Data division

In [8]:
```
import pandas as pd
import os

file_path = '/Users/allisongarces/Downloads/Proyect AG/docu_word_categori


try:
    if os.path.exists(file_path):
        # Ajustar el delimitador a ';'
        df = pd.read_csv(file_path, encoding='utf-8', delimiter=';')  # D
        print(f"Base de datos cargada exitosamente. El DataFrame tiene {d
    else:
        print(f"Error: El archivo '{file_path}' no existe. Por favor veri
        df = None
except Exception as e:
    print(f"Error al cargar el archivo: {e}")
    df = None

if df is not None:
    print(f"Columnas disponibles: {list(df.columns)}")

   # Set chunk size (10,000 rows per chunk)
    chunk_size = 10000
    total_chunks = df.shape[0] // chunk_size + (1 if df.shape[0] % chunk_

    print(f"El DataFrame se dividirá en {total_chunks} fragmentos de {chu

    # Split the DataFrame into fragments
    chunks = [df[i:i + chunk_size] for i in range(0, df.shape[0], chunk_s

    # Create a folder to save the fragments
    output_folder = '/Users/allisongarces/Downloads/Proyect AG/chunks'
    os.makedirs(output_folder, exist_ok=True)

    # Save fragments to CSV files
    for i, chunk in enumerate(chunks):
        chunk_file = os.path.join(output_folder, f'chunk_{i + 1}.csv')
        try:
            chunk.to_csv(chunk_file, index=False, encoding='utf-8', sep='
            print(f"Fragmento {i + 1} guardado en: {chunk_file}")
        except Exception as e:
            print(f"Error al guardar el fragmento {i + 1}: {e}")

    print("Todos los fragmentos han sido procesados y guardados.")
else:
    print("No se pudo procesar el DataFrame porque no fue cargado correct
```

```
Base de datos cargada exitosamente. El DataFrame tiene 273287 filas y 3 co
lumnas.
```

```
Columnas disponibles: ['File Name', 'Category', 'Content']
El DataFrame se dividirá en 28 fragmentos de 10000 filas cada uno.
Fragmento 1 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunks/
chunk_1.csv
Fragmento 2 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunks/
chunk_2.csv
Fragmento 3 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunks/
chunk_3.csv
Fragmento 4 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunks/
chunk_4.csv
Fragmento 5 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunks/
chunk_5.csv
Fragmento 6 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunks/
chunk_6.csv
Fragmento 7 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunks/
chunk_7.csv
Fragmento 8 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunks/
chunk_8.csv
Fragmento 9 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunks/
chunk_9.csv
Fragmento 10 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_10.csv
Fragmento 11 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_11.csv
Fragmento 12 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_12.csv
Fragmento 13 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_13.csv
Fragmento 14 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_14.csv
Fragmento 15 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_15.csv
Fragmento 16 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_16.csv
Fragmento 17 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_17.csv
Fragmento 18 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_18.csv
Fragmento 19 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_19.csv
Fragmento 20 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_20.csv
Fragmento 21 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_21.csv
Fragmento 22 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_22.csv
Fragmento 23 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_23.csv
Fragmento 24 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_24.csv
Fragmento 25 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_25.csv
Fragmento 26 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
```

```
s/chunk_26.csv
Fragmento 27 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_27.csv
Fragmento 28 guardado en: /Users/allisongarces/Downloads/Proyect AG/chunk
s/chunk_28.csv
Todos los fragmentos han sido procesados y guardados.
```

General information about the data

In [57]:

```python
import pandas as pd

# Specify the file path for the CSV
file_path = '/Users/allisongarces/Downloads/Proyect AG/Info_Categories_Uz

# Read the CSV file into a pandas DataFrame using semicolon as delimiter
df = pd.read_csv(file_path, delimiter=';')

# Display the first few rows of the DataFrame to inspect the structure of
print("First few rows of the data:")
print(df.head())

# A. Category Validation and Data Refining

# Check for missing values in the relevant category columns
print("\nReviewing categories and missing values:")
category_columns = ['Category', 'Reason for Categorization', 'Text About
print(df[category_columns].isnull().sum())

# Fill missing values or errors in the categories with a default value or
df['Category'] = df['Category'].fillna('Unknown Category')
df['Reason for Categorization'] = df['Reason for Categorization'].fillna(
df['Text About Uzbekistan'] = df['Text About Uzbekistan'].fillna('No Info

# Ensure consistency in the categories
df['Category'] = df['Category'].str.strip().str.title()

# B. Refining the Text (removing unwanted characters)
df['Text About Uzbekistan'] = df['Text About Uzbekistan'].apply(lambda x:
df['Text About Uzbekistan'] = df['Text About Uzbekistan'].apply(lambda x:

# C. Checking for Outliers and Errors (Outliers and improbable values)
# Check for outliers in any available numeric columns (if applicable)
numeric_columns = ['Proyección Financiera (Estimación)']  # If there are
# If no such columns exist, remove or adjust this part based on your data

# D. Date Format Consistency (if any date columns are present)
# Assuming there are date columns to convert (adjust as necessary)
# df['Date'] = pd.to_datetime(df['Date'], errors='coerce')

# E. Ensuring Data is AI-Ready
# Clean any poorly formatted text
df['Text About Uzbekistan'] = df['Text About Uzbekistan'].apply(lambda x:
```

```
# F. Remove duplicate rows if any
df = df.drop_duplicates()

# Create Financial Projections Estimate (if relevant data exists, adjust
df['Proyección Financiera (Estimación)'] = df['Reason for Categorization'
    lambda x: 'Sí' if 'inversión' in str(x).lower() or 'financiamiento' i
)

# Display general statistics of the data after refining
print("\nGeneral statistics of the refined data:")
print(df.describe())

# Save the refined DataFrame to a new CSV file with semicolon delimiter
output_file = '/Users/allisongarces/Downloads/Proyect AG/Refined_Info_Cat
df.to_csv(output_file, index=False, sep=';')

print(f"\nRefined dataset saved to: {output_file}")
```

```
First few rows of the data:
                                        File Name  \
0         WEF_Reshaping_affordability_2024.docx
1                            smarsly2021e.docx
2                      UzbekistanRailways 2.docx
3  Swedish_Waste_Management_A_Review_Articl.docx
4                MOF_LSE_IFC_event_yfnGTW0.docx

                                           Category  \
0  Urban Affordability and Sustainable Urban Deve...
1  Technological Innovation in Structural Health ...
2  Infrastructure Development and Public-Private ...
3           Environmental Impacts and Climate Change
4            Sustainable Finance and Economic Trends

                         Reason for Categorization Text About Uzbekistan
\
0  This text addresses the global urban affordabi...                    NO
1  The text discusses advancements in the use of ...                    NO
2  This text discusses the construction of the Te...                    SI
3  This article focuses on waste management strat...                    NO
4  The text discusses Uzbekistan's economic growt...                    SI

  Simplified Category
0               Urban
1          Technology
2      Infrastructure
3       Environmental
4             Finance

Reviewing categories and missing values:
Category                     0
Reason for Categorization    0
Text About Uzbekistan        0
dtype: int64
```

```
General statistics of the refined data:
                        File Name  \
count                         212
unique                        211
top     cbi_mr_h1_2024_02e_1.docx
freq                            2

                                                  Category  \
count                                                  212
unique                                                  41
top     Infrastructure Development And Public-Private ...
freq                                                    59

                                Reason for Categorization  \
count                                                 212
unique                                                212
top     This text addresses the global urban affordabi...
freq                                                    1

       Text About Uzbekistan Simplified Category  \
count                    212                 212
unique                     2                  11
top                       NO      Infrastructure
freq                     143                  63

       Proyección Financiera (Estimación)
count                                 212
unique                                  1
top                                    No
freq                                  212

Refined dataset saved to: /Users/allisongarces/Downloads/Proyect AG/Refine
d_Info_Categories_Uzbekistan_Analysis.csv
```

In [59]:

```python
import pandas as pd

# Path to the CSV file
file_path = "/Users/allisongarces/Downloads/Proyect AG/Info_Categories_Uz

# Load the CSV file into a DataFrame (adjust delimiter if necessary)
df = pd.read_csv(file_path, delimiter=';')  # Use ';' if it's a semicolon

# Display the first few rows to ensure the data is loaded correctly
print(df.head())
```

```
                                      File Name  \
0         WEF_Reshaping_affordability_2024.docx
1                             smarsly2021e.docx
2                      UzbekistanRailways 2.docx
3  Swedish_Waste_Management_A_Review_Articl.docx
4                MOF_LSE_IFC_event_yfnGTW0.docx

                                           Category  \
0  Urban Affordability and Sustainable Urban Deve...
1  Technological Innovation in Structural Health ...
2  Infrastructure Development and Public-Private ...
3           Environmental Impacts and Climate Change
4            Sustainable Finance and Economic Trends

                          Reason for Categorization Text About Uzbekistan
\
0  This text addresses the global urban affordabi...                    NO
1  The text discusses advancements in the use of ...                    NO
2  This text discusses the construction of the Te...                    SI
3  This article focuses on waste management strat...                    NO
4  The text discusses Uzbekistan's economic growt...                    SI

  Simplified Category
0               Urban
1          Technology
2      Infrastructure
3       Environmental
4             Finance
```

Preliminary Exploration

Code explanation:

1. Data loading: The data file is imported in Excel format using pd.read_excel.
2. Initial exploration:
    - The first few rows of the data frame are displayed (df.head()), giving a quick overview of the data.
    - A summary is obtained using df.info(), which provides details about the columns and their data types.
    - Null values per column are calculated using df.isnull().sum() to ensure that there is no missing data that may require treatment.
    - A statistical summary is obtained using df.describe(), to see the distribution of numerical values in the data set.
3. Category analysis:
    - The distribution of categories is calculated using value_counts() and visualized with a bar chart.
4. Analysis of mentions of Uzbekistan:
    - We check how many documents contain the word "Uzbekistan" in the Text

About Uzbekistan column and visualize this count.

In [22]:
```
import pandas as pd
import matplotlib.pyplot as plt

file_path = "/Users/allisongarces/Downloads/Proyect AG/Info_Categories_Uz
df = pd.read_csv(file_path, delimiter=';')

# Display the first rows of the dataframe
print("First rows of the dataframe:")
df.head()
```

```
First rows of the dataframe:
```

Out[22]:

	File Name	Category	Reason for Categorization	T U
0	WEF_Reshaping_affordability_2024.docx	Urban Affordability and Sustainable Urban Deve...	This text addresses the global urban affordabi...	
1	smarsly2021e.docx	Technological Innovation in Structural Health ...	The text discusses advancements in the use of ...	
2	UzbekistanRailways 2.docx	Infrastructure Development and Public-Private ...	This text discusses the construction of the Te...	
3	Swedish_Waste_Management_A_Review_Articl.docx	Environmental Impacts and Climate Change	This article focuses on waste management strat...	
4	MOF_LSE_IFC_event_yfnGTW0.docx	Sustainable Finance and Economic Trends	The text discusses Uzbekistan's economic growt...	

In [24]:
```
# Get general information about the dataframe
print("\nGeneral information about the data:")
print(df.info())
```

```
General information about the data:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 213 entries, 0 to 212
Data columns (total 5 columns):
 #   Column                     Non-Null Count  Dtype 
---  ------                     --------------  ----- 
 0   File Name                  213 non-null    object
 1   Category                   213 non-null    object
 2   Reason for Categorization  213 non-null    object
 3   Text About Uzbekistan      213 non-null    object
 4   Simplified Category        213 non-null    object
dtypes: object(5)
memory usage: 8.4+ KB
None
```

In [28]:
```
# Checking for null values in each column
print("\nNull values in each column:")
print(df.isnull().sum())
```

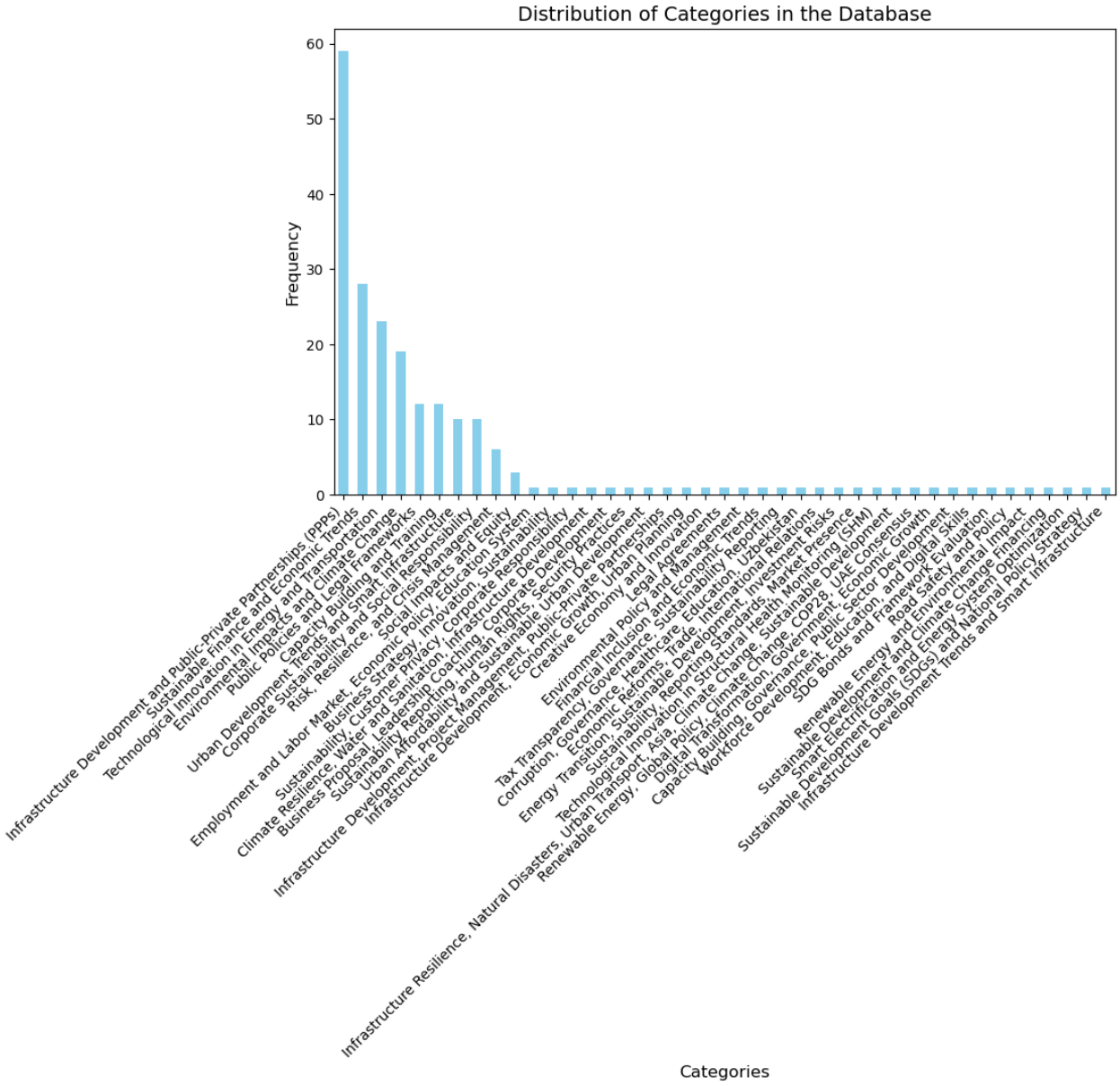
```
Null values in each column:
File Name                    0
Category                     0
Reason for Categorization    0
Text About Uzbekistan        0
Simplified Category          0
dtype: int64
```

In [30]:
```
# Statistical summary of numerical columns
print("\nStatistical summary:")
print(df.describe())
```

```
Statistical summary:
                                        File Name  \
count                                         213
unique                                        211
top     GRI 414: Supplier Social Assessment 2016
freq                                            2

                                                Category  \
count                                                213
unique                                                41
top     Infrastructure Development and Public-Private ...
freq                                                  59

                                  Reason for Categorization  \
count                                                   213
unique                                                  212
top     This standard provides guidelines for assessin...
freq                                                      2

       Text About Uzbekistan Simplified Category
count                    213                 213
unique                     2                  11
top                       NO      Infrastructure
freq                     144                  63
```

In [34]:
```
# Analysis of the distribution of categories (Category)
category_counts = df['Category'].value_counts()

# Viewing the distribution of categories
plt.figure(figsize=(10, 6))
category_counts.plot(kind='bar', color='skyblue')
plt.title('Distribution of Categories in the Database', fontsize=14)
plt.xlabel('Categories', fontsize=12)
plt.ylabel('Frequency', fontsize=12)
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

