In [1]:
```
pip install nltk matplotlib seaborn wordcloud
```

```
Defaulting to user installation because normal site-packages is not writeable
Collecting nltk
  Using cached nltk-3.9.1-py3-none-any.whl.metadata (2.9 kB)
Requirement already satisfied: matplotlib in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (3.9.4)
Collecting seaborn
  Using cached seaborn-0.13.2-py3-none-any.whl.metadata (5.4 kB)
Collecting wordcloud
  Downloading wordcloud-1.9.4-cp39-cp39-macosx_11_0_arm64.whl.metadata (3.4 kB)
Collecting click (from nltk)
  Using cached click-8.1.8-py3-none-any.whl.metadata (2.3 kB)
Requirement already satisfied: joblib in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from nltk) (1.4.2)
Collecting regex>=2021.8.3 (from nltk)
  Downloading regex-2024.11.6-cp39-cp39-macosx_11_0_arm64.whl.metadata (40 kB)
Requirement already satisfied: tqdm in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from nltk) (4.67.1)
Requirement already satisfied: contourpy>=1.0.1 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from matplotlib) (1.3.0)
Requirement already satisfied: cycler>=0.10 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from matplotlib) (4.55.3)
Requirement already satisfied: kiwisolver>=1.3.1 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from matplotlib) (1.4.7)
Requirement already satisfied: numpy>=1.23 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from matplotlib) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from matplotlib) (24.2)
Requirement already satisfied: pillow>=8 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from matplotlib) (11.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from matplotlib) (3.2.1)
Requirement already satisfied: python-dateutil>=2.7 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from matplotlib) (2.9.0.post0)
Requirement already satisfied: importlib-resources>=3.2.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from matplotlib) (6.5.2)
Requirement already satisfied: pandas>=1.2 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from seaborn) (2.2.3)
Requirement already satisfied: zipp>=3.1.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from importlib-resources>=3.2.0->matplotlib) (3.21.0)
Requirement already satisfied: pytz>=2020.1 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from pandas>=1.2->seaborn) (2024.2)
Requirement already satisfied: tzdata>=2022.7 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from pandas>=1.2->seaborn) (2024.2)
Requirement already satisfied: six>=1.5 in /Library/Developer/CommandLineTools/Library/Frameworks/Python3.framework/Versions/3.9/lib/python3.9/site-packages (from python-dateutil>=2.7->matplotlib) (1.15.0)
Using cached nltk-3.9.1-py3-none-any.whl (1.5 MB)
Using cached seaborn-0.13.2-py3-none-any.whl (294 kB)
Downloading wordcloud-1.9.4-cp39-cp39-macosx_11_0_arm64.whl (168 kB)
Downloading regex-2024.11.6-cp39-cp39-macosx_11_0_arm64.whl (284 kB)
Using cached click-8.1.8-py3-none-any.whl (98 kB)
Installing collected packages: regex, click, nltk, wordcloud, seaborn
Successfully installed click-8.1.8 nltk-3.9.1 regex-2024.11.6 seaborn-0.13.2 wordcloud-1.9.4
Note: you may need to restart the kernel to use updated packages.
```

In [2]:
```
import os
import nltk
from collections import Counter
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud


%matplotlib inline


nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/mmadhusudan/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[2]:
```
True
```

In [3]:
```
def process_file(file_path):
    """
    Reads a text file, removes Gutenberg boilerplate (header and footer), tokenizes the text,
    converts tokens to lowercase, and filters out non-alphabetic tokens.

    Returns:
        tokens (list): A list of cleaned, alphabetic tokens.
    """

    with open(file_path, "r", encoding="utf-8") as f:
        raw_text = f.read()


    start_marker = "*** START OF THIS PROJECT GUTENBERG EBOOK"
    end_marker   = "*** END OF THIS PROJECT GUTENBERG EBOOK"


    start_idx = raw_text.find(start_marker)
    if start_idx != -1:

        text = raw_text[start_idx + len(start_marker):]
    else:
        text = raw_text  # If no marker found, use entire text


    end_idx = text.find(end_marker)
    if end_idx != -1:
        text = text[:end_idx]


    text = text.strip()


    tokens = nltk.word_tokenize(text.lower())

    tokens = [token for token in tokens if token.isalpha()]

    return tokens


test_file = os.path.join("Gutenberg_Books", "1.txt.txt")
print("Test file tokens (first 20):", process_file(test_file)[:20])
```

```
Test file tokens (first 20): ['start', 'of', 'the', 'project', 'gutenberg', 'ebook', 'note', 'this', 'file', 'combines', 'the', 'first', 'two', 'project', 'gutenberg', 'files', 'both', 'of', 'which', 'were']
```

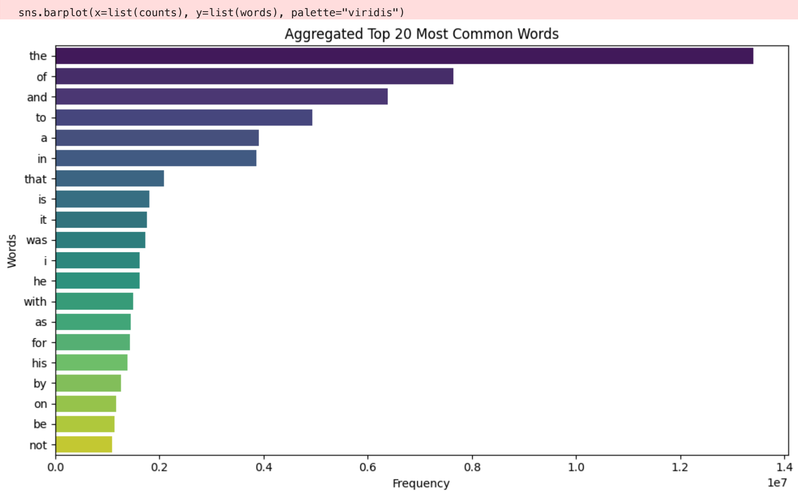
In [4]:
```
folder = "Gutenberg_Books"


all_files = [f for f in os.listdir(folder) if f.endswith(".txt.txt")]
print(f"Found {len(all_files)} files.")


aggregated_counter = Counter()
file_stats = []  # This will hold stats for each file


for filename in all_files:
    file_path = os.path.join(folder, filename)
    tokens = process_file(file_path)

    aggregated_counter.update(tokens)

    file_stats.append({
        "filename": filename,
        "num_tokens": len(tokens),
        "unique_tokens": len(set(tokens))
    })


df_stats = pd.DataFrame(file_stats)
print("Per-file statistics (first 5 rows):")
display(df_stats.head())
```

```
Found 2475 files.
Per-file statistics (first 5 rows):
```

	filename	num_tokens	unique_tokens
0	4658.txt.txt	161293	15819
1	37009.txt.txt	77551	4991
2	14609.txt.txt	89507	11561
3	5342.txt.txt	87301	6824
4	17.txt.txt	268340	5539

In [5]:
```
most_common_all = aggregated_counter.most_common(20)
print("Aggregated Top 20 words:")
print(most_common_all)


words, counts = zip(*most_common_all)


plt.figure(figsize=(10, 6))
sns.barplot(x=list(counts), y=list(words), palette="viridis")
plt.title("Aggregated Top 20 Most Common Words")
plt.xlabel("Frequency")
plt.ylabel("Words")
plt.tight_layout()
plt.show()
```

```
Aggregated Top 20 words:
[('the', 13395724), ('of', 7642031), ('and', 6378760), ('to', 4934479), ('a', 3908621), ('in', 3857791), ('that', 2092342), ('is', 1800929), ('it', 1767951), ('was', 1723232), ('i', 1623000), ('he', 1620363), ('with', 1496503), ('as', 1448598), ('for', 1433725), ('his', 1393159), ('by', 1266249), ('on', 1171581), ('be', 1141795), ('not', 1094746)]
/var/folders/7j/rv3w77nj6kb6kw_ssltcqpkr0000gp/T/ipykernel_22400/2636935665.py:11: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=list(counts), y=list(words), palette="viridis")
```

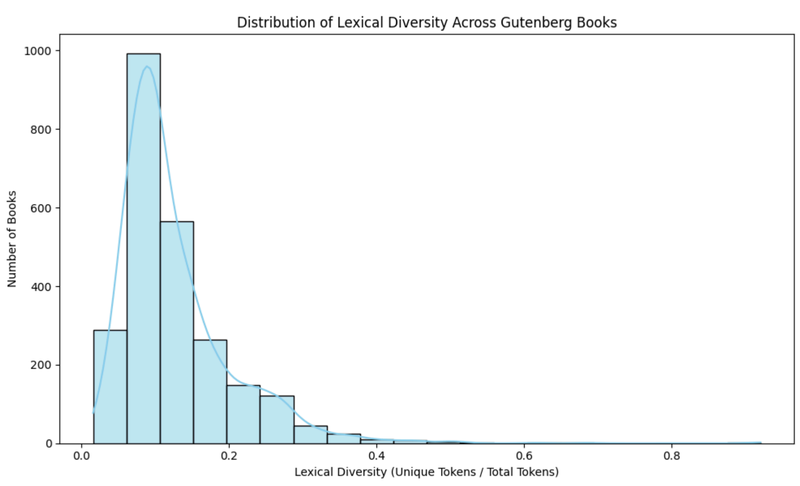


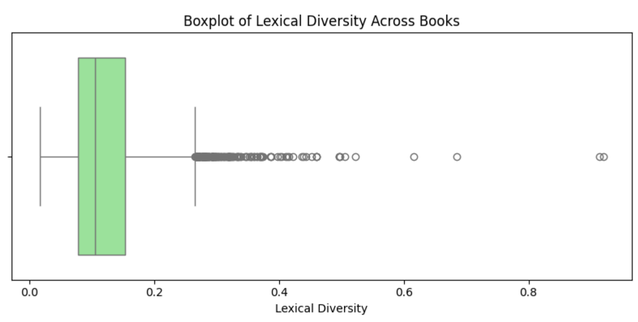
In [6]:
```
wordcloud = WordCloud(width=800, height=400, background_color='white').generate_from_frequencies(aggregated_counter)


plt.figure(figsize=(15, 8))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.title("Aggregated Word Cloud from All Gutenberg Books")
plt.show()
```

In [7]:
```
df_stats['lexical_diversity'] = df_stats['unique_tokens'] / df_stats['num_tokens']


display(df_stats.head())
```

```
plt.figure(figsize=(10, 6))
sns.histplot(df_stats['lexical_diversity'], kde=True, bins=20, color='skyblue')
plt.title("Distribution of Lexical Diversity Across Gutenberg Books")
plt.xlabel("Lexical Diversity (Unique Tokens / Total Tokens)")
plt.ylabel("Number of Books")
plt.tight_layout()
plt.show()

plt.figure(figsize=(8, 4))
sns.boxplot(x=df_stats['lexical_diversity'], color='lightgreen')
plt.title("Boxplot of Lexical Diversity Across Books")
plt.xlabel("Lexical Diversity")
plt.tight_layout()
plt.show()
```

	filename	num_tokens	unique_tokens	lexical_diversity
0	4658.txt.txt	161293	15819	0.098076
1	37009.txt.txt	77551	4991	0.064358
2	14609.txt.txt	89507	11561	0.129163
3	5342.txt.txt	87301	6824	0.078166
4	17.txt.txt	268340	5539	0.020642





In [8]:
```
all_tokens = []
for filename in all_files:
    file_path = os.path.join(folder, filename)
    tokens = process_file(file_path)
    all_tokens.extend(tokens)

print("Total tokens collected from all files:", len(all_tokens))
```

```
Total tokens collected from all files: 209085770
```

In [9]:
```
import nltk
nltk.download('averaged_perceptron_tagger')
```

```
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /Users/mmadhusudan/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

Out[9]: True

In [10]:
```
print(nltk.data.path)
nltk.download('punkt')
```

```
['/Users/mmadhusudan/nltk_data', '/Library/Developer/CommandLineTools/Library/Frameworks/Python3.framework/Versions/3.9/nltk_data', '/Library/Developer/CommandLineTools/Library/Frameworks/Python3.framework/Versions/3.9/share/nltk_data', '/Library/Developer/CommandLineTools/Library/Frameworks/Python3.framework/Versions/3.9/lib/nltk_data', '/usr/share/nltk_data', '/usr/local/share/nltk_data', '/usr/lib/nltk_data', '/usr/local/lib/nltk_data']
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/mmadhusudan/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[10]: True

In [11]:
```
import nltk
nltk.download('averaged_perceptron_tagger')
```

```
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /Users/mmadhusudan/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

Out[11]: True

In [12]:
```
print(nltk.data.find('taggers/averaged_perceptron_tagger'))
```

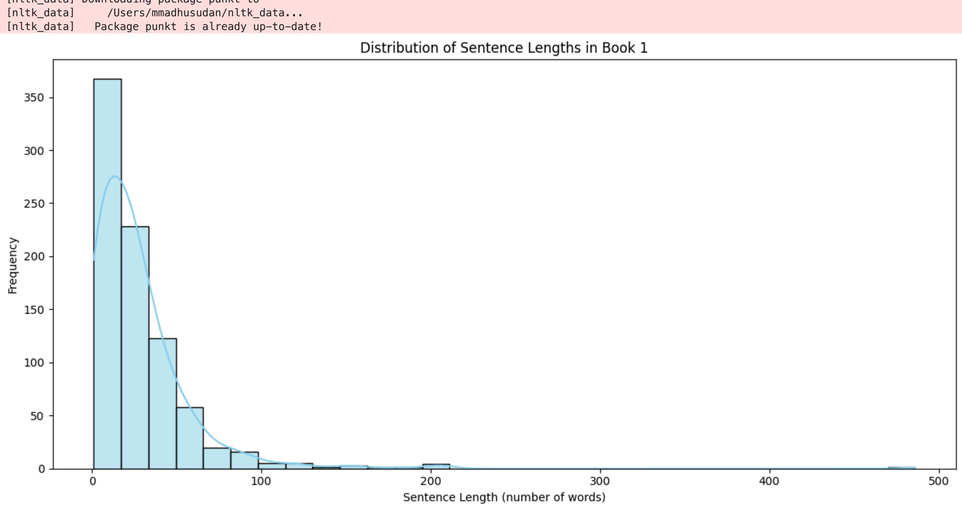
```
/Users/mmadhusudan/nltk_data/taggers/averaged_perceptron_tagger
```

In [13]:
```
import random

sample_size = min(5000, len(all_tokens))
sample_tokens = random.sample(all_tokens, sample_size)
```

In [14]:
```
file_path = os.path.join("Gutenberg_Books", "1.txt.txt")
with open(file_path, "r", encoding="utf-8") as f:
    raw_text = f.read()


start_marker = "*** START OF THIS PROJECT GUTENBERG EBOOK"
end_marker   = "*** END OF THIS PROJECT GUTENBERG EBOOK"
start_idx = raw_text.find(start_marker)
if start_idx != -1:
    text = raw_text[start_idx + len(start_marker):]
else:
    text = raw_text

end_idx = text.find(end_marker)
if end_idx != -1:
    cleaned_text = text[:end_idx]
else:
    cleaned_text = text

cleaned_text = cleaned_text.strip()
```

