week2-immigration

October 21, 2023

IMMIGRATION:

FOR CLEANING DATA ABOUT VISA

```python
[3]: import pandas as pd
     import re

     # Define a function to perform the data cleaning tasks
     def clean_text(text):
         # Convert to lowercase
         text = text.lower()
         # Remove links (URLs)
         text = re.sub(r'http\S+', '', text)

         # Remove punctuation and symbols
         text = re.sub(r'[^\w\s]', ' ', text)

         # Remove @user mentions
         text = re.sub(r'@\w+', '', text)

         # Remove hashtags
         text = re.sub(r'#\w+', '', text)



         return text

     # Load the data from an Excel file
     file_path = "/Users/harshith/Downloads/WEEK2/Immigration/visa.csv"
     df = pd.read_csv(file_path)

     # Apply the cleaning function to the specified column
     df['full_text'] = df['full_text'].apply(clean_text)
     # Remove duplicates
     df.drop_duplicates(subset=['full_text'], inplace=True)

     # Remove rows with null values
     df.dropna(subset=['full_text'], inplace=True)
```

```python
# Save the cleaned data back to a new Excel file
cleaned_file_path = "/Users/harshith/Desktop/untitled folder 3/immigration/visa.
  ↪csv"
df.to_csv(cleaned_file_path, index=False)

print(f"Cleaned data saved to {cleaned_file_path}")
```

Cleaned data saved to /Users/harshith/Desktop/untitled folder
3/immigration/visa.csv

```python
[ ]:
```

```python
[5]: import pandas as pd
     import re

     # Define a function to perform the data cleaning tasks
     def clean_text(text):
         # Convert to lowercase
         text = text.lower()
         # Remove links (URLs)
         text = re.sub(r'http\S+', '', text)

         # Remove punctuation and symbols
         text = re.sub(r'[^\w\s]', ' ', text)

         # Remove @user mentions
         text = re.sub(r'@\w+', '', text)

         # Remove hashtags
         text = re.sub(r'#\w+', '', text)



         return text

     # Load the data from an Excel file
     file_path = "/Users/harshith/Downloads/WEEK2/Immigration/passport.csv"
     df = pd.read_csv(file_path)

     # Apply the cleaning function to the specified column
     df['full_text'] = df['full_text'].apply(clean_text)
     # Remove duplicates
     df.drop_duplicates(subset=['full_text'], inplace=True)

     # Remove rows with null values
     df.dropna(subset=['full_text'], inplace=True)
```

```python
# Save the cleaned data back to a new Excel file
cleaned_file_path = "/Users/harshith/Desktop/untitled folder 3/immigration/
  ↪passport.csv"
df.to_csv(cleaned_file_path, index=False)

print(f"Cleaned data saved to {cleaned_file_path}")
```

Cleaned data saved to /Users/harshith/Desktop/untitled folder
3/immigration/passport.csv

```python
[6]: import pandas as pd
import re

# Define a function to perform the data cleaning tasks
def clean_text(text):
    # Convert to lowercase
    text = text.lower()
    # Remove links (URLs)
    text = re.sub(r'http\S+', '', text)

    # Remove punctuation and symbols
    text = re.sub(r'[^\w\s]', ' ', text)

    # Remove @user mentions
    text = re.sub(r'@\w+', '', text)

    # Remove hashtags
    text = re.sub(r'#\w+', '', text)



    return text

# Load the data from an Excel file
file_path = "/Users/harshith/Downloads/WEEK2/Immigration/security check.csv"
df = pd.read_csv(file_path)

# Apply the cleaning function to the specified column
df['full_text'] = df['full_text'].apply(clean_text)
# Remove duplicates
df.drop_duplicates(subset=['full_text'], inplace=True)

# Remove rows with null values
df.dropna(subset=['full_text'], inplace=True)

# Save the cleaned data back to a new Excel file
```

```
cleaned_file_path = "/Users/harshith/Desktop/untitled folder 3/immigration/
 ↪security.csv"
df.to_csv(cleaned_file_path, index=False)

print(f"Cleaned data saved to {cleaned_file_path}")
```

Cleaned data saved to /Users/harshith/Desktop/untitled folder
3/immigration/security.csv

[ ]: