + (df.index + 1).astype(str)

# 3) Check
print(df[['doc_id', 'Date']].head())
```

```
doc_id      Date
0  doc1 2025-03-19
1  doc2 2025-01-29
2  doc3 2024-12-18
3  doc4 2024-11-07
4  doc5 2024-09-18
```

```
import spacy
import pandas as pd
import numpy as np
```

```
import plotly.graph_objects as go

# 1) Load spaCy model for sentence splitting
nlp = spacy.load("en_core_web_sm")

# 2) Build analysis DataFrame with text and Date columns
df_analysis = pd.DataFrame({
    "text": df["cleaned_text"],
    "Date": pd.to_datetime(df["Date"]).dt.date # convert to date only
})

# 3) Function to compute sentence, word, token, and chunk metrics
def compute_bert_metrics(text):
    doc = nlp(text)
    total_sentences = len(list(doc.sents))
    total_words      = sum(len([t for t in sent if not t.is_space]) for sent in doc.sents)
    bert_tokens      = total_words * 0.75
    bert_chunks_450 = int(np.ceil(bert_tokens / 450))
    return total_sentences, total_words, round(bert_tokens, 1), bert_chunks_450

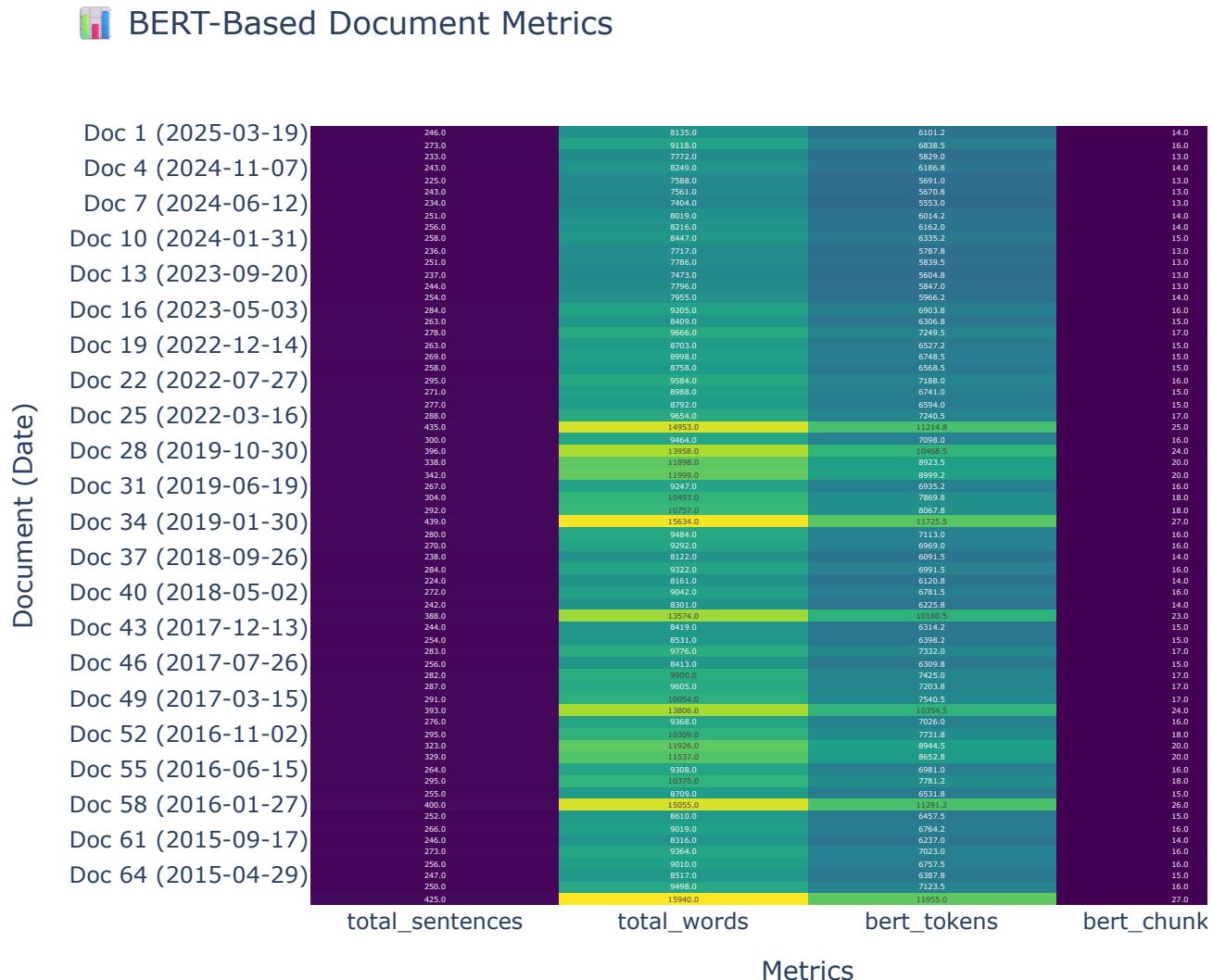
# 4) Apply metrics function and expand into separate columns
df_analysis[
    ["total_sentences", "total_words", "bert_tokens", "bert_chunks_450"]
] = pd.DataFrame(
    df_analysis["text"]
        .apply(compute_bert_metrics)
        .tolist(),
    index=df_analysis.index
)

