Lab 3: Semi-structured Data Processing with DOM and XPath

By: Christen Ford

Date: March 4th, 2019

Platform and Procedures

The platform I built this software on is Ubuntu 18.04 x64. However, it is programmed in Python 3.7.2 (which as I will explain, must be used to run the software). This said, the software makes use of the following Python packages which must be installed to run the software without error:

* Certifi
* Parsel
* Urllib3

The remaining Python packages used by the scraping application are built into Python. To install these packages in your Python installation use the following command from the command prompt or shell:

`python –m pip install --user certifi parsel urllib3` (Windows systems)

`pip3 install --user certify parsel urllib3` (Linux systems)

I do not own an apple system so I am unaware of the proper command to install these packages on said systems, however it is most likely a variant of the Linux command. Consult the documentation for pip that came with your Python distribution for instructions on how to use pip on said platforms.

The software can be ran with the following command:

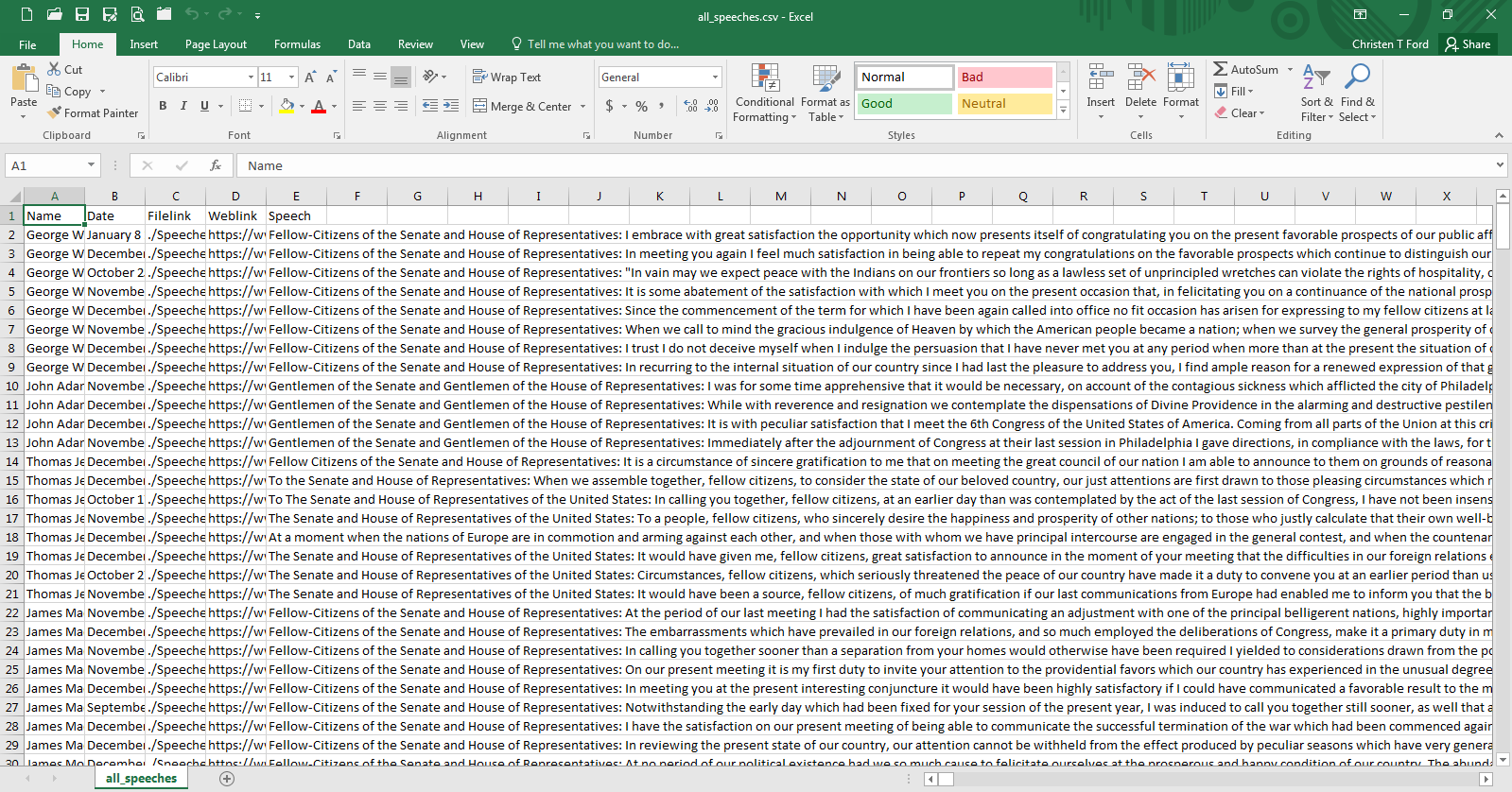
`python sotu-scraper.py`

This command will attempt to download all of the state of the union addresses from the infoplease web archive. If successful, it will create the following files in the directory of the script:

