# Lab Assignment 1

## CIS 612 Big Data and Parallel Data Processing

Sunnie S Chung

## Information Extraction Methods with Semi-Structured Data Processing: Web Scrapping with DOM and XPath

### Method 1: Webpage in Semi-Structured Data Model with DOM and XPath

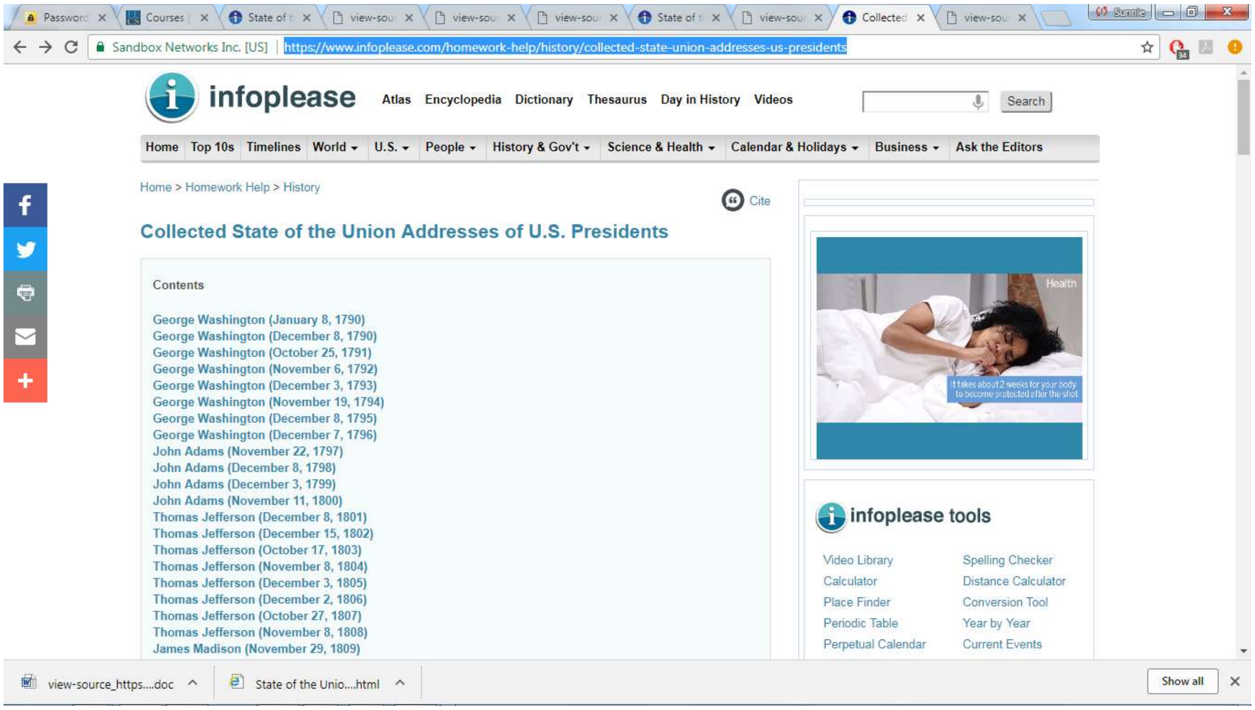
### Method 2: Webpages as Unstructured Data with Parsing

You may use any method of your choice. Process the webpages in the following sites to extract the given information below to create an index table in a SQL Server for next phases of text mining.

You can choose any script language for this lab – Python, Java Script, PHP, or any programming language to scarp the following collection of the webpages to extract information then store them in a table in a SQL Server as follow.

From **Collected State of the Union Addresses of U.S. Presidents** in the following website (in Chrome) at

https://www.infoplease.com/homework-help/history/collected-state-union-addresses-us-presidents

