Data Exploration – Load the data and show first rows

```
import pandas as pd

df = pd.read_excel("text_docs.xlsx")

print("Shape of dataset:", df.shape)
df.head()
```

```
Shape of dataset: (10, 2)
```

	document_id	text
0	1	The stock market has been experiencing volatil...
1	2	The economy is growing, and businesses are opt...
2	3	Climate change is a critical issue that needs ...
3	4	Advances in artificial intelligence have revol...
4	5	The rise of electric vehicles is shaping the f...

Next steps: Generate code with df | View recommended plots | New interactive sheet

Data Exploration – Show total rows and unique *documents*

```
print("Total rows:", len(df))
print("Unique documents:", df['document_id'].nunique())
print("Missing values per column:\n", df.isnull().sum())
```

```
Total rows: 10
Unique documents: 10
Missing values per column:
 document_id    0
text            0
dtype: int64
```

We calculate the total number of rows, the number of unique documents, and check for missing values.

Identify preprocessing steps

```
for i, text in enumerate(df['text'].head(5), 1):
    print(f"Document {i}: {text}\n")
```

```
Document 1: The stock market has been experiencing volatility due to the ec

Document 2: The economy is growing, and businesses are optimistic about the

Document 3: Climate change is a critical issue that needs immediate global

Document 4: Advances in artificial intelligence have revolutionized industr

Document 5: The rise of electric vehicles is shaping the future of the auto
```

We print a few sample texts to inspect what cleaning steps might be needed.

Task 2: Generate Topics Using LDA

Step 1: Preprocess the text

We will clean the text by: • Lowercasing • Removing punctuation and numbers • Removing stopwords • Tokenizing into words

```
import nltk
import re
from nltk.corpus import stopwords

nltk.download('stopwords')
stop_words = set(stopwords.words('english'))

def preprocess(text):
    text = text.lower()
    text = re.sub(r'[^a-z\s]', '', text)  # keep only letters
    tokens = text.split()
    tokens = [word for word in tokens if word not in stop_words]
    return tokens

df['tokens'] = df['text'].apply(preprocess)
df[['text', 'tokens']].head()
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

	text	tokens
0	The stock market has been experiencing volatil...	[stock, market, experiencing, volatility, due,...
1	The economy is growing, and businesses are opt...	[economy, growing, businesses, optimistic, fut...
2	Climate change is a critical issue that needs ...	[climate, change, critical, issue, needs, imme...
3	Advances in artificial intelligence have	[advances, artificial, intelligence,

Task 2: Generate Topics Using LDA

Step 2: Create Document–Term Matrix

We will now build a dictionary and a document-term matrix (bag-of-words) using Gensim.

```
!pip install gensim
```

```
Requirement already satisfied: gensim in /usr/local/lib/python3.12/dist-pac
Requirement already satisfied: numpy<2.0,>=1.18.5 in /usr/local/lib/python3
Requirement already satisfied: scipy<1.14.0,>=1.7.0 in /usr/local/lib/pytho
Requirement already satisfied: smart-open>=1.8.1 in /usr/local/lib/python3.
Requirement already satisfied: wrapt in /usr/local/lib/python3.12/dist-pack
```

```
from gensim import corpora

dictionary = corpora.Dictionary(df['tokens'])
corpus = [dictionary.doc2bow(text) for text in df['tokens']]

print("Number of unique words:", len(dictionary))
print("Sample document-term matrix (first document):", corpus[0])
```

```
Number of unique words: 62
Sample document-term matrix (first document): [(0, 1), (1, 1), (2, 1), (3,
```

Task 2: Generate Topics Using LDA

Step 3: Apply LDA model and extract topics

```
from gensim.models import LdaModel

# Train LDA model
lda_model = LdaModel(corpus=corpus, id2word=dictionary, num_topics=5, random_st

# Display top 5 words for each topic
for idx, topic in lda_model.print_topics(num_words=5):
    print(f"Topic {idx+1}: {topic}")
```

```
Topic 1: 0.036*"attention" + 0.036*"issue" + 0.036*"critical" + 0.036*"chan
Topic 2: 0.069*"future" + 0.069*"businesses" + 0.069*"optimistic" + 0.069*"
Topic 3: 0.045*"industry" + 0.045*"uncertainty" + 0.045*"shaping" + 0.045*"
Topic 4: 0.047*"renewable" + 0.047*"energy" + 0.047*"around" + 0.047*"techn
Topic 5: 0.047*"platforms" + 0.047*"digital" + 0.047*"become" + 0.047*"ongo
```

The LDA model extracted 5 topics from the dataset. Each topic is represented by its most important words. For example, one topic relates to economy and business growth, another to renewable energy and technology, and another to climate/global issues. This shows that the model is successfully grouping documents into meaningful themes.

By, Utkarsh Anand