Hana Bezark
SI 206

Final Project Report
Data-Oriented Programming with APIs and Visualization

My goals

According to the Project 4 - Final Project project spec, the objectives of the assignment are to demonstrate ability to create a fully working program without any scaffolding, create and modify tables in a SQLite Database, utilize APIs (including researching possible methods), utilize visualization software (including researching options) and follow proper coding and submission conventions. While these were the overarching goals that guided my procedure for the project, I also created specific goals for myself in my initial Project Plan. My Project Plan is as follows:

> The two APIs that I will be using are New York Times and Yelp. For New York Times, the data that I will be collecting for each article that contains the search term Donald Trump are as follows: headline, web url, keywords, publication date, byline and word count. For Yelp, the data I will be collecting for each review of Summer House Santa Monica restaurant in Chicago, IL are as follows: rating, user, text, time created and url. For New York Times, I will be creating a Wordle with the keywords of each article. For Yelp, I will be creating a graph using Plotly that depicts the ratings provided in the reviews.

More scores for assignments were inputted after I had submitted my Project Plan, and I deemed it feasible for me to achieve a final overall grade of an “A” without needing to use two APIs and complete two visualizations. So, I decided to use New York Times as my API and complete a corresponding visualization using Word Cloud and not use Yelp or create a visualization using Plotly.

Which goals I achieved

I successfully achieved my goal of collecting New York Times articles containing the search term Donald Trump. For each article I was successful in collecting the headline, web url, keywords, publication date, byline and word count in a json object. Additionally, I collected the news desk and score for each article as well. All of the data collected is featured in my database, except for the keywords, which I will go into more detail about in the subsequent section of my report. I successfully achieved my goal of creating a Word Cloud, however I will describe some problems that I faced in doing so in the next section.

What problems I faced

The first problem that I faced was in regards to the "100 interactions" requirement. To reach this goal, I decided that I was going to retrieve information on 100 articles that contain the search term "Donald Trump." After looking at the documentation for the New York Times API, I learned that each page has 10 results on it. So page=0 corresponds to records 0-9. I initially had the page count set to 10, so that it looped through 10 pages and I would get 100 results, but I noticed that in my SQL database I was getting fewer than 100 results. This was due to the duplicates, which were getting overwritten. With trial and error I learned to set the page count to 13 in order to get 97 interactions, which was approved by Chong as getting close to the 100 interaction mark.

Stemming from this issue was the fact that some of the article records did not have all of the data that I was looking to collect. I was collecting the print headlines, web URLs, news desk, publication date, bylines, word counts and scores for each entry. However, I learned that some of the articles did not have a publication date, for example, and other articles did not have bylines. This was causing my program to malfunction and stop running because it kept breaking when it could not find the appropriate data to write to the database. In order to solve this issue, I realized that I had to write lines to code to assign the value "EMPTY" as a placeholder if an article did not have certain information within the entry.

Another issue that I had was that I was putting a request to New York Times to access lots of information about many articles, which is a time consuming process. It was taking so long to retrieve this data that my program would move onto the next lines of code before all of the data was retrieved, causing my program to crash. After researching about this, I learned that I had to write time.sleep(1) in my code in order to implement a time delay of one second so that my program did not move onto the next line of code before all of the data had been collected.