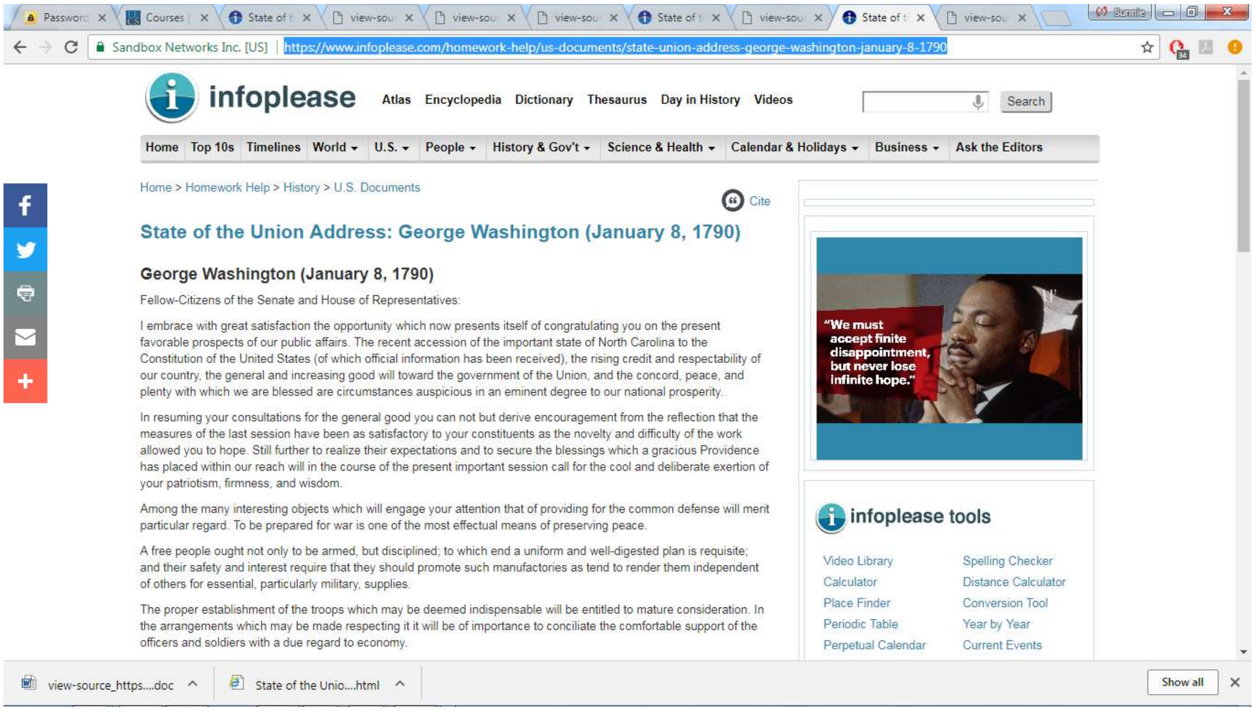


Collected State of the Union Addresses of U.S. Presidents

Contents

George Washington (January 8, 1790)
George Washington (December 8, 1790)
George Washington (October 25, 1791)
George Washington (November 6, 1792)
George Washington (December 3, 1793)
George Washington (November 19, 1794)
George Washington (December 8, 1795)
George Washington (December 7, 1796)
John Adams (November 22, 1797)
John Adams (December 8, 1798)
John Adams (December 3, 1799)
John Adams (November 11, 1800)
Thomas Jefferson (December 8, 1801)
Thomas Jefferson (December 15, 1802)
Thomas Jefferson (October 17, 1803)
Thomas Jefferson (November 8, 1804)
Thomas Jefferson (December 3, 1805)
Thomas Jefferson (December 2, 1806)
Thomas Jefferson (October 27, 1807)
Thomas Jefferson (November 8, 1808)
James Madison (November 29, 1809)

https://www.infoplease.com/homework-help/us-documents/state-union-address-george-washington-january-8-1790



Home > Homework Help > History > U.S. Documents

State of the Union Address: George Washington (January 8, 1790)

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 present favorable prospects of our public affairs. The recent accession of the important state of North Carolina to the Constitution of the United States (of which official information has been received), the rising credit and respectability of our country, the general and increasing good will toward the government of the Union, and the concord, peace, and plenty with which we are blessed are circumstances auspicious in an eminent degree to our national prosperity.

In resuming your consultations for the general good you can not but derive encouragement from the reflection that the measures of the last session have been as satisfactory to your constituents as the novelty and difficulty of the work allowed you to hope. Still further to realize their expectations and to secure the blessings which a gracious Providence has placed within our reach will in the course of the present important session call for the cool and deliberate exertion of your patriotism, firmness, and wisdom.

Among the many interesting objects which will engage your attention that of providing for the common defense will merit particular regard. To be prepared for war is one of the most effectual means of preserving peace.