In [15]:
```
# --- Sentence-Level Analysis ---


nltk.download('punkt')


sentences = nltk.sent_tokenize(cleaned_text)


sentence_lengths = [len(nltk.word_tokenize(sentence)) for sentence in sentences]


print("Number of sentences:", len(sentences))
print("Average sentence length (words):", sum(sentence_lengths) / len(sentence_lengths))


plt.figure(figsize=(12, 6))
sns.histplot(sentence_lengths, bins=30, kde=True, color='skyblue')
plt.title("Distribution of Sentence Lengths in Book 1")
plt.xlabel("Sentence Length (number of words)")
plt.ylabel("Frequency")
plt.tight_layout()
plt.show()
```

```
Number of sentences: 833
Average sentence length (words): 27.613445378151262
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/mmadhusudan/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```



In [24]:
```
!pip install spacy
!python -m spacy download en_core_web_sm
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: spacy in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (3.8.3)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (1.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (1.0.12)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (2.0.11)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (3.0.9)
Requirement already satisfied: thinc<8.4.0,>=8.3.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (8.3.4)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (1.1.3)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (2.5.1)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (2.0.10)
Requirement already satisfied: weasel<0.5.0,>=0.1.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (0.4.1)
Requirement already satisfied: typer<1.0.0,>=0.3.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (0.15.1)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (4.67.1)
Requirement already satisfied: numpy>=1.19.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (2.0.2)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (2.32.3)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (2.10.6)
Requirement already satisfied: jinja2 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (3.1.5)
Requirement already satisfied: setuptools in /Library/Developer/CommandLineTools/Library/Frameworks/Python3.framework/Versions/3.9/lib/python3.9/site-packages (from spacy) (58.0.4)
Requirement already satisfied: packaging>=20.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (24.2)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from spacy) (3.5.0)
Requirement already satisfied: language-data>=1.2 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from langcodes<4.0.0,>=3.2.0->spacy) (1.3.0)
Requirement already satisfied: annotated-types>=0.6.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy) (0.7.0)
Requirement already satisfied: pydantic-core==2.27.2 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy) (2.27.2)
Requirement already satisfied: typing-extensions>=4.12.2 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy) (4.12.2)
Requirement already satisfied: charset-normalizer<4,>=2 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from requests<3.0.0,>=2.13.0->spacy) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from requests<3.0.0,>=2.13.0->spacy) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from requests<3.0.0,>=2.13.0->spacy) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from requests<3.0.0,>=2.13.0->spacy) (2024.12.14)
Requirement already satisfied: blis<1.3.0,>=1.2.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from thinc<8.4.0,>=8.3.0->spacy) (1.2.0)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from thinc<8.4.0,>=8.3.0->spacy) (0.1.5)
Requirement already satisfied: click>=8.0.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from typer<1.0.0,>=0.3.0->spacy) (8.1.8)
Requirement already satisfied: shellingham>=1.3.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from typer<1.0.0,>=0.3.0->spacy) (1.5.4)
Requirement already satisfied: rich>=10.11.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from typer<1.0.0,>=0.3.0->spacy) (13.9.4)
Requirement already satisfied: cloudpathlib<1.0.0,>=0.7.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from weasel<0.5.0,>=0.1.0->spacy) (0.20.0)
Requirement already satisfied: smart-open<8.0.0,>=5.2.1 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from weasel<0.5.0,>=0.1.0->spacy) (7.1.0)
Requirement already satisfied: MarkupSafe>=2.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from jinja2->spacy) (3.0.2)
Requirement already satisfied: marisa-trie>=1.1.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from language-data>=1.2->langcodes<4.0.0,>=3.2.0->spacy) (1.2.1)
Requirement already satisfied: markdown-it-py>=2.2.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy) (2.19.1)
Requirement already satisfied: wrapt in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from smart-open<8.0.0,>=5.2.1->weasel<0.5.0,>=0.1.0->spacy) (1.17.2)
Requirement already satisfied: mdurl~=0.1 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from markdown-it-py>=2.2.0->rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy) (0.1.2)
zsh:1: command not found: python
```

In [25]:
```
!python3 -m spacy download en_core_web_sm
```

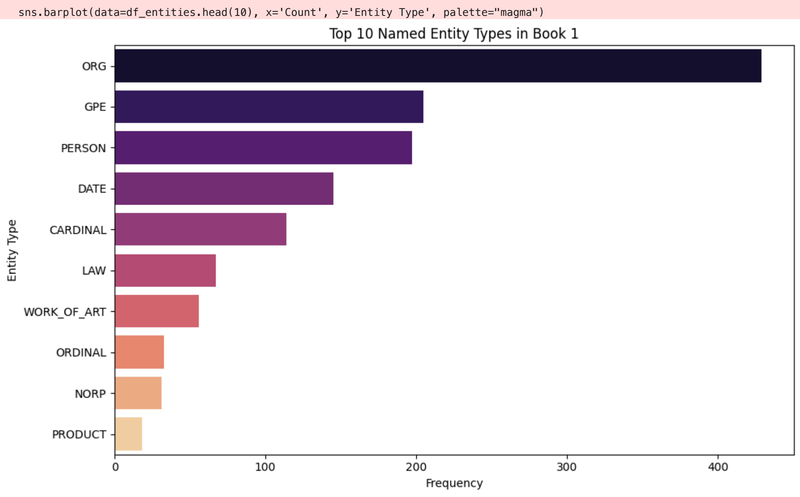
```
/Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages/urllib3/__init__.py:35: NotOpenSSLWarning: urllib3 v2 only supports OpenSSL 1.1.1+, currently the 'ssl' module is compiled with 'LibreSSL 2.8.3'. See: https://github.com/urllib3/urllib3/issues/3020
  warnings.warn(
Defaulting to user installation because normal site-packages is not writeable
Collecting en-core-web-sm==3.8.0
  Using cached https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.8.0/en_core_web_sm-3.8.0-py3-none-any.whl (12.8 MB)
✔ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
```

In [26]:
```
import spacy
from collections import Counter


nlp = spacy.load("en_core_web_sm")


doc = nlp(cleaned_text)


entities = [(ent.text, ent.label_) for ent in doc.ents]


entity_counts = Counter([ent.label_ for ent in doc.ents])


df_entities = pd.DataFrame(entity_counts.items(), columns=['Entity Type', 'Count']).sort_values(by='Count', ascending=False)

print("Named Entity counts:")
display(df_entities)


plt.figure(figsize=(10, 6))
sns.barplot(data=df_entities.head(10), x='Count', y='Entity Type', palette="magma")
plt.title("Top 10 Named Entity Types in Book 1")
plt.xlabel("Frequency")
plt.ylabel("Entity Type")
plt.tight_layout()
plt.show()
```

```
Named Entity counts:
```

	Entity Type	Count
4	ORG	429
5	GPE	205
3	PERSON	197
2	DATE	145
1	CARDINAL	114
6	LAW	67
8	WORK_OF_ART	56
0	ORDINAL	33
7	NORP	31
14	PRODUCT	18
15	LOC	16
9	MONEY	14
13	EVENT	8
12	FAC	7
16	TIME	6
11	PERCENT	5
10	QUANTITY	4
17	LANGUAGE	1

```
/var/folders/7j/rv3w77nj6kb6kw_ssltcqpkr0000gp/T/ipykernel_22400/556953189.py:24: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(data=df_entities.head(10), x='Count', y='Entity Type', palette="magma")
```

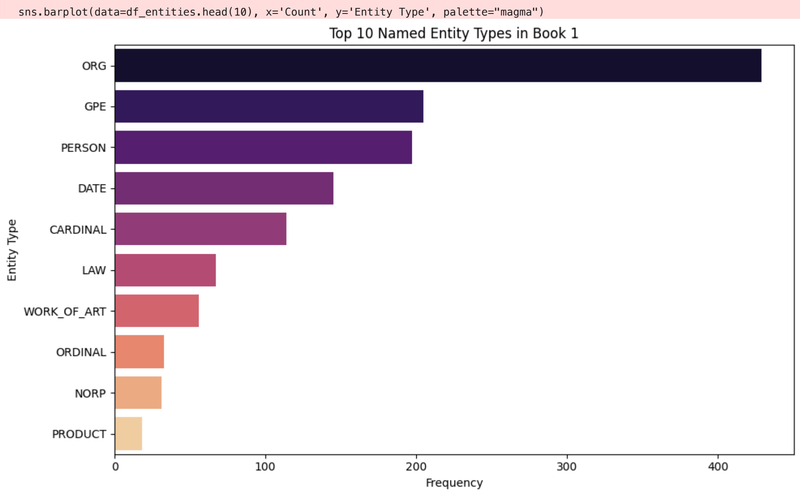


In [27]:

```
Named Entity counts:
```

	Entity Type	Count
4	ORG	429
5	GPE	205
3	PERSON	197
2	DATE	145
1	CARDINAL	114
6	LAW	67
8	WORK_OF_ART	56
0	ORDINAL	33
7	NORP	31
14	PRODUCT	18
15	LOC	16
9	MONEY	14
13	EVENT	8
12	FAC	7
16	TIME	6
11	PERCENT	5
10	QUANTITY	4
17	LANGUAGE	1

```
/var/folders/7j/rv3w77nj6kb6kw_ssltcqpkr0000gp/T/ipykernel_22400/556953189.py:24: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(data=df_entities.head(10), x='Count', y='Entity Type', palette="magma")
```



In [28]:
```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation


def get_cleaned_text(file_path):
    with open(file_path, "r", encoding="utf-8") as f:
        raw_text = f.read()

    start_marker = "*** START OF THIS PROJECT GUTENBERG EBOOK"
    end_marker   = "*** END OF THIS PROJECT GUTENBERG EBOOK"

    start_idx = raw_text.find(start_marker)
    if start_idx != -1:
        text = raw_text[start_idx + len(start_marker):]
    else:
        text = raw_text

    end_idx = text.find(end_marker)
    if end_idx != -1:
        text = text[:end_idx]

    # Remove extra whitespace and return
    return text.strip()
```

```
corpus = []
doc_names = []  # Keep track of file names (optional)
for filename in all_files:
    file_path = os.path.join(folder, filename)
    doc = get_cleaned_text(file_path)
    corpus.append(doc)
    doc_names.append(filename)

print(f"Collected {len(corpus)} documents.")


vectorizer = CountVectorizer(stop_words='english', max_df=0.95, min_df=2)
dtm = vectorizer.fit_transform(corpus)
print("DTM shape:", dtm.shape)


n_topics = 5
lda_model = LatentDirichletAllocation(n_components=n_topics, random_state=42)
lda_model.fit(dtm)


def print_top_words(model, feature_names, n_top_words):
    for topic_idx, topic in enumerate(model.components_):
        top_words = [feature_names[i] for i in topic.argsort()[:-n_top_words - 1:-1]]
        print(f"Topic #{topic_idx}: {' '.join(top_words)}")

n_top_words = 10
feature_names = vectorizer.get_feature_names_out()
print("\nTop words per topic:")
print_top_words(lda_model, feature_names, n_top_words)
```

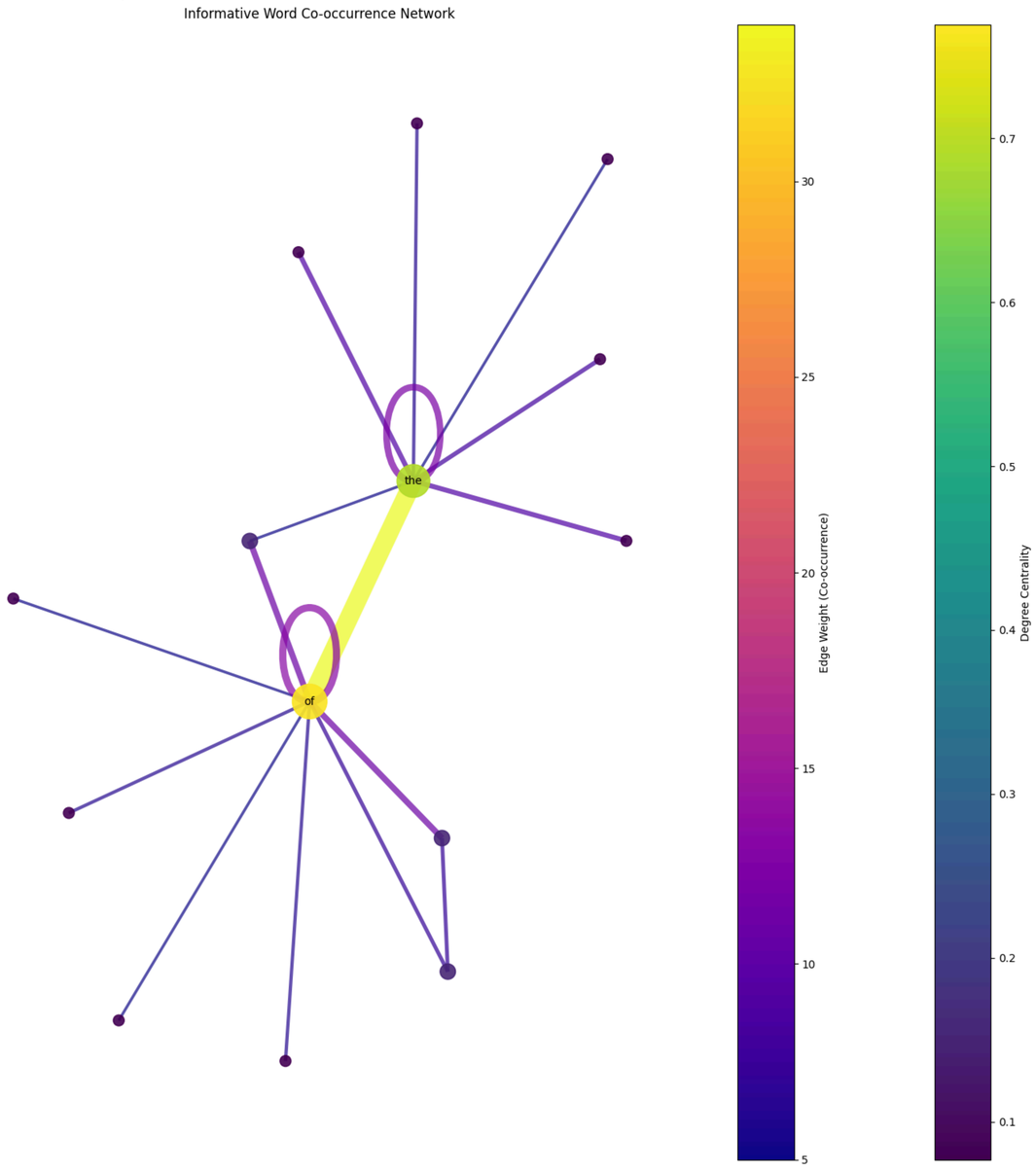
```
Collected 2475 documents.
DTM shape: (2475, 526829)

Top words per topic:
Topic #0: water small time great form species large long work used
Topic #1: said man time great men did like day good little
Topic #2: la et le les il que en des qui est
Topic #3: die der en que la und el den se los
Topic #4: est 000 km years na total male female population rate
```

In [30]:
```
!pip install networkx
```

```
Defaulting to user installation because normal site-packages is not writeable
Collecting networkx
  Downloading networkx-3.2.1-py3-none-any.whl.metadata (5.2 kB)
Downloading networkx-3.2.1-py3-none-any.whl (1.6 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.6/1.6 MB 5.2 MB/s eta 0:00:00a 0:00:01
Installing collected packages: networkx
Successfully installed networkx-3.2.1
```

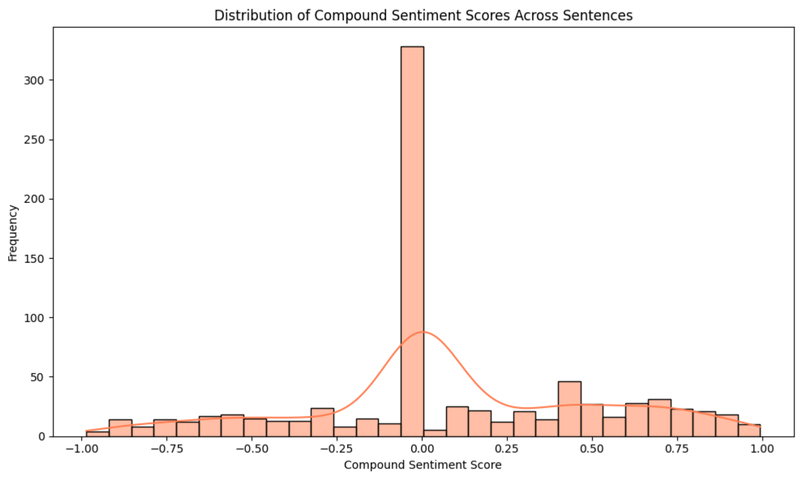
In [41]:
```
import networkx as nx
import matplotlib.pyplot as plt
import matplotlib.cm as cm
import numpy as np


window_size = 5  # Sliding window size
co_occurrence = {}


for i, token in enumerate(tokens):
    for j in range(i+1, min(i+window_size, len(tokens))):
        pair = tuple(sorted([token, tokens[j]]))
        co_occurrence[pair] = co_occurrence.get(pair, 0) + 1


G = nx.Graph()
threshold = 5
for pair, weight in co_occurrence.items():
    if weight >= threshold:
        G.add_edge(pair[0], pair[1], weight=weight)

print(f"Graph has {G.number_of_nodes()} nodes and {G.number_of_edges()} edges.")