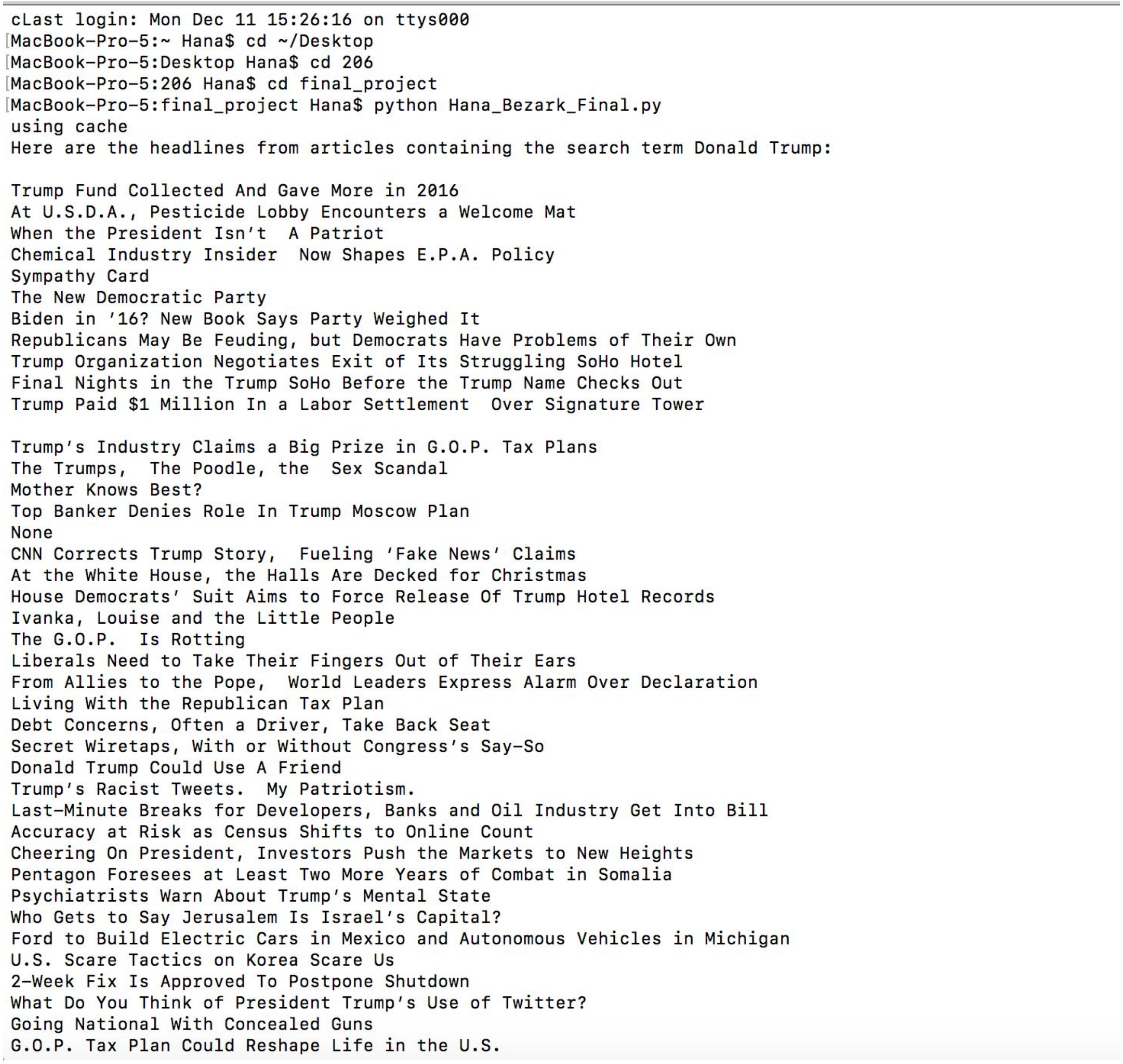
I initially wrote in my project plan that one of the data points I was going to be collecting for each article were keywords. Upon further research into the API documentation and nested dictionaries I was looking at during the data collection process, I learned that keywords were stored in many different dictionaries and that there was no way to access them neatly to place them in a SQL database. I realized I needed to make the decision to no longer collect keywords, so I instead added some other data fields into my data collection efforts as a replacement.

In my project plan I also wrote that I was going to be creating a word cloud based on the keywords from each article. As described in the previous paragraph, collecting keywords was extremely difficult to do so I had to change my course of action for the word cloud. I instead decided to save all of the headlines I retrieved into a text file and create a word cloud based on the headlines, which proved successful.

I have never created a program of this extent before in my programming career so I knew that debugging was going to prove challenging for me. I learned that I could import pdb and then write pdb.set_trace() above the lines that I wanted to debug. This was very helpful for me and I will definitely continue to use this Python debugger throughout the rest of my career.

