**Homework 5: Building a Text-based Dataset from an API**

**Student:** Patrick Walsh

**Professor:** Dr. Ami Gates

**School:** Syracuse University

**Date:** 2/14/2024

**INTRODUCTION**

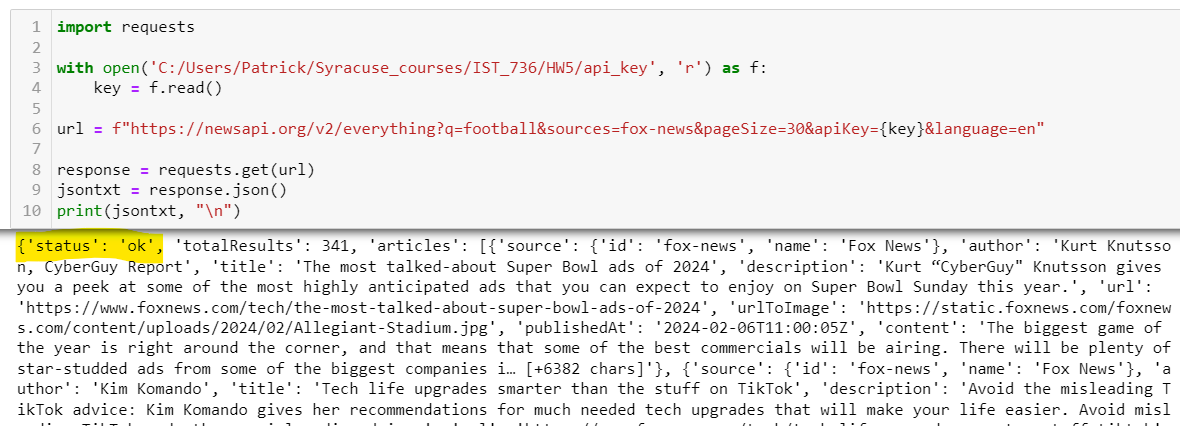
Application Programming Interfaces, or APIs, as they are known by data scientists and software engineers, are a critical technique used in data science to interact with various data sources between servers and clients. In software engineering, APIs are used to communicate and transfer data between two or more computers. They are useful in the world of text mining for obtaining various data sources in a semi-structured format for further processing. An API call is made from one computer to another and the response is returned in a semi-structured format called JSON. This JSON format is similar to Python dictionaries that have key/value pairs to associate various pieces of data.

This assignment will leverage the NewsAPI API to obtain semi-structured data on various news topics from multiple sources, all with a few simple lines of code. The goal will be to use NewsAPI to obtain news headlines from specific topics and use Python to put the data into a structured format, namely a Pandas dataframe, which is similar to an Excel spreadsheet with rows and columns. For the purposes of this exercise, we are also interested in seeing the content of the text-based information as well, so we will create several word clouds to get a visual representation of the news headlines and their word content.

**ANALYSIS**

**Testing the API key**

The first step is to ensure that our API key is valid and returns data. Below is a simple test to obtain the API key from a text file, set it to a variable, and then use that key to validate the API call in Python.



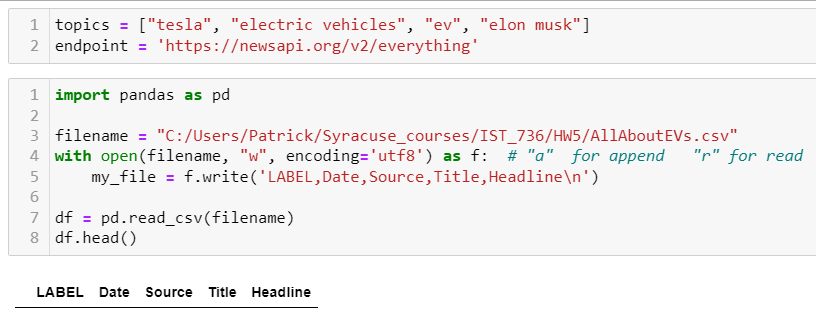
The status of ‘ok’ means that the API call was successful. We can also tell that the API call worked because it returned data, as seen in the screenshot.

**Security**

It is good practice to store the API key in a file rather than having it saved directly in the Python code since the latter method exposes the API key and makes it so that anyone with the source code can directly read the API key. The API key is sensitive and should be treated as a password, so protecting it from outside exposure is an important security practice.

**Connecting to the API and building a CSV file**

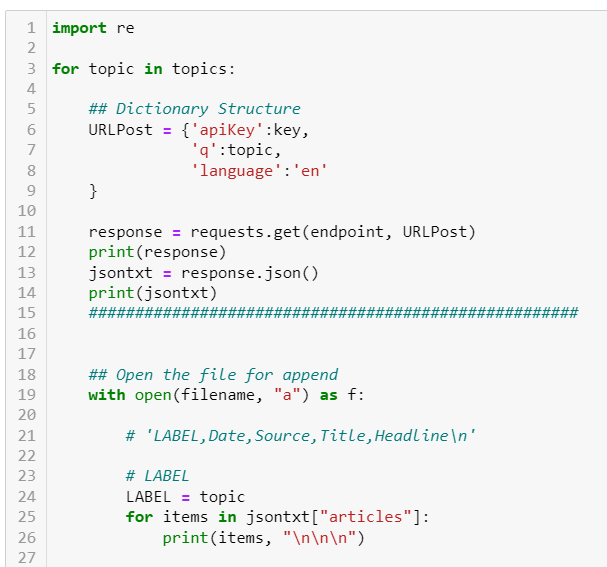
There are two parameters we need to set for the API call to work programmatically. The first is the set of topics we want to query on, and the second is the API endpoint. The endpoint is a URL where the API call will be sent, and the topics are the keywords used in the search when querying NewsAPI. Below you can see the topics and endpoint setup, as well as code to create a CSV file with five columns. The CSV file is then converted to a dataframe.

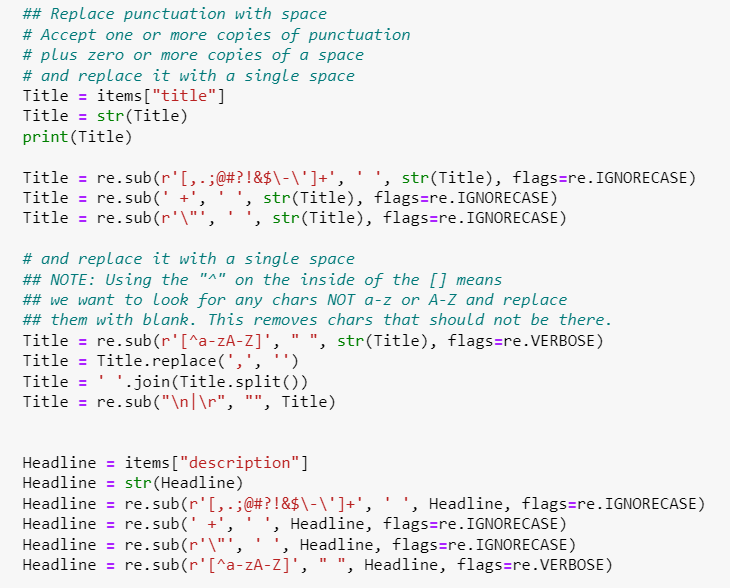
****

The CSV file and resulting dataframe are empty for now, as we have not added data to the file yet.

**Adding Data**

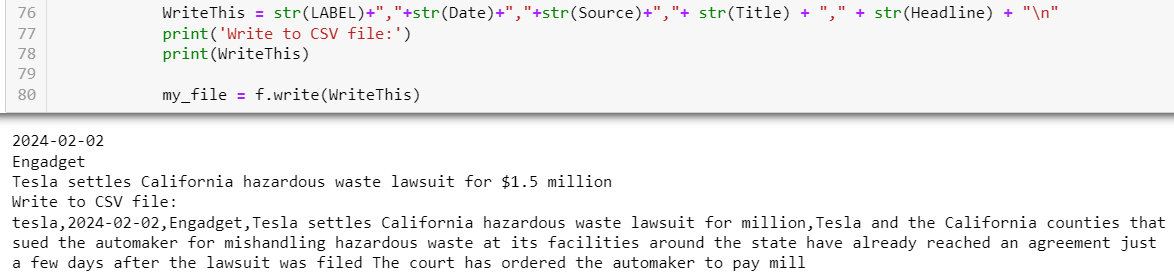
The code below is a partial screenshot showing how data is added to the corresponding CSV file. First, a for loop is setup iterating through each of the topics in the call, which are "tesla", "electric vehicles", "ev", and "elon musk". For each topic, an API call is sent to NewsAPI and the JSON is returned and parsed. Within the JSON object, the LABEL, Date, Source, Title, and Headline are extracted and processed using Python and regular expressions to put the text into a format that is easy for the CSV file to handle. Note that the file is opened in append mode so that the data is appended to the labeled columns created in the previous cell.

****

****

Some of the regular expressions are seen in the screenshot above to remove special characters, spaces, and other characters which can affect the CSV file’s ability to hold the text data.

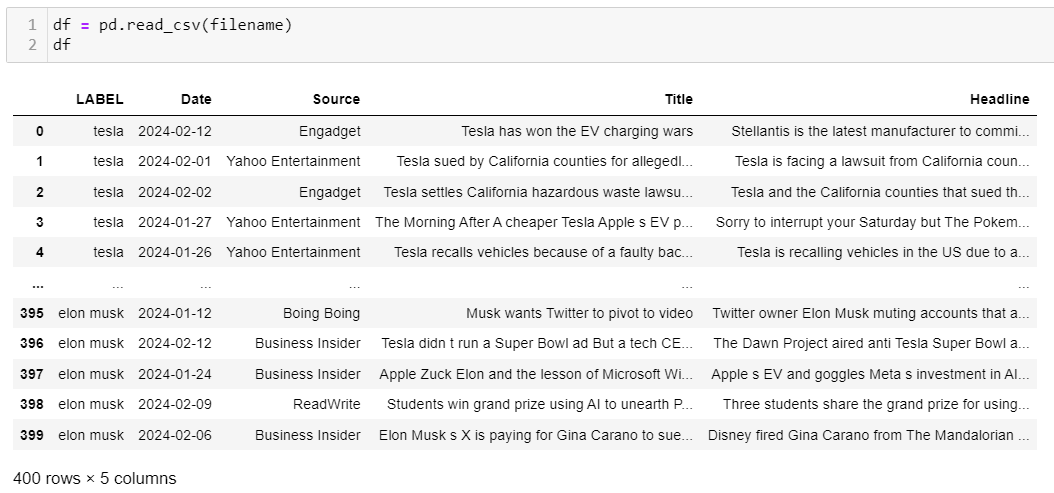
After parsing, the cleaned data is written to each row of the CSV file, with the LABEL, Date, Source, Title, and Headline for each news article returned from NewsAPI. Below is an example.



**RESULTS**

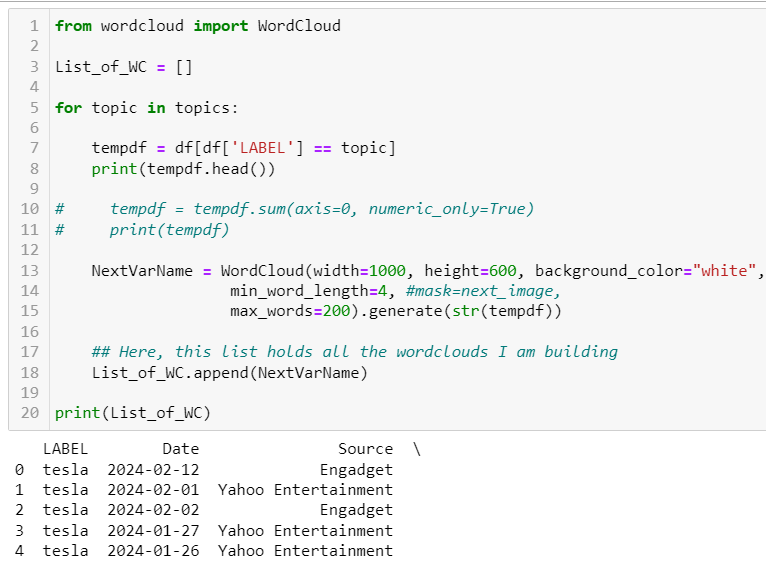
**Converting the CSV file to a Dataframe**

Once the data is done being written to the CSV file, the file must then be converted to a Pandas dataframe for additional processing. Below is a screenshot of what the converted dataframe looks like.

