Lab Assignment 1

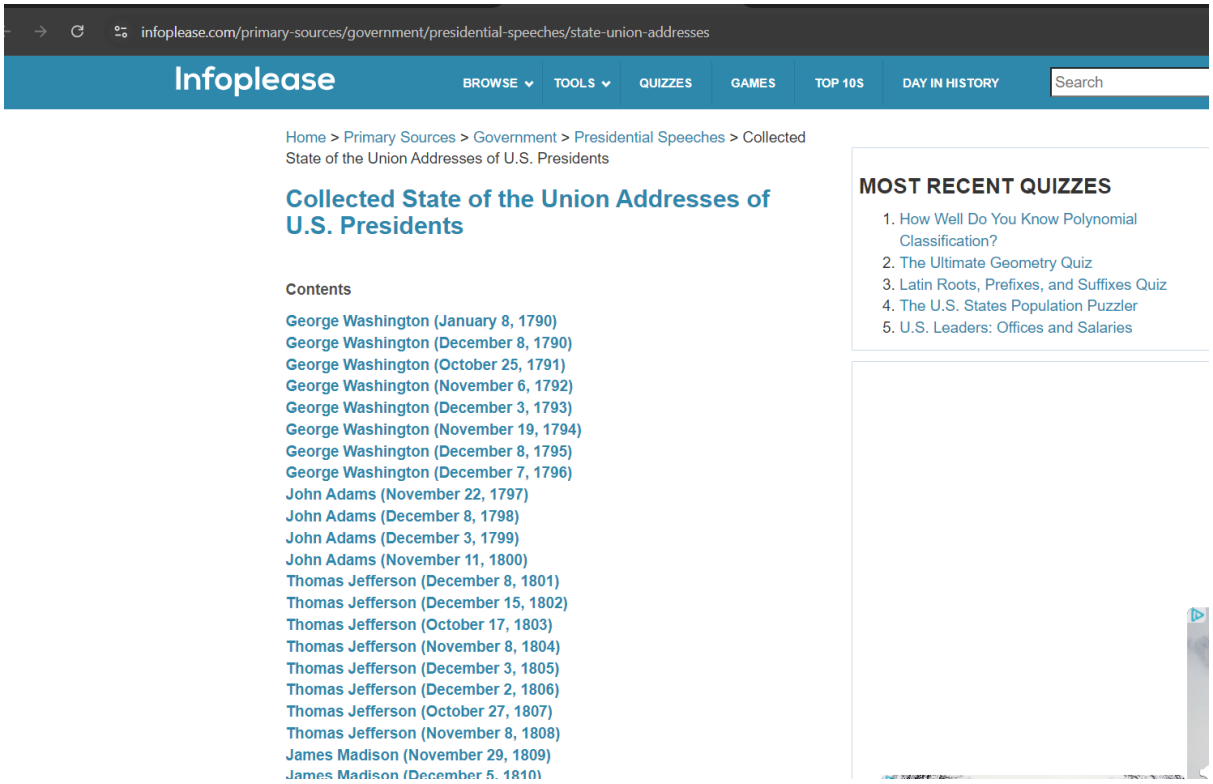
CIS 612 Big Data and Parallel Data Processing

Big data processing methods for Information Extraction from a collection of web documents as Semi-Structured Data or Unstructured Data

Method 1: Webpage as Semi-Structured Data Model with a DOM Parser and XPath

Method 2: Webpages as Unstructured Data with Parsing and Searching

Collected State of the Union Addresses of U.S. Presidents in the following website at https://www.infoplease.com/homework-help/history/collected-state-union-addresses-us-presidents



Part 1: For each item (line) in the contents in the HTML file of the page, extract the following 5 information:

- Name of President
- Date of Union Address
- Link to Address
- Filename of Address
- Text of Address

Extract the five information to write them in a table structure in CSV/TSV file with each Address per line or insert a record of the five column values into a table in your SQL server. Each value of "Name of President" and "Date of Union Address" are displayed as String and create a Link with each URL

to Address as a value of “Link to Address” column. Store the text of each Address in Full Address Text column. You may create it as Blob if necessary. For Example, your script extracts three information from the following html code of the site and display the three columns in a table format in the page as below.