My “report”

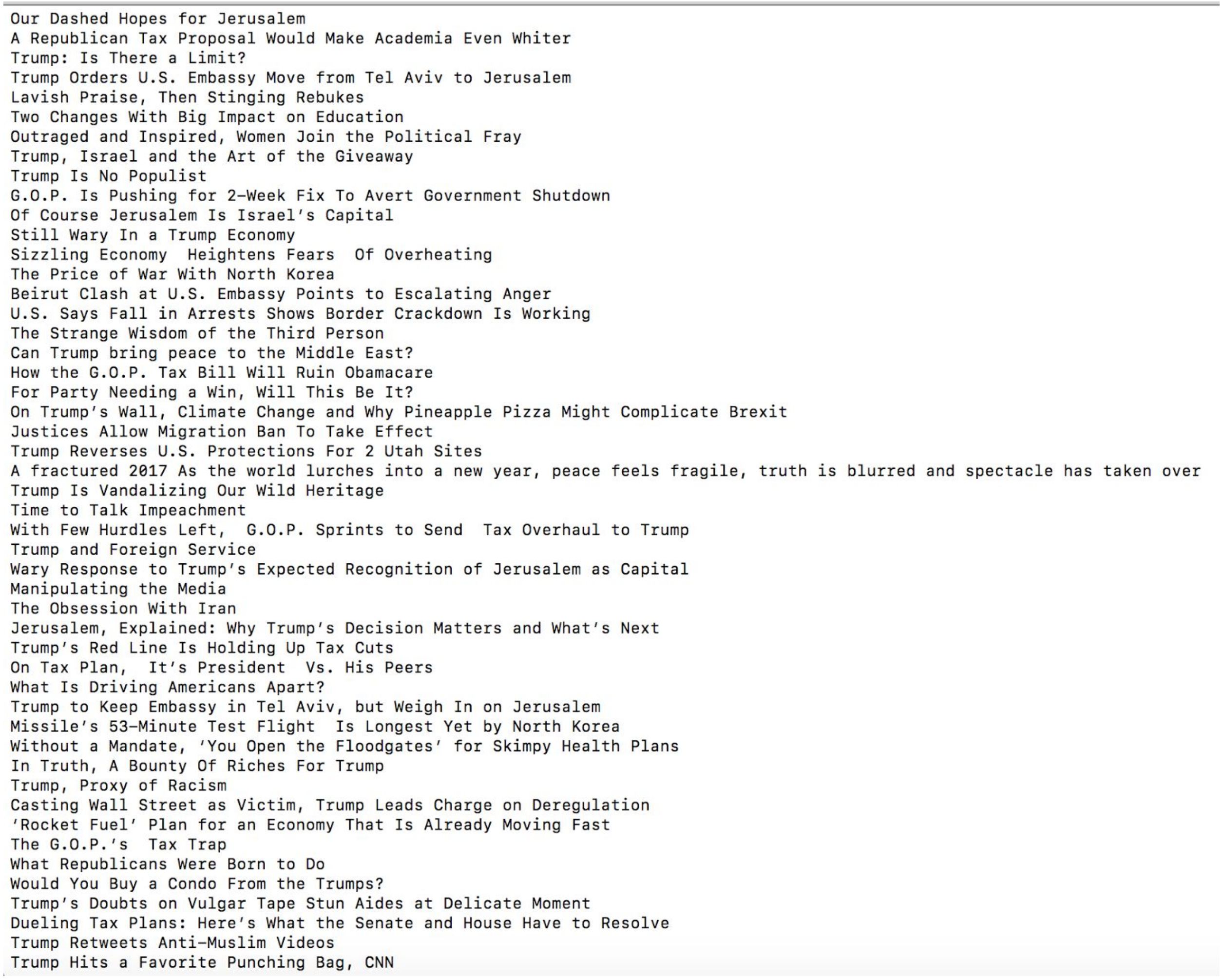
The report that I created was a print out of all the headlines collected in my terminal screen. This report can be viewed in the screenshots below (there were 97 headlines collected which is why multiple screenshots were necessary).