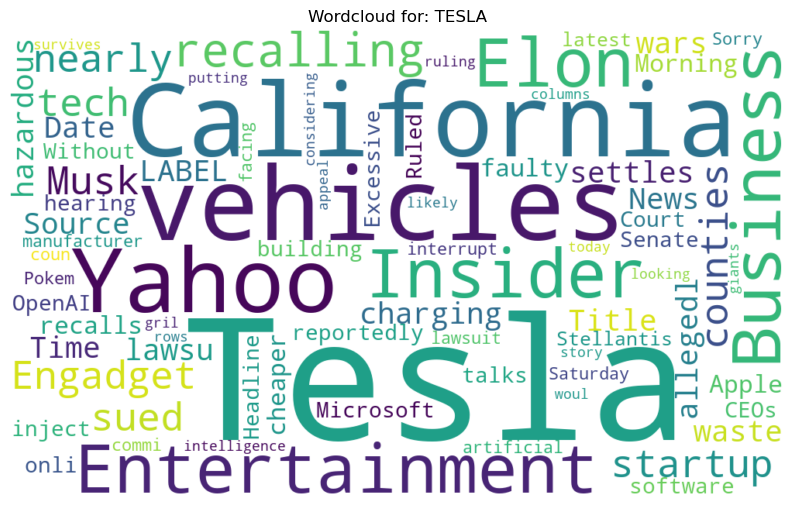
****

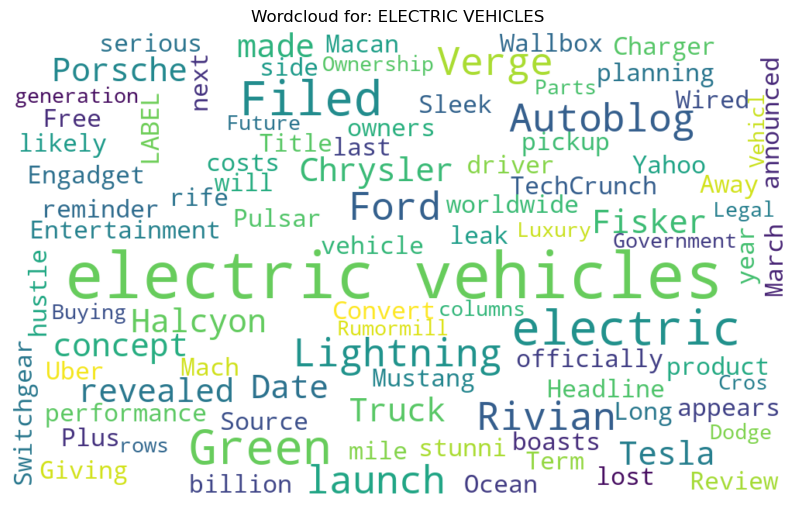
**Creating Wordclouds**

The final step in this exercise is to visually examine the dataframe by putting the Headline text for each topic into a wordcloud in order to get a sense on the content of each topic. The screenshot below shows the code for creating a list of wordcloud objects which can then be plotted to see the word content of each of the four topics.

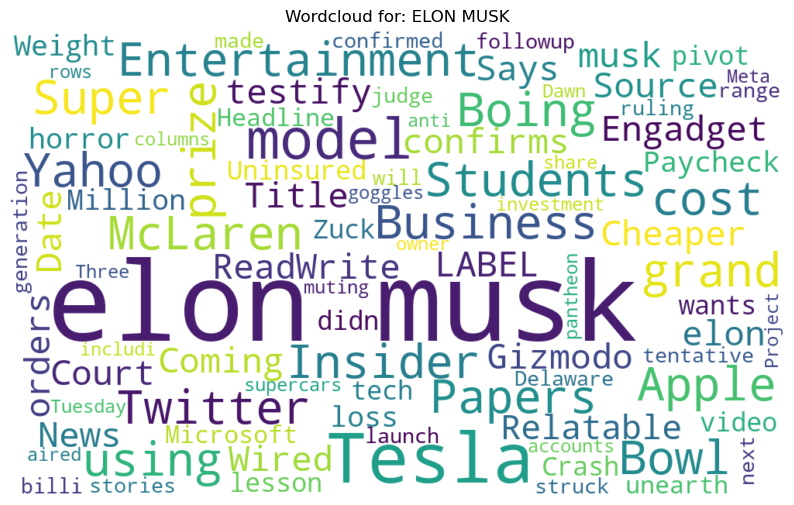
****

Below are the finalized wordclouds for each of the topics, starting with Tesla and ending with Elon Musk.









**CONCLUSIONS**

The NewsAPI API was used to obtain news information on specific topics and convert that information from a semi-structured format into a structured format for additional processing and visualization. From here, the finalized dataframes could be tokenized, vectorized, and used to create various machine learning models or to carry out additional text mining tasks. APIs are a powerful way to gather data for various data science projects and can be used to great affect in the world of machine learning, software engineering, data analysis, and computer science.