deg_centrality = nx.degree_centrality(G)
degrees = dict(G.degree())
node_color = [deg_centrality[node] for node in G.nodes()]
node_size = [degrees[node] * 100 for node in G.nodes()]
edge_weights = [G[u][v]['weight'] for u, v in G.edges()]
edge_width = [w / 2 for w in edge_weights]


fig, ax = plt.subplots(figsize=(15, 15))
pos = nx.spring_layout(G, k=0.15, seed=42)


nodes = nx.draw_networkx_nodes(
    G, pos, ax=ax,
    node_size=node_size,
    node_color=node_color,
    cmap=cm.viridis,
    alpha=0.9
)


sm = plt.cm.ScalarMappable(cmap=cm.viridis, norm=plt.Normalize(vmin=min(node_color), vmax=max(node_color)))
sm.set_array(np.array(node_color))
fig.colorbar(sm, ax=ax, label="Degree Centrality")


edges = nx.draw_networkx_edges(
    G, pos, ax=ax,
    width=edge_width,
    edge_color=edge_weights,
    edge_cmap=cm.plasma,
    alpha=0.7
)


ecolor = np.array(edge_weights)
sm2 = plt.cm.ScalarMappable(cmap=cm.plasma, norm=plt.Normalize(vmin=min(ecolor), vmax=max(ecolor)))
sm2.set_array(ecolor)
fig.colorbar(sm2, ax=ax, label="Edge Weight (Co-occurrence)")


centrality_values = np.array(list(deg_centrality.values()))
threshold_label = np.percentile(centrality_values, 90)
high_central_nodes = {node: node for node in G.nodes() if deg_centrality[node] >= threshold_label}
nx.draw_networkx_labels(G, pos, labels=high_central_nodes, font_size=10, font_color='black', ax=ax)

ax.set_title("Informative Word Co-occurrence Network")
ax.axis("off")
plt.tight_layout()
plt.show()
```

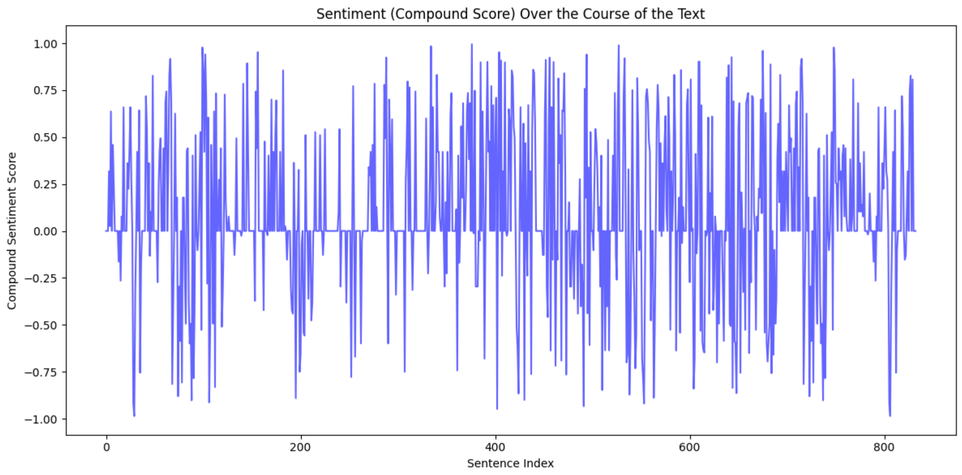
```
Graph has 14 nodes and 17 edges.
```



In [42]:
```
import nltk

nltk.download('vader_lexicon')
from nltk.sentiment.vader import SentimentIntensityAnalyzer

sentences = nltk.sent_tokenize(cleaned_text)


sia = SentimentIntensityAnalyzer()

sentiment_scores = [sia.polarity_scores(sentence)['compound'] for sentence in sentences]


print("Number of sentences:", len(sentences))
print("Average compound sentiment score:", sum(sentiment_scores)/len(sentiment_scores))


plt.figure(figsize=(10, 6))
sns.histplot(sentiment_scores, bins=30, kde=True, color='coral')
plt.title("Distribution of Compound Sentiment Scores Across Sentences")
plt.xlabel("Compound Sentiment Score")
plt.ylabel("Frequency")
plt.tight_layout()
plt.show()

plt.figure(figsize=(12, 6))
plt.plot(sentiment_scores, color='blue', alpha=0.6)
plt.title("Sentiment (Compound Score) Over the Course of the Text")
plt.xlabel("Sentence Index")
plt.ylabel("Compound Sentiment Score")
plt.tight_layout()
plt.show()
```

```
Number of sentences: 833
Average compound sentiment score: 0.08476578631452589
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     /Users/mmadhusudan/nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
```





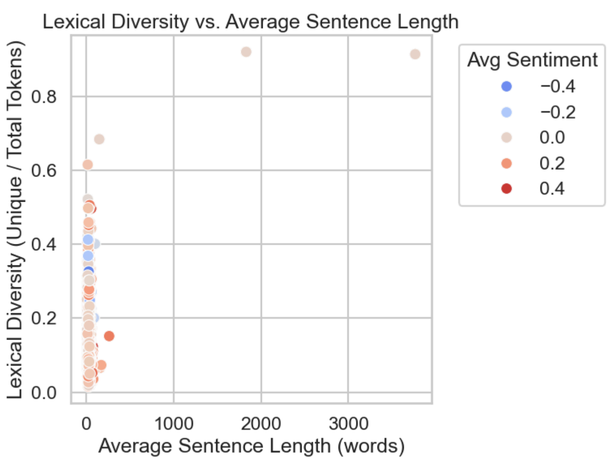
In [43]:

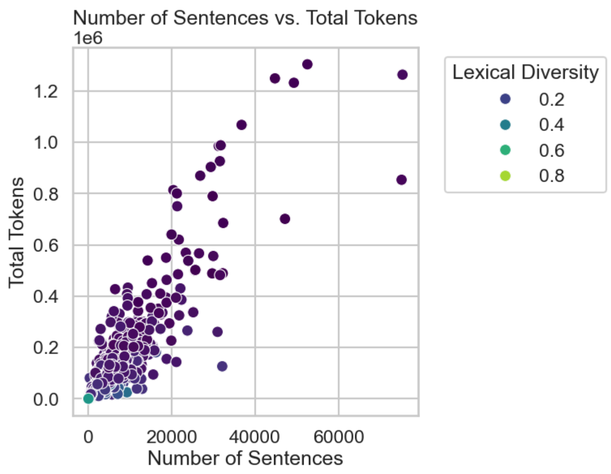
```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import os
import nltk
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


nltk.download('punkt')
nltk.download('vader_lexicon')

def get_cleaned_text(file_path):
    with open(file_path, "r", encoding="utf-8") as f:
        raw_text = f.read()
    start_marker = "*** START OF THIS PROJECT GUTENBERG EBOOK"
    end_marker = "*** END OF THIS PROJECT GUTENBERG EBOOK"
    start_idx = raw_text.find(start_marker)
    if start_idx != -1:
        text = raw_text[start_idx + len(start_marker):]
    else:
        text = raw_text
    end_idx = text.find(end_marker)
    if end_idx != -1:
        text = text[:end_idx]
    return text.strip()


def compute_file_metrics(file_path):
    text = get_cleaned_text(file_path)
    # Word-level analysis:
    tokens = nltk.word_tokenize(text.lower())
    tokens = [t for t in tokens if t.isalpha()]
    num_tokens = len(tokens)
    num_unique = len(set(tokens))
    lexical_diversity = num_unique / num_tokens if num_tokens > 0 else 0

    # Sentence-level analysis:
    sentences = nltk.sent_tokenize(text)
    num_sentences = len(sentences)
    sentence_lengths = [len(nltk.word_tokenize(s)) for s in sentences]
    avg_sentence_length = sum(sentence_lengths)/num_sentences if num_sentences > 0 else 0

    # Sentiment analysis using VADER:
    sia = SentimentIntensityAnalyzer()
    sentiment_scores = [sia.polarity_scores(s)['compound'] for s in sentences]
    avg_sentiment = sum(sentiment_scores)/num_sentences if num_sentences > 0 else 0

    return {
        'file': os.path.basename(file_path),
        'num_tokens': num_tokens,
        'num_unique': num_unique,
        'lexical_diversity': lexical_diversity,
        'num_sentences': num_sentences,
        'avg_sentence_length': avg_sentence_length,
        'avg_sentiment': avg_sentiment
    }


folder = "Gutenberg_Books"
all_files = [f for f in os.listdir(folder) if f.endswith(".txt.txt")]


metrics_list = []
for filename in all_files:
    file_path = os.path.join(folder, filename)
    try:
        metrics = compute_file_metrics(file_path)
        metrics_list.append(metrics)
    except Exception as e:
        print(f"Error processing {filename}: {e}")


df_metrics = pd.DataFrame(metrics_list)
print("Per-file Metrics:")
display(df_metrics)


sns.set(style="whitegrid", context="talk")
plt.figure(figsize=(8,6))
sns.scatterplot(data=df_metrics, x='avg_sentence_length', y='lexical_diversity',
                hue='avg_sentiment', palette='coolwarm', s=100)
plt.title("Lexical Diversity vs. Average Sentence Length")
plt.xlabel("Average Sentence Length (words)")
plt.ylabel("Lexical Diversity (Unique / Total Tokens)")
plt.legend(title='Avg Sentiment', bbox_to_anchor=(1.05, 1), loc=2)
plt.tight_layout()
plt.show()