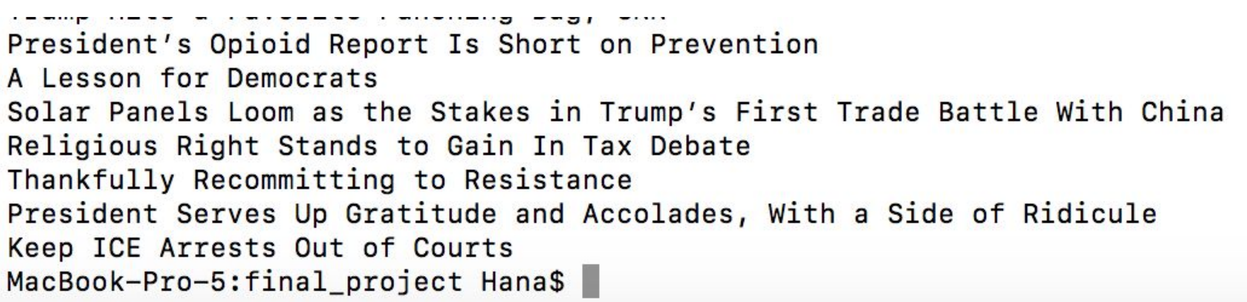
```
cLast login: Mon Dec 11 15:26:16 on ttys000
[MacBook-Pro-5:~ Hana$ cd ~/Desktop
[MacBook-Pro-5:Desktop Hana$ cd 206
[MacBook-Pro-5:206 Hana$ cd final_project
[MacBook-Pro-5:final_project Hana$ python Hana_Bezark_Final.py
using cache
Here are the headlines from articles containing the search term Donald Trump:

Trump Fund Collected And Gave More in 2016
At U.S.D.A., Pesticide Lobby Encounters a Welcome Mat
When the President Isn’t  A Patriot
Chemical Industry Insider  Now Shapes E.P.A. Policy
Sympathy Card
The New Democratic Party
Biden in ’16? New Book Says Party Weighed It
Republicans May Be Feuding, but Democrats Have Problems of Their Own
Trump Organization Negotiates Exit of Its Struggling SoHo Hotel
Final Nights in the Trump SoHo Before the Trump Name Checks Out
Trump Paid $1 Million In a Labor Settlement  Over Signature Tower

Trump’s Industry Claims a Big Prize in G.O.P. Tax Plans
The Trumps,  The Poodle, the  Sex Scandal
Mother Knows Best?
Top Banker Denies Role In Trump Moscow Plan
None
CNN Corrects Trump Story,  Fueling ‘Fake News’ Claims
At the White House, the Halls Are Decked for Christmas
House Democrats’ Suit Aims to Force Release Of Trump Hotel Records
Ivanka, Louise and the Little People
The G.O.P.  Is Rotting
Liberals Need to Take Their Fingers Out of Their Ears
From Allies to the Pope,  World Leaders Express Alarm Over Declaration
Living With the Republican Tax Plan
Debt Concerns, Often a Driver, Take Back Seat
Secret Wiretaps, With or Without Congress’s Say-So
Donald Trump Could Use A Friend
Trump’s Racist Tweets.  My Patriotism.
Last-Minute Breaks for Developers, Banks and Oil Industry Get Into Bill
Accuracy at Risk as Census Shifts to Online Count
Cheering On President, Investors Push the Markets to New Heights
Pentagon Foresees at Least Two More Years of Combat in Somalia
Psychiatrists Warn About Trump’s Mental State
Who Gets to Say Jerusalem Is Israel’s Capital?
Ford to Build Electric Cars in Mexico and Autonomous Vehicles in Michigan
U.S. Scare Tactics on Korea Scare Us
2-Week Fix Is Approved To Postpone Shutdown
What Do You Think of President Trump’s Use of Twitter?
Going National With Concealed Guns
G.O.P. Tax Plan Could Reshape Life in the U.S.
```