George Washington (January 8, 1790)

George Washington (December 8, 1790)

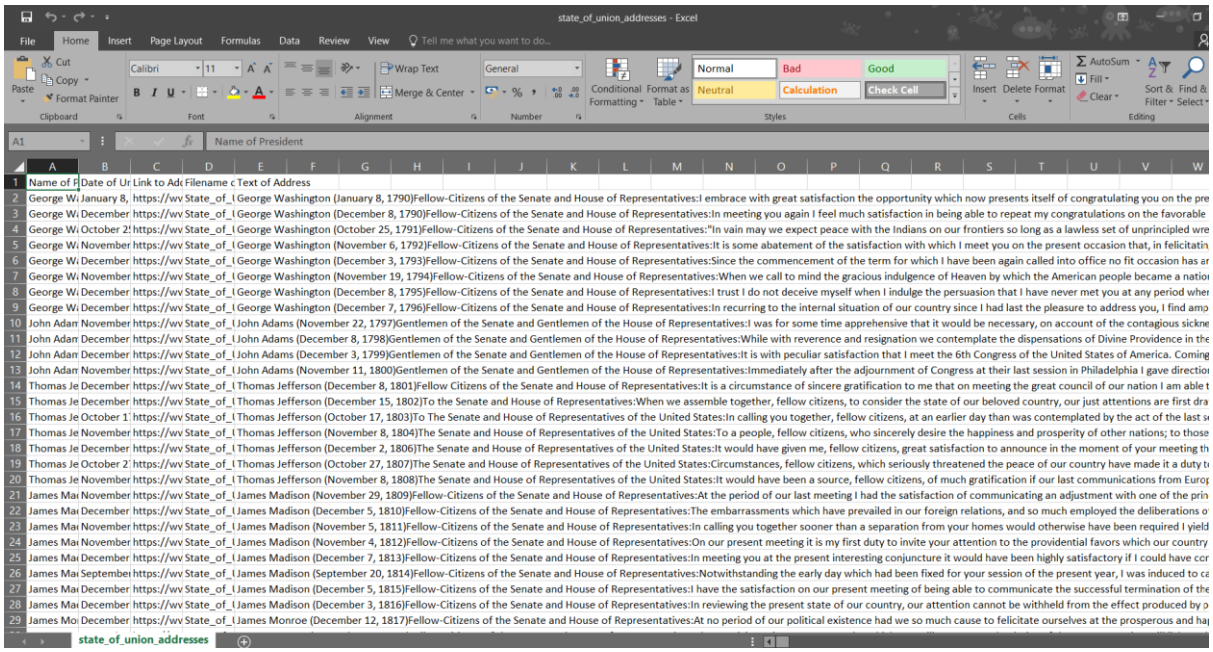
Save the full text of each address in a file and store the file path as pointer to the file in a column so that you can create a table from it to retrieve them anytime later and create another column to save the full text of address. The structured text file (for example, CSV, TSV) created would look like the below. Note your transformed CSV/TSV text file wouldn’t have the heading with the column names. They are for a table scheme (You may change your column names when creating a table in your SQL Server). For some Big Data API, it may require to insert the heading with column titles in the first line of your CSV file to be created as a table in a database server.

Source code for WebScraping which shows the output in CSV file:

```
import requests
from bs4 import BeautifulSoup
import csv
import os

# The URL of the website to scrape
URL = "https://www.infoplease.com/primary-sources/government/presidential-speeches/state-union-addresses"

# Create a directory to save the address text files
os.makedirs("State_of_Union_Addresses", exist_ok=True)

# Function to extract data from the webpage
def extract_state_of_union_addresses(url):
    try:
        # Send a GET request to the URL
        response = requests.get(url)

        # Check if the request was successful (status code 200)
        if response.status_code == 200:
            # Parse the HTML content using BeautifulSoup
            soup = BeautifulSoup(response.content, 'html.parser')

            # Find all links to the addresses
            address_links = soup.select("a[href*='state-union-address']")

            # Open a CSV file to write data
            with open('state_of_union_addresses.csv', mode='w', newline='', encoding='utf-8') as csv_file:
                writer = csv.writer(csv_file)
                writer.writerow(['Name of President', 'Date of Union Address', 'Link to Address', 'Filename of Address', 'Text of Address'])
```

```
            # Iterate over each link and extract required information
            for link in address_links:
                address_url = "https://www.infoplease.com" + link['href']
                address_response = requests.get(address_url)

                if address_response.status_code == 200:
                    address_soup = BeautifulSoup(address_response.content, 'html.parser')

                    # Extract President name and date from link text
                    address_text = link.text.strip()
                    try:
                        name, date = address_text.split(' (')
                        date = date.rstrip(')')
                    except ValueError:
                        continue  # If the text doesn't match the expected pattern, skip it

                    # Extract the specific part of the address using the equivalent CSS selector
                    address_div = address_soup.select_one("#mainaside > article > div:nth-of-type(2) >
div:nth-of-type(2)")
                    if address_div:
                        address_full_text = address_div.get_text(strip=True)

                        # Save the text of the address to a file
                        filename = f"State_of_Union_Addresses/{name.replace(' ', '_')}_{date.replace(' ',
'_').replace(',', '')}.txt"
                        with open(filename, 'w', encoding='utf-8') as text_file:
                            text_file.write(address_full_text)

                        # Write data to the CSV file
                        writer.writerow([name, date, address_url, filename, address_full_text])

            print("Data extraction complete. Check 'state_of_union_addresses.csv' and
'State_of_Union_Addresses' directory for results.")

        else:
            print(f"Failed to retrieve the webpage. Status code: {response.status_code}")

    except requests.exceptions.RequestException as e:
        print(f"An error occurred: {e}")

# Call the function
extract_state_of_union_addresses(URL)
```