# 5) Prepare heatmap data
metrics  = ["total_sentences", "total_words", "bert_tokens", "bert_chunks_450"]
z        = df_analysis[metrics].values
z_text   = [[str(val) for val in row] for row in z]
x_labels = metrics
y_labels = [f"Doc {i+1} ({date})" for i, date in enumerate(df_analysis["Date"])] 

# 6) Build and display the heatmap with values and clean dates
fig = go.Figure(data=go.Heatmap(
    z=z,
    x=x_labels,
    y=y_labels,
    text=z_text,           # display each cell's value
    texttemplate="%{text}", # template for cell text
    colorscale="Viridis",
    colorbar=dict(title="Value")
))

# 7) Reverse y-axis so that Doc1 appears at the top
fig.update_yaxes(autorange="reversed")
```

```
# 8) Layout tweaks
fig.update_layout(
    title="📊 BERT-Based Document Metrics",
    xaxis_title="Metrics",
    yaxis_title="Document (Date)",
    height=600,
    width=800,
    margin=dict(l=150) # make space for longer y-labels
)
fig.show()
```



df

URL	Date	Year	Month	Day	Content	date
-----	------	------	-------	-----	---------	------

0	https://www.federalreserve.gov/monetarypolicy/...	2025-03-19	2025	2025-03-01	19	Home Monetary Policy Federal Open Market Committ...	2025-03-31
1	https://www.federalreserve.gov/monetarypolicy/...	2025-01-29	2025	2025-01-01	29	Home Monetary Policy Federal Open Market Committ...	2025-01-31
2	https://www.federalreserve.gov/monetarypolicy/...	2024-12-18	2024	2024-12-01	18	Home Monetary Policy Federal Open Market Committ...	2024-12-31
3	https://www.federalreserve.gov/monetarypolicy/...	2024-11-07	2024	2024-11-01	7	Home Monetary Policy Federal Open Market Committ...	2024-11-30
4	https://www.federalreserve.gov/monetarypolicy/...	2024-09-18	2024	2024-09-01	18	Home Monetary Policy Federal Open Market Committ...	2024-09-30
...
61	https://www.federalreserve.gov/monetarypolicy/...	2015-07-29	2015	2015-07-01	29	Home Monetary Policy Federal Open Market Committ...	2015-07-31
62	https://www.federalreserve.gov/monetarypolicy/...	2015-06-17	2015	2015-06-01	17	Home Monetary Policy Federal Open Market Committ...	2015-06-30
63	https://www.federalreserve.gov/monetarypolicy/...	2015-04-29	2015	2015-04-01	29	Home Monetary Policy Federal Open Market Committ...	2015-04-30
64	https://www.federalreserve.gov/monetarypolicy/...	2015-03-18	2015	2015-03-01	18	Home Monetary Policy Federal Open Market Committ...	2015-03-31
65	https://www.federalreserve.gov/monetarypolicy/...	2015-01-28	2015	2015-01-01	28	Home Monetary Policy Federal Open Market Committ...	2015-01-31

66 rows × 21 columns

!pip install --upgrade transformers

```
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (f
Requirement already satisfied: huggingface-hub<1.0,>=0.30.0 in /usr/local/lib/python3
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-pa
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (f
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/di
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.1
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-package
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-p
```

```
import transformers
print(transformers.__version__)
```

4.51.3

```
!pip install --upgrade spacy
```

```
Requirement already satisfied: spacy in /usr/local/lib/python3.11/dist-packages (3.8.
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /usr/local/lib/python3.
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /usr/local/lib/python3.
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/python3.11
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.11/dist-
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/python3.11/dis
Requirement already satisfied: thinc<8.4.0,>=8.3.4 in /usr/local/lib/python3.11/dist-
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /usr/local/lib/python3.11/dist
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /usr/local/lib/python3.11/dist-
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /usr/local/lib/python3.11/d
Requirement already satisfied: weasel<0.5.0,>=0.1.0 in /usr/local/lib/python3.11/dist
Requirement already satisfied: typer<1.0.0,>=0.3.0 in /usr/local/lib/python3.11/dist-
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /usr/local/lib/python3.11/dist-
Requirement already satisfied: numpy>=1.19.0 in /usr/local/lib/python3.11/dist-packag
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/local/lib/python3.11/d
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in /usr/local/lib
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (fro
Requirement already satisfied: setuptools in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /usr/local/lib/python3.11/d
Requirement already satisfied: language-data>=1.2 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.11/di
Requirement already satisfied: pydantic-core==2.33.1 in /usr/local/lib/python3.11/dis
Requirement already satisfied: typing-extensions>=4.12.2 in /usr/local/lib/python3.11
Requirement already satisfied: typing-inspection>=0.4.0 in /usr/local/lib/python3.11/
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-package
```

```
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: blis<1.4.0,>=1.3.0 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: click>=8.0.0 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: shellingham>=1.3.0 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: rich>=10.11.0 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: cloudpathlib<1.0.0,>=0.7.0 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: smart-open<8.0.0,>=5.2.1 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: marisa-trie>=1.1.0 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: markdown-it-py>=2.2.0 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: wrapt in /usr/local/lib/python3.11/dist-packages (from -r requirements.txt (line 1))
Requirement already satisfied: mdurl~0.1 in /usr/local/lib/python3.11/dist-packages
```