plt.figure(figsize=(8,6))
sns.scatterplot(data=df_metrics, x='num_sentences', y='num_tokens',
                hue='lexical_diversity', palette='viridis', s=100)
plt.title("Number of Sentences vs. Total Tokens")
plt.xlabel("Number of Sentences")
plt.ylabel("Total Tokens")
plt.legend(title='Lexical Diversity', bbox_to_anchor=(1.05, 1), loc=2)
plt.tight_layout()
plt.show()
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/mmadhusudan/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     /Users/mmadhusudan/nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
Per-file Metrics:
```

	file	num_tokens	num_unique	lexical_diversity	num_sentences	avg_sentence_length	avg_sentiment
0	4658.txt.txt	161293	15819	0.098076	8086	26.728172	0.040154
1	37009.txt.txt	77551	4991	0.064358	3295	27.069803	0.119815
2	14609.txt.txt	89507	11561	0.129163	4501	25.416130	0.013161
3	5342.txt.txt	87301	6824	0.078166	6239	17.282898	0.054297
4	17.txt.txt	268340	5539	0.020642	7676	40.448801	0.062850
...	...	...	...	...	...	...	...
2470	55836.txt.txt	57012	7572	0.132814	2200	31.482727	0.023423
2471	1452.txt.txt	132776	10999	0.082839	5061	30.670223	0.032598
2472	10061.txt.txt	3575	1081	0.302378	143	31.643357	-0.012249
2473	8395.txt.txt	18416	2276	0.123588	626	34.236422	0.035615
2474	31011.txt.txt	462	230	0.497835	33	18.515152	0.074618

2475 rows × 7 columns

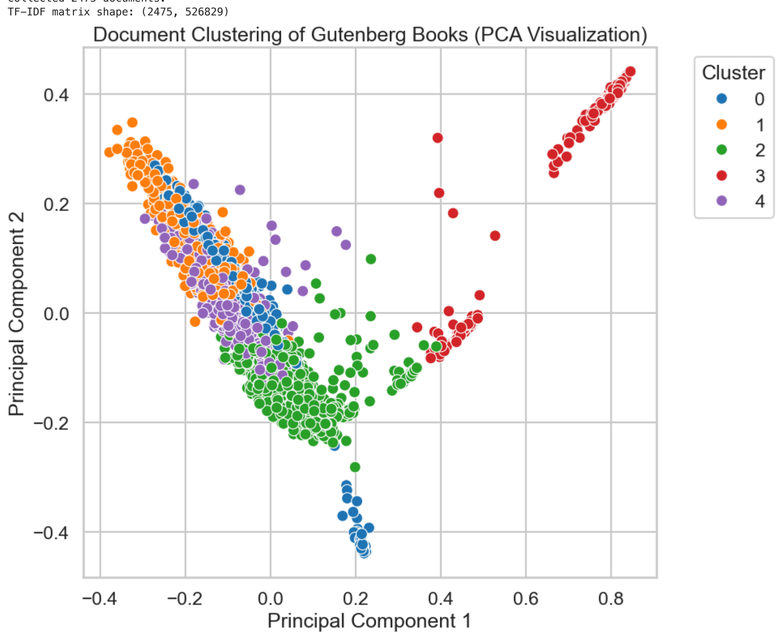




In [44]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import os


corpus = []
doc_names = []

for filename in all_files:
    file_path = os.path.join(folder, filename)

    text = get_cleaned_text(file_path)
    corpus.append(text)
    doc_names.append(filename)

print(f"Collected {len(corpus)} documents.")


vectorizer = TfidfVectorizer(stop_words='english', max_df=0.95, min_df=2)
tfidf_matrix = vectorizer.fit_transform(corpus)
print("TF-IDF matrix shape:", tfidf_matrix.shape)

k = 5
kmeans = KMeans(n_clusters=k, random_state=42)
clusters = kmeans.fit_predict(tfidf_matrix)


pca = PCA(n_components=2, random_state=42)
tfidf_pca = pca.fit_transform(tfidf_matrix.toarray())

df_plot = pd.DataFrame({
    'PC1': tfidf_pca[:, 0],
    'PC2': tfidf_pca[:, 1],
    'Cluster': clusters,
    'Document': doc_names
})

plt.figure(figsize=(10, 8))
sns.scatterplot(data=df_plot, x='PC1', y='PC2', hue='Cluster', palette='tab10', s=100, legend='full')
plt.title("Document Clustering of Gutenberg Books (PCA Visualization)")
plt.xlabel("Principal Component 1")
plt.ylabel("Principal Component 2")
plt.legend(title="Cluster", bbox_to_anchor=(1.05, 1), loc=2)
plt.tight_layout()
plt.show()

for i in range(k):
    docs_in_cluster = df_plot[df_plot['Cluster'] == i]['Document'].tolist()
    print(f"Cluster {i} ({len(docs_in_cluster)} documents): {docs_in_cluster}")
```

```
Collected 2475 documents.
TF-IDF matrix shape: (2475, 526829)
```



```
Cluster 0 (475 documents): ['5342.txt.txt', '24737.txt.txt', '36020.txt.txt', '14407.txt.txt', '7849.txt.txt', '23393.txt.txt', '24852.txt.txt', '507.txt.txt', '53372.txt.txt', '389.txt.txt', '21736.txt.txt', '6927.txt.txt', '2392.txt.txt', '28650.txt.txt', '215.txt.txt', '982.txt.txt', '173.txt.txt', '15767.txt.txt', '2770.txt.txt', '53386.txt.txt', '24025.txt.txt', '13602.t
xt.txt', '12116.txt.txt', '8492.txt.txt', '5817.txt.txt', '8565.txt.txt', '28572.txt.txt', '137.txt.txt', '807.txt.txt', '23672.txt.txt', '18937.txt.txt', '4633.txt.txt', '17412.txt.txt', '13646.txt.txt', '4014.txt.txt', '8914.txt.txt', '21316.txt.txt', '2756.txt.txt', '145.txt.txt', '17811.txt.txt', '155.txt.txt', '2833.txt.txt', '6879.txt.txt', '1947.txt.txt', '223.txt.tx
t', '17396.txt.txt', '17362.txt.txt', '2350.txt.txt', '43.txt.txt', '10607.txt.txt', '108.txt.txt', '316.txt.txt', '1872.txt.txt', '98.txt.txt', '9479.txt.txt', '16251.txt.txt', '11171.txt.txt', '35993.txt.txt', '2735.txt.txt', '24108.txt.txt', '21656.txt.txt', '7841.txt.txt', '2488.txt.txt', '12225.txt.txt', '1897.txt.txt', '13647.txt.txt', '1300.txt.txt', '34165.txt.txt',
'19533.txt.txt', '42.txt.txt', '12545.txt.txt', '28693.txt.txt', '25887.txt.txt', '5061.txt.txt', '9903.txt.txt', '1257.txt.txt', '23756.txt.txt', '599.txt.txt', '14833.txt.txt', '520.txt.txt', '24875.txt.txt', '1188.txt.txt', '646.txt.txt', '11592.txt.txt', '21948.txt.txt', '21078.txt.txt', '5343.txt.txt', '16.txt.txt', '4731.txt.txt', '3777.txt.txt', '9925.txt.txt', '2892.t
xt.txt', '10007.txt.txt', '6133.txt.txt', '7439.txt.txt', '28198.txt.txt', '12573.txt.txt', '398.txt.txt', '564.txt.txt', '24821.txt.txt', '19673.txt.txt', '2609.txt.txt', '1563.txt.txt', '2407.txt.txt', '14893.txt.txt', '204.txt.txt', '58866.txt.txt', '2393.txt.txt', '172.txt.txt', '24389.txt.txt', '2097.txt.txt', '30003.txt.txt', '16718.txt.txt', '2465.txt.txt', '1354.txt.t
xt', '24858.txt.txt', '10586.txt.txt', '146.txt.txt', '30576.txt.txt', '5348.txt.txt', '4017.txt.txt', '7205.txt.txt', '289.txt.txt', '2441.txt.txt', '15213.txt.txt', '10743.txt.txt', '2852.txt.txt', '31856.txt.txt', '17860.txt.txt', '28617.txt.txt', '58820.txt.txt', '4699.txt.txt', '6984.txt.txt', '4715.txt.txt', '2057.txt.txt', '31963.txt.txt', '644.txt.txt', '24877.txt.tx
t', '22285.txt.txt', '27190.txt.txt', '9963.txt.txt', '3457.txt.txt', '1472.txt.txt', '24618.txt.txt', '31219.txt.txt', '10554.txt.txt', '14668.txt.txt', '10886.txt.txt', '22014.txt.txt', '13499.txt.txt', '768.txt.txt', '514.txt.txt', '1825.txt.txt', '12297.txt.txt', '20606.txt.txt', '917.txt.txt', '16264.txt.txt', '19389.txt.txt', '5341.txt.txt', '32069.txt.txt', '23922.txt.
txt', '559.txt.txt', '421.txt.txt', '19726.txt.txt', '50661.txt.txt', '696.txt.txt', '3536.txt.txt', '5747.txt.txt', '832.txt.txt', '3005.txt.txt', '14060.txt.txt', '23625.txt.txt', '4357.txt.txt', '24350.txt.txt', '2126.txt.txt', '17824.txt.txt', '7875.txt.txt', '2391.txt.txt', '1183.txt.txt', '22925.txt.txt', '53370.txt.txt', '14874.txt.txt', '5322.txt.txt', '974.txt.txt',
'10066.txt.txt', '37215.txt.txt', '2070.txt.txt', '36022.txt.txt', '11668.txt.txt', '22544.txt.txt', '5340.txt.txt', '31547.txt.txt', '21687.txt.txt', '14375.txt.txt', '499.txt.txt', '2233.txt.txt', '16663.txt.txt', '17379.txt.txt', '51854.txt.txt', '420.txt.txt', '558.txt.txt', '113.txt.txt', '31861.txt.txt', '5697.txt.txt', '12680.txt.txt', '19755.txt.txt', '8178.txt.txt',
'15387.txt.txt', '1182.txt.txt', '583.txt.txt', '9909.txt.txt', '21322.txt.txt', '7144.txt.txt', '3268.txt.txt', '18485.txt.txt', '24644.txt.txt', '7204.txt.txt', '13921.txt.txt', '4792.txt.txt', '157.txt.txt', '770.txt.txt', '61262.txt.txt', '57323.txt.txt', '7423.txt.txt', '22659.txt.txt', '12122.txt.txt', '25783.txt.txt', '24876.txt.txt', '533.txt.txt', '20547.txt.txt', '6
985.txt.txt', '794.txt.txt', '3424.txt.txt', '21359.txt.txt', '12630.txt.txt', '37329.txt.txt', '51.txt.txt', '5304.txt.txt', '6753.txt.txt', '17782.txt.txt', '53356.txt.txt', '8164.txt.txt', '24873.txt.txt', '8649.txt.txt', '18253.txt.txt', '1251.txt.txt', '781.txt.txt', '2408.txt.txt', '25449.txt.txt', '6980.txt.txt', '44.txt.txt', '5311.txt.txt', '1155.txt.txt', '54.txt.tx
t', '2347.txt.txt', '31516.txt.txt', '12108.txt.txt', '9977.txt.txt', '2098.txt.txt', '15281.txt.txt', '24811.txt.txt', '544.txt.txt', '21279.txt.txt', '2834.txt.txt', '707.txt.txt', '7847.txt.txt', '471.txt.txt', '16957.txt.txt', '120.txt.txt', '24887.txt.txt', '21728.txt.txt', '8190.txt.txt', '12267.txt.txt', '2226.txt.txt', '11758.txt.txt', '11620.txt.txt', '1661.txt.txt',
'4047.txt.txt', '836.txt.txt', '2028.txt.txt', '1342.txt.txt', '9806.txt.txt', '18505.txt.txt', '308.txt.txt', '1399.txt.txt', '174.txt.txt', '5164.txt.txt', '24022.txt.txt', '18019.txt.txt', '3289.txt.txt', '6382.txt.txt', '23462.txt.txt', '21721.txt.txt', '18833.txt.txt', '14766.txt.txt', '13648.txt.txt', '13213.txt.txt', '1685.txt.txt', '6877.txt.txt', '271.txt.txt', '3266
4.txt.txt', '3533.txt.txt', '837.txt.txt', '7028.txt.txt', '22761.txt.txt', '1259.txt.txt', '23065.txt.txt', '14075.txt.txt', '446.txt.txt', '12590.txt.txt', '203.txt.txt', '175.txt.txt', '17959.txt.txt', '11255.txt.txt', '1441.txt.txt', '7065.txt.txt', '14028.txt.txt', '73.txt.txt', '6124.txt.txt', '11.txt.txt', '3760.txt.txt', '2557.txt.txt', '20613.txt.txt', '902.txt.txt',
'2759.txt.txt', '23569.txt.txt', '9932.txt.txt', '2607.txt.txt', '2042.txt.txt', '22566.txt.txt', '5780.txt.txt', '2183.txt.txt', '3499.txt.txt', '5946.txt.txt', '20796.txt.txt', '1095.txt.txt', '1874.txt.txt', '24968.txt.txt', '7335.txt.txt', '13058.txt.txt', '55.txt.txt', '6112.txt.txt', '45.txt.txt', '2020.txt.txt', '10601.txt.txt', '706.txt.txt', '13650.txt.txt', '2273.tx
t.txt', '33407.txt.txt', '33619.txt.txt', '21300.txt.txt', '863.txt.txt', '22160.txt.txt', '30667.txt.txt', '6768.txt.txt', '5396.txt.txt', '47338.txt.txt', '2857.txt.txt', '121.txt.txt', '34465.txt.txt', '7471.txt.txt', '19826.txt.txt', '53416.txt.txt', '30486.txt.txt', '996.txt.txt', '22816.txt.txt', '85.txt.txt', '1965.txt.txt', '1184.txt.txt', '95.txt.txt', '3796.txt.tx
t', '18343.txt.txt', '17161.txt.txt', '24560.txt.txt', '1239.txt.txt', '8486.txt.txt', '14640.txt.txt', '2225.txt.txt', '47696.txt.txt', '19596.txt.txt', '19337.txt.txt', '15839.txt.txt', '910.txt.txt', '21986.txt.txt', '2641.txt.txt', '7896.txt.txt', '34063.txt.txt', '15787.txt.txt', '19033.txt.txt', '13690.txt.txt', '419.txt.txt', '6684.txt.txt', '47.txt.txt', '486.txt.tx
t', '20366.txt.txt', '1703.txt.txt', '20533.txt.txt', '17314.txt.txt', '16255.txt.txt', '1252.txt.txt', '10914.txt.txt', '30691.txt.txt', '6852.txt.txt', '78.txt.txt', '49131.txt.txt', '38777.txt.txt', '19142.txt.txt', '766.txt.txt', '13937.txt.txt', '41562.txt.txt', '6343.txt.txt', '5141.txt.txt', '4376.txt.txt', '53489.txt.txt', '2752.txt.txt', '20532.txt.txt', '22517.txt.t
xt', '46.txt.txt', '1298.txt.txt', '20785.txt.txt', '22976.txt.txt', '209.txt.txt', '8176.txt.txt', '6440.txt.txt', '4358.txt.txt', '4230.txt.txt', '2788.txt.txt', '17866.txt.txt', '19706.txt.txt', '51233.txt.txt', '19460.txt.txt', '244.txt.txt', '20348.txt.txt', '1429.txt.txt', '37363.txt.txt', '4532.txt.txt', '6342.txt.txt', '1952.txt.txt', '21699.txt.txt', '12288.txt.txt',
'236.txt.txt', '18891.txt.txt', '1480.txt.txt', '24878.txt.txt', '3829.txt.txt', '5200.txt.txt', '584.txt.txt', '25452.txt.txt', '21053.txt.txt', '20757.txt.txt', '31619.txt.txt', '53299.txt.txt', '834.txt.txt', '11544.txt.txt', '24571.txt.txt', '21031.txt.txt', '1036.txt.txt', '1026.txt.txt', '20840.txt.txt', '21286.txt.txt', '25823.txt.txt', '12.txt.txt', '2554.txt.txt', '6
1168.txt.txt', '502.txt.txt', '22492.txt.txt', '159.txt.txt', '20194.txt.txt', '2781.txt.txt', '560.txt.txt', '5325.txt.txt', '4747.txt.txt', '1980.txt.txt', '60.txt.txt', '2005.txt.txt']
Cluster 1 (509 documents): ['17.txt.txt', '21111.txt.txt', '4693.txt.txt', '2848.txt.txt', '21528.txt.txt', '23475.txt.txt', '29839.txt.txt', '3704.txt.txt', '1995.txt.txt', '4912.txt.txt', '28536.txt.txt', '2760.txt.txt', '11366.txt.txt', '17309.txt.txt', '7242.txt.txt', '11637.txt.txt', '1666.txt.txt', '23169.txt.txt', '20203.txt.txt', '19850.txt.txt', '2147.txt.txt', '831.
txt.txt', '12753.txt.txt', '24388.txt.txt', '18674.txt.txt', '17587.txt.txt', '12058.txt.txt', '4300.txt.txt', '1015.txt.txt', '13003.txt.txt', '34844.txt.txt', '13797.txt.txt', '3177.txt.txt', '6104.txt.txt', '1152.txt.txt', '31469.txt.txt', '5891.txt.txt', '5625.txt.txt', '940.txt.txt', '18767.txt.txt', '34294.txt.txt', '10148.txt.txt', '14431.txt.txt', '13711.txt.txt', '16
772.txt.txt', '5449.txt.txt', '17310.txt.txt', '2981.txt.txt', '11323.txt.txt', '6052.txt.txt', '1004.txt.txt', '3031.txt.txt', '7499.txt.txt', '14916.txt.txt', '3296.txt.txt', '2850.txt.txt', '711.txt.txt', '12873.txt.txt', '19608.txt.txt', '1484.txt.txt', '1494.txt.txt', '38233.txt.txt', '13796.txt.txt', '46521.txt.txt', '16941.txt.txt', '15263.txt.txt', '30249.txt.txt', '2
2089.txt.txt', '29056.txt.txt', '37814.txt.txt', '5307.txt.txt', '11112.txt.txt', '13127.txt.txt', '18909.txt.txt', '15675.txt.txt', '13241.txt.txt', '11431.txt.txt', '22675.txt.txt', '2610.txt.txt', '7026.txt.txt', '2600.txt.txt', '2323.txt.txt', '22409.txt.txt', '42622.txt.txt', '6585.txt.txt', '23499.txt.txt', '21381.txt.txt', '353.txt.txt', '1837.txt.txt', '21899.txt.tx
t', '11348.txt.txt', '37206.txt.txt', '2882.txt.txt', '7999.txt.txt', '18857.txt.txt', '1597.txt.txt', '9886.txt.txt', '10150.txt.txt', '14314.txt.txt', '5529.txt.txt', '18193.txt.txt', '6782.txt.txt', '441.txt.txt', '529.txt.txt', '4294.txt.txt', '5266.txt.txt', '18910.txt.txt', '14366.txt.txt', '11438.txt.txt', '830.txt.txt', '23533.txt.txt', '18297.txt.txt', '3534.txt.tx
t', '37787.txt.txt', '348.txt.txt', '22693.txt.txt', '10985.txt.txt', '22117.txt.txt', '3327.txt.txt', '1312.txt.txt', '2276.txt.txt', '7440.txt.txt', '1653.txt.txt', '6615.txt.txt', '22456.txt.txt', '53368.txt.txt', '124.txt.txt', '18362.txt.txt', '3284.txt.txt', '26196.txt.txt', '6818.txt.txt', '12384.txt.txt', '4028.txt.txt', '20748.txt.txt', '25016.txt.txt', '5998.txt.tx
t', '15040.txt.txt', '169.txt.txt', '18247.txt.txt', '654.txt.txt', '20955.txt.txt', '5921.txt.txt', '2388.txt.txt', '14591.txt.txt', '34607.txt.txt', '1039.txt.txt', '3721.txt.txt', '5305.txt.txt', '12561.txt.txt', '13113.txt.txt', '6764.txt.txt', '16160.txt.txt', '14466.txt.txt', '30230.txt.txt', '54337.txt.txt', '7838.txt.txt', '23875.txt.txt', '341.txt.txt', '3810.txt.tx
t', '8220.txt.txt', '14241.txt.txt', '8299.txt.txt', '2874.txt.txt', '16672.txt.txt', '12426.txt.txt', '8795.txt.txt', '33434.txt.txt', '5160.txt.txt', '13291.txt.txt', '36844.txt.txt', '82.txt.txt', '443.txt.txt', '12595.txt.txt', '22396.txt.txt', '29349.txt.txt', '7951.txt.txt', '36970.txt.txt', '601.txt.txt', '37959.txt.txt', '18721.txt.txt', '16939.txt.txt', '6130.txt.tx
t', '22373.txt.txt', '13102.txt.txt', '25325.txt.txt', '11772.txt.txt', '10897.txt.txt', '41085.txt.txt', '16452.txt.txt', '964.txt.txt', '30221.txt.txt', '27887.txt.txt', '11241.txt.txt', '13559.txt.txt', '2881.txt.txt', '14328.txt.txt', '22886.txt.txt', '30760.txt.txt', '15.txt.txt', '228.txt.txt', '16697.txt.txt', '42324.txt.txt', '28497.txt.txt', '2145.txt.txt', '6886.tx
t.txt', '19529.txt.txt', '22083.txt.txt', '18500.txt.txt', '19068.txt.txt', '7864.txt.txt', '6791.txt.txt', '207.txt.txt', '37904.txt.txt', '39674.txt.txt', '14880.txt.txt', '7477.txt.txt', '3623.txt.txt', '17766.txt.txt', '5157.txt.txt', '2744.txt.txt', '12473.txt.txt', '3050.txt.txt', '19240.txt.txt', '25918.txt.txt', '10845.txt.txt', '135.txt.txt', '31847.txt.txt', '2853.t
xt.txt', '26046.txt.txt', '7889.txt.txt', '11565.txt.txt', '36131.txt.txt', '15659.txt.txt', '15164.txt.txt', '1371.txt.txt', '11330.txt.txt', '19061.txt.txt', '3837.txt.txt', '6788.txt.txt', '13097.txt.txt', '7353.txt.txt', '30103.txt.txt', '23.txt.txt', '3434.txt.txt', '20885.txt.txt', '3261.txt.txt', '7025.txt.txt', '12758.txt.txt', '11277.txt.txt', '22500.txt.txt', '2554
6.txt.txt', '17208.txt.txt', '23580.txt.txt', '21189.txt.txt', '19312.txt.txt', '7099.txt.txt', '10662.txt.txt', '36.txt.txt', '3725.txt.txt', '37122.txt.txt', '12543.txt.txt', '11741.txt.txt', '3160.txt.txt', '20656.txt.txt', '2149.txt.txt', '12060.txt.txt', '8743.txt.txt', '6687.txt.txt', '22420.txt.txt', '34736.txt.txt', '2680.txt.txt', '15697.txt.txt', '12380.txt.txt', '1
4297.txt.txt', '2846.txt.txt', '11408.txt.txt', '13268.txt.txt', '31162.txt.txt', '36124.txt.txt', '2166.txt.txt', '1002.txt.txt', '14020.txt.txt', '22217.txt.txt', '11894.txt.txt', '12422.txt.txt', '7245.txt.txt', '17162.txt.txt', '14360.txt.txt', '38827.txt.txt', '2150.txt.txt', '12888.txt.txt', '12898.txt.txt', '7766.txt.txt', '22096.txt.txt', '45619.txt.txt', '22382.txt.t
xt', '24902.txt.txt', '86.txt.txt', '33687.txt.txt', '12982.txt.txt', '21614.txt.txt', '6168.txt.txt', '139.txt.txt', '15794.txt.txt', '7885.txt.txt', '19900.txt.txt', '16985.txt.txt', '62.txt.txt', '19287.txt.txt', '8800.txt.txt', '6702.txt.txt', '3761.txt.txt', '23245.txt.txt', '903.txt.txt', '345.txt.txt', '11870.txt.txt', '1831.txt.txt', '21692.txt.txt', '4389.txt.txt',
'29246.txt.txt', '9380.txt.txt', '7401.txt.txt', '35554.txt.txt', '2151.txt.txt', '7777.txt.txt', '4330.txt.txt', '17163.txt.txt', '21138.txt.txt', '2400.txt.txt', '24676.txt.txt', '456.txt.txt', '213.txt.txt', '17034.txt.txt', '14226.txt.txt', '808.txt.txt', '17024.txt.txt', '1727.txt.txt', '8801.txt.txt', '12259.txt.txt', '16984.txt.txt', '11123.txt.txt', '18274.txt.txt',
'33965.txt.txt', '677.txt.txt', '6087.txt.txt', '10376.txt.txt', '45858.txt.txt', '13202.txt.txt', '11462.txt.txt', '2064.txt.txt', '25718.txt.txt', '14411.txt.txt', '10026.txt.txt', '11369.txt.txt', '1250.txt.txt', '16131.txt.txt', '8789.txt.txt', '8799.txt.txt', '18134.txt.txt', '12797.txt.txt', '4380.txt.txt', '41445.txt.txt', '11740.txt.txt', '2503.txt.txt', '23997.txt.tx
t', '14445.txt.txt', '2148.txt.txt', '4339.txt.txt', '1467.txt.txt', '14368.txt.txt', '1493.txt.txt', '5225.txt.txt', '13015.txt.txt', '2750.txt.txt', '873.txt.txt', '26346.txt.txt', '35345.txt.txt', '31111.txt.txt', '25763.txt.txt', '1365.txt.txt', '30511.txt.txt', '16537.txt.txt', '5402.txt.txt', '23031.txt.txt', '1003.txt.txt', '10430.txt.txt', '19226.txt.txt', '131.txt.tx
t', '16042.txt.txt', '35534.txt.txt', '2131.txt.txt', '21324.txt.txt', '12928.txt.txt', '39341.txt.txt', '23495.txt.txt', '22381.txt.txt', '24869.txt.txt', '10380.txt.txt', '4974.txt.txt', '19721.txt.txt', '17851.txt.txt', '20851.txt.txt', '6317.txt.txt', '11310.txt.txt', '11887.txt.txt', '3657.txt.txt', '13053.txt.txt', '23650.txt.txt', '11863.txt.txt', '503.txt.txt', '2248
3.txt.txt', '148.txt.txt', '10002.txt.txt', '60908.txt.txt', '12351.txt.txt', '17567.txt.txt', '12333.txt.txt', '16986.txt.txt', '1008.txt.txt', '2527.txt.txt', '3155.txt.txt', '972.txt.txt', '10070.txt.txt', '25600.txt.txt', '7947.txt.txt', '18214.txt.txt', '10838.txt.txt', '8578.txt.txt', '7190.txt.txt', '944.txt.txt', '13579.txt.txt', '25428.txt.txt', '3726.txt.txt', '1534
3.txt.txt', '3163.txt.txt', '2563.txt.txt', '35.txt.txt', '10671.txt.txt', '792.txt.txt', '3422.txt.txt', '12169.txt.txt', '1001.txt.txt', '1998.txt.txt', '12242.txt.txt', '17221.txt.txt', '6574.txt.txt', '38767.txt.txt', '12814.txt.txt', '22694.txt.txt', '17770.txt.txt', '14572.txt.txt', '31103.txt.txt', '699.txt.txt', '32518.txt.txt', '689.txt.txt', '6422.txt.txt', '7128.tx
t.txt', '16140.txt.txt', '15399.txt.txt', '37872.txt.txt', '28019.txt.txt', '21254.txt.txt', '13944.txt.txt', '32934.txt.txt', '36405.txt.txt', '2743.txt.txt', '32300.txt.txt', '22101.txt.txt', '26500.txt.txt', '7446.txt.txt', '13643.txt.txt', '18828.txt.txt', '8159.txt.txt', '10842.txt.txt', '23494.txt.txt', '2199.txt.txt', '2800.txt.txt', '6932.txt.txt', '8642.txt.txt', '8
4.txt.txt', '21973.txt.txt', '13665.txt.txt', '23545.txt.txt', '3317.txt.txt', '20102.txt.txt', '15145.txt.txt', '680.txt.txt', '13334.txt.txt', '7386.txt.txt', '41537.txt.txt', '1906.txt.txt', '44896.txt.txt', '10806.txt.txt', '20778.txt.txt', '4018.txt.txt', '16751.txt.txt', '512.txt.txt', '25281.txt.txt', '27933.txt.txt', '15950.txt.txt', '9882.txt.txt', '20907.txt.txt',
'45542.txt.txt', '15932.txt.txt', '11387.txt.txt', '36564.txt.txt', '11246.txt.txt', '14460.txt.txt', '24263.txt.txt']
Cluster 2 (823 documents): ['4658.txt.txt', '37009.txt.txt', '25023.txt.txt', '22065.txt.txt', '15620.txt.txt', '22748.txt.txt', '22620.txt.txt', '34781.txt.txt', '20191.txt.txt', '8700.txt.txt', '18751.txt.txt', '14814.txt.txt', '38516.txt.txt', '26867.txt.txt', '19514.txt.txt', '37823.txt.txt', '3462.txt.txt', '34554.txt.txt', '18846.txt.txt', '17354.txt.txt', '33502.txt.tx
t', '30155.txt.txt', '38398.txt.txt', '12327.txt.txt', '19274.txt.txt', '61142.txt.txt', '34093.txt.txt', '33629.txt.txt', '24506.txt.txt', '5173.txt.txt', '35448.txt.txt', '25731.txt.txt', '37350.txt.txt', '17742.txt.txt', '23626.txt.txt', '35013.txt.txt', '6630.txt.txt', '28957.txt.txt', '35542.txt.txt', '29816.txt.txt', '15091.txt.txt', '17855.txt.txt', '20587.txt.txt', '2
77.txt.txt', '35829.txt.txt', '30293.txt.txt', '13493.txt.txt', '2939.txt.txt', '20153.txt.txt', '2009.txt.txt', '21531.txt.txt', '31175.txt.txt', '27638.txt.txt', '14600.txt.txt', '42198.txt.txt', '2746.txt.txt', '10986.txt.txt', '24448.txt.txt', '33767.txt.txt', '5710.txt.txt', '20774.txt.txt', '16543.txt.txt', '22114.txt.txt', '18556.txt.txt', '24923.txt.txt', '20386.txt.t
xt', '18884.txt.txt', '13923.txt.txt', '24787.txt.txt', '26113.txt.txt', '23991.txt.txt', '26558.txt.txt', '19921.txt.txt', '34604.txt.txt', '20924.txt.txt', '46773.txt.txt', '19998.txt.txt', '35830.txt.txt', '35842.txt.txt', '19550.txt.txt', '8215.txt.txt', '39471.txt.txt', '38003.txt.txt', '27868.txt.txt', '34903.txt.txt', '38013.txt.txt', '31050.txt.txt', '27675.txt.txt',
'19953.txt.txt', '24222.txt.txt', '11734.txt.txt', '30754.txt.txt', '32376.txt.txt', '26323.txt.txt', '12774.txt.txt', '33029.txt.txt', '10751.txt.txt', '3674.txt.txt', '22379.txt.txt', '23673.txt.txt', '43375.txt.txt', '38404.txt.txt', '10834.txt.txt', '2938.txt.txt', '18525.txt.txt', '39372.txt.txt', '9457.txt.txt', '19406.txt.txt', '11662.txt.txt', '26672.txt.txt', '27378.
txt.txt', '58008.txt.txt', '1887.txt.txt', '16119.txt.txt', '30321.txt.txt', '1268.txt.txt', '34175.txt.txt', '48007.txt.txt', '34984.txt.txt', '33766.txt.txt', '24449.txt.txt', '11498.txt.txt', '28434.txt.txt', '18237.txt.txt', '12956.txt.txt', '8423.txt.txt', '35744.txt.txt', '34326.txt.txt', '38482.txt.txt', '14990.txt.txt', '27560.txt.txt', '39396.txt.txt', '15665.txt.tx
t', '15020.txt.txt', '14558.txt.txt', '35450.txt.txt', '34501.txt.txt', '34479.txt.txt', '6986.txt.txt', '375.txt.txt', '8172.txt.txt', '15193.txt.txt', '27911.txt.txt', '51547.txt.txt', '39235.txt.txt', '9666.txt.txt', '33972.txt.txt', '9650.txt.txt', '19769.txt.txt', '23770.txt.txt', '7010.txt.txt', '27748.txt.txt', '16441.txt.txt', '34076.txt.txt', '14474.txt.txt', '15535.
txt.txt', '13111.txt.txt', '33287.txt.txt', '28764.txt.txt', '28402.txt.txt', '26393.txt.txt', '53373.txt.txt', '28247.txt.txt', '29635.txt.txt', '22766.txt.txt', '2124.txt.txt', '20116.txt.txt', '12648.txt.txt', '18183.txt.txt', '1615.txt.txt', '19499.txt.txt', '3620.txt.txt', '29233.txt.txt', '32974.txt.txt', '38440.txt.txt', '19420.txt.txt', '19275.txt.txt', '19103.txt.tx