CSV file output:

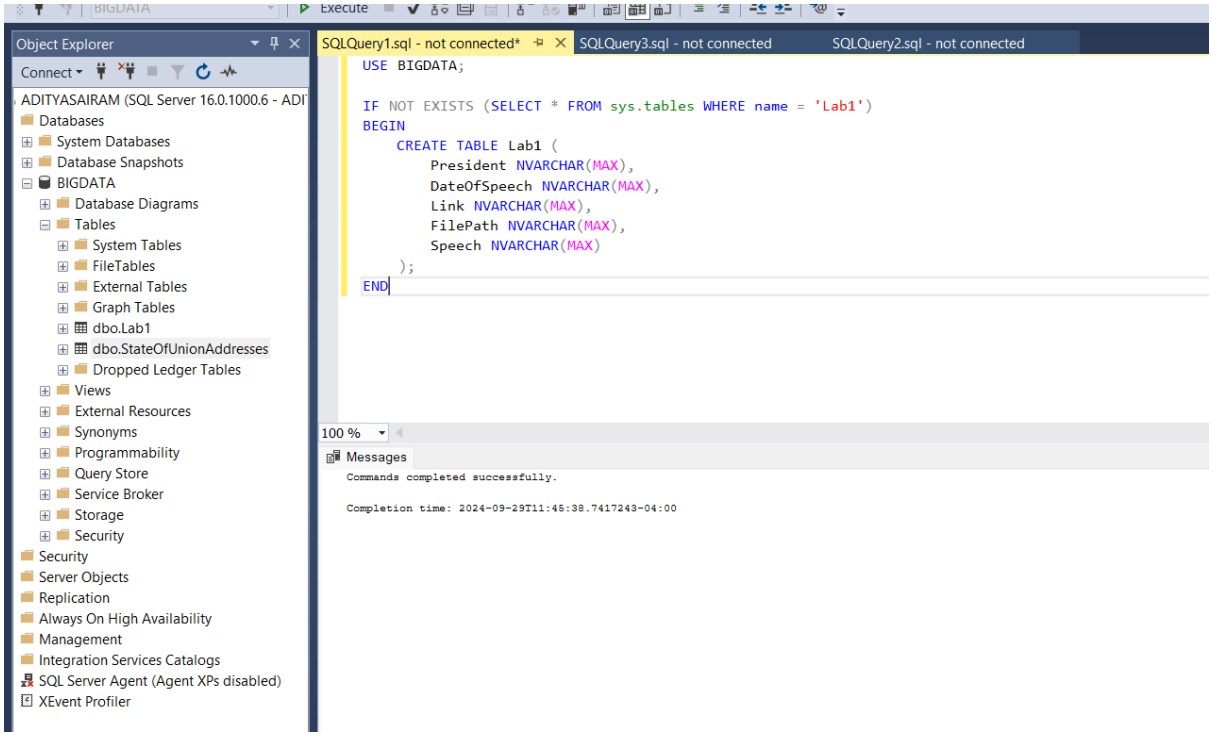


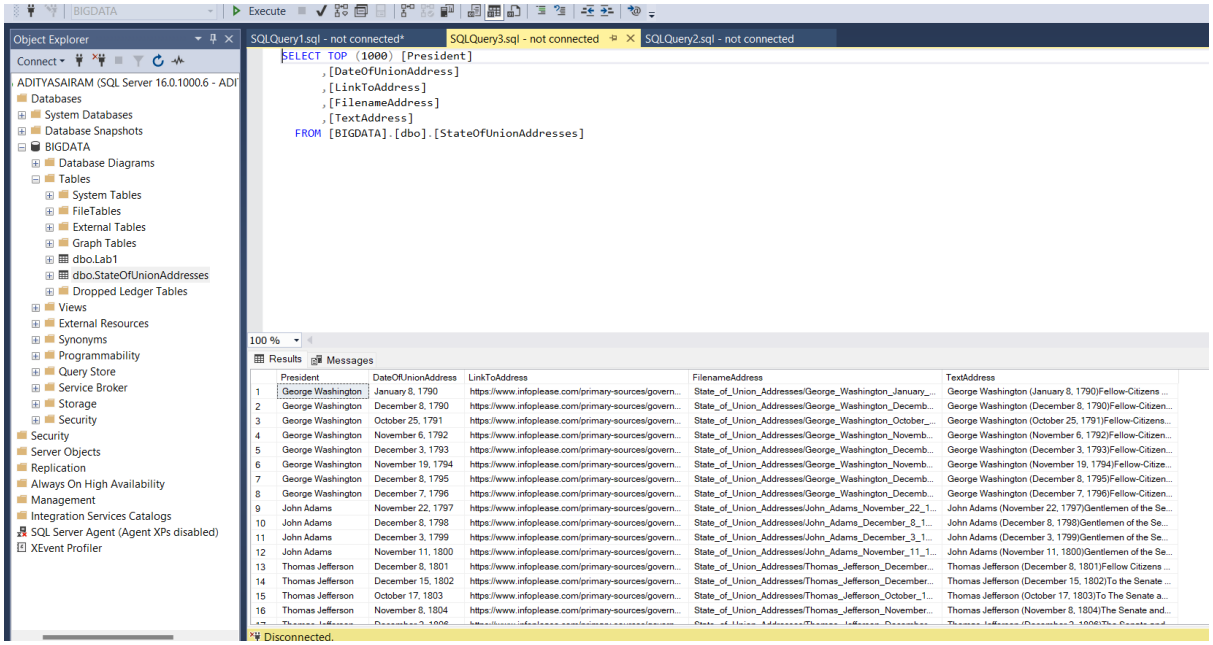
Name of P	Date of Ur	Link to Ad	Filename c	Text of Address
George Wa	January 8,	https://wv	State_of_U	George Washington (January 8, 1790)Fellow-Citizens of the Senate and House of Representatives:I embrace with great satisfaction the opportunity which now presents itself of congratulating you on the pre
George Wa	December	https://wv	State_of_U	George Washington (December 8, 1790)Fellow-Citizens of the Senate and House of Representatives:In meeting you again I feel much satisfaction in being able to repeat my congratulations on the favorable
George Wa	October 25	https://wv	State_of_U	George Washington (October 25, 1791)Fellow-Citizens of the Senate and House of Representatives:"In vain may we expect peace with the Indians on our frontiers so long as a lawless set of unprincipled wre
George Wa	November	https://wv	State_of_U	George Washington (November 6, 1792)Fellow-Citizens of the Senate and House of Representatives:It is some abatement of the satisfaction with which I meet you on the present occasion that, in felicitatin
George Wa	December	https://wv	State_of_U	George Washington (December 3, 1793)Fellow-Citizens of the Senate and House of Representatives:Since the commencement of the term for which I have been again called into office no fit occasion has ar
George Wa	November	https://wv	State_of_U	George Washington (November 19, 1794)Fellow-Citizens of the Senate and House of Representatives:When we call to mind the gracious indulgence of Heaven by which the American people became a natio
George Wa	December	https://wv	State_of_U	George Washington (December 8, 1795)Fellow-Citizens of the Senate and House of Representatives:I trust I do not deceive myself when I indulge the persuasion that I have never met you at any period wher
George Wa	December	https://wv	State_of_U	George Washington (December 7, 1796)Fellow-Citizens of the Senate and House of Representatives:In recurring to the internal situation of our country since I had last the pleasure to address you, I find amp
John Adam	November	https://wv	State_of_U	John Adams (November 22, 1797)Gentlemen of the Senate and Gentlemen of the House of Representatives:I was for some time apprehensive that it would be necessary, on account of the contagious sickne
John Adam	December	https://wv	State_of_U	John Adams (December 8, 1798)Gentlemen of the Senate and Gentlemen of the House of Representatives:While with reverence and resignation we contemplate the dispensations of Divine Providence in the
John Adam	December	https://wv	State_of_U	John Adams (December 3, 1799)Gentlemen of the Senate and Gentlemen of the House of Representatives:It is with peculiar satisfaction that I meet the 6th Congress of the United States of America. Coming
John Adam	November	https://wv	State_of_U	John Adams (November 11, 1800)Gentlemen of the Senate and Gentlemen of the House of Representatives:Immediately after the adjournment of Congress at their last session in Philadelphia I gave directio
Thomas Je	December	https://wv	State_of_U	Thomas Jefferson (December 8, 1801)Fellow Citizens of the Senate and House of Representatives:It is a circumstance of sincere gratification to me that on meeting the great council of our nation I am able t
Thomas Je	December	https://wv	State_of_U	Thomas Jefferson (December 15, 1802)To the Senate and House of Representatives:When we assemble together, fellow citizens, to consider the state of our beloved country, our just attentions are first dra