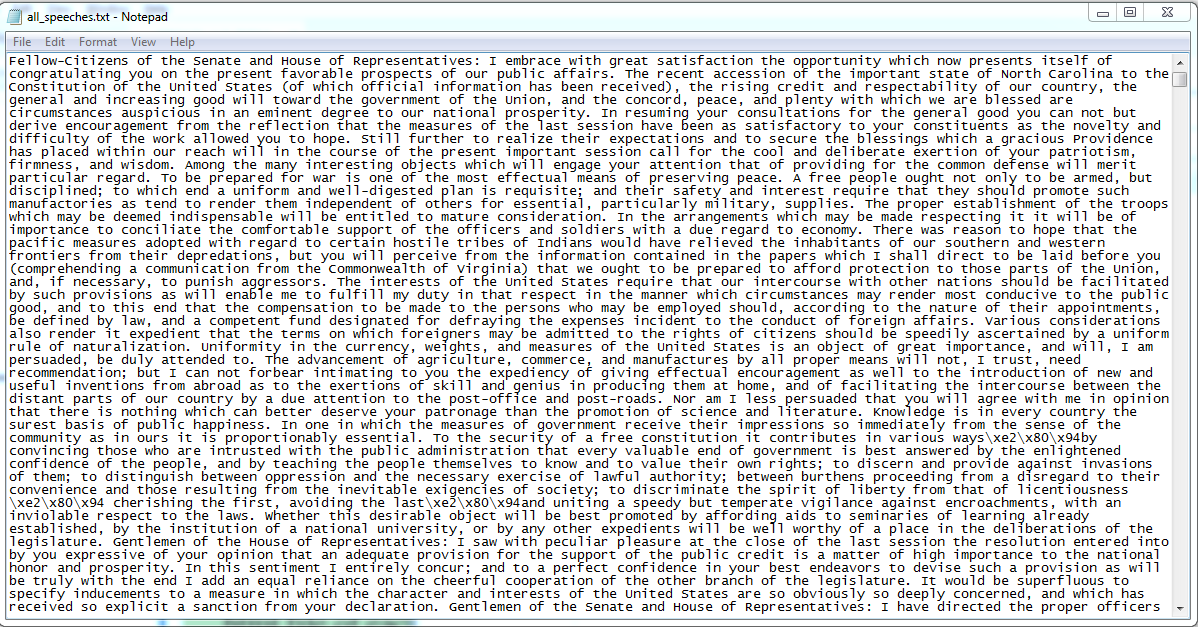
* ./Speeches/\*.txt – The Speeches directory will contain all of the state of the union addresses labeled by the President and date of the address. The name of each file is always in the form `[firstname]-[optional middle]-[lastname]-[month]-[optional date]-[year].txt`
* ./all\_speeches.txt – This file contains all of the state of the union addresses by each President where each speech is separated by a single newline character.
* ./all\_speeches.csv – This file contains all of the state of the union addresses by each President stored in a comma-delimited format. This makes the addresses easy to import into a SQL database. The first row of the file is a header row, while every proceeding row is a data row of the format: `name, date, filelink, weblink, speech`

Note: If you have any issues running the software, it is most likely because you are not running it using the latest version of Python (v 3.7.2). Python 2 is no longer officially supported by the Python Software Foundation. There are several language features that I utilize in the scraper that are dependent on running a Python installation at a certain version level of Python 3 and above (and I did not code it to be backwards compatible). To be safe, just use the latest version.