A free people ought not only to be armed, but disciplined; to which end a uniform and well-digested plan is requisite; and their safety and interest require that they should promote such manufactories as tend to render them independent of others for essential, particularly military, supplies.

The proper establishment of the troops which may be deemed indispensable will be entitled to mature consideration. In the arrangements which may be made respecting it it will be of importance to conciliate the comfortable support of the officers and soldiers with a due regard to economy.

## 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.

```
<dt><span xmlns="" class="article"><a href="/homework-help/us-documents/state-union-address-george-washington-january-8-1790">George Washington (January 8, 1790)</a></span></dt>

<dt><span xmlns="" class="article"><a href="/homework-help/us-documents/state-union-address-george-washington-december-8-1790">George Washington (December 8, 1790)</a></span></dt>
```

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.

| Name of President | Date of Union Address | Link to Address | Filename_Address | Text of Address |
|---|---|---|---|---|
| George Washington | January 8, 1790 | https://www.infoplease.com/homework-help/us-documents/state-union-address-george-washington-january-8-1790 | C:\Users\Sunnie\Work\WebPage\CIS612\Lab3\InfoUnionAddress_1.txt | Fellow-Citizens of the Senate and Hous opportunity which now presents itself |
| George Washington | December 8, 1790 | https://www.infoplease.com/homework-help/us-documents/state-union-address-george-washington-december-8-1790 | C:\Users\Sunnie\Work\WebPage\CIS612\Lab3\InfoUnionAddress_2.txt | Fellow-Citizens of the Senate and Hous In meeting you again I feel much satisf prospects which continue to distinguis |
| … | … | … | … | … |

The five column names as a heading here will be your table schema when you create a Table in a SQL Server later from your CSV or text file.  Five columns per row have the following schema:

(President, Date of Union Address, Link to Address, Filename_Address, Text_Address).

For each saved text in a file and store the file path as a pointer to the text in a column for further data processing in next phase so that you can extract information from each file to store them in a table in a SQL server to retrieve them later as needed for next phase of data processing.

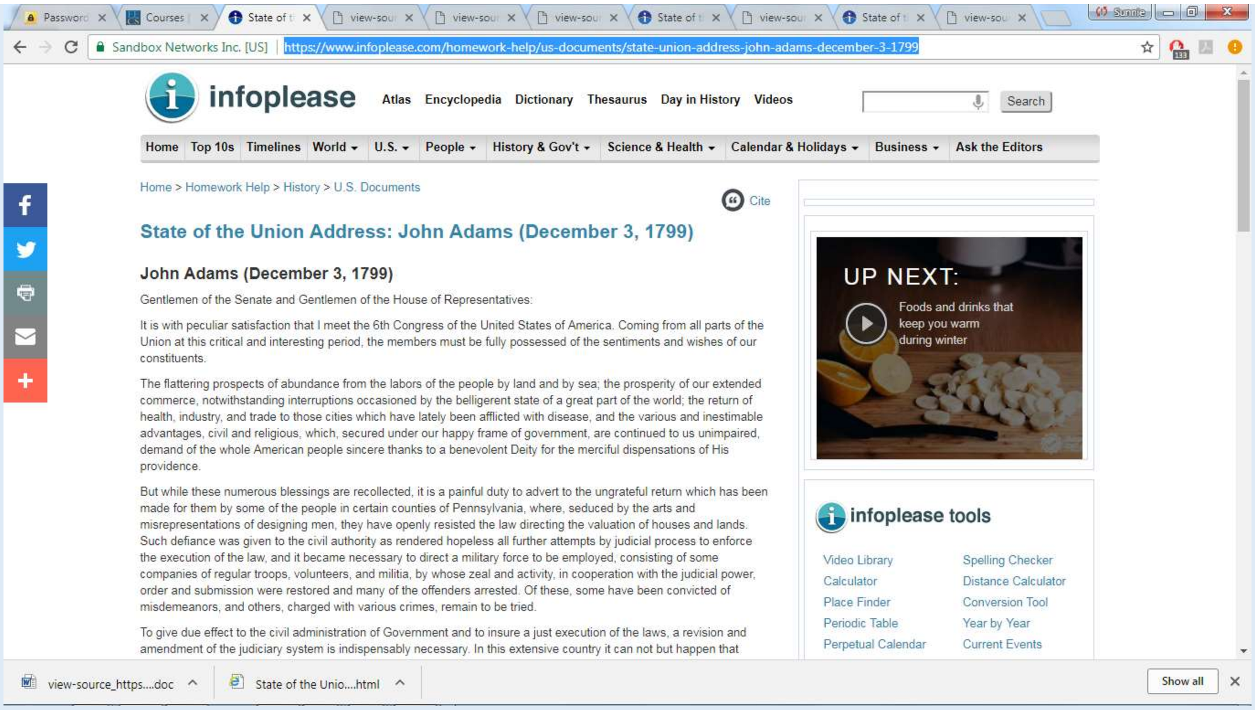
Note that some links of the infoplease site is hijacked to be redirected to another page. To get each correct Union Address text, you will have to go to each address page in the following format:

https://www.infoplease.com/homework-help/us-documents/state-union-address-firstname-lastname-month-day-year

Example:

For **John Adams (December 3, 1799)**, to reach the address site, go to :

https://www.infoplease.com/homework-help/us-documents/state-union-address-john-adams-december-3-1799



State of the Union Address: John Adams (December 3, 1799)

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 from all parts of the Union at this critical and interesting period, the members must be fully possessed of the sentiments and wishes of our constituents.

The flattering prospects of abundance from the labors of the people by land and by sea; the prosperity of our extended commerce, notwithstanding interruptions occasioned by the belligerent state of a great part of the world; the return of health, industry, and trade to those cities which have lately been afflicted with disease, and the various and inestimable advantages, civil and religious, which, secured under our happy frame of government, are continued to us unimpaired, demand of the whole American people sincere thanks to a benevolent Deity for the merciful dispensations of His providence.

But while these numerous blessings are recollected, it is a painful duty to advert to the ungrateful return which has been made for them by some of the people in certain counties of Pennsylvania, where, seduced by the arts and misrepresentations of designing men, they have openly resisted the law directing the valuation of houses and lands. Such defiance was given to the civil authority as rendered hopeless all further attempts by judicial process to enforce the execution of the law, and it became necessary to direct a military force to be employed, consisting of some companies of regular troops, volunteers, and militia, by whose zeal and activity, in cooperation with the judicial power, order and submission were restored and many of the offenders arrested. Of these, some have been convicted of misdemeanors, and others, charged with various crimes, remain to be tried.

To give due effect to the civil administration of Government and to insure a just execution of the laws, a revision and amendment of the judiciary system is indispensably necessary. In this extensive country it can not but happen that

**Make it Automated for the Table Creation in your Database Server from your script/program:**

**Either with ODBC/JDBC calls for insertions of each record directly from your script/program to a table in your SQL Server**

**or**

**Creating a Stored Procedure in your SQL Server to create a Table from your output file (.CSV) with Bulk Insert.**

**You may use any other methods such as CLR Table Valued Function to create a Table as well.**

## 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.

See the Simplified HTML Source below.

## HTML (in SimplifiedInfoUnionAddress.htm) of the Simplified State Union Address

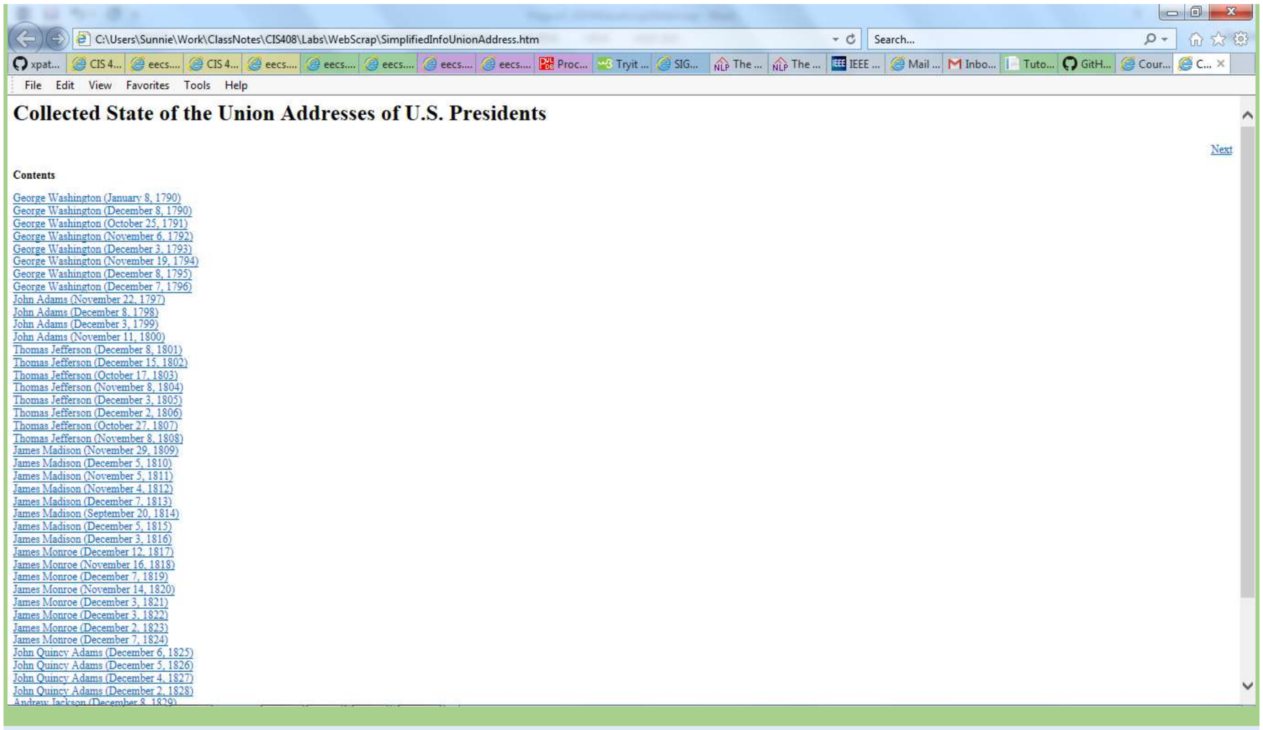
```
<html><head><meta http-equiv="Content-Type" content="text/html; charset=UTF-8"></head>

<body>

<h1 class="page-title"><span><a id="idm6504896"></a>Collected State of the Union Addresses of U.S.
Presidents</span></h1> <section class = "main-section col-sm-12" id="mainaside"> <article data-history-
node-id="83028" role="article" about="/homework-help/history/collected-state-union-addresses-us-
presidents"> <!-- AddThis Button BEGIN --> <!-- AddThis Button END --> <div style="position:relative"><div
class="navheader"><table width="100%" class="PageLinearNav" summary="Navigation header"><tr><td
width="20%" align="left" class="PLN-prev"> </td><th align="center"> </th><td width="20%" align="right"
class="PLN-next"> <a accesskey="n" href="/homework-help/us-documents/state-union-address-george-
washington-january-8-1790">Next</a></td></tr></table></div><div class="book" lang="en" xml:lang="en"
xml:lang="en"><div class="titlepage"><div><div></div></div></div><div
class="toc"><p><b>Contents</b></p><dl><dt><span xmlns="" class="article"><a href="/homework-help/us-
documents/state-union-address-george-washington-january-8-1790">George Washington (January 8,
1790)</a></span></dt><dt><span xmlns="" class="article"><a href="/homework-help/us-documents/state-union-
address-george-washington-december-8-1790">George Washington (December 8, 1790)</a></span></dt><dt><span
xmlns="" class="article"><a href="/homework-help/us-documents/state-union-address-george-washington-
october-25-1791">George Washington (October 25, 1791)</a></span></dt><dt><span xmlns="" class="article"><a
href="/homework-help/us-documents/state-union-address-george-washington-november-6-1792">George Washington
(November 6, 1792)</a></span></dt><dt><span xmlns="" class="article"><a href="/homework-help/us-
documents/state-union-address-george-washington-december-3-1793">George Washington (December 3,
1793)</a></span></dt><dt><span xmlns="" class="article"><a href="/homework-help/us-documents/state-union-
address-george-washington-november-19-1794">George Washington (November 19, 1794)</a></span></dt><dt><span
xmlns="" class="article"><a href="/homework-help/us-documents/state-union-address-george-washington-
december-8-1795">George Washington (December 8, 1795)</a></span></dt><dt><span xmlns="" class="article"><a
href="/homework-help/us-documents/state-union-address-george-washington-december-7-1796">George Washington
(December 7, 1796)</a></span></dt><dt><span xmlns="" class="article"><a href="/homework-help/us-
documents/state-union-address-john-adams-november-22-1797">John Adams (November 22,
1797)</a></span></dt><dt><span xmlns="" class="article"><a href="/homework-help/us-documents/state-union-
address-john-adams-december-8-1798">John Adams (December 8, 1798)</a></span></dt>
<dt><span xmlns="" class="article"><a href="/t/hist/state-of-the-union/1/node/4838">John Adams (December
3, 1799)</a></span></dt><dt><span xmlns="" class="article"><a href="/t/hist/state-of-the-
union/1/node/4949">John Adams (November 11, 1800)</a></span></dt><dt><span xmlns="" class="article"><a
href="/t/hist/state-of-the-union/1/node/4985">Thomas Jefferson (December 8, 1801)</a></span></dt><dt><span
xmlns="" class="article"><a href="/t/hist/state-of-the-union/1/node/4996">Thomas Jefferson (December 15,
1802)</a></span></dt><dt><span xmlns="" class="article"><a href="/t/hist/state-of-the-
union/1/node/5007">Thomas Jefferson (October 17, 1803)</a></span></dt><dt><span xmlns=""
class="article"><a href="/t/hist/state-of-the-union/1/node/5018">Thomas Jefferson (November 8,
1804)</a></span></dt><dt><span xmlns="" class="article"><a href="/t/hist/state-of-the-
union/1/node/5029">Thomas Jefferson (December 3, 1805)</a></span></dt><dt><span xmlns=""
class="article"><a href="/t/hist/state-of-the-union/1/node/5040">Thomas Jefferson (December 2,
1806)</a></span></dt><dt><span xmlns="" class="article"><a href="/t/hist/state-of-the-
union/1/node/5051">Thomas Jefferson (October 27, 1807)</a></span></dt><dt><span xmlns=""
class="article"><a href="/homework-help/us-documents/state-union-address-thomas-jefferson-november-8-
1808">Thomas Jefferson (November 8, 1808)</a></span></dt><dt><span xmlns="" class="article"><a
href="/t/hist/state-of-the-union/2/node/4838">James Madison (November 29, 1809)</a></span></dt><dt><span
xmlns="" class="article"><a href="/t/hist/state-of-the-union/2/node/4949">James Madison (December 5,
1810)</a></span></dt><dt><span xmlns="" class="article"><a href="/t/hist/state-of-the-
union/2/node/4985">James Madison (November 5, 1811)</a></span></dt><dt><span xmlns="" class="article"><a
href="/t/hist/state-of-the-union/2/node/4996">James Madison (November 4, 1812)</a></span></dt><dt><span
xmlns="" class="article"><a href="/t/hist/state-of-the-union/2/node/5007">James Madison (December 7,
1813)</a></span></dt><dt><span xmlns="" class="article"><a href="/t/hist/state-of-the-
union/2/node/5018">James Madison (September 20, 1814)</a></span></dt><dt><span xmlns="" class="article"><a
href="/t/hist/state-of-the-union/2/node/5029">James Madison (December 5, 1815)</a></span></dt><dt><span
xmlns="" class="article"><a href="/t/hist/state-of-the-union/2/node/5040">James Madison (December 3,
1816)</a></span></dt><dt><span xmlns="" class="article"><a href="/t/hist/state-of-the-
```

```
union/2/node/5051">James Monroe (December 12, 1817)</a></span></dt><dt><span xmlns="" class="article"><a
href="/homework-help/us-documents/state-union-address-james-monroe-november-16-1818">James Monroe
(November 16, 1818)</a></span></dt><dt><span xmlns="" class="article"><a href="/t/hist/state-of-the-
union/3/node/4838">James Monroe (December 7, 1819)</a></span></dt><dt><span xmlns="" class="article"><a
href="/t/hist/state-of-the-union/3/node/4949">James Monroe (November 14, 1820)</a></span></dt><dt><span
xmlns="" class="article"><a href="/t/hist/state-of-the-union/3/node/4985">James Monroe (December 3,
1821)</a></span></dt><dt><span xmlns="" class="article"><a href="/t/hist/state-of-the-
union/3/node/4996">James Monroe (December 3, 1822)</a></span></dt><dt><span xmlns="" class="article"><a
href="/t/hist/state-of-the-union/3/node/5007">James Monroe (December 2, 1823)</a></span></dt><dt><span
xmlns="" class="article"><a href="/t/hist/state-of-the-union/3/node/5018">James Monroe (December 7,
1824)</a></span></dt><dt><span xmlns="" class="article"><a href="/t/hist/state-of-the-
union/3/node/5029">John Quincy Adams (December 6, 1825)</a></span></dt><dt><span xmlns=""
class="article"><a href="/t/hist/state-of-the-union/3/node/5040">John Quincy Adams (December 5,
1826)</a></span></dt><dt><span xmlns="" class="article"><a href="/t/hist/state-of-the-
union/3/node/5051">John Quincy Adams (December 4, 1827)</a></span></dt><dt><span xmlns=""
class="article"><a href="/homework-help/us-documents/state-union-address-john-quincy-adams-december-2-
1828">John Quincy Adams (December 2, 1828)</a></span></dt><dt><span xmlns="" class="article"><a
href="/t/hist/state-of-the-union/4/node/4838">Andrew Jackson (December 8, 1829)</a></span></dt><dt><span
xmlns="" class="article"><a href="/t/hist/state-of-the-union/4/node/4949">Andrew Jackson (December 6,
1830)</a></span></dt><dt><span xmlns="" class="article"><a href="/t/hist/state-of-the-
union/4/node/4985">Andrew Jackson (December 6, 1831)</a></span></dt><dt><span xmlns="" class="article"><a
href="/t/hist/state-of-the-union/4/node/4996">Andrew Jackson (December 4, 1832)</a></span></dt><dt><span
xmlns="" class="article"><a href="/t/hist/state-of-the-union/4/node/5007">Andrew Jackson (December 3,
1833)</a></span></dt><dt><span xmlns="" class="article"><a href="/t/hist/state-of-the-
union/4/node/5018">Andrew Jackson (December 1, 1834)</a></span></dt><dt><span xmlns="" class="article"><a
href="/t/hist/state-of-the-union/4/node/5029">Andrew Jackson (December 7, 1835)</a></span></dt><dt><span
xmlns="" class="article"><a href="/t/hist/state-of-the-union/4/node/5040">Andrew Jackson (December 5,
1836)</a></span></dt>

</body>
</html>
```



C:\Users\Sunnie\Work\ClassNotes\CIS408\Labs\WebScrap\SimplifiedInfoUnionAddress.htm

**Collected State of the Union Addresses of U.S. Presidents**

**Contents**

George Washington (January 8, 1790)
George Washington (December 8, 1790)
George Washington (October 25, 1791)
George Washington (November 6, 1792)
George Washington (December 3, 1793)
George Washington (November 19, 1794)
George Washington (December 8, 1795)
George Washington (December 7, 1796)
John Adams (November 22, 1797)
John Adams (December 8, 1798)
John Adams (December 3, 1799)
John Adams (November 11, 1800)
Thomas Jefferson (December 8, 1801)
Thomas Jefferson (December 15, 1802)
Thomas Jefferson (October 17, 1803)
Thomas Jefferson (November 8, 1804)
Thomas Jefferson (December 3, 1805)
Thomas Jefferson (December 2, 1806)
Thomas Jefferson (October 27, 1807)
Thomas Jefferson (November 8, 1808)
James Madison (November 29, 1809)
James Madison (December 5, 1810)
James Madison (November 5, 1811)
James Madison (November 4, 1812)
James Madison (December 7, 1813)
James Madison (September 20, 1814)
James Madison (December 5, 1815)
James Madison (December 3, 1816)
James Monroe (December 12, 1817)
James Monroe (November 16, 1818)
James Monroe (December 7, 1819)
James Monroe (November 14, 1820)
James Monroe (December 3, 1821)
James Monroe (December 3, 1822)
James Monroe (December 2, 1823)
James Monroe (December 7, 1824)
John Quincy Adams (December 6, 1825)
John Quincy Adams (December 5, 1826)
John Quincy Adams (December 4, 1827)
John Quincy Adams (December 2, 1828)
Andrew Jackson (December 8, 1829)

For Python Script:

> http://docs.python-guide.org/en/latest/scenarios/scrape/
>
> Submit BOTH in the followings:
>
> On Blackboard:

> 1. Lab Report that shows your set up/platform procedure, the execution to generate the output file in a structured any text file format (or in CSV, TSV), your output in text and as a SQL table and your source codes/scripts.
> 2. All your codes/scripts, input and all the output files in one zip file.