Thomas Je	October 17	https://wv	State_of_U	Thomas Jefferson (October 17, 1803)To The Senate and House of Representatives of the United States:In calling you together, fellow citizens, at an earlier day than was contemplated by the act of the last s
Thomas Je	November	https://wv	State_of_U	Thomas Jefferson (November 8, 1804)The Senate and House of Representatives of the United States:To a people, fellow citizens, who sincerely desire the happiness and prosperity of other nations; to those
Thomas Je	December	https://wv	State_of_U	Thomas Jefferson (December 2, 1806)The Senate and House of Representatives of the United States:It would have given me, fellow citizens, great satisfaction to announce in the moment of your meeting th
Thomas Je	October 27	https://wv	State_of_U	Thomas Jefferson (October 27, 1807)The Senate and House of Representatives of the United States:Circumstances, fellow citizens, which seriously threatened the peace of our country have made it a duty t
Thomas Je	November	https://wv	State_of_U	Thomas Jefferson (November 8, 1808)The Senate and House of Representatives of the United States:It would have been a source, fellow citizens, of much gratification if our last communications from Europ
James Ma	November	https://wv	State_of_U	James Madison (November 29, 1809)Fellow-Citizens of the Senate and House of Representatives:At the period of our last meeting I had the satisfaction of communicating an adjustment with one of the prin
James Ma	December	https://wv	State_of_U	James Madison (December 5, 1810)Fellow-Citizens of the Senate and House of Representatives:The embarrassments which have prevailed in our foreign relations, and so much employed the deliberations o
James Ma	November	https://wv	State_of_U	James Madison (November 5, 1811)Fellow-Citizens of the Senate and House of Representatives:In calling you together sooner than a separation from your homes would otherwise have been required I yield
James Ma	November	https://wv	State_of_U	James Madison (November 4, 1812)Fellow-Citizens of the Senate and House of Representatives:On our present meeting it is my first duty to invite your attention to the providential favors which our country
James Ma	December	https://wv	State_of_U	James Madison (December 7, 1813)Fellow-Citizens of the Senate and House of Representatives:In meeting you at the present interesting conjuncture it would have been highly satisfactory if I could have cor
James Ma	Septembe	https://wv	State_of_U	James Madison (September 20, 1814)Fellow-Citizens of the Senate and House of Representatives:Notwithstanding the early day which had been fixed for your session of the present year, I was induced to ca
James Ma	December	https://wv	State_of_U	James Madison (December 5, 1815)Fellow-Citizens of the Senate and House of Representatives:I have the satisfaction on our present meeting of being able to communicate the successful termination of the
James Ma	December	https://wv	State_of_U	James Madison (December 3, 1816)Fellow-Citizens of the Senate and House of Representatives:In reviewing the present state of our country, our attention cannot be withheld from the effect produced by p
James Mo	December	https://wv	State_of_U	James Monroe (December 12, 1817)Fellow-Citizens of the Senate and House of Representatives:At no period of our political existence had we so much cause to felicitate ourselves at the prosperous and ha