```
nlp = spacy.load("en_core_web_sm")
```

```
!pip install bert-score
```

```
Collecting bert-score
  Downloading bert_score-0.3.13-py3-none-any.whl.metadata (15 kB)
Requirement already satisfied: torch>=1.0.0 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: pandas>=1.0.1 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: transformers>=3.0.0 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from bert-score)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from bert-score)
Requirement already satisfied: tqdm>=4.31.1 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: matplotlib in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: packaging>=20.9 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from bert-score)
Requirement already satisfied: typing-extensions>=4.10.0 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from bert-score)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (from bert-score)
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages (from bert-score)
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch>=1.0.0->bert-score)
  Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (15 kB)
Collecting nvidia-cuda-runtime-cu12==12.4.127 (from torch>=1.0.0->bert-score)
  Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (15 kB)
Collecting nvidia-cuda-cupti-cu12==12.4.127 (from torch>=1.0.0->bert-score)
  Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (15 kB)
Collecting nvidia-cudnn-cu12==9.1.0.70 (from torch>=1.0.0->bert-score)
  Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl.metadata (15 kB)
Collecting nvidia-cublas-cu12==12.4.5.8 (from torch>=1.0.0->bert-score)
  Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl.metadata (15 kB)
Collecting nvidia-cufft-cu12==11.2.1.3 (from torch>=1.0.0->bert-score)
  Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl.metadata (15 kB)
Collecting nvidia-curand-cu12==10.3.5.147 (from torch>=1.0.0->bert-score)
  Downloading nvidia_curand_cu12-10.3.5.147-py3-none-manylinux2014_x86_64.whl.metadata (15 kB)
Collecting nvidia-cusolver-cu12==11.6.1.9 (from torch>=1.0.0->bert-score)
  Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-manylinux2014_x86_64.whl.metadata (15 kB)
```

```
Collecting nvidia-cusparse-cu12==12.3.1.170 (from torch>=1.0.0->bert-score)
  Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-manylinux2014_x86_64.whl.metadata
Requirement already satisfied: nvidia-cusparseelt-cu12==0.6.2 in /usr/local/lib/python
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.1
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch>=1.0.0->bert-score)
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: huggingface-hub<1.0,>=0.30.0 in /usr/local/lib/python3
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-pa
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/di
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.11/dist-package
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.11/dist-pa
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.11/dist-pa
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.11/dist-packages (
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.11/dist-pac
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-package
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-p
```

```
# # STEP 1: UNINSTALL OLD VERSIONS (if needed)
# !pip uninstall -y torch torchvision

# # STEP 2: INSTALL STABLE COMPATIBLE VERSIONS
# !pip install torch==2.1.0 torchvision==0.16.0
```

>Loading the model and the tokenizer

```
# 🚨 Force reinstall PyTorch & NumPy without cache to fix compatibility
!pip install --force-reinstall --no-cache-dir torch==2.1.0 torchvision==0.16.0 numpy==1.26.
```

```
Collecting torch==2.1.0
  Downloading torch-2.1.0-cp311-cp311-manylinux1_x86_64.whl.metadata (25 kB)
Collecting torchvision==0.16.0
  Downloading torchvision-0.16.0-cp311-cp311-manylinux1_x86_64.whl.metadata (6.6 kB)
Collecting numpy==1.26.4
  Downloading numpy-1.26.4-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
                                                 61.0/61.0 kB 185.1 MB/s eta 0:00:00
Collecting filelock (from torch==2.1.0)
  Downloading filelock-3.18.0-py3-none-any.whl.metadata (2.9 kB)
Collecting typing-extensions (from torch==2.1.0)
  Downloading typing_extensions-4.13.2-py3-none-any.whl.metadata (3.0 kB)
Collecting sympy (from torch==2.1.0)
  Downloading sympy-1.14.0-py3-none-any.whl.metadata (12 kB)
```

```
Collecting networkx (from torch==2.1.0)
  Downloading networkx-3.4.2-py3-none-any.whl.metadata (6.3 kB)
Collecting jinja2 (from torch==2.1.0)
  Downloading jinja2-3.1.6-py3-none-any.whl.metadata (2.9 kB)
Collecting fsspec (from torch==2.1.0)
  Downloading fsspec-2025.3.2-py3-none-any.whl.metadata (11 kB)
Collecting nvidia-cuda-nvrtc-cu12==12.1.105 (from torch==2.1.0)
  Downloading nvidia_cuda_nvrtc_cu12-12.1.105-py3-none-manylinux1_x86_64.whl.metadata
Collecting nvidia-cuda-runtime-cu12==12.1.105 (from torch==2.1.0)
  Downloading nvidia_cuda_runtime_cu12-12.1.105-py3-none-manylinux1_x86_64.whl.metadata
Collecting nvidia-cuda-cupti-cu12==12.1.105 (from torch==2.1.0)
  Downloading nvidia_cuda_cupti_cu12-12.1.105-py3-none-manylinux1_x86_64.whl.metadata
Collecting nvidia-cudnn-cu12==8.9.2.26 (from torch==2.1.0)
  Downloading nvidia_cudnn_cu12-8.9.2.26-py3-none-manylinux1_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cublas-cu12==12.1.3.1 (from torch==2.1.0)
  Downloading nvidia_cublas_cu12-12.1.3.1-py3-none-manylinux1_x86_64.whl.metadata (1.1 kB)
Collecting nvidia-cufft-cu12==11.0.2.54 (from torch==2.1.0)
  Downloading nvidia_cufft_cu12-11.0.2.54-py3-none-manylinux1_x86_64.whl.metadata (1.1 kB)
Collecting nvidia-curand-cu12==10.3.2.106 (from torch==2.1.0)
  Downloading nvidia_curand_cu12-10.3.2.106-py3-none-manylinux1_x86_64.whl.metadata (1.1 kB)
Collecting nvidia-cusolver-cu12==11.4.5.107 (from torch==2.1.0)
  Downloading nvidia_cusolver_cu12-11.4.5.107-py3-none-manylinux1_x86_64.whl.metadata (1.1 kB)
Collecting nvidia-cusparse-cu12==12.1.0.106 (from torch==2.1.0)
  Downloading nvidia_cusparse_cu12-12.1.0.106-py3-none-manylinux1_x86_64.whl.metadata (1.1 kB)
Collecting nvidia-nccl-cu12==2.18.1 (from torch==2.1.0)
  Downloading nvidia_nccl_cu12-2.18.1-py3-none-manylinux1_x86_64.whl.metadata (1.8 kB)
Collecting nvidia-nvtx-cu12==12.1.105 (from torch==2.1.0)
  Downloading nvidia_nvtx_cu12-12.1.105-py3-none-manylinux1_x86_64.whl.metadata (1.7 kB)
Collecting triton==2.1.0 (from torch==2.1.0)
  Downloading triton-2.1.0-0-cp311-cp311-manylinux2014_x86_64.manylinux_2_17_x86_64.whl
Collecting requests (from torchvision==0.16.0)
  Downloading requests-2.32.3-py3-none-any.whl.metadata (4.6 kB)
Collecting pillow!=8.3.*,>=5.3.0 (from torchvision==0.16.0)
  Downloading pillow-11.2.1-cp311-cp311-manylinux_2_28_x86_64.whl.metadata (8.9 kB)
Collecting nvidia-nvjitlink-cu12 (from nvidia-cusolver-cu12==11.4.5.107->torch==2.1.0)
  Downloading nvidia_nvjitlink_cu12-12.9.41-py3-none-manylinux2010_x86_64.manylinux_2_20_x86_64.whl
Collecting MarkupSafe>=2.0 (from jinja2->torch==2.1.0)
  Downloading MarkupSafe-3.0.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
Collecting charset-normalizer<4,>=2 (from requests->torchvision==0.16.0)
  Downloading charset_normalizer-3.4.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
Collecting idna<4,>=2.5 (from requests->torchvision==0.16.0)
  Downloading idna-3.10-py3-none-any.whl.metadata (10 kB)
Collecting urllib3<3,>=1.21.1 (from requests->torchvision==0.16.0)
  Downloading urllib3-2.4.0-py3-none-any.whl.metadata (6.5 kB)
Collecting certifi>=2017.4.17 (from requests->torchvision==0.16.0)
  Downloading certifi-2025.4.26-py3-none-any.whl.metadata (2.5 kB)
Collecting mpmath<1.4,>=1.1.0 (from sympy->torch==2.1.0)
  Downloading mpmath-1.3.0-py3-none-any.whl.metadata (8.6 kB)
Downloading torch-2.1.0-cp311-cp311-manylinux1_x86_64.whl (670.2 MB)
  670.2/670.2 MB 76.7 MB/s eta 0:00:00
Downloading torchvision-0.16.0-cp311-cp311-manylinux1_x86_64.whl (6.9 MB)
  6.9/6.9 MB 106.6 MB/s eta 0:00:00
Downloading numpy-1.26.4-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (
  18.3/18.3 MB 53.0 MB/s eta 0:00:00
```