t', '18334.txt.txt', '20788.txt.txt', '41533.txt.txt', '28897.txt.txt', '42649.txt.txt', '12261.txt.txt', '15237.txt.txt', '43282.txt.txt', '35070.txt.txt', '18900.txt.txt', '37566.txt.txt', '32505.txt.txt', '13575.txt.txt', '34131.txt.txt', '38959.txt.txt', '24258.txt.txt', '19251.txt.txt', '34585.txt.txt', '32950.txt.txt', '29019.txt.txt', '22107.txt.txt', '17132.txt.txt',
'4524.txt.txt', '6344.txt.txt', '31751.txt.txt', '19380.txt.txt', '34211.txt.txt', '14056.txt.txt', '60281.txt.txt', '20411.txt.txt', '33508.txt.txt', '24637.txt.txt', '35601.txt.txt', '18790.txt.txt', '4748.txt.txt', '24072.txt.txt', '5761.txt.txt', '31987.txt.txt', '19145.txt.txt', '14293.txt.txt', '16921.txt.txt', '22.txt.txt', '31570.txt.txt', '40818.txt.txt', '20556.txt.
txt', '27676.txt.txt', '11204.txt.txt', '17748.txt.txt', '39472.txt.txt', '8018.txt.txt', '14959.txt.txt', '32817.txt.txt', '11684.txt.txt', '8102.txt.txt', '19449.txt.txt', '30850.txt.txt', '19364.txt.txt', '3788.txt.txt', '8559.txt.txt', '53347.txt.txt', '20927.txt.txt', '16220.txt.txt', '31674.txt.txt', '19786.txt.txt', '31149.txt.txt', '49119.txt.txt', '22344.txt.txt', '3
3527.txt.txt', '17209.txt.txt', '28681.txt.txt', '38480.txt.txt', '13618.txt.txt', '32982.txt.txt', '31534.txt.txt', '20159.txt.txt', '17149.txt.txt', '22790.txt.txt', '34954.txt.txt', '48836.txt.txt', '13489.txt.txt', '18203.txt.txt', '37842.txt.txt', '39996.txt.txt', '36903.txt.txt', '17192.txt.txt', '24583.txt.txt', '37529.txt.txt', '26477.txt.txt', '31630.txt.txt', '1760
6.txt.txt', '17408.txt.txt', '14.txt.txt', '35157.txt.txt', '8866.txt.txt', '4973.txt.txt', '9411.txt.txt', '274.txt.txt', '24964.txt.txt', '35490.txt.txt', '20846.txt.txt', '33648.txt.txt', '17987.txt.txt', '38687.txt.txt', '18285.txt.txt', '35062.txt.txt', '34568.txt.txt', '14012.txt.txt', '15460.txt.txt', '19591.txt.txt', '17455.txt.txt', '2232.txt.txt', '30001.txt.txt',
'14625.txt.txt', '26656.txt.txt', '14070.txt.txt', '30626.txt.txt', '7223.txt.txt', '27509.txt.txt', '29086.txt.txt', '30550.txt.txt', '11365.txt.txt', '21566.txt.txt', '22764.txt.txt', '7700.txt.txt', '38207.txt.txt', '6934.txt.txt', '18050.txt.txt', '8001.txt.txt', '216.txt.txt', '14838.txt.txt', '25973.txt.txt', '17910.txt.txt', '28569.txt.txt', '779.txt.txt', '19506.txt.t
xt', '1176.txt.txt', '37101.txt.txt', '20346.txt.txt', '6.txt.txt', '21724.txt.txt', '20665.txt.txt', '13274.txt.txt', '22728.txt.txt', '33129.txt.txt', '30243.txt.txt', '31621.txt.txt', '15622.txt.txt', '13008.txt.txt', '38290.txt.txt', '2306.txt.txt', '34371.txt.txt', '17571.txt.txt', '36285.txt.txt', '35097.txt.txt', '32958.txt.txt', '19564.txt.txt', '25286.txt.txt', '778
7.txt.txt', '46981.txt.txt', '31141.txt.txt', '33721.txt.txt', '12629.txt.txt', '33659.txt.txt', '15707.txt.txt', '30523.txt.txt', '48.txt.txt', '30000.txt.txt', '34579.txt.txt', '8952.txt.txt', '27386.txt.txt', '548.txt.txt', '4962.txt.txt', '21916.txt.txt', '19737.txt.txt', '103.txt.txt', '26598.txt.txt', '22397.txt.txt', '19423.txt.txt', '83.txt.txt', '11647.txt.txt', '189
71.txt.txt', '20426.txt.txt', '17740.txt.txt', '53384.txt.txt', '9090.txt.txt', '20750.txt.txt', '4346.txt.txt', '30541.txt.txt', '2888.txt.txt', '23259.txt.txt', '6478.txt.txt', '20776.txt.txt', '25140.txt.txt', '15435.txt.txt', '22260.txt.txt', '15270.txt.txt', '51058.txt.txt', '24931.txt.txt', '19250.txt.txt', '23186.txt.txt', '18544.txt.txt', '1487.txt.txt', '54460.txt.tx
t', '28570.txt.txt', '34834.txt.txt', '4759.txt.txt', '24063.txt.txt', '1642.txt.txt', '34437.txt.txt', '16523.txt.txt', '31732.txt.txt', '8452.txt.txt', '34098.txt.txt', '168.txt.txt', '36924.txt.txt', '19634.txt.txt', '10912.txt.txt', '23755.txt.txt', '31334.txt.txt', '40819.txt.txt', '24156.txt.txt', '15139.txt.txt', '5000.txt.txt', '12017.txt.txt', '41958.txt.txt', '5508
4.txt.txt', '23433.txt.txt', '24407.txt.txt', '31513.txt.txt', '41.txt.txt', '14993.txt.txt', '15284.txt.txt', '38315.txt.txt', '38428.txt.txt', '23691.txt.txt', '7297.txt.txt', '50133.txt.txt', '19885.txt.txt', '20839.txt.txt', '17829.txt.txt', '3421.txt.txt', '26378.txt.txt', '20317.txt.txt', '16130.txt.txt', '31574.txt.txt', '3008.txt.txt', '27600.txt.txt', '27778.txt.tx
t', '31035.txt.txt', '18298.txt.txt', '32472.txt.txt', '13402.txt.txt', '1233.txt.txt', '38956.txt.txt', '5067.txt.txt', '23585.txt.txt', '9612.txt.txt', '53343.txt.txt', '27975.txt.txt', '279.txt.txt', '62579.txt.txt', '8106.txt.txt', '1881.txt.txt', '20763.txt.txt', '31830.txt.txt', '10840.txt.txt', '24790.txt.txt', '15489.txt.txt', '519.txt.txt', '18637.txt.txt', '21781.tx
t.txt', '14868.txt.txt', '19029.txt.txt', '34949.txt.txt', '5765.txt.txt', '22009.txt.txt', '22990.txt.txt', '34101.txt.txt', '19090.txt.txt', '6322.txt.txt', '11335.txt.txt', '32653.txt.txt', '15203.txt.txt', '15207.txt.txt', '3672.txt.txt', '27559.txt.txt', '30666.txt.txt', '20467.txt.txt', '35128.txt.txt', '17451.txt.txt', '15464.txt.txt', '25267.txt.txt', '14006.txt.txt', '14370.txt.txt',
'14218.txt.txt', '15147.txt.txt', '28553.txt.txt', '19722.txt.txt', '18350.txt.txt', '19444.txt.txt', '53669.txt.txt', '22288.txt.txt', '32962.txt.txt', '18779.txt.txt', '37901.txt.txt', '19115.txt.txt', '34094.txt.txt', '164.txt.txt', '38658.txt.txt', '53381.txt.txt', '25992.txt.txt', '27713.txt.txt', '37512.txt.txt', '31240.txt.txt', '27348.txt.txt', '11385.txt.txt', '1845
2.txt.txt', '5180.txt.txt', '22784.txt.txt', '180.txt.txt', '56482.txt.txt', '31458.txt.txt', '59417.txt.txt', '13117.txt.txt', '18929.txt.txt', '34449.txt.txt', '24730.txt.txt', '30775.txt.txt', '36504.txt.txt', '14776.txt.txt', '16116.txt.txt', '14400.txt.txt', '33044.txt.txt', '31624.txt.txt', '32426.txt.txt', '2884.txt.txt', '48991.txt.txt', '4204.txt.txt', '10011.txt.tx
t', '45532.txt.txt', '38032.txt.txt', '28233.txt.txt', '19042.txt.txt', '39205.txt.txt', '31951.txt.txt', '28466.txt.txt', '26908.txt.txt', '29362.txt.txt', '28710.txt.txt', '8142.txt.txt', '37856.txt.txt', '12293.txt.txt', '24855.txt.txt', '16295.txt.txt', '15097.txt.txt', '2871.txt.txt', '19723.txt.txt', '38308.txt.txt', '19368.txt.txt', '7254.txt.txt', '32159.txt.txt', '54
24.txt.txt', '19856.txt.txt', '48769.txt.txt', '4248.txt.txt', '25983.txt.txt', '22771.txt.txt', '30310.txt.txt', '30066.txt.txt', '25529.txt.txt', '18184.txt.txt', '19262.txt.txt', '87.txt.txt', '994.txt.txt', '18333.txt.txt', '14473.txt.txt', '2016.txt.txt', '35413.txt.txt', '34634.txt.txt', '5605.txt.txt', '33504.txt.txt', '18928.txt.txt', '2.txt.txt', '59416.txt.txt', '14
091.txt.txt', '18735.txt.txt', '19031.txt.txt', '53364.txt.txt', '18206.txt.txt', '37595.txt.txt', '34787.txt.txt', '19606.txt.txt', '17966.txt.txt', '19053.txt.txt', '29728.txt.txt', '15491.txt.txt', '59298.txt.txt', '17575.txt.txt', '33566.txt.txt', '36036.txt.txt', '22636.txt.txt', '25646.txt.txt', '22600.txt.txt', '36645.txt.txt', '9914.txt.txt', '27238.txt.txt', '34353.t
xt.txt', '37151.txt.txt', '12787.txt.txt', '35596.txt.txt', '38077.txt.txt', '31293.txt.txt', '19209.txt.txt', '14987.txt.txt', '37776.txt.txt', '18866.txt.txt', '34259.txt.txt', '2030.txt.txt', '25874.txt.txt', '30181.txt.txt', '13791.txt.txt', '20390.txt.txt', '49080.txt.txt', '4657.txt.txt', '13640.txt.txt', '3332.txt.txt', '34737.txt.txt', '34110.txt.txt', '17700.txt.tx
t', '38189.txt.txt', '30677.txt.txt', '18931.txt.txt', '27558.txt.txt', '24077.txt.txt', '14664.txt.txt', '16780.txt.txt', '24485.txt.txt', '19715.txt.txt', '12443.txt.txt', '26014.txt.txt', '7234.txt.txt', '17275.txt.txt', '5201.txt.txt', '20113.txt.txt', '25735.txt.txt', '1323.txt.txt', '13347.txt.txt', '19116.txt.txt', '33830.txt.txt', '10084.txt.txt', '35937.txt.txt', '11
924.txt.txt', '19270.txt.txt', '16972.txt.txt', '29122.txt.txt', '8547.txt.txt', '33852.txt.txt', '13325.txt.txt', '31147.txt.txt', '27137.txt.txt', '20019.txt.txt', '1027.txt.txt', '7413.txt.txt', '14015.txt.txt', '26076.txt.txt', '1662.txt.txt', '21020.txt.txt', '31221.txt.txt', '7256.txt.txt', '3807.txt.txt', '45647.txt.txt', '20195.txt.txt', '33967.txt.txt', '22829.txt.tx
t', '14403.txt.txt', '2066.txt.txt', '19180.txt.txt', '20769.txt.txt', '53480.txt.txt', '23403.txt.txt', '33574.txt.txt', '34532.txt.txt', '49032.txt.txt', '3772.txt.txt', '1112.txt.txt', '61.txt.txt', '34550.txt.txt', '18852.txt.txt', '37742.txt.txt', '20663.txt.txt', '13272.txt.txt', '19913.txt.txt', '15003.txt.txt', '33915.txt.txt', '39421.txt.txt', '571.txt.txt', '16847.t
xt.txt', '14872.txt.txt', '38418.txt.txt', '22282.txt.txt', '20848.txt.txt', '48807.txt.txt', '25063.txt.txt', '22035.txt.txt', '25550.txt.txt', '27517.txt.txt', '25.txt.txt', '3754.txt.txt', '34672.txt.txt', '926.txt.txt', '16410.txt.txt', '33941.txt.txt', '28216.txt.txt', '36922.txt.txt', '29031.txt.txt', '8177.txt.txt', '14826.txt.txt', '32189.txt.txt', '28018.txt.txt', '1
9279.txt.txt', '17474.txt.txt', '53929.txt.txt', '11615.txt.txt', '23666.txt.txt', '16360.txt.txt', '48070.txt.txt', '16370.txt.txt', '18458.txt.txt', '8419.txt.txt', '33874.txt.txt', '18013.txt.txt', '46303.txt.txt', '17382.txt.txt', '32947.txt.txt', '34582.txt.txt', '39988.txt.txt', '12299.txt.txt', '21688.txt.txt', '10843.txt.txt', '11344.txt.txt', '31756.txt.txt', '10726.
txt.txt', '15100.txt.txt', '2742.txt.txt', '60088.txt.txt', '53499.txt.txt', '17135.txt.txt', '13007.txt.txt', '31558.txt.txt', '40820.txt.txt', '34044.txt.txt', '24776.txt.txt', '15352.txt.txt', '21081.txt.txt', '21918.txt.txt', '18155.txt.txt', '16081.txt.txt', '27977.txt.txt', '19729.txt.txt', '18223.txt.txt', '22914.txt.txt', '19651.txt.txt', '18251.txt.txt', '14837.txt.t
xt', '8997.txt.txt', '19494.txt.txt', '33543.txt.txt', '11649.txt.txt', '12655.txt.txt', '23434.txt.txt', '26457.txt.txt', '23319.txt.txt', '12815.txt.txt', '17289.txt.txt', '20299.txt.txt', '21007.txt.txt', '37632.txt.txt', '12406.txt.txt', '3708.txt.txt', '21534.txt.txt', '29691.txt.txt', '1533.txt.txt', '2753.txt.txt', '20771.txt.txt', '17124.txt.txt', '30560.txt.txt', '36
669.txt.txt', '15888.txt.txt', '15069.txt.txt', '35158.txt.txt', '48530.txt.txt', '4011.txt.txt', '29444.txt.txt', '16955.txt.txt', '10852.txt.txt', '15831.txt.txt', '8502.txt.txt', '1974.txt.txt', '19271.txt.txt', '4907.txt.txt', '19424.txt.txt', '40817.txt.txt', '23066.txt.txt', '35006.txt.txt', '25990.txt.txt', '3307.txt.txt', '15137.txt.txt', '9097.txt.txt', '1228.txt.tx
t', '10773.txt.txt', '17170.txt.txt', '10136.txt.txt', '24409.txt.txt', '32032.txt.txt', '3003.txt.txt', '4778.txt.txt', '32677.txt.txt', '13962.txt.txt', '37809.txt.txt', '2872.txt.txt', '7188.txt.txt', '36830.txt.txt', '7014.txt.txt', '12238.txt.txt', '34376.txt.txt', '12350.txt.txt', '34523.txt.txt', '13177.txt.txt', '6875.txt.txt', '38356.txt.txt', '2485.txt.txt', '14969.
txt.txt', '19138.txt.txt', '34848.txt.txt', '33966.txt.txt', '33914.txt.txt', '5192.txt.txt', '19354.txt.txt', '20239.txt.txt', '6329.txt.txt', '9943.txt.txt', '22657.txt.txt', '29688.txt.txt', '16388.txt.txt', '16378.txt.txt', '8395.txt.txt', '31011.txt.txt']
Cluster 3 (147 documents): ['14609.txt.txt', '54020.txt.txt', '320.txt.txt', '18723.txt.txt', '2000.txt.txt', '26608.txt.txt', '5258.txt.txt', '60656.txt.txt', '43389.txt.txt', '17203.txt.txt', '23654.txt.txt', '42648.txt.txt', '13951.txt.txt', '18783.txt.txt', '37989.txt.txt', '32315.txt.txt', '42131.txt.txt', '56327.txt.txt', '62406.txt.txt', '49168.txt.txt', '17707.txt.tx
t', '5126.txt.txt', '41322.txt.txt', '4791.txt.txt', '42765.txt.txt', '32298.txt.txt', '4968.txt.txt', '28827.txt.txt', '4717.txt.txt', '4649.txt.txt', '46111.txt.txt', '21282.txt.txt', '17798.txt.txt', '5097.txt.txt', '2820.txt.txt', '40827.txt.txt', '20401.txt.txt', '37951.txt.txt', '43761.txt.txt', '45468.txt.txt', '62405.txt.txt', '44958.txt.txt', '57687.txt.txt', '35103.
txt.txt', '39201.txt.txt', '15066.txt.txt', '22268.txt.txt', '10775.txt.txt', '62196.txt.txt', '24515.txt.txt', '27278.txt.txt', '18798.txt.txt', '13216.txt.txt', '17419.txt.txt', '53749.txt.txt', '34783.txt.txt', '60882.txt.txt', '17691.txt.txt', '54873.txt.txt', '20394.txt.txt', '27752.txt.txt', '13703.txt.txt', '62404.txt.txt', '35498.txt.txt', '42525.txt.txt', '57648.txt.
txt', '19496.txt.txt', '36315.txt.txt', '17013.txt.txt', '19643.txt.txt', '18157.txt.txt', '15113.txt.txt', '59037.txt.txt', '24924.txt.txt', '14287.txt.txt', '27757.txt.txt', '18784.txt.txt', '14154.txt.txt', '14158.txt.txt', '11748.txt.txt', '49437.txt.txt', '10814.txt.txt', '20852.txt.txt', '26699.txt.txt', '59859.txt.txt', '39101.txt.txt', '60198.txt.txt', '16826.txt.txt', '39331.txt.tx
t', '11302.txt.txt', '26370.txt.txt', '799.txt.txt', '16827.txt.txt', '31881.txt.txt', '16059.txt.txt', '2419.txt.txt', '39328.txt.txt', '25317.txt.txt', '18289.txt.txt', '10841.txt.txt', '24536.txt.txt', '18921.txt.txt', '13622.txt.txt', '58706.txt.txt', '19106.txt.txt', '44300.txt.txt', '58801.txt.txt', '62615.txt.txt', '13846.txt.txt', '52123.txt.txt', '38674.txt.txt', '55
855.txt.txt', '14765.txt.txt', '16105.txt.txt', '4548.txt.txt', '53523.txt.txt', '35802.txt.txt', '15353.txt.txt', '49887.txt.txt', '18864.txt.txt', '1619.txt.txt', '19234.txt.txt', '26818.txt.txt', '51338.txt.txt', '34008.txt.txt', '12230.txt.txt', '16885.txt.txt', '24766.txt.txt', '14799.txt.txt', '12533.txt.txt', '17073.txt.txt', '35444.txt.txt', '55554.txt.txt', '45590.tx
t.txt', '28281.txt.txt', '13525.txt.txt', '13792.txt.txt', '52894.txt.txt', '15127.txt.txt', '25756.txt.txt', '43901.txt.txt', '49836.txt.txt', '57547.txt.txt', '36708.txt.txt', '48529.txt.txt', '53540.txt.txt', '55836.txt.txt', '10061.txt.txt']
Cluster 4 (521 documents): ['13172.txt.txt', '23609.txt.txt', '2439.txt.txt', '17326.txt.txt', '7001.txt.txt', '11464.txt.txt', '15699.txt.txt', '15955.txt.txt', '16831.txt.txt', '14866.txt.txt', '6812.txt.txt', '53527.txt.txt', '16528.txt.txt', '8390.txt.txt', '3029.txt.txt', '2376.txt.txt', '1961.txt.txt', '16399.txt.txt', '23428.txt.txt', '4511.txt.txt', '24708.txt.txt',
'14988.txt.txt', '17213.txt.txt', '13612.txt.txt', '2157.txt.txt', '13249.txt.txt', '1468.txt.txt', '754.txt.txt', '2096.txt.txt', '26490.txt.txt', '17845.txt.txt', '11952.txt.txt', '13941.txt.txt', '24951.txt.txt', '16598.txt.txt', '16531.txt.txt', '4278.txt.txt', '2442.txt.txt', '19289.txt.txt', '14672.txt.txt', '10444.txt.txt', '15262.txt.txt', '13656.txt.txt', '4362.txt.t
xt', '17774.txt.txt', '17188.txt.txt', '14975.txt.txt', '31511.txt.txt', '20014.txt.txt', '4386.txt.txt', '36496.txt.txt', '449.txt.txt', '28659.txt.txt', '17897.txt.txt', '14264.txt.txt', '20897.txt.txt', '22606.txt.txt', '4716.txt.txt', '10114.txt.txt', '25049.txt.txt', '13635.txt.txt', '27262.txt.txt', '17461.txt.txt', '26600.txt.txt', '2443.txt.txt', '14291.txt.txt', '134
82.txt.txt', '2832.txt.txt', '28677.txt.txt', '19243.txt.txt', '477.txt.txt', '3043.txt.txt', '45846.txt.txt', '30047.txt.txt', '12235.txt.txt', '31425.txt.txt', '7452.txt.txt', '14577.txt.txt', '4363.txt.txt', '23593.txt.txt', '624.txt.txt', '11102.txt.txt', '16467.txt.txt', '22607.txt.txt', '24700.txt.txt', '30755.txt.txt', '26981.txt.txt', '15561.txt.txt', '6167.txt.txt',
'99.txt.txt', '448.txt.txt', '60736.txt.txt', '12902.txt.txt', '18273.txt.txt', '19322.txt.txt', '24726.txt.txt', '17327.txt.txt', '22557.txt.txt', '14760.txt.txt', '1271.txt.txt', '22631.txt.txt', '6462.txt.txt', '22885.txt.txt', '10065.txt.txt', '20666.txt.txt', '20023.txt.txt', '5.txt.txt', '5321.txt.txt', '16983.txt.txt', '13888.txt.txt', '13831.txt.txt', '45303.txt.txt',
'28257.txt.txt', '2849.txt.txt', '16331.txt.txt', '22776.txt.txt', '17753.txt.txt', '852.txt.txt', '15776.txt.txt', '26377.txt.txt', '15255.txt.txt', '47730.txt.txt', '15803.txt.txt', '983.txt.txt', '5694.txt.txt', '11953.txt.txt', '19217.txt.txt', '28066.txt.txt', '13613.txt.txt', '17212.txt.txt', '13056.txt.txt', '31234.txt.txt', '6312.txt.txt', '28020.txt.txt', '14976.txt.
txt', '13858.txt.txt', '13376.txt.txt', '37530.txt.txt', '4361.txt.txt', '1302.txt.txt', '28297.txt.txt', '17579.txt.txt', '13893.txt.txt', '12400.txt.txt', '2529.txt.txt', '6605.txt.txt', '24062.txt.txt', '25568.txt.txt', '2162.txt.txt', '20715.txt.txt', '22994.txt.txt', '9104.txt.txt', '7959.txt.txt', '37839.txt.txt', '3743.txt.txt', '4705.txt.txt', '4908.txt.txt', '8112.tx
t.txt', '9408.txt.txt', '636.txt.txt', '626.txt.txt', '22677.txt.txt', '15479.txt.txt', '40780.txt.txt', '18845.txt.txt', '19719.txt.txt', '21840.txt.txt', '28039.txt.txt', '3800.txt.txt', '4210.txt.txt', '10738.txt.txt', '3775.txt.txt', '11951.txt.txt', '8438.txt.txt', '22716.txt.txt', '20987.txt.txt', '22153.txt.txt', '59.txt.txt', '14657.txt.txt', '13611.txt.txt', '22460.t
xt.txt', '22257.txt.txt', '1561.txt.txt', '22131.txt.txt', '24505.txt.txt', '3279.txt.txt', '16965.txt.txt', '20910.txt.txt', '53360.txt.txt', '17287.txt.txt', '47192.txt.txt', '2449.txt.txt', '10639.txt.txt', '10633.txt.txt', '17607.txt.txt', '17324.txt.txt', '4732.txt.txt', '16350.txt.txt', '13610.txt.txt', '13600.txt.txt', '5430.txt.txt', '28075.txt.txt', '1911.txt.txt', '20847.txt.txt',
'19172.txt.txt', '3300.txt.txt', '3310.txt.txt', '15106.txt.txt', '11352.txt.txt', '35514.txt.txt', '24654.txt.txt', '24001.txt.txt', '47091.txt.txt', '14977.txt.txt', '34856.txt.txt', '45456.txt.txt', '18269.txt.txt', '1497.txt.txt', '27827.txt.txt', '8567.txt.txt', '47132.txt.txt', '59573.txt.txt', '25254.txt.txt', '35333.txt.txt', '34901.txt.txt', '19317.txt.txt', '30756.t
xt.txt', '2330.txt.txt', '11538.txt.txt', '5637.txt.txt', '13407.txt.txt', '16702.txt.txt', '15478.txt.txt', '22345.txt.txt', '14825.txt.txt', '6841.txt.txt', '47449.txt.txt', '14426.txt.txt', '11376.txt.txt', '11378.txt.txt', '24518.txt.txt', '1404.txt.txt', '27341.txt.txt', '11114.txt.txt', '15663.txt.txt', '12625.txt.txt', '25828.txt.txt', '21077.txt.txt', '22430.txt.txt', '52819.txt.txt',
'31078.txt.txt', '34853.txt.txt', '36866.txt.txt', '19400.txt.txt', '36299.txt.txt', '24780.txt.txt', '19237.txt.txt', '28056.txt.txt', '28613.txt.txt', '50535.txt.txt', '14900.txt.txt', '130.txt.txt', '13300.txt.txt', '11560.txt.txt', '2176.txt.txt', '16363.txt.txt', '18794.txt.txt', '31529.txt.txt', '2445.txt.txt', '16728.txt.txt', '4602.txt.txt', '18281.txt.txt', '26556.tx
t.txt', '20842.txt.txt', '19211.txt.txt', '11955.txt.txt', '20580.txt.txt', '11689.txt.txt', '202.txt.txt', '22608.txt.txt', '1320.txt.txt', '4035.txt.txt', '14064.txt.txt', '14080.txt.txt', '20660.txt.txt', '22010.txt.txt', '2017.txt.txt', '14472.txt.txt', '24107.txt.txt', '22523.txt.txt', '12565.txt.txt', '17321.txt.txt', '7370.txt.txt', '12342.txt.txt', '35470.txt.txt', '1
2661.txt.txt', '20967.txt.txt', '27785.txt.txt', '15480.txt.txt', '28668.txt.txt', '18113.txt.txt', '28678.txt.txt', '60235.txt.txt', '17388.txt.txt', '13789.txt.txt', '5827.txt.txt', '41017.txt.txt', '11954.txt.txt', '19200.txt.txt', '19594.txt.txt', '12423.txt.txt', '13614.txt.txt', '21677.txt.txt', '19846.txt.txt', '14209.txt.txt', '4352.txt.txt', '7140.txt.txt', '13468.tx
t.txt', '15134.txt.txt', '21615.txt.txt', '18194.txt.txt', '97.txt.txt', '16960.txt.txt', '28458.txt.txt', '984.txt.txt', '27867.txt.txt', '19098.txt.txt', '4069.txt.txt', '19296.txt.txt', '28126.txt.txt', '23860.txt.txt', '20293.txt.txt', '15921.txt.txt', '36974.txt.txt', '24586.txt.txt', '8508.txt.txt', '18757.txt.txt', '47708.txt.txt', '1694.txt.txt', '25282.txt.txt', '195
60.txt.txt', '28669.txt.txt', '11029.txt.txt', '33333.txt.txt', '7524.txt.txt', '20248.txt.txt', '10366.txt.txt', '10000.txt.txt', '3253.txt.txt', '19192.txt.txt', '24461.txt.txt', '24519.txt.txt', '19699.txt.txt', '17306.txt.txt', '8855.txt.txt', '22323.txt.txt', '23680.txt.txt', '2900.txt.txt', '2099.txt.txt', '47109.txt.txt', '18087.txt.txt', '20521.txt.txt', '5268.txt.tx
t', '31671.txt.txt', '1232.txt.txt', '10611.txt.txt', '14378.txt.txt', '3207.txt.txt', '22832.txt.txt', '22822.txt.txt', '16494.txt.txt', '9173.txt.txt', '1951.txt.txt', '470.txt.txt', '26030.txt.txt', '3567.txt.txt', '20137.txt.txt', '1656.txt.txt', '14674.txt.txt', '29270.txt.txt', '2847.txt.txt', '3291.txt.txt', '15250.txt.txt', '22136.txt.txt', '18564.txt.txt', '201.txt.t
xt', '6933.txt.txt', '444.txt.txt', '11956.txt.txt', '21900.txt.txt', '5681.txt.txt', '28540.txt.txt', '16287.txt.txt', '2873.txt.txt', '19164.txt.txt', '17009.txt.txt', '26278.txt.txt', '10251.txt.txt', '15483.txt.txt', '18755.txt.txt', '4583.txt.txt', '17611.txt.txt', '10657.txt.txt', '12027.txt.txt', '1549.txt.txt', '7015.txt.txt', '8909.txt.txt', '17322.txt.txt', '11716.t
xt.txt', '17332.txt.txt', '22542.txt.txt', '16996.txt.txt', '19284.txt.txt', '6763.txt.txt', '51292.txt.txt', '21201.txt.txt', '14461.txt.txt', '27118.txt.txt', '42238.txt.txt', '5183.txt.txt', '20906.txt.txt', '11198.txt.txt', '18993.txt.txt', '26198.txt.txt', '8115.txt.txt', '18477.txt.txt', '15718.txt.txt', '16462.txt.txt', '14721.txt.txt', '24777.txt.txt', '26117.txt.tx
t', '20523.txt.txt', '22259.txt.txt', '3069.txt.txt', '20439.txt.txt', '2040.txt.txt', '10661.txt.txt', '2808.txt.txt', '360.txt.txt', '23692.txt.txt', '5255.txt.txt', '23033.txt.txt', '16653.txt.txt', '3034.txt.txt', '16643.txt.txt', '13065.txt.txt', '26912.txt.txt', '47703.txt.txt', '28672.txt.txt', '18880.txt.txt', '5039.txt.txt', '9841.txt.txt', '10163.txt.txt', '8910.tx
t.txt', '17253.txt.txt', '13642.txt.txt', '13568.txt.txt', '22352.txt.txt', '6888.txt.txt', '10574.txt.txt', '21091.txt.txt', '15293.txt.txt', '3580.txt.txt', '7960.txt.txt', '9662.txt.txt', '18127.txt.txt', '13029.txt.txt', '53453.txt.txt', '17087.txt.txt', '28601.txt.txt', '41479.txt.txt', '6603.txt.txt', '2447.txt.txt', '18932.txt.txt', '1989.txt.txt', '20220.txt.txt', '35
409.txt.txt', '2995.txt.txt', '2649.txt.txt', '5767.txt.txt', '16534.txt.txt', '16546.txt.txt', '4367.txt.txt', '21630.txt.txt', '10580.txt.txt', '28026.txt.txt', '28673.txt.txt', '4280.txt.txt', '4290.txt.txt', '200.txt.txt', '11015.txt.txt', '10477.txt.txt', '4341.txt.txt', '4351.txt.txt', '14555.txt.txt', '12019.txt.txt', '2130.txt.txt', '22700.txt.txt', '14004.txt.txt',
'10338.txt.txt', '27250.txt.txt', '38750.txt.txt', '22108.txt.txt', '13444.txt.txt', '11224.txt.txt', '1265.txt.txt', '21427.txt.txt', '54905.txt.txt', '13722.txt.txt', '17323.txt.txt', '30802.txt.txt', '18440.txt.txt', '26978.txt.txt', '24681.txt.txt', '17280.txt.txt', '418.txt.txt', '18843.txt.txt', '6762.txt.txt', '7300.txt.txt', '2526.txt.txt', '19285.txt.txt', '1.txt.tx
t', '1452.txt.txt']
```

In [45]:

```
import math
from wordcloud import WordCloud


cluster_texts = {i: "" for i in range(k)}
for text, cluster in zip(corpus, clusters):
    cluster_texts[cluster] += " " + text
```

```
cols = 2
rows = math.ceil(k / cols)

fig, axs = plt.subplots(rows, cols, figsize=(cols * 6, rows * 4))
axs = axs.flatten()

for i in range(k):

    wc = WordCloud(stopwords='english', background_color='white',
                   max_words=100, width=400, height=300)
    wc.generate(cluster_texts[i])
    axs[i].imshow(wc, interpolation='bilinear')
    axs[i].axis("off")
    axs[i].set_title(f"Cluster {i}")


for j in range(k, len(axs)):
    fig.delaxes(axs[j])

plt.suptitle("Word Clouds for Each Document Cluster", fontsize=16)
plt.tight_layout(rect=[0, 0.03, 1, 0.95])
plt.show()
```

In [46]:
```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation

print("Topic Modeling for each Cluster:")


for cluster_id, text in cluster_texts.items():
    print(f"\n--- Cluster {cluster_id} ---")


    paragraphs = [para.strip() for para in text.split("\n\n") if len(para.split()) > 20]

    if len(paragraphs) < 5:
        print("Not enough paragraphs for robust topic modeling. Skipping this cluster.")
        continue


    vectorizer_cluster = CountVectorizer(stop_words='english', max_df=0.95, min_df=2)
    dtm_cluster = vectorizer_cluster.fit_transform(paragraphs)


    n_topics_cluster = 2
    lda_cluster = LatentDirichletAllocation(n_components=n_topics_cluster, random_state=42)
    lda_cluster.fit(dtm_cluster)

    feature_names = vectorizer_cluster.get_feature_names_out()

    for topic_idx, topic in enumerate(lda_cluster.components_):
        top_words = [feature_names[i] for i in topic.argsort()[:-11:-1]]
        print(f"Topic {topic_idx}: {', '.join(top_words)}")
```

```
Topic Modeling for each Cluster:

--- Cluster 0 ---
Topic 0: die, der, und, den, zu, von, das, des, dem, sie
Topic 1: said, man, little, like, time, did, know, mr, good, old

--- Cluster 1 ---
Topic 0: thou, said, god, shall, thy, man, thee, king, did, lord
Topic 1: time, little, great, man, day, like, old, men, long, came

--- Cluster 2 ---
Topic 0: est, 000, party, years, male, female, 00, 15, km, president
Topic 1: di, time, water, great, en, little, la, like, work, small

--- Cluster 3 ---
Topic 0: et, la, le, les, que, il, en, des, qui, est
Topic 1: que, la, el, en, se, los, por, las, su, del

--- Cluster 4 ---
Topic 0: man, god, life, men, time, great, shall, good, world, things
Topic 1: great, time, war, general, king, men, new, government, country, years
```

In [49]:
```
!pip install gensim
!pip install usd-core
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: gensim in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (4.3.3)
Requirement already satisfied: numpy<2.0,>=1.18.5 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from gensim) (1.26.4)
Requirement already satisfied: scipy<1.14.0,>=1.7.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from gensim) (1.13.1)
Requirement already satisfied: smart-open>=1.8.1 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from gensim) (7.1.0)
Requirement already satisfied: wrapt in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from smart-open>=1.8.1->gensim) (1.17.2)
Defaulting to user installation because normal site-packages is not writeable
Collecting usd-core
  Downloading usd_core-25.2.post1-cp39-none-macosx_10_9_universal2.whl.metadata (1.6 kB)
Downloading usd_core-25.2.post1-cp39-none-macosx_10_9_universal2.whl (37.8 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 37.8/37.8 MB 2.4 MB/s eta 0:00:00
Installing collected packages: usd-core
Successfully installed usd-core-25.2.post1
```

In [51]:
```
!pip install nltk
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: nltk in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (3.9.1)
Requirement already satisfied: click in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from nltk) (8.1.8)
Requirement already satisfied: joblib in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from nltk) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from nltk) (2024.11.6)
Requirement already satisfied: tqdm in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from nltk) (4.67.1)
```

In [55]:
```
!pip install --upgrade cython gensim
!pip install --force-reinstall gensim
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: cython in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (3.0.11)
Requirement already satisfied: gensim in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (4.3.3)
Requirement already satisfied: numpy<2.0,>=1.18.5 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from gensim) (1.26.4)
Requirement already satisfied: scipy<1.14.0,>=1.7.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from gensim) (1.13.1)
Requirement already satisfied: smart-open>=1.8.1 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from gensim) (7.1.0)
Requirement already satisfied: wrapt in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from smart-open>=1.8.1->gensim) (1.17.2)
Defaulting to user installation because normal site-packages is not writeable
Collecting gensim
  Using cached gensim-4.3.3-cp39-cp39-macosx_11_0_arm64.whl.metadata (8.3 kB)
Collecting numpy<2.0,>=1.18.5 (from gensim)
  Using cached numpy-1.26.4-cp39-cp39-macosx_11_0_arm64.whl.metadata (61 kB)
Collecting scipy<1.14.0,>=1.7.0 (from gensim)
  Using cached scipy-1.13.1-cp39-cp39-macosx_12_0_arm64.whl.metadata (60 kB)
Collecting smart-open>=1.8.1 (from gensim)
  Using cached smart_open-7.1.0-py3-none-any.whl.metadata (24 kB)
Collecting wrapt (from smart-open>=1.8.1->gensim)
  Using cached wrapt-1.17.2-cp39-cp39-macosx_11_0_arm64.whl.metadata (6.4 kB)
Using cached gensim-4.3.3-cp39-cp39-macosx_11_0_arm64.whl (24.0 MB)
Using cached numpy-1.26.4-cp39-cp39-macosx_11_0_arm64.whl (14.0 MB)
Using cached scipy-1.13.1-cp39-cp39-macosx_12_0_arm64.whl (30.3 MB)
Using cached smart_open-7.1.0-py3-none-any.whl (61 kB)
Using cached wrapt-1.17.2-cp39-cp39-macosx_11_0_arm64.whl (38 kB)
Installing collected packages: wrapt, numpy, smart-open, scipy, gensim
  Attempting uninstall: wrapt
    Found existing installation: wrapt 1.17.2
    Uninstalling wrapt-1.17.2:
      Successfully uninstalled wrapt-1.17.2
  Attempting uninstall: numpy
    Found existing installation: numpy 1.26.4
    Uninstalling numpy-1.26.4:
      Successfully uninstalled numpy-1.26.4
  Attempting uninstall: smart-open
    Found existing installation: smart-open 7.1.0
    Uninstalling smart-open-7.1.0:
      Successfully uninstalled smart-open-7.1.0
  Attempting uninstall: scipy
    Found existing installation: scipy 1.13.1
    Uninstalling scipy-1.13.1:
      Successfully uninstalled scipy-1.13.1
  Attempting uninstall: gensim
    Found existing installation: gensim 4.3.3
    Uninstalling gensim-4.3.3:
      Successfully uninstalled gensim-4.3.3
Successfully installed gensim-4.3.3 numpy-1.26.4 scipy-1.13.1 smart-open-7.1.0 wrapt-1.17.2
```

In [63]:
```
!pip install --upgrade pip setuptools wheel
!pip install gensim==4.3.0
```