Source code for database server connection and retrieving the data:

```
import pyodbc
import csv

# Set up the connection to SQL Server
conn = pyodbc.connect(
    'DRIVER={SQL Server};'
    'SERVER=ADITYASAIRAM;'
    'DATABASE=BIGDATA;'
)

cursor = conn.cursor()

# Create a table if it doesn't already exist
cursor.execute('''
    IF OBJECT_ID('dbo.StateOfUnionAddresses', 'U') IS NULL
    CREATE TABLE StateOfUnionAddresses (
        President NVARCHAR(255),
        DateOfUnionAddress NVARCHAR(255),
        LinkToAddress NVARCHAR(1024),
        FilenameAddress NVARCHAR(1024),
        TextAddress TEXT
    )
''')

# Read the data from CSV file
csv_filename = "state_of_union_addresses.csv"
with open(csv_filename, 'r', encoding='utf-8') as csvfile:
    csvreader = csv.reader(csvfile)
```

```
    next(csvreader)  # Skip the header

    # Insert each row into the SQL Server table
    for row in csvreader:
        cursor.execute('''
            INSERT INTO StateOfUnionAddresses (President, DateOfUnionAddress, LinkToAddress,
FilenameAddress, TextAddress)
            VALUES (?, ?, ?, ?, ?)
        ''', row)

# Commit the transaction
conn.commit()

# Close the connection
cursor.close()
conn.close()

print("Data successfully inserted into SQL Server.")
```





Part 2: Create one big text file Create one big text file to include all the Address text that were extracted from each html file of each Address. Combine (Append) each address text into one big text file that has all the texts of the Union Addresses (each text starts with its name of President followed by date) so that you can perform further data processing for Big data analytics on the big, combined text of your entire collection such as word count of your entire collection to do text mining process. For some phases of data processing later, you will need to process all the texts, so it is easier to have them in one file.
You may use XPath library call for this Lab. See the instruction of How to Debug and Set Up XPath Library in the Lab section.

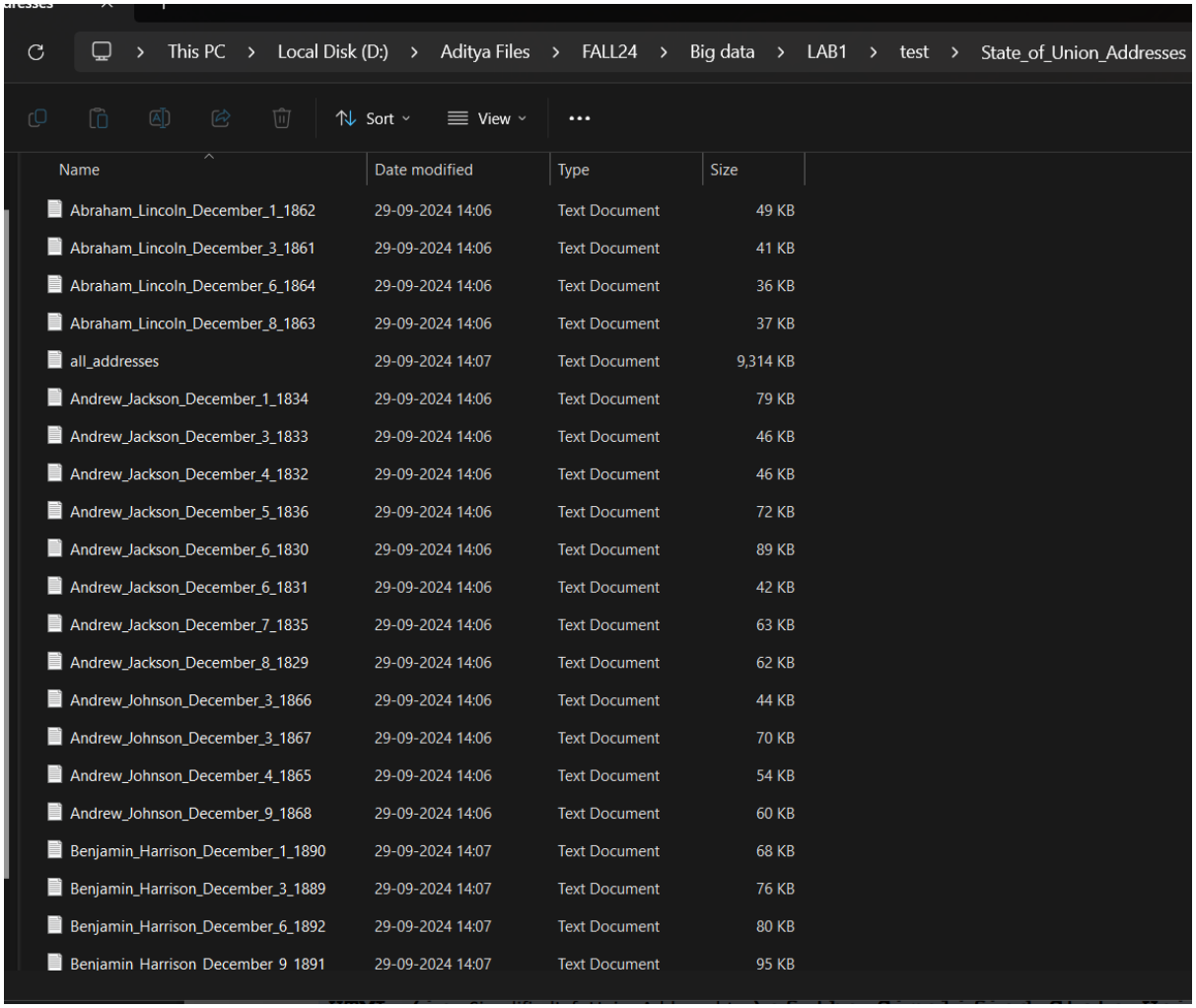
Source code :

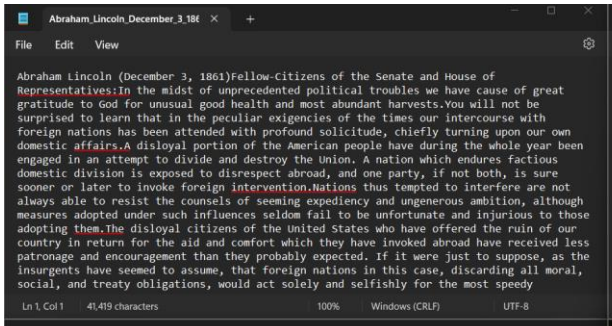
```
import requests
from bs4 import BeautifulSoup
import csv
import os

# The URL of the website to scrape
URL = "https://www.infoplease.com/primary-sources/government/presidential-speeches/state-union-addresses"

# Create a directory to save the address text files
os.makedirs("State_of_Union_Addresses", exist_ok=True)

# Function to extract data from the webpage
def extract_state_of_union_addresses(url):
    try:
        # Send a GET request to the URL
        response = requests.get(url)

        # Check if the request was successful (status code 200)
```

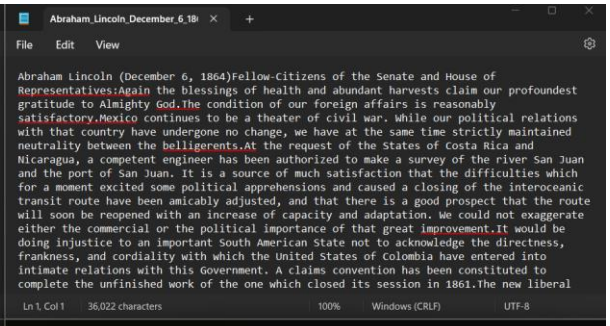
```
    if response.status_code == 200:
        # Parse the HTML content using BeautifulSoup
        soup = BeautifulSoup(response.content, 'html.parser')

        # Find all links to the addresses
        address_links = soup.select("a[href*='state-union-address']")

        # Open a CSV file to write data
        with open('state_of_union_addresses.csv', mode='w', newline='', encoding='utf-8') as csv_file, \
             open('State_of_Union_Addresses/all_addresses.txt', 'w', encoding='utf-8') as big_file:

            writer = csv.writer(csv_file)
            writer.writerow(['Name of President', 'Date of Union Address', 'Link to Address', 'Filename of Address', 'Text of Address'])

            # Iterate over each link and extract required information
            for link in address_links:
                address_url = "https://www.infoplease.com" + link['href']
                address_response = requests.get(address_url)

                if address_response.status_code == 200:
                    address_soup = BeautifulSoup(address_response.content, 'html.parser')

                    # Extract President name and date from link text
                    address_text = link.text.strip()
                    try:
                        name, date = address_text.split(' (')
                        date = date.rstrip(')')
                    except ValueError:
                        continue  # If the text doesn't match the expected pattern, skip it

                    # Extract the specific part of the address using the equivalent CSS selector
                    address_div = address_soup.select_one("#mainaside > article > div:nth-of-type(2) > div:nth-of-type(2)")
                    if address_div:
                        address_full_text = address_div.get_text(strip=True)

                        # Save the text of the address to an individual file
                        filename = f"State_of_Union_Addresses/{name.replace(' ', '_')}_{date.replace(' ', '_').replace(',', '')}.txt"
                        with open(filename, 'w', encoding='utf-8') as text_file:
                            text_file.write(address_full_text)

                        # Write the address text to the big file
                        big_file.write(f"State of the Union Address by {name} on {date}\n")
                        big_file.write(f"{address_full_text}\n")
                        big_file.write("="*80 + "\n\n")
```

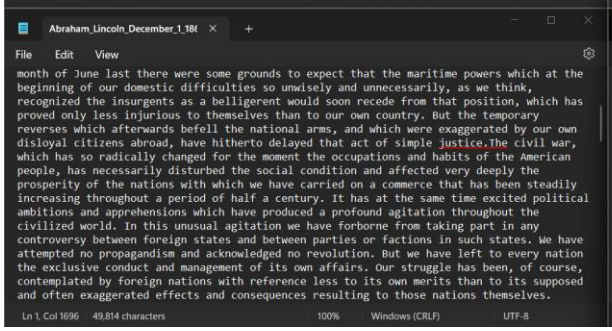
```
                    # Write data to the CSV file
                    writer.writerow([name, date, address_url, filename, address_full_text])

        print("Data extraction complete. Check 'state_of_union_addresses.csv',
'State_of_Union_Addresses' directory, and 'all_addresses.txt' for results.")

    else:
        print(f"Failed to retrieve the webpage. Status code: {response.status_code}")

except requests.exceptions.RequestException as e:
    print(f"An error occurred: {e}")

# Call the function
extract_state_of_union_addresses(URL)
```

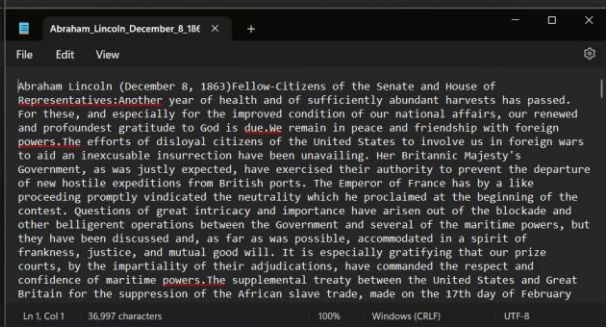
output:




Abraham_Lincoln_December_3_18€
File Edit View
Abraham Lincoln (December 3, 1861)Fellow-Citizens of the Senate and House of Representatives:In the midst of unprecedented political troubles we have cause of great gratitude to God for unusual good health and most abundant harvests.You will not be surprised to learn that in the peculiar exigencies of the times our intercourse with foreign nations has been attended with profound solicitude, chiefly turning upon our own domestic affairs.A disloyal portion of the American people have during the whole year been engaged in an attempt to divide and destroy the Union. A nation which endures factious domestic division is exposed to disrespect abroad, and one party, if not both, is sure sooner or later to invoke foreign intervention.Nations thus tempted to interfere are not always able to resist the counsels of seeming expediency and ungenerous ambition, although measures adopted under such influences seldom fail to be unfortunate and injurious to those adopting them.The disloyal citizens of the United States who have offered the ruin of our country in return for the aid and comfort which they have invoked abroad have received less patronage and encouragement than they probably expected. If it were just to suppose, as the insurgents have seemed to assume, that foreign nations in this case, discarding all moral, social, and treaty obligations, would act solely and selfishly for the most speedy
Ln 1, Col 1 41,419 characters 100% Windows (CRLF) UTF-8


Abraham_Lincoln_December_6_18
File Edit View
Abraham Lincoln (December 6, 1864)Fellow-Citizens of the Senate and House of Representatives:Again the blessings of health and abundant harvests claim our profoundest gratitude to Almighty God.The condition of our foreign affairs is reasonably satisfactory.Mexico continues to be a theater of civil war. While our political relations with that country have undergone no change, we have at the same time strictly maintained neutrality between the belligerents.At the request of the States of Costa Rica and Nicaragua, a competent engineer has been authorized to make a survey of the river San Juan and the port of San Juan. It is a source of much satisfaction that the difficulties which for a moment excited some political apprehensions and caused a closing of the interoceanic transit route have been amicably adjusted, and that there is a good prospect that the route will soon be reopened with an increase of capacity and adaptation. We could not exaggerate either the commercial or the political importance of that great improvement.It would be doing injustice to an important South American State not to acknowledge the directness, frankness, and cordiality with which the United States of Colombia have entered into intimate relations with this Government. A claims convention has been constituted to complete the unfinished work of the one which closed its session in 1861.The new liberal
Ln 1, Col 1 36,022 characters 100% Windows (CRLF) UTF-8


Abraham_Lincoln_December_1_18€
File Edit View
month of June last there were some grounds to expect that the maritime powers which at the beginning of our domestic difficulties so unwisely and unnecessarily, as we think, recognized the insurgents as a belligerent would soon recede from that position, which has proved only less injurious to themselves than to our own country. But the temporary reverses which afterwards befell the national arms, and which were exaggerated by our own disloyal citizens abroad, have hitherto delayed that act of simple justice.The civil war, which has so radically changed for the moment the occupations and habits of the American people, has necessarily disturbed the social condition and affected very deeply the prosperity of the nations with which we have carried on a commerce that has been steadily increasing throughout a period of half a century. It has at the same time excited political ambitions and apprehensions which have produced a profound agitation throughout the civilized world. In this unusual agitation we have forborne from taking part in any controversy between foreign states and between parties or factions in such states. We have attempted no propagandism and acknowledged no revolution. But we have left to every nation the exclusive conduct and management of its own affairs. Our struggle has been, of course, contemplated by foreign nations with reference less to its own merits than to its supposed and often exaggerated effects and consequences resulting to those nations themselves.
Ln 1, Col 1696 49,814 characters 100% Windows (CRLF) UTF-8


Abraham_Lincoln_December_8_18€
File Edit View
Abraham Lincoln (December 8, 1863)Fellow-Citizens of the Senate and House of Representatives:Another year of health and of sufficiently abundant harvests has passed. For these, and especially for the improved condition of our national affairs, our renewed and profoundest gratitude to God is due.We remain in peace and friendship with foreign powers.The efforts of disloyal citizens of the United States to involve us in foreign wars to aid an inexcusable insurrection have been unavailing. Her Britannic Majesty's Government, as was justly expected, have exercised their authority to prevent the departure of new hostile expeditions from British ports. The Emperor of France has by a like proceeding promptly vindicated the neutrality which he proclaimed at the beginning of the contest. Questions of great intricacy and importance have arisen out of the blockade and other belligerent operations between the Government and several of the maritime powers, but they have been discussed and, as far as was possible, accommodated in a spirit of frankness, justice, and mutual good will. It is especially gratifying that our prize courts, by the impartiality of their adjudications, have commanded the respect and confidence of maritime powers.The supplemental treaty between the United States and Great Britain for the suppression of the African slave trade, made on the 17th day of February
Ln 1, Col 1 36,997 characters 100% Windows (CRLF) UTF-8