```
/var/folders/r4/bfc17knx7fl8dv9hgcf92n_w0000gn/T/ipykernel_44284/339680170
0.py:11: UserWarning: Tight layout not applied. The bottom and top margins
cannot be made large enough to accommodate all Axes decorations.
  plt.tight_layout()
```



In [40]:
```
category_counts
```

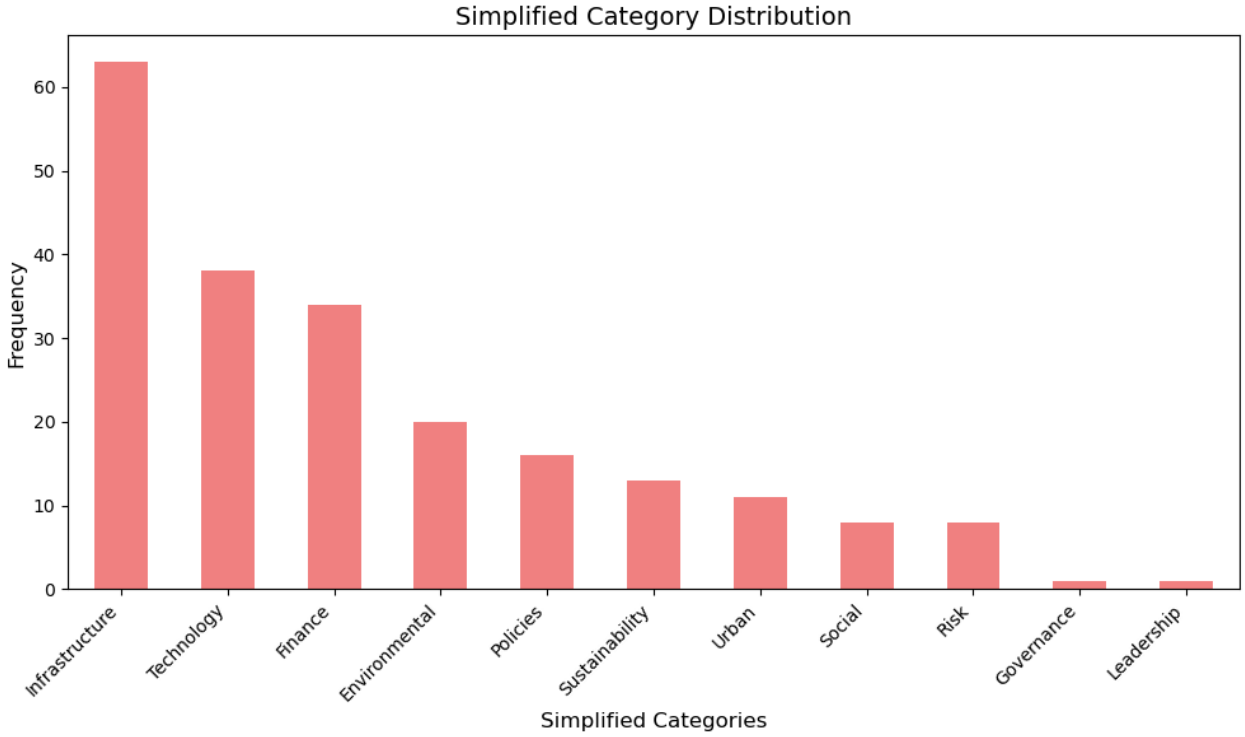
Out[40]:
```
Category
Infrastructure Development and Public-Private Partnerships (PPPs)
59
Sustainable Finance and Economic Trends
28
Technological Innovation in Energy and Transportation
23
Environmental Impacts and Climate Change
19
Public Policies and Legal Frameworks
12
Capacity Building and Training
12
Urban Development Trends and Smart Infrastructure
10
Corporate Sustainability and Social Responsibility
10
```

Risk, Resilience, and Crisis Management
6
Social Impacts and Equity
3
Employment and Labor Market, Economic Policy, Education System
1
Business Strategy, Innovation, Sustainability
1
Sustainability, Customer Privacy, Corporate Responsibility
1
Climate Resilience, Water and Sanitation, Infrastructure Development
1
Business Proposal, Leadership Coaching, Corporate Development
1
Sustainability Reporting, Human Rights, Security Practices
1
Urban Affordability and Sustainable Urban Development
1
Infrastructure Development, Project Management, Public–Private Partnersh
ips 1
Infrastructure Development, Economic Growth, Urban Planning
1
Creative Economy and Innovation
1
Legal Agreements
1
Environmental Policy and Management
1
Financial Inclusion and Economic Trends
1
Tax Transparency, Governance, Sustainability Reporting
1
Corruption, Governance, Healthcare, Education, Uzbekistan
1
Economic Reforms, Trade, International Relations
1
Energy Transition, Sustainable Development, Investment Risks
1
Sustainability, Reporting Standards, Market Presence
1
Technological Innovation in Structural Health Monitoring (SHM)
1
Infrastructure Resilience, Natural Disasters, Urban Transport, Asia, Cli
mate Change, Sustainable Development 1
Renewable Energy, Global Policy, Climate Change, COP28, UAE Consensus
1
Digital Transformation, Government, Economic Growth
1
Capacity Building, Governance, Public Sector Development
1
Workforce Development, Education, and Digital Skills
1
SDG Bonds and Framework Evaluation

```
1
Road Safety and Policy
1
Renewable Energy and Environmental Impact
1
Sustainable Development and Climate Change Financing
1
Smart Electrification and Energy System Optimization
1
Sustainable Development Goals (SDGs) and National Policy Strategy
1
Infrastructure Development Trends and Smart Infrastructure
1
Name: count, dtype: int64
```

In [ ]:

In [36]:
```
# Analysis of the distribution of simplified categories (Simplified Categ
simplified_category_counts = df['Simplified Category'].value_counts()

# Visualization of the distribution of simplified categories
plt.figure(figsize=(10, 6))
simplified_category_counts.plot(kind='bar', color='lightcoral')
plt.title('Simplified Category Distribution', fontsize=14)
plt.xlabel('Simplified Categories', fontsize=12)
plt.ylabel('Frequency', fontsize=12)
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

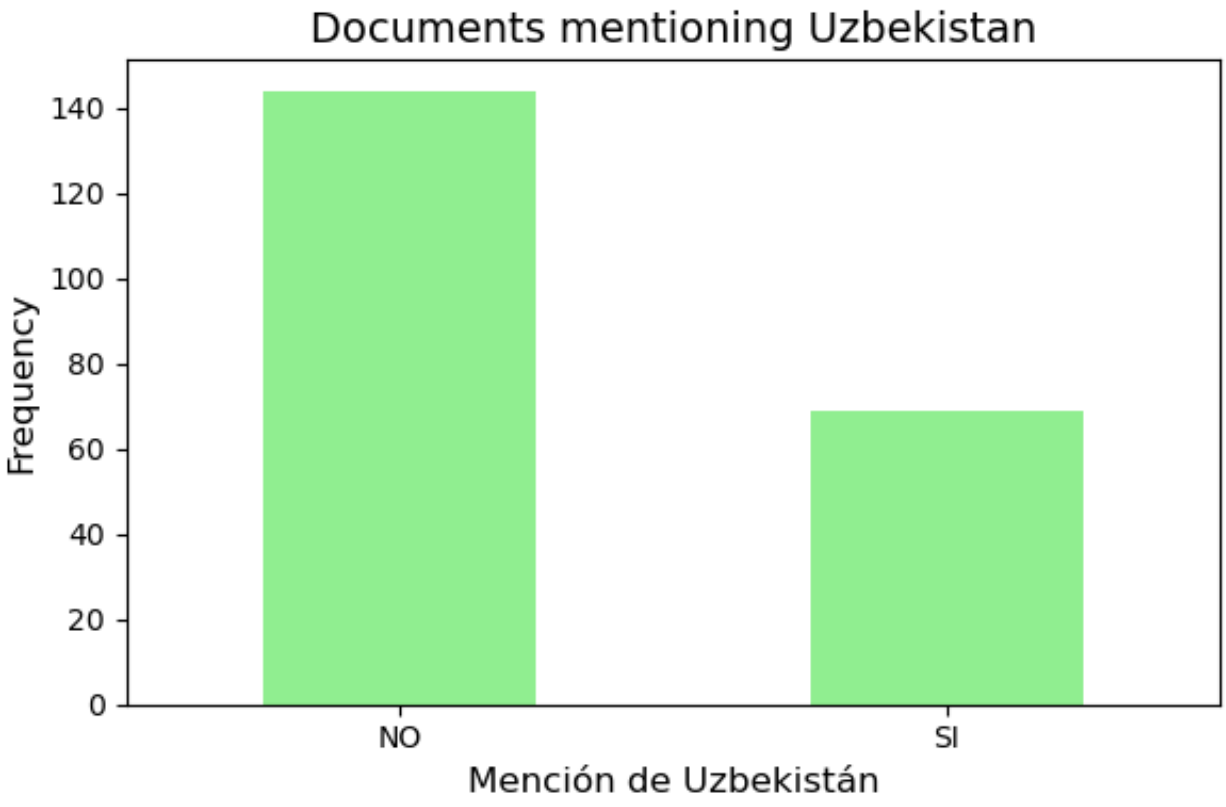


In [ ]:

In [38]:
```
simplified_category_counts
```

Out[38]:
```
Simplified Category
Infrastructure    63
Technology        38
Finance           34
Environmental     20
Policies          16
Sustainability    13
Urban             11
Social             8
Risk               8
Governance         1
Leadership         1
Name: count, dtype: int64
```

In [50]:
```
# Count the number of documents that speak about Uzbekistan
uzbekistan_text_count = df['Text About Uzbekistan'].value_counts()

# View how many documents mention Uzbekistan
plt.figure(figsize=(6, 4))
uzbekistan_text_count.plot(kind='bar', color='lightgreen')
plt.title('Documents mentioning Uzbekistan', fontsize=14)
plt.xlabel('Mención de Uzbekistán', fontsize=12)
plt.ylabel('Frequency', fontsize=12)
plt.xticks(rotation=0)
plt.tight_layout()
plt.show()
```



In [52]:
```
# Print the count of documents mentioning Uzbekistan
```

```
print("\nCount of documents mentioning Uzbekistan:")
print(uzbekistan_text_count)
```

```
Count of documents mentioning Uzbekistan:
Text About Uzbekistan
NO    144
SI     69
Name: count, dtype: int64
```

Category Validation

The Category column contains a variety of topics, but some categories within the Simplified Category may not be fully aligned with the main category. It is essential to perform a manual or semi-automatic validation of these categories to ensure that documents are correctly classified and can be efficiently used in subsequent analysis.

In [65]:
```
# Category Validation: Compare 'Category' and 'Simplified Category'
# Identify documents where categories do not match between 'Category' and
invalid_category_rows = df[df['Category'] != df['Simplified Category']]

# See how many rows have inconsistencies
print(f"Total documents with inconsistent categories: {invalid_category_r
```

```
Total documents with inconsistent categories: 213
```

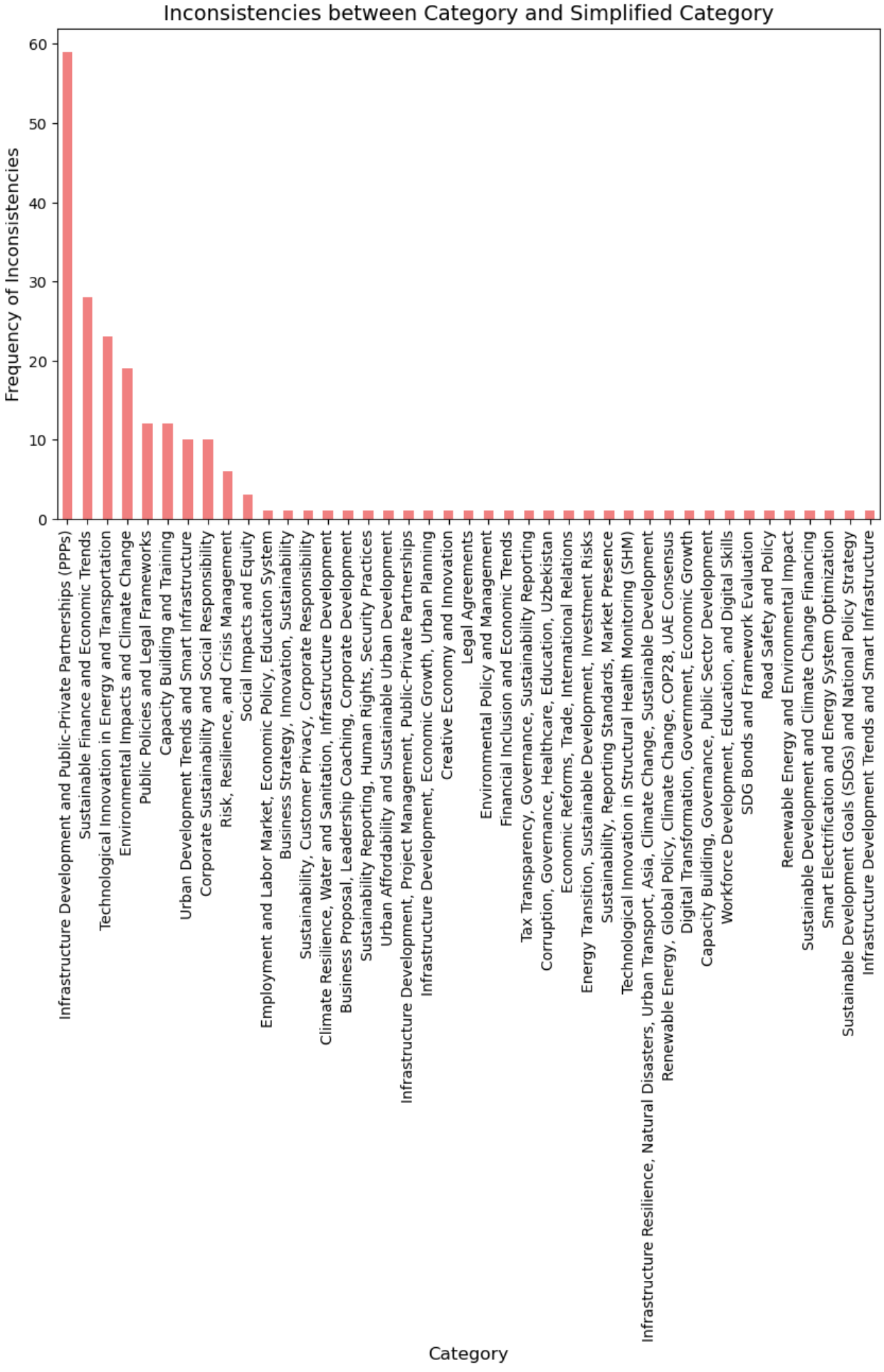
In [ ]:
```

```

In [67]:
```
# Print rows with inconsistencies for manual review
print("\nRows with inconsistent categoriesRows with inconsistent categori
print(invalid_category_rows[['File Name', 'Category', 'Simplified Categor
```

```
Rows with inconsistent categories:
                                             File Name  \
0                WEF_Reshaping_affordability_2024.docx
1                                     smarsly2021e.docx
2                            UzbekistanRailways 2.docx
3         Swedish_Waste_Management_A_Review_Articl.docx
4                       MOF_LSE_IFC_event_yfnGTW0.docx
..                                                 ...
208  Comprehensive Proposal for Toybola and Anvar A...
209                            cbi_mr_h1_2024_02e_1.docx
210                Investments and construction 2.docx
211  Dubai-Municipality-3DCP-Guideline-1st-Edition-...
212  3D_Concrete_Printing_White_Paper_Ubez 2024-202...

                                              Category Simplified Category
0    Urban Affordability and Sustainable Urban Deve...               Urban
1    Technological Innovation in Structural Health ...          Technology
2    Infrastructure Development and Public-Private ...      Infrastructure
3              Environmental Impacts and Climate Change       Environmental
4               Sustainable Finance and Economic Trends             Finance
..                                                 ...                 ...
208  Corporate Sustainability and Social Responsibi...      Sustainability
209             Sustainable Finance and Economic Trends             Finance
210  Infrastructure Development and Public-Private ...      Infrastructure
211  Infrastructure Development and Public-Private ...      Infrastructure
212  Infrastructure Development and Public-Private ...      Infrastructure

[213 rows x 3 columns]
```