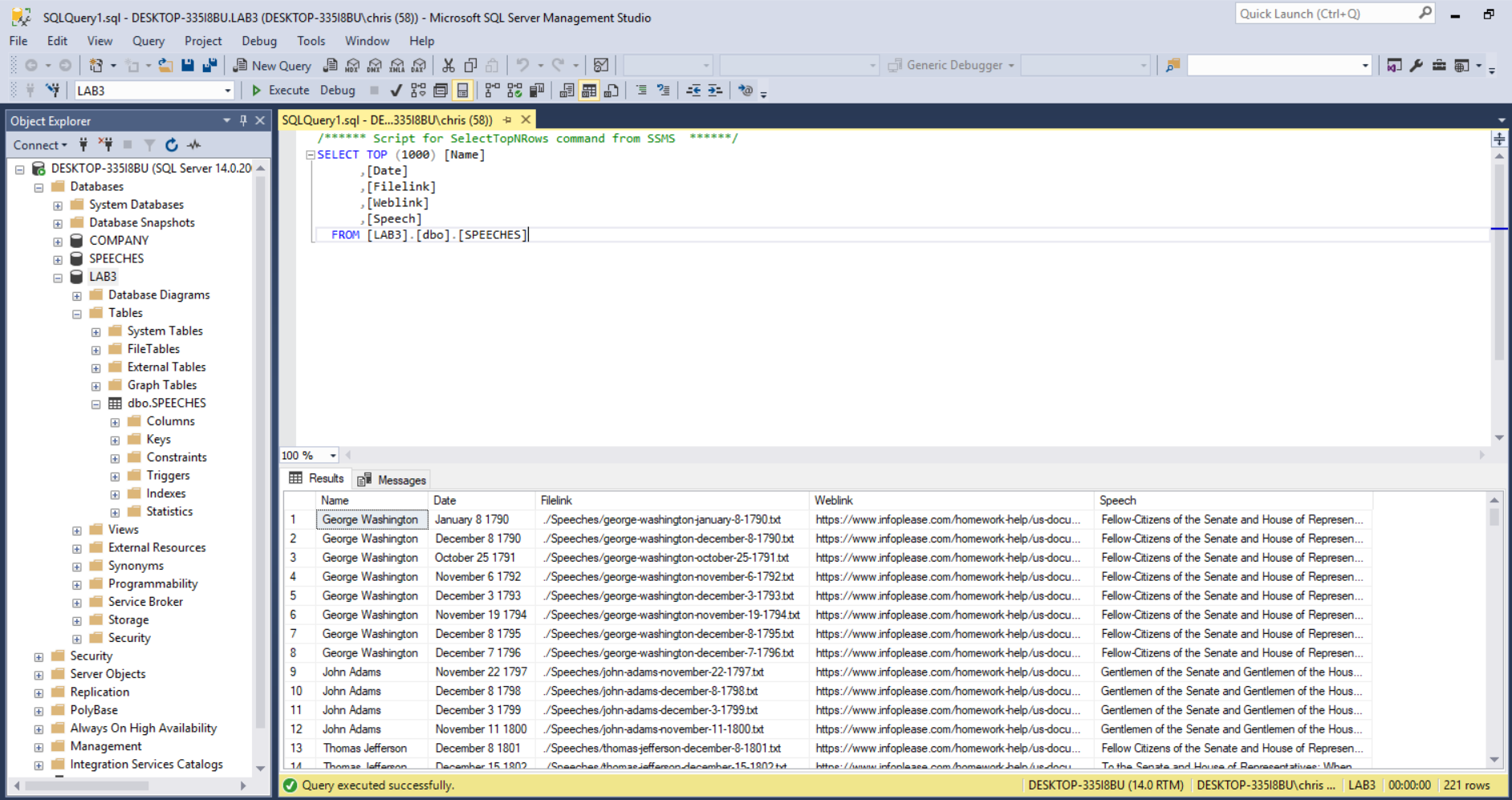
Artifacts



This image shows the raw data that the scraper obtained written to a CSV file (generated by the scraper).



This image shows the raw speech data scraped by the tool as one large file, the entire contents of which cannot be displayed here.



This image shows the results of importing the scraper generated CSV file into SQL Server. In addition, this table was exported to the file `./export\_from\_ssms.txt` as further proof that the table was created successfully from the CSV file.

Source Code for the Scraper is on the Following Pages