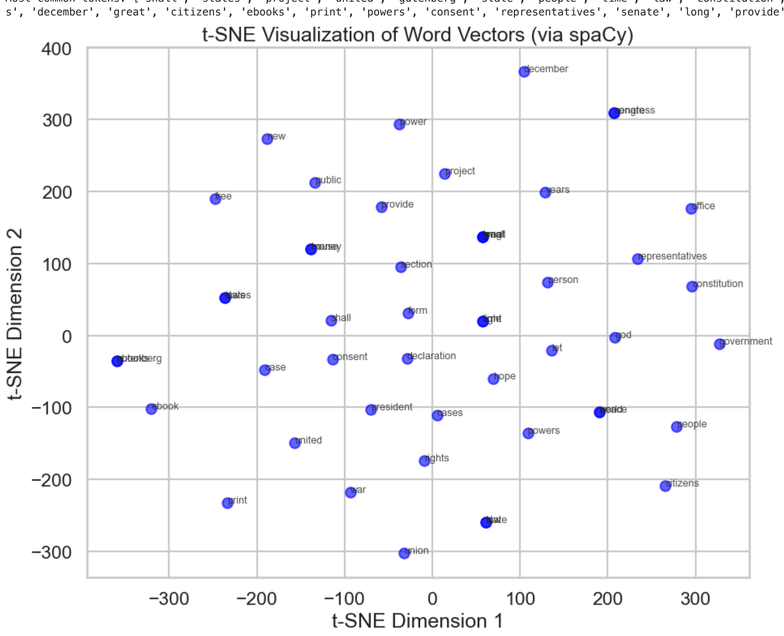
```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: pip in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (25.0)
Requirement already satisfied: setuptools in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (75.8.0)
Requirement already satisfied: wheel in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (0.45.1)
Defaulting to user installation because normal site-packages is not writeable
Collecting gensim==4.3.0
  Downloading gensim-4.3.0.tar.gz (23.3 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 23.3/23.3 MB 2.3 MB/s eta 0:00:0000:0100:01
  Preparing metadata (setup.py) ... done
Requirement already satisfied: numpy>=1.18.5 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from gensim==4.3.0) (1.26.4)
Requirement already satisfied: scipy>=1.7.0 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from gensim==4.3.0) (1.13.1)
Requirement already satisfied: smart_open>=1.8.1 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from gensim==4.3.0) (7.1.0)
Collecting FuzzyTM>=0.4.0 (from gensim==4.3.0)
  Downloading FuzzyTM-2.0.9-py3-none-any.whl.metadata (7.9 kB)
Requirement already satisfied: pandas in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from FuzzyTM>=0.4.0->gensim==4.3.0) (2.2.3)
Collecting pyfume (from FuzzyTM>=0.4.0->gensim==4.3.0)
  Downloading pyFUME-0.3.4-py3-none-any.whl.metadata (9.7 kB)
Requirement already satisfied: wrapt in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from smart_open>=1.8.1->gensim==4.3.0) (1.17.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from pandas->FuzzyTM>=0.4.0->gensim==4.3.0) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from pandas->FuzzyTM>=0.4.0->gensim==4.3.0) (2024.2)
Requirement already satisfied: tzdata>=2022.7 in /Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages (from pandas->FuzzyTM>=0.4.0->gensim==4.3.0) (2024.2)
Collecting scipy>=1.7.0 (from gensim==4.3.0)
  Downloading scipy-1.10.1-cp39-cp39-macosx_12_0_arm64.whl.metadata (53 kB)
Collecting numpy>=1.18.5 (from gensim==4.3.0)
  Downloading numpy-1.24.4-cp39-cp39-macosx_11_0_arm64.whl.metadata (5.6 kB)
Collecting simpful==2.12.0 (from pyfume->FuzzyTM>=0.4.0->gensim==4.3.0)
  Downloading simpful-2.12.0-py3-none-any.whl.metadata (4.8 kB)
Collecting fst-pso==1.8.1 (from pyfume->FuzzyTM>=0.4.0->gensim==4.3.0)
  Downloading fst-pso-1.8.1.tar.gz (18 kB)
  Preparing metadata (setup.py) ... done
Collecting pandas (from FuzzyTM>=0.4.0->gensim==4.3.0)
  Downloading pandas-1.5.3-cp39-cp39-macosx_11_0_arm64.whl.metadata (11 kB)
Collecting miniful (from fst-pso==1.8.1->pyfume->FuzzyTM>=0.4.0->gensim==4.3.0)
  Downloading miniful-0.0.6.tar.gz (2.8 kB)
  Preparing metadata (setup.py) ... done
Requirement already satisfied: six>=1.5 in /Library/Developer/CommandLineTools/Library/Frameworks/Python3.framework/Versions/3.9/lib/python3.9/site-packages (from python-dateutil>=2.8.2->pandas->FuzzyTM>=0.4.0->gensim==4.3.0) (1.15.0)
Downloading FuzzyTM-2.0.9-py3-none-any.whl (31 kB)
Downloading pyFUME-0.3.4-py3-none-any.whl (60 kB)
Downloading numpy-1.24.4-cp39-cp39-macosx_11_0_arm64.whl (13.9 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.9/13.9 MB 3.8 MB/s eta 0:00:00a 0:00:01
Downloading scipy-1.10.1-cp39-cp39-macosx_12_0_arm64.whl (28.9 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 28.9/28.9 MB 6.1 MB/s eta 0:00:0000:0100:01
Downloading pandas-1.5.3-cp39-cp39-macosx_11_0_arm64.whl (11.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 11.0/11.0 MB 6.7 MB/s eta 0:00:00 0:00:01
Downloading simpful-2.12.0-py3-none-any.whl (24 kB)
Building wheels for collected packages: gensim, fst-pso, miniful
  Building wheel for gensim (setup.py) ... done
  Created wheel for gensim: filename=gensim-4.3.0-cp39-cp39-macosx_10_9_universal2.whl size=24457501 sha256=bd02df765d212a4a856c85b33fe883236359b773ef620d627b39b6d70071d13c
  Stored in directory: /Users/mmadhusudan/Library/Caches/pip/wheels/f4/88/4d/7bef8c2e7a9e0bd4d8882e33aea52c9c577a1f94a362290191
  Building wheel for fst-pso (setup.py) ... done
  Created wheel for fst-pso: filename=fst_pso-1.8.1-py3-none-any.whl size=20478 sha256=10c887fe80944bd599ac995626e85d575135d5d9da3404236c859de4e213d070
  Stored in directory: /Users/mmadhusudan/Library/Caches/pip/wheels/99/66/48/d7ce0c6927f6abf167bbcdee537affc7b92c03632f78028411
  Building wheel for miniful (setup.py) ... done
  Created wheel for miniful: filename=miniful-0.0.6-py3-none-any.whl size=3554 sha256=5cb4d85ba076b077a00aa957e03bc5bc242b2de31a296ac1f513f4765789318e
  Stored in directory: /Users/mmadhusudan/Library/Caches/pip/wheels/d9/c7/71/db1d4646d963b34c530667501d3d6f34c0825eaffae2f0f2cb
Successfully built gensim fst-pso miniful
Installing collected packages: numpy, scipy, pandas, simpful, miniful, fst-pso, pyfume, FuzzyTM, gensim
  Attempting uninstall: numpy
    Found existing installation: numpy 1.26.4
    Uninstalling numpy-1.26.4:
      Successfully uninstalled numpy-1.26.4
  Attempting uninstall: scipy
    Found existing installation: scipy 1.13.1
    Uninstalling scipy-1.13.1:
      Successfully uninstalled scipy-1.13.1
  Attempting uninstall: pandas
    Found existing installation: pandas 2.2.3
    Uninstalling pandas-2.2.3:
      Successfully uninstalled pandas-2.2.3
  Attempting uninstall: gensim
    Found existing installation: gensim 4.3.3
    Uninstalling gensim-4.3.3:
      Successfully uninstalled gensim-4.3.3
Successfully installed FuzzyTM-2.0.9 fst-pso-1.8.1 gensim-4.3.0 miniful-0.0.6 numpy-1.24.4 pandas-1.5.3 pyfume-0.3.4 scipy-1.10.1 simpful-2.12.0
```

In [71]:
```
!python3 -m spacy download en_core_web_md
```

```
9442.04s - pydevd: Sending message related to process being replaced timed-out after 5 seconds
/Users/mmadhusudan/Library/Python/3.9/lib/python/site-packages/urllib3/__init__.py:35: NotOpenSSLWarning: urllib3 v2 only supports OpenSSL 1.1.1+, currently the 'ssl' module is compiled with 'LibreSSL 2.8.3'. See: https://github.com/urllib3/urllib3/issues/3020
  warnings.warn(
Defaulting to user installation because normal site-packages is not writeable
Collecting en-core-web-md==3.8.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_md-3.8.0/en_core_web_md-3.8.0-py3-none-any.whl (33.5 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 33.5/33.5 MB 8.8 MB/s eta 0:00:00a 0:00:01
Installing collected packages: en-core-web-md
Successfully installed en-core-web-md-3.8.0
✔ Download and installation successful
You can now load the package via spacy.load('en_core_web_md')
```

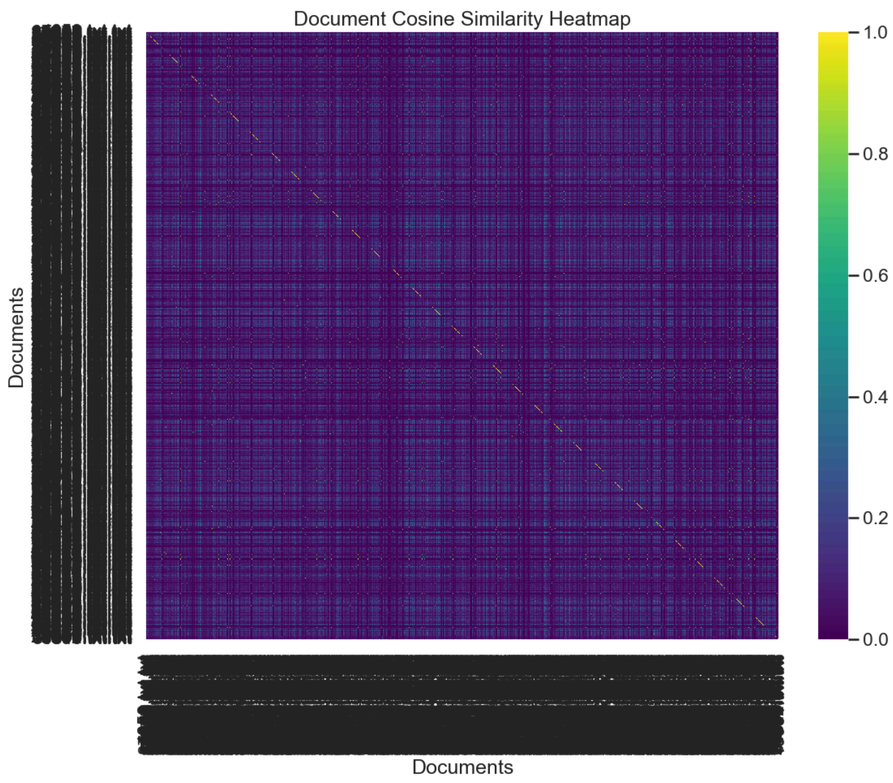
In [72]:
```
import spacy
from sklearn.manifold import TSNE
import matplotlib.pyplot as plt
import numpy as np
import nltk
from collections import Counter

nltk.download('punkt')

nlp = spacy.load("en_core_web_md")


doc = nlp(cleaned_text)


tokens = [token.text.lower() for token in doc if token.is_alpha and not token.is_stop and token.has_vector]

freq = Counter(tokens)


most_common_tokens = [word for word, count in freq.most_common(50)]
print("Most common tokens:", most_common_tokens)


word_vectors = np.array([nlp.vocab[word].vector for word in most_common_tokens])

tsne = TSNE(n_components=2, random_state=42)
tsne_results = tsne.fit_transform(word_vectors)

plt.figure(figsize=(10, 8))
plt.scatter(tsne_results[:, 0], tsne_results[:, 1], color='blue', alpha=0.6)
for i, word in enumerate(most_common_tokens):
    plt.annotate(word, (tsne_results[i, 0], tsne_results[i, 1]), fontsize=9, alpha=0.8)
plt.title("t-SNE Visualization of Word Vectors (via spaCy)")
plt.xlabel("t-SNE Dimension 1")
plt.ylabel("t-SNE Dimension 2")
plt.tight_layout()
plt.show()
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/mmadhusudan/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
Most common tokens: ['shall', 'states', 'project', 'united', 'gutenberg', 'state', 'people', 'time', 'law', 'constitution', 'laws', 'congress', 'government', 'president', 'right', 'new', 'war', 'public', 'house', 'union', 'free', 'power', 'ebook', 'person', 'let', 'section', 'years', 'cases', 'world', 'form', 'case', 'god', 'office', 'peace', 'hope', 'small', 'money', 'rights', 'december', 'great', 'citizens', 'ebooks', 'print', 'powers', 'consent', 'representatives', 'senate', 'long', 'provide', 'declaration']
```



In [73]:
```
import os
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

def get_cleaned_text(file_path):
    with open(file_path, "r", encoding="utf-8") as f:
        raw_text = f.read()
    # Use common markers to strip Gutenberg header and footer.
    start_marker = "*** START OF THIS PROJECT GUTENBERG EBOOK"
    end_marker = "*** END OF THIS PROJECT GUTENBERG EBOOK"
    start_idx = raw_text.find(start_marker)
    if start_idx != -1:
        text = raw_text[start_idx + len(start_marker):]
    else:
        text = raw_text
    end_idx = text.find(end_marker)
    if end_idx != -1:
        text = text[:end_idx]
    return text.strip()


folder = "Gutenberg_Books"
all_files = [f for f in os.listdir(folder) if f.endswith(".txt.txt")]


corpus = []
doc_names = []
for filename in all_files:
    file_path = os.path.join(folder, filename)
    try:
        text = get_cleaned_text(file_path)
        corpus.append(text)
        doc_names.append(filename)
    except Exception as e:
        print(f"Error processing {filename}: {e}")

print(f"Collected {len(corpus)} documents.")


vectorizer = TfidfVectorizer(stop_words='english', max_df=0.95, min_df=2)
tfidf_matrix = vectorizer.fit_transform(corpus)
print("TF-IDF matrix shape:", tfidf_matrix.shape)


cos_sim_matrix = cosine_similarity(tfidf_matrix)


df_sim = pd.DataFrame(cos_sim_matrix, index=doc_names, columns=doc_names)

# --- Visualization: Document Similarity Heatmap ---
plt.figure(figsize=(12, 10))
sns.heatmap(df_sim, cmap='viridis', xticklabels=True, yticklabels=True)
plt.title("Document Cosine Similarity Heatmap")
plt.xlabel("Documents")
plt.ylabel("Documents")
plt.tight_layout()
plt.show()
```

```
Collected 2475 documents.
TF-IDF matrix shape: (2475, 526829)
```



In [74]:
```
import nltk
from nltk.collocations import BigramCollocationFinder, BigramAssocMeasures
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


all_tokens = []
for doc in corpus:
    # Tokenize each document into words, convert to lowercase, and filter out non-alphabetic tokens.
    tokens = nltk.word_tokenize(doc.lower())
    tokens = [token for token in tokens if token.isalpha()]
    all_tokens.extend(tokens)

print(f"Total tokens aggregated from the corpus: {len(all_tokens)}")


bigram_measures = BigramAssocMeasures()
finder = BigramCollocationFinder.from_words(all_tokens)


finder.apply_freq_filter(3)


scored_bigrams = finder.score_ngrams(bigram_measures.pmi)
# Convert to DataFrame for easy handling
df_bigrams = pd.DataFrame(scored_bigrams, columns=["bigram", "PMI"]).sort_values(by="PMI", ascending=False).head(20)
print("Top 20 bigrams by PMI:")
display(df_bigrams)


df_bigrams["bigram_str"] = df_bigrams["bigram"].apply(lambda x: " ".join(x))

plt.figure(figsize=(10, 6))
sns.barplot(data=df_bigrams, x="PMI", y="bigram_str", palette="Blues_d")
plt.xlabel("PMI Score")
plt.ylabel("Bigram")
plt.title("Top 20 Bigrams by PMI in the Corpus")
plt.tight_layout()
plt.show()
```

```
Total tokens aggregated from the corpus: 209085770
Top 20 bigrams by PMI:
```

	bigram	PMI
0	(abgeruehrter, kugelhopf)	26.054557
104	(khandu, wangchuk)	26.054557
132	(napao, wetikoo)	26.054557
133	(ndeh, ntumazah)	26.054557
134	(nerbia, espartafilardo)	26.054557
135	(nikica, valentic)	26.054557
136	(nurzhan, subkhanberdin)	26.054557
137	(nuzas, rocabertis)	26.054557
138	(ochthodromus, wilsonius)	26.054557
139	(odjo, tankpinon)	26.054557
140	(ojasta, allikkoon)	26.054557
141	(olaudah, equiano)	26.054557
142	(orhan, ucok)	26.054557
143	(otinielu, tausi)	26.054557
144	(oudom, khattiya)	26.054557
145	(palafoxes, nuzas)	26.054557
146	(papeis, avulsos)	26.054557
147	(pastissons, giraumous)	26.054557
148	(paucás, hórás)	26.054557
149	(paucís, annís)	26.054557

```
/var/folders/7j/rv3w77nj6kb6kw_ssltcqpkr0000gp/T/ipykernel_22400/2160407253.py:37: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(data=df_bigrams, x="PMI", y="bigram_str", palette="Blues_d")
```