```
Downloading nvidia_cublas_cu12-12.1.3.1-py3-none-manylinux1_x86_64.whl (410.6 MB)
                                                 410.6/410.6 MB 121.7 MB/s eta 0:00:00
Downloading nvidia_cuda_cupti_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (14.1 MB)
                                                 14.1/14.1 MB 203.4 MB/s eta 0:00:00
Downloading nvidia_cuda_nvrtc_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (23.7 MB)
                                                 23.7/23.7 MB 89.6 MB/s eta 0:00:00
Downloading nvidia_cuda_runtime_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (823 kB)
                                                 823.6/823.6 kB 245.1 MB/s eta 0:00:00
Downloading nvidia_cudnn_cu12-8.9.2.26-py3-none-manylinux1_x86_64.whl (731.7 MB)
                                                 731.7/731.7 MB 211.6 MB/s eta 0:00:00
Downloading nvidia_cufft_cu12-11.0.2.54-py3-none-manylinux1_x86_64.whl (121.6 MB)
                                                 121.6/121.6 MB 291.2 MB/s eta 0:00:00
Downloading nvidia_curand_cu12-10.3.2.106-py3-none-manylinux1_x86_64.whl (56.5 MB)
                                                 56.5/56.5 MB 225.9 MB/s eta 0:00:00
Downloading nvidia_cusolver_cu12-11.4.5.107-py3-none-manylinux1_x86_64.whl (124.2 MB)
                                                 124.2/124.2 MB 252.1 MB/s eta 0:00:00
Downloading nvidia_cusparse_cu12-12.1.0.106-py3-none-manylinux1_x86_64.whl (196.0 MB)
                                                 196.0/196.0 MB 250.4 MB/s eta 0:00:00
Downloading nvidia_nccl_cu12-2.18.1-py3-none-manylinux1_x86_64.whl (209.8 MB)
                                                 209.8/209.8 MB 186.0 MB/s eta 0:00:00
Downloading nvidia_nvtx_cu12-12.1.105-py3-none-manylinux1_x86_64.whl (99 kB)
                                                 99.1/99.1 kB 196.6 MB/s eta 0:00:00
Downloading triton-2.1.0-0-cp311-cp311-manylinux2014_x86_64.manylinux_2_17_x86_64.whl
                                                 89.2/89.2 MB 168.1 MB/s eta 0:00:00
Downloading pillow-11.2.1-cp311-cp311-manylinux_2_28_x86_64.whl (4.6 MB)
                                                 4.6/4.6 MB 102.6 MB/s eta 0:00:00
Downloading filelock-3.18.0-py3-none-any.whl (16 kB)
Downloading fsspec-2025.3.2-py3-none-any.whl (194 kB)
                                                 194.4/194.4 kB 257.9 MB/s eta 0:00:00
Downloading jinja2-3.1.6-py3-none-any.whl (134 kB)
                                                 134.9/134.9 kB 247.5 MB/s eta 0:00:00
Downloading networkx-3.4.2-py3-none-any.whl (1.7 MB)
                                                 1.7/1.7 MB 197.5 MB/s eta 0:00:00
Downloading requests-2.32.3-py3-none-any.whl (64 kB)
                                                 64.9/64.9 kB 234.0 MB/s eta 0:00:00
Downloading sympy-1.14.0-py3-none-any.whl (6.3 MB)
                                                 6.3/6.3 MB 161.0 MB/s eta 0:00:00
Downloading typing_extensions-4.13.2-py3-none-any.whl (45 kB)
                                                 45.8/45.8 kB 213.5 MB/s eta 0:00:00
Downloading certifi-2025.4.26-py3-none-any.whl (159 kB)
                                                 159.6/159.6 kB 374.9 MB/s eta 0:00:00
Downloading charset_normalizer-3.4.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_
                                                 147.3/147.3 kB 252.3 MB/s eta 0:00:00
Downloading idna-3.10-py3-none-any.whl (70 kB)
                                                 70.4/70.4 kB 228.9 MB/s eta 0:00:00
Downloading MarkupSafe-3.0.2-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.w
Downloading mpmath-1.3.0-py3-none-any.whl (536 kB)
                                                 536.2/536.2 kB 217.2 MB/s eta 0:00:00
Downloading urllib3-2.4.0-py3-none-any.whl (128 kB)
                                                 128.7/128.7 kB 254.9 MB/s eta 0:00:00
Downloading nvidia_nvjitlink_cu12-12.9.41-py3-none-manylinux2010_x86_64.manylinux_2_1
                                                 39.7/39.7 MB 253.7 MB/s eta 0:00:00
Installing collected packages: mpmath, urllib3, typing-extensions, sympy, pillow, nvi
Attempting uninstall: mpmath
```

```
Found existing installation: mpmath 1.3.0
Uninstalling mpmath-1.3.0:
    Successfully uninstalled mpmath-1.3.0
Attempting uninstall: urllib3
    Found existing installation: urllib3 2.4.0
    Uninstalling urllib3-2.4.0:
        Successfully uninstalled urllib3-2.4.0
Attempting uninstall: typing-extensions
    Found existing installation: typing_extensions 4.13.2
    Uninstalling typing_extensions-4.13.2:
        Successfully uninstalled typing_extensions-4.13.2
Attempting uninstall: sympy
    Found existing installation: sympy 1.13.1
    Uninstalling sympy-1.13.1:
        Successfully uninstalled sympy-1.13.1
Attempting uninstall: pillow
    Found existing installation: pillow 11.2.1
    Uninstalling pillow-11.2.1:
        Successfully uninstalled pillow-11.2.1
Attempting uninstall: nvidia-nvtx-cu12
    Found existing installation: nvidia-nvtx-cu12 12.4.127
    Uninstalling nvidia-nvtx-cu12-12.4.127:
        Successfully uninstalled nvidia-nvtx-cu12-12.4.127
Attempting uninstall: nvidia-nvjitlink-cu12
    Found existing installation: nvidia-nvjitlink-cu12 12.4.127
    Uninstalling nvidia-nvjitlink-cu12-12.4.127:
        Successfully uninstalled nvidia-nvjitlink-cu12-12.4.127
Attempting uninstall: nvidia-nccl-cu12
    Found existing installation: nvidia-nccl-cu12 2.21.5
    Uninstalling nvidia-nccl-cu12-2.21.5:
        Successfully uninstalled nvidia-nccl-cu12-2.21.5
Attempting uninstall: nvidia-curand-cu12
    Found existing installation: nvidia-curand-cu12 10.3.5.147
    Uninstalling nvidia-curand-cu12-10.3.5.147:
        Successfully uninstalled nvidia-curand-cu12-10.3.5.147
Attempting uninstall: nvidia-cufft-cu12
    Found existing installation: nvidia-cufft-cu12 11.2.1.3
    Uninstalling nvidia-cufft-cu12-11.2.1.3:
        Successfully uninstalled nvidia-cufft-cu12-11.2.1.3
Attempting uninstall: nvidia-cuda-runtime-cu12
    Found existing installation: nvidia-cuda-runtime-cu12 12.4.127
    Uninstalling nvidia-cuda-runtime-cu12-12.4.127:
        Successfully uninstalled nvidia-cuda-runtime-cu12-12.4.127
Attempting uninstall: nvidia-cuda-nvrtc-cu12
    Found existing installation: nvidia-cuda-nvrtc-cu12 12.4.127
    Uninstalling nvidia-cuda-nvrtc-cu12-12.4.127:
        Successfully uninstalled nvidia-cuda-nvrtc-cu12-12.4.127
Attempting uninstall: nvidia-cuda-cupti-cu12
    Found existing installation: nvidia-cuda-cupti-cu12 12.4.127
    Uninstalling nvidia-cuda-cupti-cu12-12.4.127:
        Successfully uninstalled nvidia-cuda-cupti-cu12-12.4.127
Attempting uninstall: nvidia-cublas-cu12
    Found existing installation: nvidia-cublas-cu12 12.4.5.8
    Uninstalling nvidia-cublas-cu12-12.4.5.8:
        Successfully uninstalled nvidia-cublas-cu12-12.4.5.8
```