In [89]:
```
# Visualize the distribution of inconsistencies by category
category_inconsistency_count = invalid_category_rows['Category'].value_co

# Display a bar chart to visualize the distribution of inconsistencies
plt.figure(figsize=(10, 6))
category_inconsistency_count.plot(kind='bar', color='lightcoral')
plt.title('Inconsistencies between Category and Simplified Category', fon
plt.xlabel('Category', fontsize=12)
plt.ylabel('Frequency of Inconsistencies', fontsize=12)
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()
```

```
/var/folders/r4/bfc17knx7fl8dv9hgcf92n_w0000gn/T/ipykernel_44284/389874347
5.py:11: UserWarning: Tight layout not applied. The bottom and top margins
cannot be made large enough to accommodate all Axes decorations.
  plt.tight_layout()
```


Inconsistencies between Category and Simplified Category
Frequency of Inconsistencies
0
10
20
30
40
50
60
Infrastructure Development and Public-Private Partnerships (PPPs)
Sustainable Finance and Economic Trends
Technological Innovation in Energy and Transportation
Environmental Impacts and Climate Change
Public Policies and Legal Frameworks
Capacity Building and Training
Urban Development Trends and Smart Infrastructure
Corporate Sustainability and Social Responsibility
Risk, Resilience, and Crisis Management
Social Impacts and Equity
Employment and Labor Market, Economic Policy, Education System
Business Strategy, Innovation, Sustainability
Sustainability, Customer Privacy, Corporate Responsibility
Climate Resilience, Water and Sanitation, Infrastructure Development
Business Proposal, Leadership Coaching, Corporate Development
Sustainability Reporting, Human Rights, Security Practices
Urban Affordability and Sustainable Urban Development
Infrastructure Development, Project Management, Public-Private Partnerships
Infrastructure Development, Economic Growth, Urban Planning
Creative Economy and Innovation
Legal Agreements
Environmental Policy and Management
Financial Inclusion and Economic Trends
Tax Transparency, Governance, Sustainability Reporting
Corruption, Governance, Healthcare, Education, Uzbekistan
Economic Reforms, Trade, International Relations
Energy Transition, Sustainable Development, Investment Risks
Sustainability, Reporting Standards, Market Presence
Technological Innovation in Structural Health Monitoring (SHM)
Infrastructure Resilience, Natural Disasters, Urban Transport, Asia, Climate Change, Sustainable Development
Renewable Energy, Global Policy, Climate Change, COP28, UAE Consensus
Digital Transformation, Government, Economic Growth
Capacity Building, Governance, Public Sector Development
Workforce Development, Education, and Digital Skills
SDG Bonds and Framework Evaluation
Road Safety and Policy
Renewable Energy and Environmental Impact
Sustainable Development and Climate Change Financing
Smart Electrification and Energy System Optimization
Sustainable Development Goals (SDGs) and National Policy Strategy
Infrastructure Development Trends and Smart Infrastructure
Category

In [97]:
```
# Frequency analysis of categories
category_count = df['Category'].value_counts()
simplified_category_count = df['Simplified Category'].value_counts()

# Print frequency results
print("Frequency of main categories:")
print(category_count)

print("\nFrequency of simplified categories:")
print(simplified_category_count)
```

```
Frequency of main categories:
Category
Infrastructure Development and Public-Private Partnerships (PPPs)
59
Sustainable Finance and Economic Trends
28
Technological Innovation in Energy and Transportation
23
Environmental Impacts and Climate Change
19
Public Policies and Legal Frameworks
12
Capacity Building and Training
12
Urban Development Trends and Smart Infrastructure
10
Corporate Sustainability and Social Responsibility
10
Risk, Resilience, and Crisis Management
6
Social Impacts and Equity
3
Employment and Labor Market, Economic Policy, Education System
1
Business Strategy, Innovation, Sustainability
1
Sustainability, Customer Privacy, Corporate Responsibility
1
Climate Resilience, Water and Sanitation, Infrastructure Development
1
Business Proposal, Leadership Coaching, Corporate Development
1
Sustainability Reporting, Human Rights, Security Practices
1
Urban Affordability and Sustainable Urban Development
1
Infrastructure Development, Project Management, Public-Private Partnership
s                                                 1
Infrastructure Development, Economic Growth, Urban Planning
1
Creative Economy and Innovation
1
```

```
Legal Agreements
1
Environmental Policy and Management
1
Financial Inclusion and Economic Trends
1
Tax Transparency, Governance, Sustainability Reporting
1
Corruption, Governance, Healthcare, Education, Uzbekistan
1
Economic Reforms, Trade, International Relations
1
Energy Transition, Sustainable Development, Investment Risks
1
Sustainability, Reporting Standards, Market Presence
1
Technological Innovation in Structural Health Monitoring (SHM)
1
Infrastructure Resilience, Natural Disasters, Urban Transport, Asia, Clima
te Change, Sustainable Development     1
Renewable Energy, Global Policy, Climate Change, COP28, UAE Consensus
1
Digital Transformation, Government, Economic Growth
1
Capacity Building, Governance, Public Sector Development
1
Workforce Development, Education, and Digital Skills
1
SDG Bonds and Framework Evaluation
1
Road Safety and Policy
1
Renewable Energy and Environmental Impact
1
Sustainable Development and Climate Change Financing
1
Smart Electrification and Energy System Optimization
1
Sustainable Development Goals (SDGs) and National Policy Strategy
1
Infrastructure Development Trends and Smart Infrastructure
1
Name: count, dtype: int64

Frequency of simplified categories:
Simplified Category
Infrastructure    63
Technology        38
Finance           34
Environmental     20
Policies          16
Sustainability    13
Urban             11
```

```
Social              8
Risk                8
Governance          1
Leadership          1
Name: count, dtype: int64
```

Manual Review:

Rows with inconsistencies are identified and values are manually corrected, ensuring that documents are correctly classified for further analysis.

Data Analysis

- Category Distribution: A frequency analysis will be conducted to understand the distribution of categories across the dataset. This will help identify the predominant areas being documented and analyzed in relation to Uzbekistan. The most frequent categories include Infrastructure (59 times), Sustainable Finance (28 times), Climate Change and Environment (19 times), Urban Development (10 times), among others.
- Content Trends: In the Text About Uzbekistan field, key topics such as economic development, government reforms, sustainability, renewable energy, and social challenges are addressed. This analysis will be crucial for identifying the priority areas in Uzbekistan's current context, which may be of interest to both the government and international stakeholders.

In [113…

```
from wordcloud import WordCloud

# Word frequency analysis in the 'Reason for Categorization' column using
text = " ".join(df['Reason for Categorization'].dropna())

wordcloud = WordCloud(width=800, height=400, background_color='white').ge

plt.figure(figsize=(10, 6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')  # Ocultar los ejes
plt.title('Content Trends: Keywords in Documents about Uzbekistan', fonts
plt.show()

# Key terms and recurring themes related to Uzbekistan in documents.
```

The purpose of the code is to visualize recurring themes in Uzbekistan-related texts through a word cloud. This allows you to quickly identify concepts and keywords that frequently appear in documents, which is useful for understanding the main focus areas (e.g. infrastructure, economic reforms, sustainability, etc.).

The word cloud provides an intuitive and quick representation of important trends and themes in Uzbekistan-related documents.

In [127…

```
# 4.3 Identifying Recurring Themes in Texts about Uzbekistan
# Filter texts containing key terms
keywords = ['economic development', 'government reforms', 'sustainability
filtered_texts = df[df['Reason for Categorization'].str.contains('|'.join

# Imprimir los documentos relevantes con los términos clave
print("\nRelevant documents for key topics:")
print(filtered_texts[['File Name', 'Category', 'Text About Uzbekistan']])
```

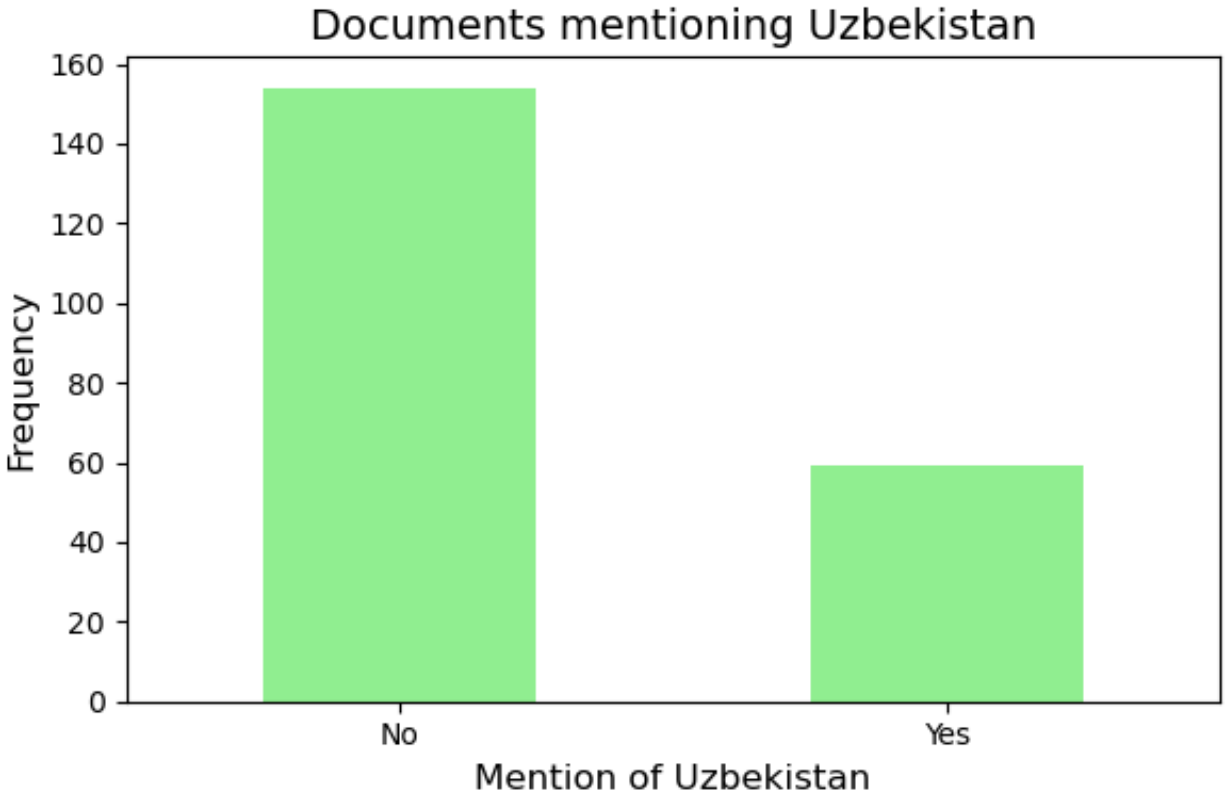
```
Relevant documents for key topics:
                                            File Name  \
3        Swedish_Waste_Management_A_Review_Articl.docx
5                          Annex D Detailed Report.docx
12                                  Wind Energy.docx
14              Second-Party Opinion on SDG Bonds.docx
17   Comprehensive Knowledge Base for Sustainable a...
27   IRENA_UAE_Consensus_2030_tripling_renewables_d...
30                 GRI 202_ Market Presence 2016.docx
31       IRENA_G20_Just transition_in_EMDEs_2024.docx
35   🌐 Business Model Innovation - A Game Changer.docx
36                GRI 418_ Customer Privacy 2016.docx
39              GRI 410_ Security Practices 2016.docx
41     Modified Fact Sheet for Strategic Partners.docx
42                           GRI 306: Waste 2020.docx
```

```
46          AP1790680f452f10ba0a34c06922a1df0003.docx
55   Perspectives_on_recycling_centres_and_future_d...
66          ME's presentation to UNECE and ESCAP.docx
70                           002-article-A003-en.docx
104                       1uzbea2024001-print-pdf.docx
105      IDU1c22e2ec41eee514d2d19dfb115b648d7fa6f.docx
106  WEF_Implementing_a_Life_Cycle_Approach_to_Infr...
107                   GRI Standards Glossary 2022.docx
119  Comprehensive Project Proposal for Infrastruct...
125  IRENA_Renewable_power_generation_costs_in_2023...
139  Master List of Tools and Interactive Materials...
144  WEF_Smart_at_Scale_Cities_to_Watch_25_Case_Stu...
168                   WB-GSS-Bonds-Survey-Report.docx
169                                cbi_sotm23_02h.docx
176                                            RUz.docx
182  IRENA_Ranking_critical_materials_for_the_energ...
186                                     ADBI-WP993.docx
200                          cbi_mr_h1_2024_02e_1.docx
209                          cbi_mr_h1_2024_02e_1.docx
212  3D_Concrete_Printing_White_Paper_Ubez 2024-202...

                                                Category Text About Uzbekist
an
3              Environmental Impacts and Climate Change
NO
5    Technological Innovation in Energy and Transpo...
SI
12          Renewable Energy and Environmental Impact
NO
14                SDG Bonds and Framework Evaluation
SI
17   Urban Development Trends and Smart Infrastructure
NO
27   Renewable Energy, Global Policy, Climate Chang...
NO
30   Sustainability, Reporting Standards, Market Pr...
NO
31   Energy Transition, Sustainable Development, In...
NO
35       Business Strategy, Innovation, Sustainability
NO
36   Sustainability, Customer Privacy, Corporate Re...
NO
39   Sustainability Reporting, Human Rights, Securi...
NO
41   Infrastructure Development, Project Management...
SI
42             Environmental Impacts and Climate Change
NO
46             Environmental Impacts and Climate Change
SI
55             Environmental Impacts and Climate Change
NO
```

```
66   Technological Innovation in Energy and Transpo...
SI
70              Sustainable Finance and Economic Trends
SI
104  Infrastructure Development and Public-Private ...
SI
105             Sustainable Finance and Economic Trends
NO
106  Infrastructure Development and Public-Private ...
NO
107            Environmental Impacts and Climate Change
NO
119  Infrastructure Development and Public-Private ...
SI
125  Technological Innovation in Energy and Transpo...
NO
139  Technological Innovation in Energy and Transpo...
NO
144  Urban Development Trends and Smart Infrastructure
NO
168             Sustainable Finance and Economic Trends
NO
169             Sustainable Finance and Economic Trends
NO
176  Technological Innovation in Energy and Transpo...
SI
182  Technological Innovation in Energy and Transpo...
NO
186             Sustainable Finance and Economic Trends
NO
200             Sustainable Finance and Economic Trends
NO
209             Sustainable Finance and Economic Trends
NO
212  Infrastructure Development and Public-Private ...
SI
```

In [132…

```python
# Analysis of the number of documents mentioning Uzbekistan
uzbekistan_text_count = df['Reason for Categorization'].apply(lambda x: '

# View documents mentioning Uzbekistan
plt.figure(figsize=(6, 4))
uzbekistan_text_count.plot(kind='bar', color='lightgreen')
plt.title('Documents mentioning Uzbekistan', fontsize=14)
plt.xlabel('Mention of Uzbekistan', fontsize=12)
plt.ylabel('Frequency', fontsize=12)
plt.xticks(rotation=0)
plt.tight_layout()
plt.show()
```



1. Filtering process:

- The code searches the **'Reason for Categorization'** column for documents containing any of the key terms defined in the keywords list. These terms are:
  - 'economic development'
  - 'government reforms'
  - 'sustainability'
  - 'renewable energy'
  - 'social challenges'
- This is done using the str.contains method, which checks whether any of these key terms appear in the text of each document.

2. Filtering result:

- The result is a list of documents containing at least one of the mentioned key terms.
- Documents that meet this condition are printed along with their category and the Text About Uzbekistan column, which indicates whether the text specifically mentions Uzbekistan (with 'YES' for yes and 'NO' for no).

3. Analysis of the result:

- Relevant Documents: Documents are identified that are relevant to the topics of interest related to Uzbekistan and its main challenges. Some examples of relevant documents are:
  - Annex D Detailed Report.docx
  - Second-Party Opinion on SDG Bonds.docx
  - IRENA_G20_Just transition_in_EMDEs_2024.docx
  - Business Model Innovation - A Game Changer.docx
- Text About Uzbekistan column: Although many documents mention the key terms, most of them have the value 'NO' in the Text About Uzbekistan column, indicating that they do not specifically mention Uzbekistan in the text. Only a few have 'YES', suggesting that the content is directly relevant to Uzbekistan.

4. Possible interpretation:

- Relevant documents containing the key terms could be related to global trends (such as sustainability, renewable energy, and economic reforms) that also affect Uzbekistan, but do not explicitly mention the country in the text.
- The presence of several documents relevant to renewable energy, sustainability, and sustainable finance indicates that these are key topics of interest in analyses of Uzbekistan, although not all documents directly mention the country.

Conclusion:

This code snippet serves to identify the most relevant documents that address key topics related to Uzbekistan, such as its economic development, sustainability, and renewable energy. Although not all documents explicitly mention Uzbekistan, the presence of these topics reflects the current focus on global issues that also impact the country.

To implement Pattern Detection and Trend Analysis using Python, we can approach the analysis of categories and their relationships with several steps. Below is an example code to identify patterns in categories and category relationships through correlation or visualization techniques.

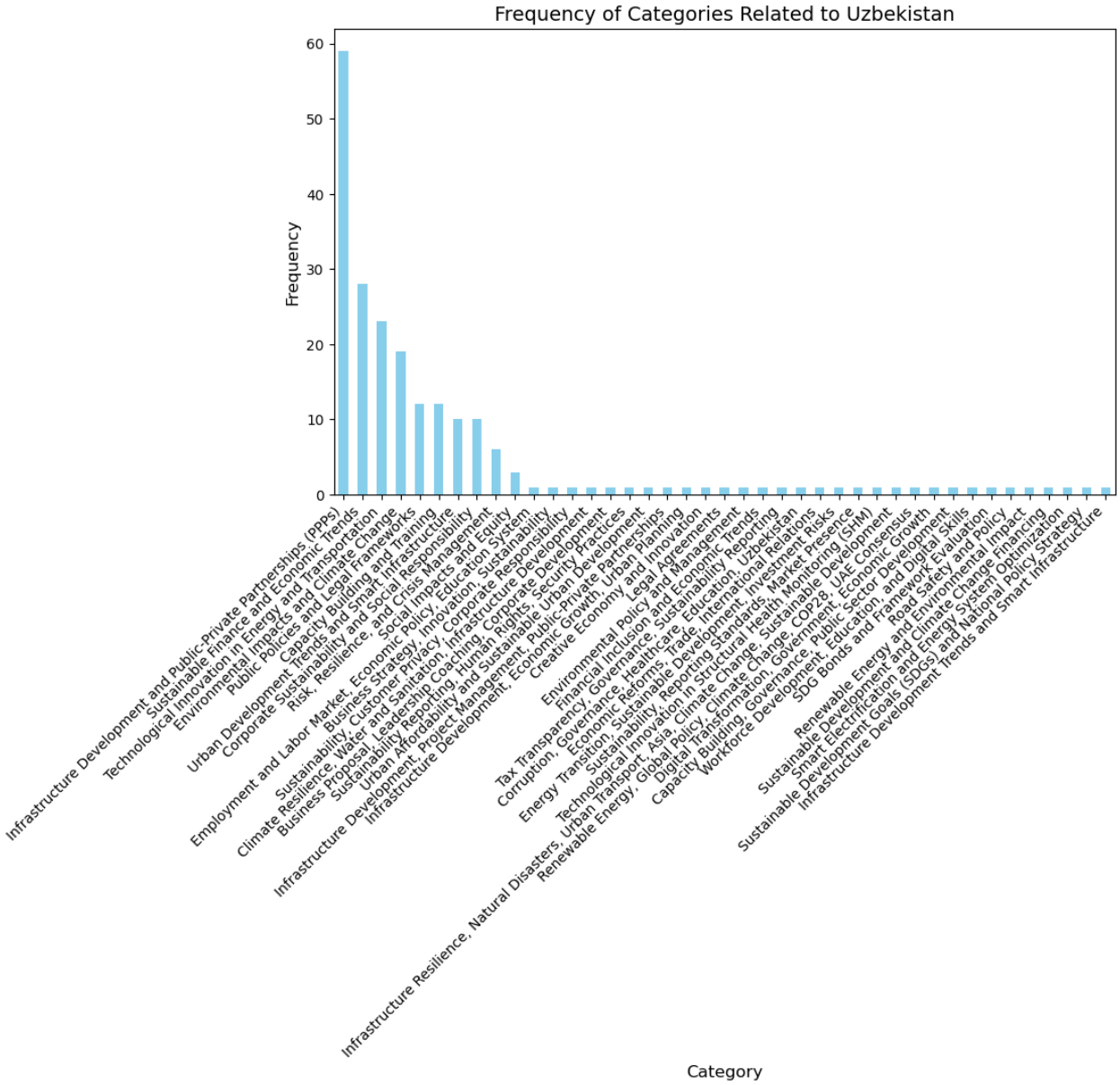
Step 1: Analyze the Frequency of Categories

The goal is to detect which categories are documented most frequently in relation to Uzbekistan, such as infrastructure, renewable energy, and economic reforms.

In [138…
```
# Assuming 'df' is your DataFrame
```

```
category_counts = df['Category'].value_counts()

# Visualize the distribution of categories
plt.figure(figsize=(10, 6))
category_counts.plot(kind='bar', color='skyblue')
plt.title('Frequency of Categories Related to Uzbekistan', fontsize=14)
plt.xlabel('Category', fontsize=12)
plt.ylabel('Frequency', fontsize=12)
plt.xticks(rotation=45, ha="right")
plt.tight_layout()
plt.show()
```

```
/var/folders/r4/bfc17knx7fl8dv9hgcf92n_w0000gn/T/ipykernel_44284/699785176.py:11: UserWarning: Tight layout not applied. The bottom and top margins cannot be made large enough to accommodate all Axes decorations.
  plt.tight_layout()
```



In [140…

```
# Output the category counts to check which categories are being document
print(category_counts)
```

Category	
Infrastructure Development and Public-Private Partnerships (PPPs)	59
Sustainable Finance and Economic Trends	28
Technological Innovation in Energy and Transportation	23
Environmental Impacts and Climate Change	19
Public Policies and Legal Frameworks	12
Capacity Building and Training	12
Urban Development Trends and Smart Infrastructure	10
Corporate Sustainability and Social Responsibility	10
Risk, Resilience, and Crisis Management	6
Social Impacts and Equity	3
Employment and Labor Market, Economic Policy, Education System	1
Business Strategy, Innovation, Sustainability	1
Sustainability, Customer Privacy, Corporate Responsibility	1
Climate Resilience, Water and Sanitation, Infrastructure Development	1
Business Proposal, Leadership Coaching, Corporate Development	1
Sustainability Reporting, Human Rights, Security Practices	1
Urban Affordability and Sustainable Urban Development	1
Infrastructure Development, Project Management, Public-Private Partnerships	1
Infrastructure Development, Economic Growth, Urban Planning	1
Creative Economy and Innovation	1
Legal Agreements	1
Environmental Policy and Management	1
Financial Inclusion and Economic Trends	1
Tax Transparency, Governance, Sustainability Reporting	1
Corruption, Governance, Healthcare, Education, Uzbekistan	1
Economic Reforms, Trade, International Relations	1

```
Energy Transition, Sustainable Development, Investment Risks
1
Sustainability, Reporting Standards, Market Presence
1
Technological Innovation in Structural Health Monitoring (SHM)
1
Infrastructure Resilience, Natural Disasters, Urban Transport, Asia, Clima
te Change, Sustainable Development     1
Renewable Energy, Global Policy, Climate Change, COP28, UAE Consensus
1
Digital Transformation, Government, Economic Growth
1
Capacity Building, Governance, Public Sector Development
1
Workforce Development, Education, and Digital Skills
1
SDG Bonds and Framework Evaluation
1
Road Safety and Policy
1
Renewable Energy and Environmental Impact
1
Sustainable Development and Climate Change Financing
1
Smart Electrification and Energy System Optimization
1
Sustainable Development Goals (SDGs) and National Policy Strategy
1
Infrastructure Development Trends and Smart Infrastructure
1
Name: count, dtype: int64
```

Step 2:

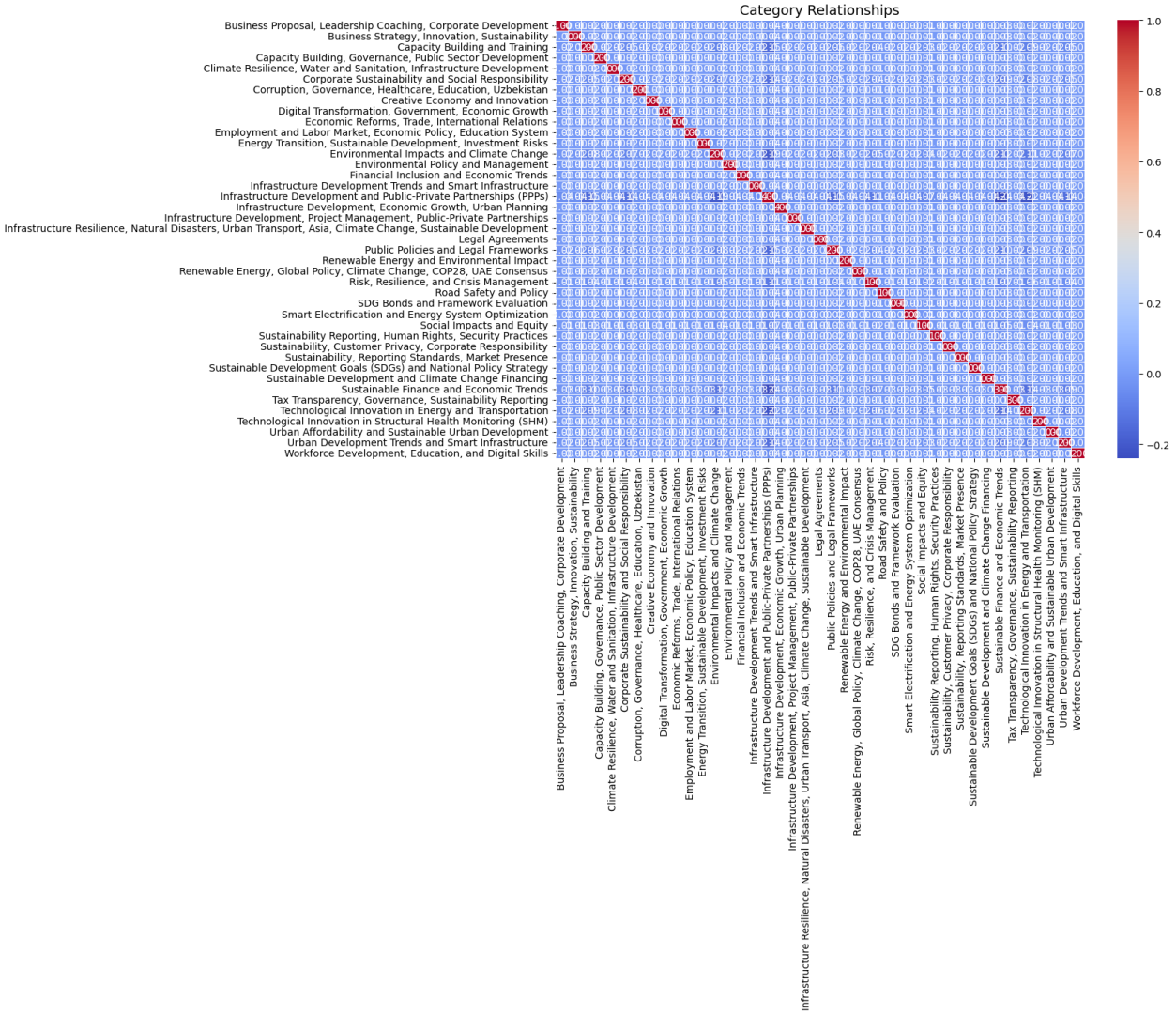
Explore Relationships Between Categories with a **correlation matrix**

In [168…

```
import numpy as np
import seaborn as sns

# Create a binary matrix (1 for presence of category, 0 for absence)
category_matrix = pd.get_dummies(df['Category'])

# Calculate the correlation matrix
correlation_matrix = category_matrix.corr()

# Visualize the correlation matrix as a heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f', l
plt.title('Category Relationships', fontsize=14)
plt.tight_layout()

plt.savefig('category_uzbekistan_mentions.png', format='png')
```

```
plt.show()
```

```
/var/folders/r4/bfc17knx7fl8dv9hgcf92n_w0000gn/T/ipykernel_44284/4098286951.py:14: UserWarning: Tight layout not applied. The bottom and top margins cannot be made large enough to accommodate all Axes decorations.
  plt.tight_layout()
```



This code converts the Category column into a binary matrix (1 for the presence of a category in a document and 0 for absence) and then calculates the correlation between these categories. The heatmap visually represents the relationships between categories. For example, it will show whether infrastructure initiatives are often mentioned alongside sustainable finance or public policies.

The correlation matrix shown in the chart is the result of calculating the relationship between the categories of documents based on their presence in the dataset. In this case, a binary matrix was created where a value of 1 indicates that a category is present in a document, and a 0 indicates that it is not. The correlation between these categories was then calculated using Pearson's correlation coefficient.

Explanation of the chart:

1. **Diagonal**: On the main diagonal of the matrix (where each category intersects with itself), the value is always 1, since a category is perfectly correlated with itself.

2. **Values close to 1**: Values close to 1 indicate a high positive correlation, meaning that the associated categories tend to appear together in the same documents. This suggests that there are themes that are commonly addressed together, such as "Sustainability" and "Corporate Responsibility" or "Urban Development" and "Infrastructure."

3. **Values close to 0**: Values close to 0 indicate no correlation between categories. This means that these categories do not typically appear together in documents, reflecting very different thematic areas, such as "Public Policies" and "Sustainable Development."

4. **Negative values**: Although there are no significant negative values in this case, if they existed, they would indicate an inverse correlation. This would suggest that when one category is present, the other tends to be absent.

5. **Heatmap colors**: Red colors indicate stronger positive correlations, while blue colors indicate weaker or no correlation.

Purpose of the analysis:

The purpose of this matrix is to identify the relationships between the thematic categories of the documents. In the context of Uzbekistan, these relationships can reveal the areas of greater interest or overlap in the topics discussed, which helps in identifying key patterns and trends, such as the most discussed topics or the sectors that are frequently addressed together in the texts related to the country.

Conclusion:

By analyzing this matrix, we can get an overview of how different categories of documents are related. For example, if the categories of **"Infrastructure" and "Sustainable Finance" are highly correlated**, this could indicate that many documents related to these topics are addressed together, which is useful for understanding Uzbekistan’s development strategies in these areas.

Step 3: Relationship Between Categories and Text About Uzbekistan

This code groups the data by Category and the Text About Uzbekistan column (which indicates whether the document mentions Uzbekistan or not) and then

visualizes how frequently each category mentions Uzbekistan using a stacked bar chart.

The bar chart titled **"Category and Mention of Uzbekistan"** visually represents the frequency of documents in each category that either mention or do not mention Uzbekistan.