```
S U C C E S S F U L L Y  U N I N S T A L L E D  n i v i d i a - c u s p a r s e - c u 1 2 - 1 2 . 3 . 1 . 1 7 0
```

Attempting uninstall: numpy
Found existing installation: numpy 2.0.2
Uninstalling numpy-2.0.2:
Successfully uninstalled numpy-2.0.2

Attempting uninstall: networkx
Found existing installation: networkx 3.4.2
Uninstalling networkx-3.4.2:
Successfully uninstalled networkx-3.4.2

Attempting uninstall: MarkupSafe
Found existing installation: MarkupSafe 3.0.2
Uninstalling MarkupSafe-3.0.2:
Successfully uninstalled MarkupSafe-3.0.2

Attempting uninstall: idna
Found existing installation: idna 3.10
Uninstalling idna-3.10:
Successfully uninstalled idna-3.10

Attempting uninstall: fsspec
Found existing installation: fsspec 2025.3.2
Uninstalling fsspec-2025.3.2:
Successfully uninstalled fsspec-2025.3.2

Attempting uninstall: filelock
Found existing installation: filelock 3.18.0
Uninstalling filelock-3.18.0:
Successfully uninstalled filelock-3.18.0

Attempting uninstall: charset-normalizer
Found existing installation: charset-normalizer 3.4.1
Uninstalling charset-normalizer-3.4.1:
Successfully uninstalled charset-normalizer-3.4.1

Attempting uninstall: certifi
Found existing installation: certifi 2025.4.26
Uninstalling certifi-2025.4.26:
Successfully uninstalled certifi-2025.4.26

Attempting uninstall: triton
Found existing installation: triton 3.2.0
Uninstalling triton-3.2.0:
Successfully uninstalled triton-3.2.0

Attempting uninstall: requests
Found existing installation: requests 2.32.3
Uninstalling requests-2.32.3:
Successfully uninstalled requests-2.32.3

Attempting uninstall: nvidia-cusparse-cu12
Found existing installation: nvidia-cusparse-cu12 12.3.1.170
Uninstalling nvidia-cusparse-cu12-12.3.1.170:
Successfully uninstalled nvidia-cusparse-cu12-12.3.1.170

Attempting uninstall: nvidia-cudnn-cu12
Found existing installation: nvidia-cudnn-cu12 9.1.0.70
Uninstalling nvidia-cudnn-cu12-9.1.0.70:
Successfully uninstalled nvidia-cudnn-cu12-9.1.0.70

Attempting uninstall: jinja2
Found existing installation: Jinja2 3.1.6
Uninstalling Jinja2-3.1.6:
Successfully uninstalled Jinja2-3.1.6

Attempting uninstall: nvidia-cusolver-cu12
Found existing installation: nvidia-cusolver-cu12 11.6.1.9

```
Uninstalling nvidia-cusolver-cu12-11.6.1.9:
  Successfully uninstalled nvidia-cusolver-cu12-11.6.1.9
Attempting uninstall: torch
  Found existing installation: torch 2.6.0+cu124
Uninstalling torch-2.6.0+cu124:
  Successfully uninstalled torch-2.6.0+cu124
Attempting uninstall: torchvision
  Found existing installation: torchvision 0.21.0+cu124
Uninstalling torchvision-0.21.0+cu124:
  Successfully uninstalled torchvision-0.21.0+cu124
ERROR: pip's dependency resolver does not currently take into account all the package
torchaudio 2.6.0+cu124 requires torch==2.6.0, but you have torch 2.1.0 which is incom
thinc 8.3.6 requires numpy<3.0.0,>=2.0.0, but you have numpy 1.26.4 which is incompat
Successfully installed MarkupSafe-3.0.2 certifi-2025.4.26 charset-normalizer-3.4.2 fi
WARNING: The following packages were previously imported in this runtime:
  [PIL,certifi,charset_normalizer,markupsafe,requests,torch,torchgen]
You must restart the runtime in order to use newly installed versions.
```

RESTART SESSION

```
# _____
# Block 0: Imports & Model Loading (yiyanghkust/finbert-tone)
# _____
import spacy
import torch
from transformers import AutoTokenizer, AutoModel