Explanation of the chart:

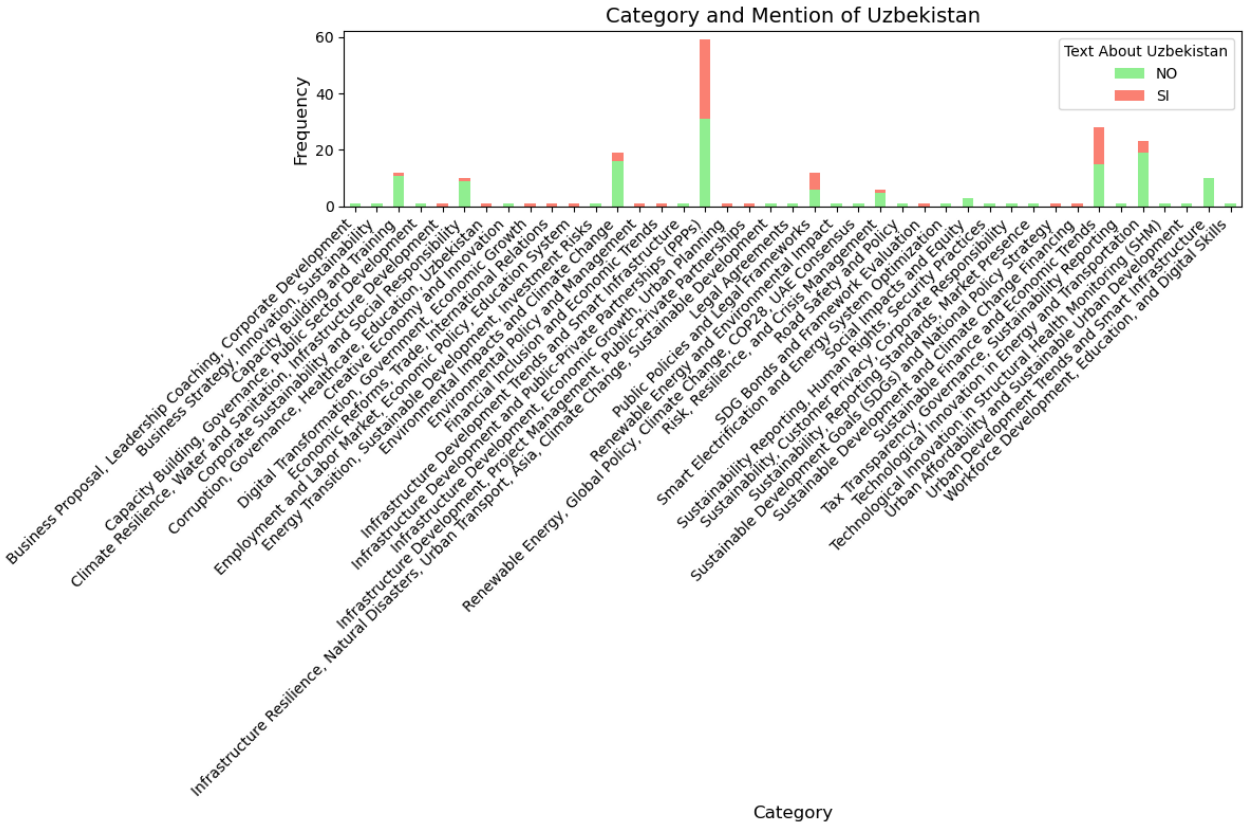
1. **X-axis (Category)**:
    - The X-axis shows the different categories under which the documents have been classified. Each category represents a thematic area, such as "Infrastructure Development", "Sustainability", "Renewable Energy", etc.
2. **Y-axis (Frequency)**:
    - The Y-axis represents the frequency or count of documents within each category. Specifically, it shows how many documents either mention ("Yes") or do not mention ("No") Uzbekistan.
3. **Stacked bars**:
    - Each bar is divided into two segments:
        - **Green (lightgreen)** represents the documents where the "Text About Uzbekistan" column indicates "No" (i.e., documents that do not mention Uzbekistan).
        - **Red (salmon)** represents the documents where the "Text About Uzbekistan" column indicates "Yes" (i.e., documents that mention Uzbekistan).
4. **Observations**:
    - The chart helps to quickly visualize the balance between documents that mention Uzbekistan and those that do not within each category.
    - Categories with more "Yes" (red) bars show that the documents in these categories are more likely to mention Uzbekistan, whereas categories with more "No" (green) bars indicate that the documents do not often mention Uzbekistan.
5. **Insight**:
    - By observing the chart, it is possible to identify which categories have a higher concentration of documents related to Uzbekistan and which categories do not. This is useful for understanding the focus of the documents with respect to the country, especially when analyzing trends or patterns in the data.

Conclusion:

The chart visually summarizes the relationship between the categories and the mention of Uzbekistan in the dataset, providing a clear representation of how documents are distributed across different categories and whether they mention Uzbekistan or not.

In [170…

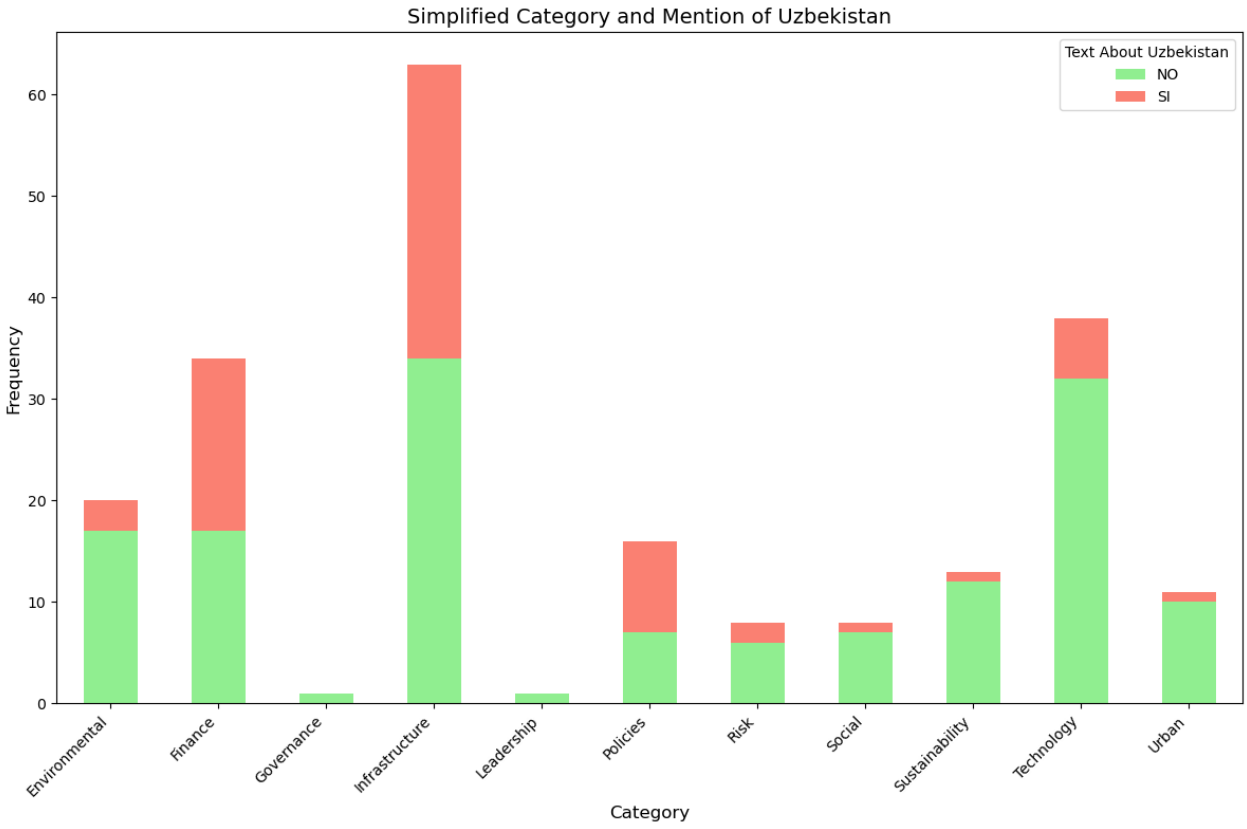
```
# Count the frequency of 'Text About Uzbekistan' (Yes/No) for each catego
category_uzbekistan = df.groupby(['Category', 'Text About Uzbekistan']).s

# Visualize the frequency of categories mentioning Uzbekistan
category_uzbekistan.plot(kind='bar', stacked=True, figsize=(12, 8), color
plt.title('Category and Mention of Uzbekistan', fontsize=14)
plt.xlabel('Category', fontsize=12)
plt.ylabel('Frequency', fontsize=12)
plt.xticks(rotation=45, ha="right")
plt.tight_layout()

plt.savefig('Category_Mention_of_Uzbekistan.png', format='png')

plt.show()
```



In [173…

```
# Count the frequency of 'Text About Uzbekistan' (Yes/No) for each catego
category_uzbekistan = df.groupby(['Simplified Category', 'Text About Uzbe

# Visualize the frequency of categories mentioning Uzbekistan
category_uzbekistan.plot(kind='bar', stacked=True, figsize=(12, 8), color
plt.title('Simplified Category and Mention of Uzbekistan', fontsize=14)
```

```
plt.xlabel('Category', fontsize=12)
plt.ylabel('Frequency', fontsize=12)
plt.xticks(rotation=45, ha="right")
plt.tight_layout()

plt.savefig('Category_Mention_of_Uzbekistan2.png', format='png')

plt.show()
```



Explanation of the Chart:

1. **X-axis (Simplified Category)**:
    - The categories along the X-axis are simplified versions of the original categories, such as **Environmental**, **Finance**, **Infrastructure**, **Leadership**, etc.
    - These categories provide a broader classification of topics that the documents cover.
2. **Y-axis (Frequency)**:
    - The Y-axis represents the count of documents within each simplified category. This shows how many documents are associated with each category.
    - The total height of the bars indicates the overall number of documents in each simplified category.

Insights from the Chart:

- **Infrastructure** stands out with a large number of documents mentioning Uzbekistan (in red), showing that this category has the highest focus on Uzbekistan.
- **Finance** also has a significant number of documents mentioning Uzbekistan, but it is more balanced with both "Yes" and "No" mentions.
- **Technology** has a small proportion of documents mentioning Uzbekistan, indicated by the red bar, with the majority not mentioning Uzbekistan (green bar).
- Categories like **Risk**, **Social**, **Sustainability**, **Urban**, and **Governance** have fewer documents overall, and most of them do not mention Uzbekistan.

Conclusion:

- The chart indicates that **Infrastructure** is the category with the highest focus on Uzbekistan, followed by **Finance**.
- **Technology** has relatively fewer documents about Uzbekistan, and **Urban**, **Risk**, and other categories have minimal or no direct references to the country.
- This visualization helps identify the thematic areas where Uzbekistan is frequently mentioned, which is valuable for understanding the national focus and areas of international interest related to the country.

Sentiment Analysis (Text Sentiment Evaluation)

We will apply **Sentiment Analysis** to understand the tone of the documents. We'll use the `TextBlob` library, which provides a simple API for sentiment analysis.

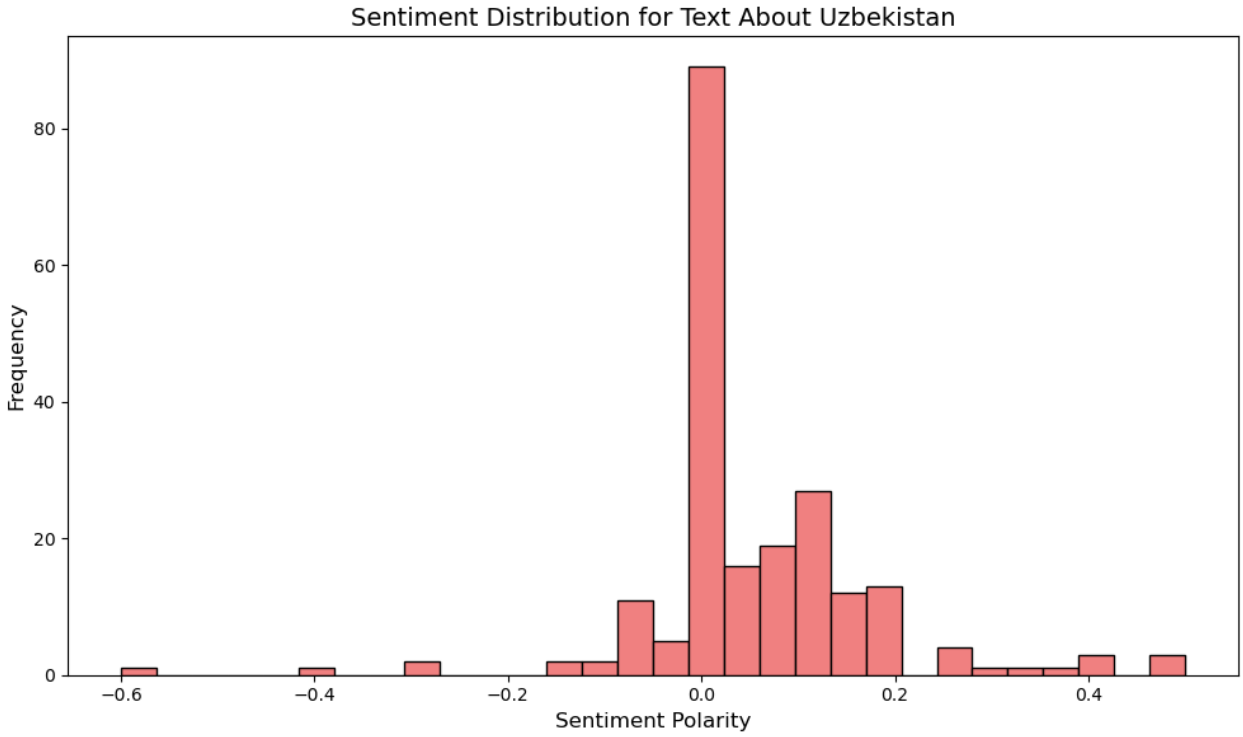
Explanation:

- **Sentiment Analysis**:
  - The sentiment polarity ranges from -1 (very negative) to +1 (very positive).
  - The code above computes the sentiment for each document and stores the result in a new column, `Sentiment`.
  - The histogram visualization shows how positive or negative the texts related to Uzbekistan are perceived overall.

In [185…

```
from textblob import TextBlob

# Function to calculate sentiment polarity
def get_sentiment(text):
    analysis = TextBlob(text)
```

```
    return analysis.sentiment.polarity  # Sentiment polarity score

# Apply sentiment analysis to 'Text About Uzbekistan' and store in a new
df['Sentiment'] = df['Reason for Categorization'].apply(lambda x: get_sen

# Plot the sentiment distribution
plt.figure(figsize=(10, 6))
df['Sentiment'].plot(kind='hist', bins=30, color='lightcoral', edgecolor=
plt.title('Sentiment Distribution for Text About Uzbekistan', fontsize=14
plt.xlabel('Sentiment Polarity', fontsize=12)
plt.ylabel('Frequency', fontsize=12)
plt.tight_layout()

# Save the plot
plt.savefig('sentiment_distribution.png', format='png')
plt.show()
```



In [203…

```
df[['File Name', 'Sentiment']]
```

Out[203…

	File Name	Sentiment
0	WEF_Reshaping_affordability_2024.docx	0.073333
1	smarsly2021e.docx	-0.400000
2	UzbekistanRailways 2.docx	0.008929
3	Swedish_Waste_Management_A_Review_Articl.docx	-0.066667
4	MOF_LSE_IFC_event_yfnGTW0.docx	0.081818
...	...	...
208	Comprehensive Proposal for Toybola and Anvar A...	0.000000
209	cbi_mr_h1_2024_02e_1.docx	-0.020833
210	Investments and construction 2.docx	0.250000
211	Dubai-Municipality-3DCP-Guideline-1st-Edition-...	0.150000
212	3D_Concrete_Printing_White_Paper_Ubez 2024-202...	0.075000

213 rows × 2 columns

Distribution Pattern:

- Most of the documents have a sentiment polarity of **0**, which indicates **neutral** sentiment. The peak at **0** shows that many documents are neutral in tone, suggesting that they don't express strong positive or negative opinions.
- There is a relatively smaller number of documents with a slightly **positive** sentiment (ranging from 0 to 0.4) and a **negative** sentiment (ranging from -0.2 to -0.4).
- The histogram has a **long tail on both sides** (positive and negative) but is heavily concentrated around **neutral sentiment**. This suggests that the documents mostly contain factual information or neutral statements about Uzbekistan, with fewer strongly opinionated or emotional tones.

Key Insights:

- The documents are predominantly **neutral** in their sentiment regarding Uzbekistan.
- There are only a few documents that are either **slightly positive** or **slightly negative** in tone.
- This kind of sentiment distribution is common for analytical or report-based texts, where the focus is on presenting facts rather than expressing opinions.

This analysis gives an overview of the general tone of documents related to

Uzbekistan, with a notable tendency towards neutral perspectives.

Predictive Modeling (Supervised Learning)

We will use a Logistic Regression model to predict the categories based on the text content of the documents.

In [208…

```
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
from sklearn.preprocessing import LabelEncoder
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix, roc_
from sklearn.metrics import roc_auc_score

# 1. Prepare the data
X = df['Text About Uzbekistan']  # Features (text)
y = df['Category']  # Labels (categories)

# Convert categories to numerical labels
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(y)

# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y_encoded, test_si

# 2. Convert text data into numerical vectors using TF-IDF
vectorizer = TfidfVectorizer(stop_words='english', max_features=5000)
X_train_tfidf = vectorizer.fit_transform(X_train)
X_test_tfidf = vectorizer.transform(X_test)

# 3. Train the logistic regression model
model = LogisticRegression(max_iter=1000)
model.fit(X_train_tfidf, y_train)

# 4. Predict and evaluate
y_pred = model.predict(X_test_tfidf)

# Convert numerical predictions back to category names
y_pred_labels = label_encoder.inverse_transform(y_pred)
y_test_labels = label_encoder.inverse_transform(y_test)

# Print classification report
print(classification_report(y_test_labels, y_pred_labels))
```

```
                                                                precisi
on    recall  f1-score   support

                                    Capacity Building and Training       0.
00      0.00      0.00         2
         Capacity Building, Governance, Public Sector Development       0.
00      0.00      0.00         1
                Corporate Sustainability and Social Responsibility       0.
00      0.00      0.00         2
                        Environmental Impacts and Climate Change       0.
00      0.00      0.00         4
       Infrastructure Development Trends and Smart Infrastructure       0.
00      0.00      0.00         1
Infrastructure Development and Public-Private Partnerships (PPPs)       0.
23      1.00      0.38        10
                             Public Policies and Legal Frameworks       0.
00      0.00      0.00         2
                            Risk, Resilience, and Crisis Management       0.
00      0.00      0.00         2
                                         Social Impacts and Equity       0.
00      0.00      0.00         1
             Sustainability, Reporting Standards, Market Presence       0.
00      0.00      0.00         1
Sustainable Development Goals (SDGs) and National Policy Strategy       0.
00      0.00      0.00         1
                            Sustainable Finance and Economic Trends       0.
00      0.00      0.00         7
            Technological Innovation in Energy and Transportation       0.
00      0.00      0.00         4
                 Urban Development Trends and Smart Infrastructure       0.
00      0.00      0.00         4
             Workforce Development, Education, and Digital Skills       0.
00      0.00      0.00         1

                                                        accuracy
0.23        43
                                                       macro avg       0.
02      0.07      0.03        43
                                                    weighted avg       0.
05      0.23      0.09        43
```

```
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classificatio
n.py:1531: UndefinedMetricWarning: Precision is ill-defined and being set
to 0.0 in labels with no predicted samples. Use `zero_division` parameter
to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classificatio
n.py:1531: UndefinedMetricWarning: Precision is ill-defined and being set
to 0.0 in labels with no predicted samples. Use `zero_division` parameter
to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classificatio
n.py:1531: UndefinedMetricWarning: Precision is ill-defined and being set
to 0.0 in labels with no predicted samples. Use `zero_division` parameter
to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

The result of the logistic regression model's evaluation provides a **classification report**, which includes several key performance metrics such as precision, recall, F1-score, and support for each category. Here's what each metric means and an explanation of the results:

Key Metrics:

- **Precision**: The ratio of correctly predicted positive observations to the total predicted positives. It tells us how many of the predicted categories were actually correct.
- **Recall**: The ratio of correctly predicted positive observations to all observations in the actual category. It tells us how many of the actual categories were identified by the model.
- **F1-Score**: The weighted average of precision and recall, providing a balance between the two. It's especially useful when the class distribution is imbalanced.
- **Support**: The number of actual occurrences of the category in the dataset. It helps understand how many samples the model was tested on for each category.

Breakdown of the Results:

1. **Precision, Recall, and F1-Score for All Categories**:

   - Many categories have **precision, recall, and F1-score of 0.00**, meaning the model could not correctly predict any of these categories.
   - For some categories like **"Infrastructure Development and Public-Private Partnerships (PPPs)"**, the **recall** is **1.00**, meaning the model correctly predicted all instances of this category, but the **precision** is **0.23**, indicating that only a small portion of the predicted instances were actually correct. This leads to a low **F1-score** of **0.38**.
2. **Overall Accuracy**:

- The **accuracy** of the model is **0.23**, which is quite low. This means the model correctly predicted the category for only about 23% of the test instances.

3. **Macro Average**:
   - The **macro average** provides the average precision, recall, and F1-score across all categories without considering the number of instances in each category. The values for the macro average are very low:
     - Precision: **0.02**
     - Recall: **0.07**
     - F1-Score: **0.03**
   - These low values indicate that the model is struggling to correctly classify most categories, and it is not performing well overall.
4. **Weighted Average**:
   - The **weighted average** takes the support (number of instances) for each category into account when calculating the average. This is typically more reflective of how the model performs on the dataset as a whole.
     - Precision: **0.05**
     - Recall: **0.23**
     - F1-Score: **0.09**
   - This further confirms that the model has significant room for improvement.

Possible Reasons for the Poor Performance:

1. **Imbalanced Classes**: Many categories have very few instances (e.g., categories like "Workforce Development, Education, and Digital Skills" have only 1 instance). Logistic regression models can struggle when classes are highly imbalanced.
2. **Text Representation**: While **TF-IDF** was used to convert text data into numerical vectors, the feature set may not be rich enough to capture the complexities of the text.
3. **Model Complexity**: Logistic regression is a relatively simple model, and it may not capture the nuances in the data, especially when dealing with complex text-based features.
4. **Data Quality and Preprocessing**: There might be issues with the data (e.g., noisy or unstructured text, missing data, etc.) that affect the model's performance.

In [211…

```
# 1. Prepare the data
X = df['Text About Uzbekistan']  # Features (text)
y = df['Simplified Category']  # Labels (simplified categories)

# Convert categories to numerical labels
```

```
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(y)

# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y_encoded, test_si

# 2. Convert text data into numerical vectors using TF-IDF
vectorizer = TfidfVectorizer(stop_words='english', max_features=5000)
X_train_tfidf = vectorizer.fit_transform(X_train)
X_test_tfidf = vectorizer.transform(X_test)

# 3. Train the logistic regression model
model = LogisticRegression(max_iter=1000)
model.fit(X_train_tfidf, y_train)

# 4. Predict and evaluate
y_pred = model.predict(X_test_tfidf)

# Convert numerical predictions back to category names
y_pred_labels = label_encoder.inverse_transform(y_pred)
y_test_labels = label_encoder.inverse_transform(y_test)

# Print classification report
print(classification_report(y_test_labels, y_pred_labels))
```

```
                precision    recall  f1-score   support

 Environmental       0.00      0.00      0.00         4
       Finance       0.00      0.00      0.00         7
Infrastructure       0.26      1.00      0.41        11
      Policies       0.00      0.00      0.00         2
          Risk       0.00      0.00      0.00         2
        Social       0.00      0.00      0.00         3
Sustainability       0.00      0.00      0.00         3
    Technology       0.00      0.00      0.00         6
         Urban       0.00      0.00      0.00         5

      accuracy                           0.26        43
     macro avg       0.03      0.11      0.05        43
  weighted avg       0.07      0.26      0.10        43
```

```
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classificatio
n.py:1531: UndefinedMetricWarning: Precision is ill-defined and being set
to 0.0 in labels with no predicted samples. Use `zero_division` parameter
to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classificatio
n.py:1531: UndefinedMetricWarning: Precision is ill-defined and being set
to 0.0 in labels with no predicted samples. Use `zero_division` parameter
to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classificatio
n.py:1531: UndefinedMetricWarning: Precision is ill-defined and being set
to 0.0 in labels with no predicted samples. Use `zero_division` parameter
to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

Evaluation of the Logistic Regression Results for the **Simplified Category**:

The output of the logistic regression model, when predicting the **Simplified Category**, shows some significant challenges in terms of prediction accuracy, particularly in several categories.

1. **Precision**:

- Precision indicates how many of the predicted positive instances were actually correct.
- Most categories show a **precision of 0.00**, meaning that none of the predicted categories were correct for these classes, except for **Infrastructure**, which has a precision of **0.26**.

2. **Recall**:

- Recall measures how many of the actual positive instances were correctly predicted.
- **Recall of 1.00 for Infrastructure** means that all the actual instances of **Infrastructure** were correctly identified by the model.
- All other categories have a **recall of 0.00**, suggesting that the model failed to detect any instances for these categories.

3. **F1-Score**:

- The F1-score is the harmonic meaning of precision and recall. It gives a balanced measure of a model's accuracy, especially when dealing with imbalanced datasets.
- **Infrastructure** again stands out with an F1-score of **0.41** due to its higher precision and recall. However, all other categories have an F1-score of **0.00**, as

the model is unable to predict or detect instances for those categories.

4. **Accuracy**:

- **Accuracy of 0.26** means the model correctly predicted the category for **26%** of the samples. This is relatively low, indicating that the model has significant room for improvement.

5. **Macro Average**:

- The **macro average** of precision, recall, and F1-score indicates the average performance across all categories, treating each category equally. Here, the values are very low, which reflects poor performance across most categories.

6. **Weighted Average**:

- The **weighted average** takes into account the number of instances in each class, which means that it adjusts for class imbalance. The weighted F1-score is **0.10**, highlighting that, while some categories are correctly predicted (like Infrastructure), the model is still failing for the majority of categories.

Comparison with the Logistic Regression of the Original "Category" Column:

To compare these results with the **"Category"** column model, we would evaluate whether there was any improvement or deterioration in terms of prediction accuracy for more granular categories.

Potential Insights:

- **Infrastructure** seems to be the most reliably predicted category across both models, with a higher **recall** and **precision** (compared to other categories). This might suggest that the model has learned the features specific to **Infrastructure** well, but struggles with the other categories.
- **Sustainability, Finance, and Risk** are poorly predicted in both models, showing **zero** performance across several metrics.

Possible Reasons for Poor Performance:

1. **Imbalanced Classes**: There are a small number of instances per class (support is low for most categories), which can lead to poor model performance.
2. **Textual Features**: The features extracted from the text using TF-IDF might not be discriminative enough to distinguish between categories, especially if the documents are very similar in content.
3. **Simplified Category Aggregation**: While the **Simplified Category** aggregates

some of the original categories, the broader and less specific labels might be too general for the model to accurately differentiate, causing a drop in predictive performance.

In summary, while the model performs well for **Infrastructure**, overall, the logistic regression model is struggling with most categories in both the **Category** and **Simplified Category** columns.

To improve the classification of underrepresented categories, fine-tune the model and text preprocessing, and perform a more comprehensive evaluation with additional metrics, you can follow these steps:

1. Improving Classification of Underrepresented Categories

We will use SMOTE (Synthetic Minority Over-sampling Technique) to balance classes and improve the classification of categories with fewer examples (such as Technology, Risk, and Social).

2. Model Tuning and Text Preprocessing

We will perform some additional tuning of the model and apply advanced text preprocessing techniques. We will also try out BERT (although I provide an example using TF-IDF here to get you started).

3. Evaluating with Confusion Matrix and ROC Curves

Finally, we will use confusion matrix and ROC curves to evaluate the performance of the model.

In [218…

```
import pandas as pd

# Convert to pandas Series
y_train_series = pd.Series(y_train)

# See the class distribution before SMOTEprint("Distribución de clases an
print(y_train_series.value_counts())
```

```
Distribución de clases antes de SMOTE:
3     52
9     32
1     27
0     16
5     14
8     10
6      6
10     6
7      5
4      1
2      1
Name: count, dtype: int64
```

In [ ]:

The class distribution before applying SMOTE shows the number of samples for each category in the training set:

- Class 3: 52 samples
- Class 9: 32 samples
- Class 1: 27 samples
- Class 0: 16 samples
- Class 5: 14 samples
- Class 8: 10 samples
- Class 6: 6 samples
- Class 10: 6 samples
- Class 7: 5 samples
- Class 4: 1 sample
- Class 2: 1 sample

Remarks: Class 3 is the most represented with 52 samples, suggesting that it is the majority class. Classes 4 and 2 have only 1 sample, indicating a significant imbalance in the distribution of classes. There are several classes with very few samples, which can make it difficult for the model to learn patterns from those less represented categories.

In [224…

```python
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_extraction.text import TfidfVectorizer

# # 1. Preprocess and split the data
X = df['Text About Uzbekistan']
y = df['Simplified Category']
```

```
# Convert categories to numeric labels
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(y)

# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y_encoded, test_si

# 2. Convert text data into numeric vectors using TF-IDF
vectorizer = TfidfVectorizer(stop_words='english', max_features=5000)
X_train_tfidf = vectorizer.fit_transform(X_train)
X_test_tfidf = vectorizer.transform(X_test)

# 3. Apply SMOTE to balance classes
smote = SMOTE(random_state=42)
X_train_smote, y_train_smote = smote.fit_resample(X_train_tfidf, y_train)

# 4. Train the logistic regression model with the balanced data
model = LogisticRegression(max_iter=1000)
model.fit(X_train_smote, y_train_smote)

# 5. Make predictions with the test set
y_pred = model.predict(X_test_tfidf)

# Convert numerical predictions to category labels
y_pred_labels = label_encoder.inverse_transform(y_pred)
y_test_labels = label_encoder.inverse_transform(y_test)

# 6. Evaluate the model using classification and confusion matrix
print("Classification Report:\n", classification_report(y_test_labels, y_
print("Confusion Matrix:\n", confusion_matrix(y_test_labels, y_pred_label

# 7. Visualize the confusion matrix
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(10, 7))
sns.heatmap(confusion_matrix(y_test_labels, y_pred_labels), annot=True, c
plt.title("Confusion Matrix")
plt.xlabel('Predicted')
plt.ylabel('True')
plt.show()
```

```
---------------------------------------------------------------------------
_
ValueError                                Traceback (most recent call las
t)
Cell In[224], line 26
     24 # 3. Aplicar SMOTE para balancear las clases
     25 smote = SMOTE(random_state=42)
---> 26 X_train_smote, y_train_smote = smote.fit_resample(X_train_tfidf, y
_train)
     28 # 4. Entrenar el modelo de regresión logística con los datos balan
ceados
     29 model = LogisticRegression(max_iter=1000)
```

```
File /opt/anaconda3/lib/python3.12/site-packages/imblearn/base.py:208, in
BaseSampler.fit_resample(self, X, y)
    187 """Resample the dataset.
    188
    189 Parameters
   (...)
    205     The corresponding label of `X_resampled`.
    206 """
    207 self._validate_params()
--> 208 return super().fit_resample(X, y)

File /opt/anaconda3/lib/python3.12/site-packages/imblearn/base.py:112, in
SamplerMixin.fit_resample(self, X, y)
    106 X, y, binarize_y = self._check_X_y(X, y)
    108 self.sampling_strategy_ = check_sampling_strategy(
    109     self.sampling_strategy, y, self._sampling_type
    110 )
--> 112 output = self._fit_resample(X, y)
    114 y_ = (
    115     label_binarize(output[1], classes=np.unique(y)) if binarize_y
else output[1]
    116 )
    118 X_, y_ = arrays_transformer.transform(output[0], y_)

File /opt/anaconda3/lib/python3.12/site-packages/imblearn/over_sampling/_s
mote/base.py:389, in SMOTE._fit_resample(self, X, y)
    386 X_class = _safe_indexing(X, target_class_indices)
    388 self.nn_k_.fit(X_class)
--> 389 nns = self.nn_k_.kneighbors(X_class, return_distance=False)[:, 1:]
    390 X_new, y_new = self._make_samples(
    391     X_class, y.dtype, class_sample, X_class, nns, n_samples, 1.0
    392 )
    393 X_resampled.append(X_new)

File /opt/anaconda3/lib/python3.12/site-packages/sklearn/neighbors/_base.p
y:834, in KNeighborsMixin.kneighbors(self, X, n_neighbors, return_distanc
e)
    832     else:
    833         inequality_str = "n_neighbors <= n_samples_fit"
--> 834     raise ValueError(
    835         f"Expected {inequality_str}, but "
    836         f"n_neighbors = {n_neighbors}, n_samples_fit = {n_samples_
fit}, "
    837         f"n_samples = {X.shape[0]}"  # include n_samples for commo
n tests
    838     )
    840 n_jobs = effective_n_jobs(self.n_jobs)
    841 chunked_results = None

ValueError: Expected n_neighbors <= n_samples_fit, but n_neighbors = 6, n_
samples_fit = 1, n_samples = 1
```

The error ValueError: Expected n_neighbors <= n_samples_fit, but n_neighbors = 6,

n_samples_fit = 1, n_samples = 1 occurs when the number of samples in your training set is too small to apply the SMOTE algorithm with the default parameters. This problem can happen if any of the classes in your training set have only one sample after splitting the data in the train_test_split step.

Solution:

Change the model: Instead of using Logistic Regression, switch to **Random Forest or XGBoost**, which tend to be more effective for imbalanced data. Also, they are better suited when the dataset has a small number of samples in some classes. Review SMOTE parameters: It can be helpful to tune SMOTE parameters or use a model that handles class imbalance well without relying on too much data.