# 1) spaCy for sentence splitting
nlp = spacy.load("en_core_web_sm")

# 2) FinBERT-Tone for financial-domain embeddings
TOKENIZER = AutoTokenizer.from_pretrained("yiyanghkust/finbert-tone")
FINBERT    = AutoModel.from_pretrained("yiyanghkust/finbert-tone")

# 3) Device setup
DEVICE = torch.device("cuda" if torch.cuda.is_available() else "cpu")
FINBERT.to(DEVICE).eval()
```

```
pytorch_model.bin: 100%                                         439M/439M [00:01<00:00, 256MB/
                                                               s]

/usr/local/lib/python3.11/dist-packages/torch/_utils.py:831: UserWarning: TypedStorag
  return self.fget.__get__(instance, owner)()
BertModel(
  (embeddings): BertEmbeddings(
    (word_embeddings): Embedding(30873, 768, padding_idx=0)
    (position_embeddings): Embedding(512, 768)
```



```
# _____
# curr_count + n > max_tokens.
# chunks.append(" ".join(curr))
# build overlap window
overlap_sents, count = [], 0
for prev in reversed(curr):
    ptoks = TOKENIZER.tokenize(prev)
    if count + len(ptoks) > overlap:
        break
    overlap_sents.insert(0, prev)
    count += len(ptoks)
curr, curr_count = overlap_sents, count

curr.append(sent)
curr_count += n

if curr:
    chunks.append(" ".join(curr))
return chunks

# _____
# Block 2: Extract unique sentences from all chunks
# _____
def sentences_from_chunks(chunks):
    seen = set()
    sentences = []
    for chunk in chunks:
        for sent in nlp(chunk).sents:
            t = sent.text.strip()
            if t not in seen:
                seen.add(t)
                sentences.append(t)
    return sentences

# _____
# Block 3: Get [CLS] embeddings via FinBERT
# _____
@torch.no_grad()
def embed_sentences(sent_list):
    embs = []
    for s in sent_list:
        inputs = TOKENIZER(
            s,
            return_tensors="pt",
            truncation=True,
            padding="max_length",
            max_length=512
        ).to(DEVICE)
        out = FTNRFRT(**inputs)
```

```
    cls = out.last_hidden_state[:, 0, :].cpu().numpy()
    embs.append(cls.squeeze(0))
    return np.vstack(embs)

# -----
# Block 4: MMR-based selection up to ~1500 tokens
# -----
from sklearn.metrics.pairwise import cosine_similarity

def mmr_summary(sents, embs, doc_emb, max_tokens=1500, lam=0.6):
    selected, sel_embs, used = [], [], set()
    cur_toks = 0
    doc_sim = cosine_similarity(embs, doc_emb.reshape(1, -1)).flatten()

    while used != set(range(len(sents))) and cur_toks < max_tokens:
        scores = {}
        for idx in set(range(len(sents))) - used:
            rel = doc_sim[idx]
            div = max(cosine_similarity(
                embs[idx:idx+1],
                np.vstack(sel_embs) if sel_embs else embs[idx:idx+1]
            ).flatten()) if sel_embs else 0.0
            scores[idx] = lam * rel - (1 - lam) * div

        best = max(scores, key=scores.get)
        tok_len = len(TOKENIZER.tokenize(sents[best]))
        if cur_toks + tok_len > max_tokens:
            break

        selected.append(sents[best])
        sel_embs.append(embs[best])
        used.add(best)
        cur_toks += tok_len

    return " ".join(selected)
```

```
# # - Force restart-safe installation to fix NumPy incompatibility
# !pip install --force-reinstall --no-cache-dir numpy==1.26.4
```

```
import numpy as np
print(np.__version__)
np.array([1, 2, 3])
```

```
1.26.4
array([1, 2, 3])
```

```
# _____
# Block 5: Full pipeline and store in df["summary"]
# _____
import numpy as np

def summarize(text):
    chunks = chunk_text(text)
    sents = sentences_from_chunks(chunks)
    embs = embed_sentences(sents)
    doc_emb = embs.mean(axis=0)
    return mmr_summary(sents, embs, doc_emb)

# apply to your DataFrame
df["summary"] = df["cleaned_text"].apply(summarize)

df.sample(2)
```

		URL	Date	Year	Month	Day	Content	date
46	www.federalreserve.gov/monetarypolicy/...	https://	2017-06-14	2017	2017-06-01	14	HomeMonetary PolicyFederal Open Market Committ...	2017-06-30
14	www.federalreserve.gov/monetarypolicy/...	https://	2023-06-14	2023	2023-06-01	14	HomeMonetary PolicyFederal Open Market Committ...	2023-06-30

2 rows × 21 columns

```
# Assuming df is your DataFrame
df.to_csv('/content/drive/My Drive/NLP/Assignment_3/fomc_summary.csv', index=False)
```

```
print(df.columns)
```

```
Index(['URL', 'Date', 'Year', 'Month', 'Day', 'Content', 'date', 'GDP', 'CPI',
       'Unemployment', 'FedFundsRate', 'SNR_Ratio', 'StopWord_Ratio',
       'Redundancy_Ratio', 'SpecialChar_Density', 'Semantic_Density',
       'Digit_Ratio', 'Noise_Score', 'Quality_Score', 'cleaned_text',
       'summary'],
      dtype='object')
```

```
df = df.reset_index(drop=True) # Optional: ensure clean indexing
```

```
df['doc_id'] = 'doc' + (df.index + 1).astype(str)
print(df[['doc_id', 'Date']].head())
```