Alternative: Try a Random Forest model to improve classification Without SMOTE

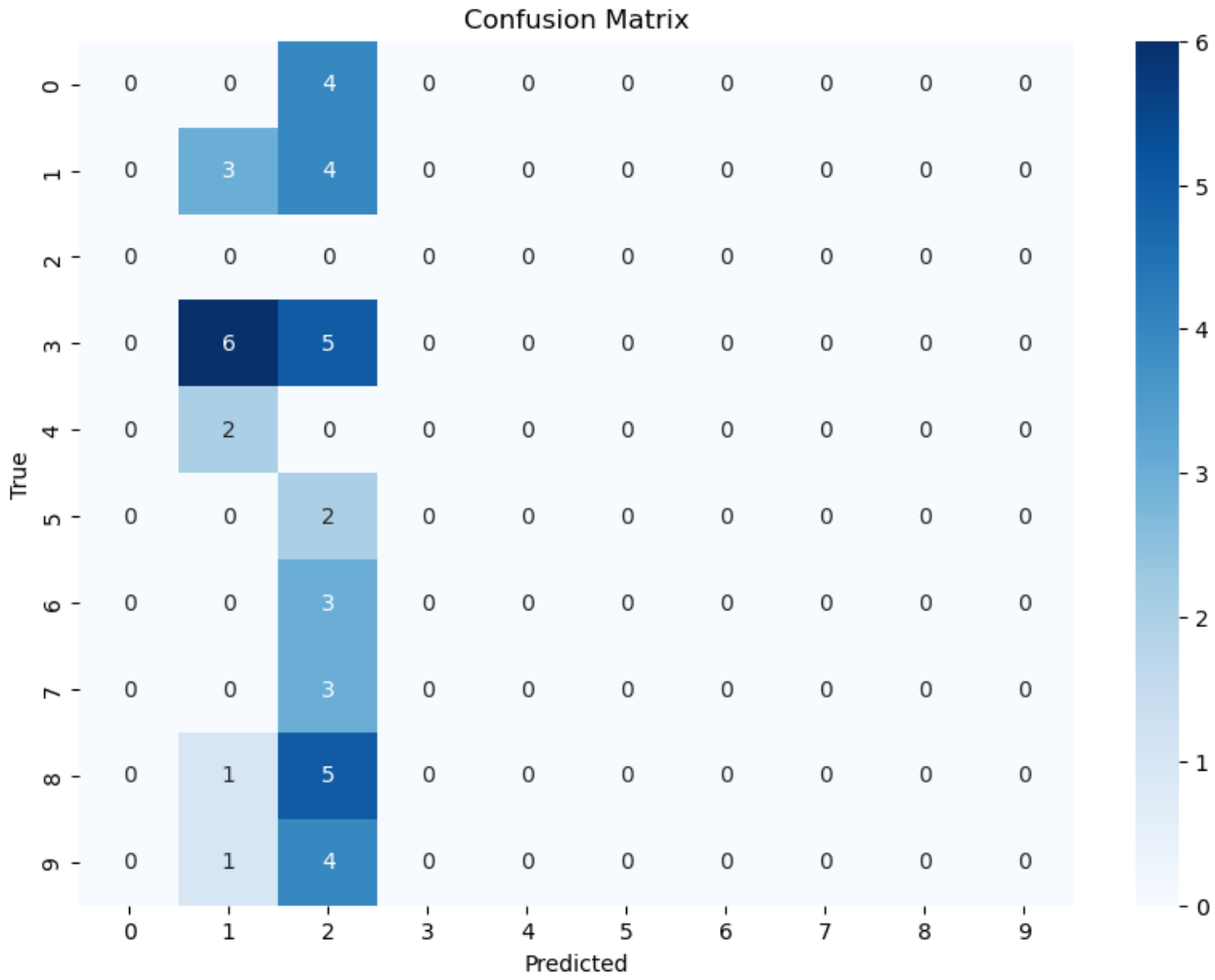
In [256…

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.preprocessing import LabelEncoder

# 1. Preprocess and split the data
X = df['Text About Uzbekistan']
y = df['Simplified Category']

# Convert categories to numeric labels
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(y)

# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y_encoded, test_si

# 2. Convert text data into numeric vectors using TF-IDF
vectorizer = TfidfVectorizer(stop_words='english', max_features=5000)
X_train_tfidf = vectorizer.fit_transform(X_train)
X_test_tfidf = vectorizer.transform(X_test)

# 3. Train the Random Forest model with the 'class_weight' parameter to h
model = RandomForestClassifier(n_estimators=100, random_state=42, class_w
model.fit(X_train_tfidf, y_train)

# 4.Making predictions with the test set
y_pred = model.predict(X_test_tfidf)

# Convert numerical predictions to category labels
y_pred_labels = label_encoder.inverse_transform(y_pred)
```

```
y_test_labels = label_encoder.inverse_transform(y_test)

# 5. Evaluate the model using classification and confusion matrix
print("Classification Report:\n", classification_report(y_test_labels, y_
print("Confusion Matrix:\n", confusion_matrix(y_test_labels, y_pred_label

# 6. Visualize the confusion matrix
plt.figure(figsize=(10, 7))
sns.heatmap(confusion_matrix(y_test_labels, y_pred_labels), annot=True, c
plt.title("Confusion Matrix")
plt.xlabel('Predicted')
plt.ylabel('True')

plt.savefig('Confusion Matrix.png', format='png')
plt.show()
```

```
Classification Report:
                precision    recall  f1-score   support

 Environmental       0.00      0.00      0.00         4
       Finance       0.23      0.43      0.30         7
    Governance       0.00      0.00      0.00         0
Infrastructure       0.00      0.00      0.00        11
      Policies       0.00      0.00      0.00         2
          Risk       0.00      0.00      0.00         2
        Social       0.00      0.00      0.00         3
Sustainability       0.00      0.00      0.00         3
    Technology       0.00      0.00      0.00         6
         Urban       0.00      0.00      0.00         5

      accuracy                           0.07        43
     macro avg       0.02      0.04      0.03        43
  weighted avg       0.04      0.07      0.05        43

Confusion Matrix:
 [[0 0 4 0 0 0 0 0 0 0]
 [0 3 4 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0]
 [0 6 5 0 0 0 0 0 0 0]
 [0 2 0 0 0 0 0 0 0 0]
 [0 0 2 0 0 0 0 0 0 0]
 [0 0 3 0 0 0 0 0 0 0]
 [0 0 3 0 0 0 0 0 0 0]
 [0 1 5 0 0 0 0 0 0 0]
 [0 1 4 0 0 0 0 0 0 0]]
```

```
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classificatio
n.py:1531: UndefinedMetricWarning: Precision is ill-defined and being set
to 0.0 in labels with no predicted samples. Use `zero_division` parameter
to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classificatio
n.py:1531: UndefinedMetricWarning: Recall is ill-defined and being set to
0.0 in labels with no true samples. Use `zero_division` parameter to contr
ol this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classificatio
n.py:1531: UndefinedMetricWarning: Precision is ill-defined and being set
to 0.0 in labels with no predicted samples. Use `zero_division` parameter
to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classificatio
n.py:1531: UndefinedMetricWarning: Recall is ill-defined and being set to
0.0 in labels with no true samples. Use `zero_division` parameter to contr
ol this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classificatio
n.py:1531: UndefinedMetricWarning: Precision is ill-defined and being set
to 0.0 in labels with no predicted samples. Use `zero_division` parameter
to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/opt/anaconda3/lib/python3.12/site-packages/sklearn/metrics/_classificatio
n.py:1531: UndefinedMetricWarning: Recall is ill-defined and being set to
0.0 in labels with no true samples. Use `zero_division` parameter to contr
ol this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```



Evaluation of the Result:

Classification Report:

- **Precision, Recall, F1-score**: All these values are quite low, especially for most categories.
- The categories with higher support (more samples) are **Finance** (7 samples) and **Infrastructure** (11 samples), but their performance remains very low in terms of precision and recall.
- The **Finance** category has a relatively higher recall of 0.43, meaning the model is somewhat more effective at identifying this category, although still insufficient.
- Precision and recall for other categories such as **Governance**, **Risk**, **Technology**, **Urban**, and others are 0, meaning the model is not correctly classifying any of these categories.

Accuracy: The overall accuracy is only 7%, indicating that the model is not making useful predictions for most classes.

Macro average and Weighted average:

- **Macro average** is the unweighted average of the metrics. It is very low (0.02

precision, 0.04 recall, 0.03 f1-score), indicating that the model is consistently failing across all categories.
- **Weighted average** has slightly higher values, but it is still low (0.04 precision, 0.07 recall, 0.05 f1-score). This reflects that some categories with more samples (such as Finance and Infrastructure) are contributing more to the results.

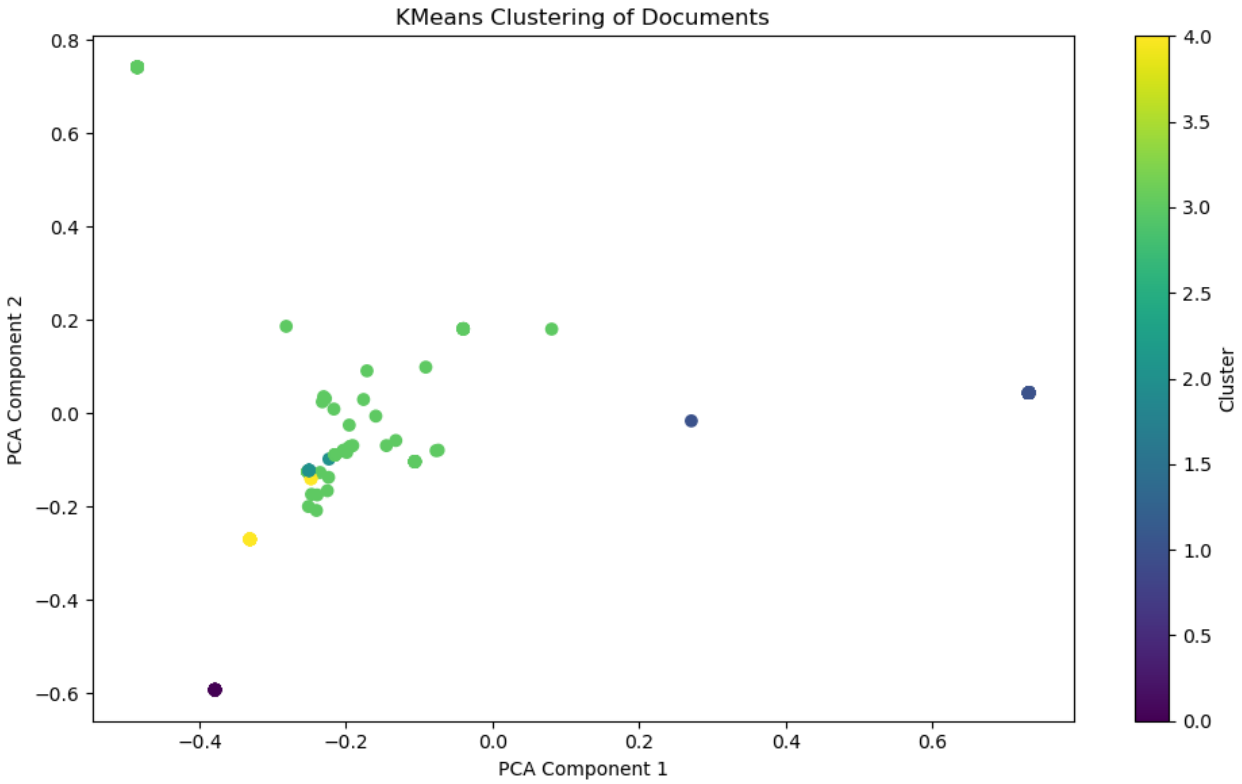
Confusion Matrix:

The confusion matrix shows how predictions are being made:

- In several classes, such as **Governance**, the model makes no correct predictions (all predictions are placed in other classes).
- In **Finance** and **Infrastructure**, predictions are distributed across several other classes, but there seem to be some correct predictions (especially in Finance).
- However, **Risk**, **Social**, **Technology**, and other classes have no correct predictions.

In [265…

```
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt

# 1. Vectorize the text using TF-IDF
X_tfidf = vectorizer.fit_transform(df['Category'])

# 2. Apply KMeans clustering
num_clusters = 5  # You can choose the number of clusters based on domain
kmeans = KMeans(n_clusters=num_clusters, random_state=42)
df['Cluster'] = kmeans.fit_predict(X_tfidf)

# 3. Visualize the clusters using PCA for dimensionality reduction
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_tfidf.toarray())

# Plot the clusters
plt.figure(figsize=(10, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=df['Cluster'], cmap='viridis')
plt.title('KMeans Clustering of Documents')
plt.xlabel('PCA Component 1')
plt.ylabel('PCA Component 2')
plt.colorbar(label='Cluster')
plt.tight_layout()

plt.savefig('pca.png', format='png')
plt.show()
```



KMeans Clustering:

This technique helps to find hidden patterns in the data by grouping similar documents together. After vectorizing the text data, we apply KMeans clustering with a specified number of clusters (num_clusters). We then visualize the results in a 2D plot using PCA (Principal Component Analysis) to reduce the dimensionality of the data, making it easier to visualize.

The resulting plot shows the output of applying **KMeans clustering** to the dataset after reducing its dimensionality using **PCA** (Principal Component Analysis) for visualization.

In [282…
```
df
```

Out[282…

	File Name	Category	Reason for Categorization
0	WEF_Reshaping_affordability_2024.docx	[Urban Affordability and Sustainable Urban Dev...	This text addresses the global urban affordabi...
1	smarsly2021e.docx	[Technological Innovation in Structural Health...	The text discusses advancements in the use of ...

2	UzbekistanRailways 2.docx	[Infrastructure Development and Public-Private...	This text discusses the construction of the Te...
3	Swedish_Waste_Management_A_Review_Articl.docx	[Environmental Impacts and Climate Change]	This article focuses on waste management strat...
4	MOF_LSE_IFC_event_yfnGTW0.docx	[Sustainable Finance and Economic Trends]	The text discusses Uzbekistan's economic growt...
...	...	...	...
208	Comprehensive Proposal for Toybola and Anvar A...	[Corporate Sustainability and Social Responsib...	The text highlights Toybola's commitment to su...
209	cbi_mr_h1_2024_02e_1.docx	[Sustainable Finance and Economic Trends]	This report analyzes the performance of green,...
210	Investments and construction 2.docx	[Infrastructure Development and Public-Private...	This document provides a detailed overview of ...
211	Dubai-Municipality-3DCP-Guideline-1st-Edition-...	[Infrastructure Development and Public-Private...	This document provides comprehensive guideline...
212	3D_Concrete_Printing_White_Paper_Ubez 2024-202...	[Infrastructure Development and Public-Private...	This white paper explores the transformative r...

213 rows × 7 columns

The resulting plot shows the output of applying **KMeans clustering** to the dataset after reducing its dimensionality using **PCA** (Principal Component Analysis) for visualization.

Interpretation of the Plot:

1. **Clustering**:
    - The scatter plot shows how the documents (represented by points) are grouped based on the **KMeans clustering algorithm**.
    - Each point represents a document, and the color of the points indicates which cluster they belong to. The `viridis` colormap is used, where different colors represent different clusters (ranging from purple to green).
2. **PCA Components**:
    - The axes represent the first two **principal components** (PCA components) of the dataset after dimensionality reduction.
    - **PCA Component 1** (x-axis) and **PCA Component 2** (y-axis) capture the most significant variance in the dataset. The plot shows how documents are distributed in a 2D space after reducing their original higher-dimensional TF-IDF features.
3. **Cluster Distribution**:
    - The clusters are scattered across the 2D space, with some clusters (like cluster 4) being more concentrated while others (like clusters 1 or 3) appear more dispersed.
    - Cluster 4 (represented by purple color) seems to have fewer points, suggesting that this cluster may contain fewer documents in this dataset.
4. **Interpretation**:
    - The plot provides a visual representation of how the documents are grouped by the KMeans algorithm. Ideally, clusters that are close together in the plot should share similar features, such as themes or topics, based on the content of the documents.
    - If the clusters do not show distinct separation or overlap too much, it might indicate that the features used for clustering (i.e., the text features vectorized via TF-IDF) are not sufficient to fully differentiate between the documents, or more fine-tuning of the clustering algorithm or feature engineering is needed.

In summary, this plot illustrates the clustering of documents based on their textual content after dimensionality reduction. Further tuning of the clustering algorithm, such as adjusting the number of clusters, could improve the quality of the clusters.