```
doc_id      Date
0  doc1 2025-03-19
1  doc2 2025-01-29
2  doc3 2024-12-18
3  doc4 2024-11-07
4  doc5 2024-09-18
```

▼ Evaluate Summary Quality & Information Retention

BERTScore (semantic similarity)

```
# !pip install bert-score

from bert_score import score

# Sample a few rows for demonstration (optional)
sample_df = df.sample(5, random_state=42)

# Compute BERTScore (reference: full text, candidate: summary)
P, R, F1 = score(cands=sample_df['summary'].tolist(),
                  refs=sample_df['cleaned_text'].tolist(),
                  lang="en", model_type="bert-base-uncased")

# Attach to DataFrame
sample_df['BERT_Precision'] = P.tolist()
sample_df['BERT_Recall'] = R.tolist()
sample_df['BERT_F1'] = F1.tolist()

# Reset index of sample_df to make 'doc_id' a column
sample_df = sample_df.reset_index()

# View scores
sample_df[['doc_id', 'BERT_Precision', 'BERT_Recall', 'BERT_F1']]
```

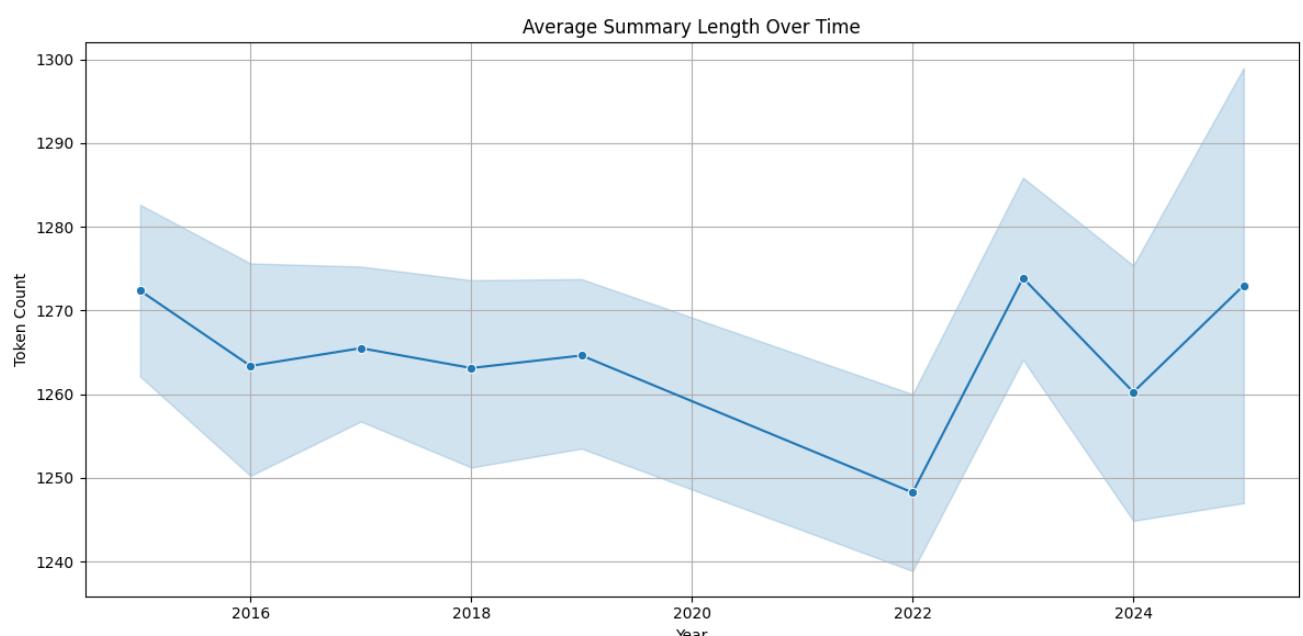
	doc_id	BERT_Precision	BERT_Recall	BERT_F1	
0	doc55	0.431291	0.412553	0.421714	
1	doc63	0.447369	0.431908	0.439503	
2	doc1	0.583355	0.584023	0.583689	
3	doc46	0.431035	0.415864	0.423314	
4	doc6	0.635131	0.621832	0.628411	

✓ Compare Summaries Over Time

```
import seaborn as sns
import matplotlib.pyplot as plt

df['Year'] = pd.to_datetime(df['Date']).dt.year
df['summary_length'] = df['summary'].str.split().str.len()

plt.figure(figsize=(12, 6))
sns.lineplot(data=df, x='Year', y='summary_length', marker='o')
plt.title("Average Summary Length Over Time")
plt.ylabel("Token Count")
plt.grid(True)
plt.tight_layout()
plt.show()
```



Word Cloud of Summaries by Year

```
from wordcloud import WordCloud

years = sorted(df['Year'].dropna().unique())
ncols = 3
nrows = (len(years) + ncols - 1) // ncols

plt.figure(figsize=(20, 5 * nrows))

for i, year in enumerate(years):
    year_text = ' '.join(df[df['Year'] == year]['summary'].astype(str))
    wordcloud = WordCloud(width=800, height=400, background_color='white').generate(year_text)

    plt.subplot(nrows, ncols, i + 1)
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis('off')
    plt.title(f'WordCloud - Summary - {year}')

plt.tight_layout()
plt.show()
```