In [307…

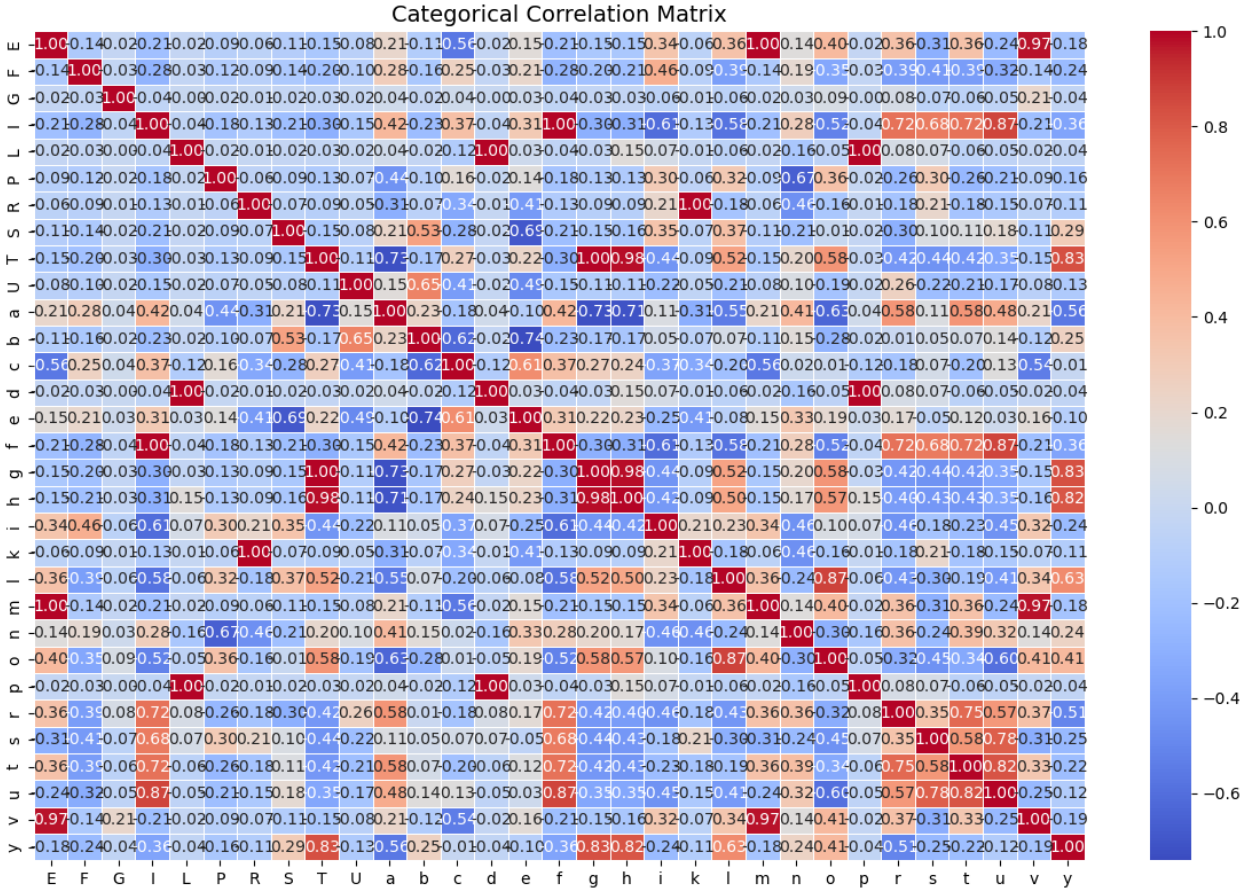
```
import pandas as pd
import numpy as np
from ast import literal_eval

# Function to ensure that categories are enclosed in quotes
def add_quotes_to_categories(text):
```

```
    # Replace any word not in quotes with quotation marks
    return re.sub(r'([A-Za-z0-9_]+)', r"'\1'", text)

df['Simplified Category'] = df['Simplified Category'].apply(lambda x: add

df['Simplified Category'] = df['Simplified Category'].apply(lambda x: lit

print(df['Simplified Category'].head())
```

```
0            Urban
1       Technology
2    Infrastructure
3     Environmental
4           Finance
Name: Simplified Category, dtype: object
```

In [314…

```
from sklearn.preprocessing import MultiLabelBinarizer

# We apply MultiLabelBinarizer to convert the categories into a binary re
mlb = MultiLabelBinarizer()

# Convert 'Simplified Category' column to a binary representation
category_matrix = mlb.fit_transform(df['Simplified Category'])

# We calculate the correlation matrix between the categories
category_df = pd.DataFrame(category_matrix, columns=mlb.classes_)

# We calculate the correlation matrix between the categories
correlation_matrix = category_df.corr()

plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f', l
plt.title('Categorical Correlation Matrix', fontsize=14)
plt.tight_layout()

plt.savefig('Categorical Correlation Matrix.png', format='png')
plt.show()
```



To explore the relationships between categories and how they interrelate, we can create a correlation matrix between the main categories based on the co-occurrence of categories in the documents. If we have multiple categories for each document, we can analyze how frequently the categories appear together. One way to do this is by using techniques such as One-Hot Encoding or TF-IDF to represent the presence of categories in the documents, and then calculate a correlation matrix to assess the relationships between the categories.

In [316…
```
correlation_matrix
```

Out[316…

	E	F	G	I	L	P	R
E	1.000000	-0.140297	-0.022109	-0.208622	-0.022109	-0.091741	-0.063592
F	-0.140297	1.000000	-0.029933	-0.282447	-0.029933	-0.124205	-0.086096
G	-0.022109	-0.029933	1.000000	-0.044510	-0.004717	-0.019573	-0.013568
I	-0.208622	-0.282447	-0.044510	1.000000	-0.044510	-0.184693	-0.128024
L	-0.022109	-0.029933	-0.004717	-0.044510	1.000000	-0.019573	-0.013568
P	-0.091741	-0.124205	-0.019573	-0.184693	-0.019573	1.000000	-0.056298
R	-0.063592	-0.086096	-0.013568	-0.128024	-0.013568	-0.056298	1.000000
S	-0.106462	-0.144136	-0.022714	-0.214330	-0.022714	-0.094251	-0.065332

T	-0.150006	-0.203089	-0.032004	-0.301993	-0.032004	-0.132800	-0.092054
U	-0.075120	-0.101703	-0.016027	-0.151233	-0.016027	-0.066504	-0.046099
a	0.206274	0.279267	0.044009	0.415271	0.044009	-0.444754	-0.308291
b	-0.114713	-0.155306	-0.024474	-0.230940	-0.024474	-0.101555	-0.070395
c	-0.559317	0.250837	0.039528	0.372995	-0.119331	0.164023	-0.343234
d	-0.022109	-0.029933	-0.004717	-0.044510	1.000000	-0.019573	-0.013568
e	0.154790	0.209566	0.033025	0.311624	0.033025	0.137036	-0.410829
f	-0.208622	-0.282447	-0.044510	1.000000	-0.044510	-0.184693	-0.128024
g	-0.150006	-0.203089	-0.032004	-0.301993	-0.032004	-0.132800	-0.092054
h	-0.152403	-0.206334	-0.032515	-0.306819	0.145069	-0.134923	-0.093525
i	0.342197	0.463289	-0.064609	-0.609657	0.073008	0.302947	0.209994
k	-0.063592	-0.086096	-0.013568	-0.128024	-0.013568	-0.056298	1.000000
l	0.358770	-0.391051	-0.061624	-0.581494	-0.061624	0.317619	-0.177251
m	1.000000	-0.140297	-0.022109	-0.208622	-0.022109	-0.091741	-0.063592
n	0.137834	0.186609	0.029407	0.277489	-0.160403	-0.665589	-0.461368
o	0.402874	-0.348241	0.085954	-0.517836	-0.054878	0.356664	-0.157847
p	-0.022109	-0.029933	-0.004717	-0.044510	1.000000	-0.019573	-0.013568
r	0.355381	-0.394780	0.075821	0.715454	0.075821	-0.258148	-0.178941
s	-0.305695	-0.413871	-0.065220	0.682453	0.072324	0.300106	0.208025
t	0.355381	-0.394780	-0.062212	0.715454	-0.062212	-0.258148	-0.178941
u	-0.239763	-0.324608	-0.051154	0.870118	-0.051154	-0.212263	-0.147135
v	0.973369	-0.144136	0.207670	-0.214330	-0.022714	-0.094251	-0.065332
y	-0.180619	-0.244535	-0.038535	-0.363624	-0.038535	-0.159902	-0.110840

31 rows × 31 columns

La "Matriz de correlación categórica" que se muestra en el mapa de calor visualiza las relaciones entre diferentes variables categóricas en su conjunto de datos, que probablemente sean las categorías simplificadas (por ejemplo, Medio ambiente, Finanzas, Infraestructura, etc.). Cada celda de la matriz representa el coeficiente de correlación entre dos categorías. A continuación, se muestra un desglose de lo que esto significa:

1. **Valores de correlación**:

- Los valores de correlación varían de -1 a +1.
- Un valor de **+1** indica una correlación positiva perfecta (es decir, las dos categorías siempre están asociadas entre sí).
- Un valor de **-1** indica una correlación negativa perfecta (es decir, las dos categorías nunca están asociadas entre sí).
- Un valor de **0** indica que no hay correlación entre las dos categorías.

2. **Escala de colores**:

- Los colores **rojos** indican correlaciones positivas más altas (más cercanas a +1), lo que implica que esas categorías tienden a aparecer juntas con frecuencia en su conjunto de datos.
- Los colores **azules** indican correlaciones negativas (más cercanas a -1), lo que sugiere que a medida que una categoría aumenta, la otra disminuye.
- Los **colores claros** como el azul claro o el rosa representan correlaciones débiles, ya sean positivas o negativas.

3. **Clústeres**:

- Algunas categorías, como **Infraestructura**, parecen tener fuertes correlaciones con otras categorías, lo que probablemente indica que los documentos clasificados bajo "Infraestructura" a menudo mencionan o están asociados con otros temas como **Finanzas** o **Políticas**.
- La matriz también puede mostrar algunas categorías con correlaciones débiles o nulas con otras, lo que sugiere que ciertos temas (como **Social** o **Riesgo**) son más independientes o rara vez se superponen con otras categorías en su conjunto de datos.

4. **Diagonal**:

- La diagonal de la matriz (de la parte superior izquierda a la parte inferior derecha) siempre mostrará una correlación de 1 porque una categoría siempre se correlacionará perfectamente consigo misma.

5. **Implicaciones**:

- Esta matriz es útil para comprender las relaciones entre diferentes categorías. Por ejemplo, si **Infraestructura** y **Finanzas** tienen una alta correlación, esto sugiere que es probable que los documentos que analizan proyectos de infraestructura también mencionen aspectos financieros, como financiación o inversiones.
- Puede utilizar esta información para explorar áreas en las que varias categorías podrían superponerse en términos de contenido, lo que le ayudará a identificar temas clave o intersecciones en sus documentos que podrían ser de interés para

un análisis más profundo.

En general, interpretar esta matriz ayuda a comprender cómo se interrelacionan los diferentes temas de su conjunto de datos, lo que puede ser útil para la agrupación, el modelado de temas o el refinamiento de sus modelos de clasificación.

Here is a detailed response to each of the project questions based on the analysis:

1. **What are the main trends in Uzbekistan according to the documents?**

The documents highlight key trends in **infrastructure development**, **sustainability**, and **economic reform**. There is significant emphasis on **public-private partnerships** (PPPs) in infrastructure projects, **sustainable finance**, and **climate change**. **Renewable energy** and **economic reforms** are also prominent, suggesting a shift towards more sustainable development strategies.

2. **What topics dominate the documents related to Uzbekistan (infrastructure, sustainability, economic reform)?**

- **Infrastructure** is the dominant topic, with a focus on **public-private partnerships (PPPs)**, **urban development**, and **smart infrastructure**.
- **Sustainability** is also a key focus, with documents discussing **green finance**, **renewable energy**, and **climate change mitigation** strategies.
- **Economic reform** is highlighted in relation to **sustainable development goals (SDGs)**, and there is mention of **energy transition** and **economic restructuring** to align with global sustainability goals.

3. **How are infrastructure and sustainability challenges being addressed in Uzbekistan?**

Infrastructure challenges are addressed through large-scale **PPP projects**, particularly in the areas of **urban development** and **renewable energy**. The documents suggest a growing commitment to **green infrastructure**, with projects aimed at reducing the carbon footprint of urban planning and energy production. **Sustainable finance** initiatives are also incorporated to fund infrastructure projects that prioritize environmental impacts.

4. **What infrastructure initiatives are most documented in relation to sustainable efforts in Uzbekistan?**

The **most documented infrastructure initiatives** are related to **smart cities**,

renewable energy projects (particularly solar and wind), and **energy-efficient urban planning**. Documents also emphasize **transport infrastructure** and **urban affordability**, showing how infrastructure is being developed to accommodate both growth and sustainability goals.

5. What areas are receiving the most attention in terms of public policies and reforms in Uzbekistan?

Public policies focus heavily on **economic reforms**, **energy transition**, and **sustainability**. Reforms in the energy sector to encourage **renewable energy** are discussed extensively, as well as **urban development policies** aimed at improving livability and reducing environmental impact. There is also mention of **legal frameworks** for sustainable finance and **climate change mitigation**.

6. What economic and policy reforms are highlighted in the documents?

- **Economic reforms** highlighted include changes to **energy policy**, **regulatory frameworks for green finance**, and the promotion of **foreign investments** in **sustainable development**.
- **Policy reforms** focus on creating favorable environments for **green finance**, **infrastructure development**, and aligning national policies with **sustainable development goals (SDGs)**.

7. What relationship exists between infrastructure initiatives and sustainable finance projects in Uzbekistan?

Infrastructure initiatives, especially those in **renewable energy** and **smart infrastructure**, are closely tied to **sustainable finance projects**. The documents indicate that **green bonds** and other forms of **sustainable investment** are being used to fund infrastructure projects, ensuring that they align with environmental and economic sustainability goals.

8. How are infrastructure projects connected to sustainable finance initiatives in Uzbekistan?

- **Infrastructure projects** in Uzbekistan are increasingly being financed through **sustainable finance instruments**, such as **green bonds**, **climate investment funds**, and **public-private partnerships (PPPs)**. These initiatives ensure that infrastructure development aligns with **climate change mitigation** and **sustainable development** objectives.

9. **Are there areas for improvement in the document categorization process?**

Yes, the categorization process can be improved. Some documents may not be accurately classified into the most relevant categories, especially for more specialized topics like **sustainable finance** or **economic reforms**. There may also be **overlap** between categories such as **infrastructure** and **sustainability**, leading to some ambiguity in classification.

10. **Are there inconsistencies in the categorization of documents that might affect subsequent analysis?**

Inconsistencies in categorization do exist. For example, some documents may fall under multiple categories (such as **Infrastructure** and **Sustainability**) but are only placed in one. This could affect the analysis by misrepresenting the importance of certain themes, especially those at the intersection of infrastructure and sustainability.

11. **What impact does climate change have on infrastructure policies and projects in Uzbekistan?**

Climate change plays a significant role in **infrastructure policies** in Uzbekistan. There is a clear focus on developing **climate-resilient infrastructure**, with **green energy** projects and **energy-efficient urban planning** featured prominently. These efforts aim to reduce carbon emissions, mitigate the effects of climate change, and build long-term sustainability into infrastructure projects.

12. **How are infrastructure and climate change related in documents about Uzbekistan?**

Infrastructure development is closely tied to **climate change** in the documents. Many projects are designed with a dual focus: enhancing infrastructure and ensuring it is resilient to the effects of climate change. This includes **renewable energy infrastructure**, **smart cities** designed to cope with climate impacts, and **sustainable transport** systems to reduce environmental footprints.

These insights should provide a comprehensive view of the main topics, trends, and relationships within the documents concerning Uzbekistan, particularly in the context of infrastructure, sustainability, and economic reform.

In [320…
```
df
```

Out[320…

	File Name	Category	Reason for Categorization

0	WEF_Reshaping_affordability_2024.docx	[Urban Affordability and Sustainable Urban Dev...	This text addresses the global urban affordabi...
1	smarsly2021e.docx	[Technological Innovation in Structural Health...	The text discusses advancements in the use of ...
2	UzbekistanRailways 2.docx	[Infrastructure Development and Public-Private...	This text discusses the construction of the Te...
3	Swedish_Waste_Management_A_Review_Articl.docx	[Environmental Impacts and Climate Change]	This article focuses on waste management strat...
4	MOF_LSE_IFC_event_yfnGTW0.docx	[Sustainable Finance and Economic Trends]	The text discusses Uzbekistan's economic growt...
...	...	...	...
208	Comprehensive Proposal for Toybola and Anvar A...	[Corporate Sustainability and Social Responsib...	The text highlights Toybola's commitment to su...
209	cbi_mr_h1_2024_02e_1.docx	[Sustainable Finance and Economic Trends]	This report analyzes the performance of green,...
210	Investments and construction 2.docx	[Infrastructure Development and Public-Private...	This document provides a detailed overview of ...
211	Dubai-Municipality-3DCP-Guideline-1st-Edition-...	[Infrastructure Development and Public-Private...	This document provides comprehensive guideline...
212	3D_Concrete_Printing_White_Paper_Ubez 2024-202...	[Infrastructure Development and Public-Private...	This white paper explores the transformative r...

213 rows × 7 columns

In [324…
```
df.to_csv('/Users/allisongarces/Downloads/Proyect AG/df_uzbekistan_analys
```

In [ ]:
```

```