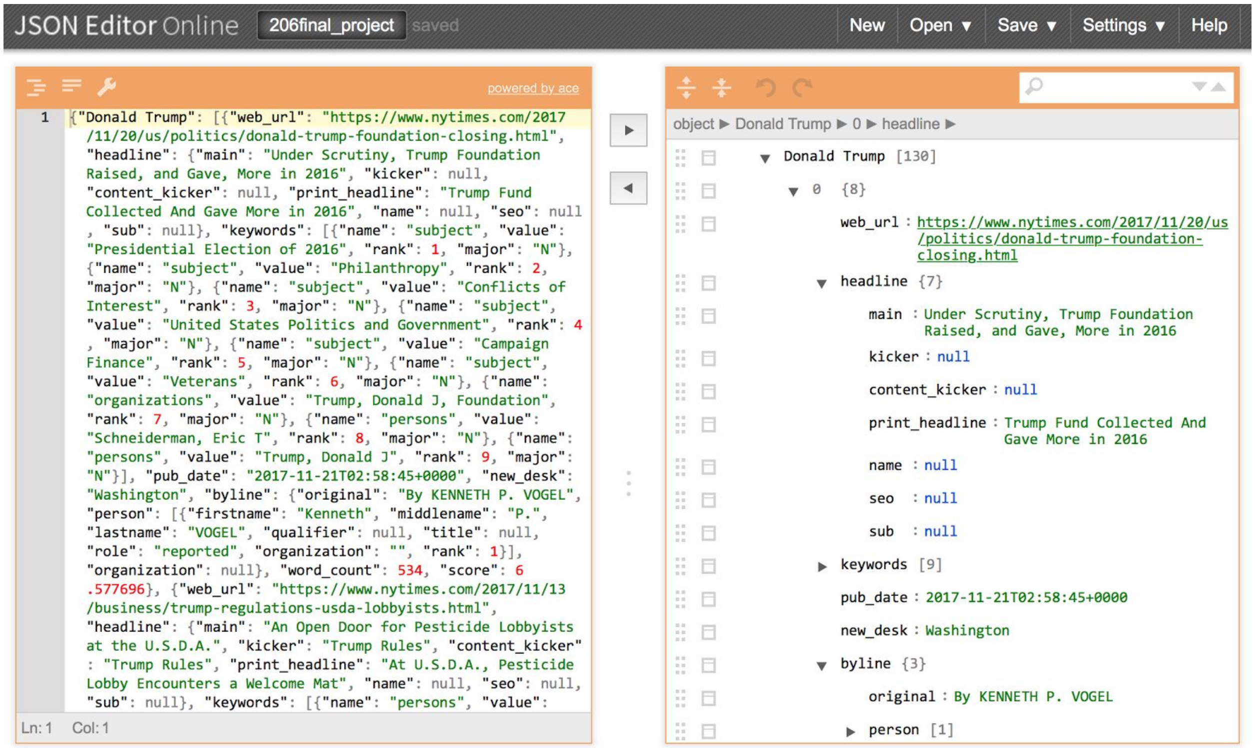


```
Our Dashed Hopes for Jerusalem
A Republican Tax Proposal Would Make Academia Even Whiter
Trump: Is There a Limit?
Trump Orders U.S. Embassy Move from Tel Aviv to Jerusalem
Lavish Praise, Then Stinging Rebukes
Two Changes With Big Impact on Education
Outraged and Inspired, Women Join the Political Fray
Trump, Israel and the Art of the Giveaway
Trump Is No Populist
G.O.P. Is Pushing for 2-Week Fix To Avert Government Shutdown
Of Course Jerusalem Is Israel’s Capital
Still Wary In a Trump Economy
Sizzling Economy  Heightens Fears  Of Overheating
The Price of War With North Korea
Beirut Clash at U.S. Embassy Points to Escalating Anger
U.S. Says Fall in Arrests Shows Border Crackdown Is Working
The Strange Wisdom of the Third Person
Can Trump bring peace to the Middle East?
How the G.O.P. Tax Bill Will Ruin Obamacare
For Party Needing a Win, Will This Be It?
On Trump’s Wall, Climate Change and Why Pineapple Pizza Might Complicate Brexit
Justices Allow Migration Ban To Take Effect
Trump Reverses U.S. Protections For 2 Utah Sites
A fractured 2017 As the world lurches into a new year, peace feels fragile, truth is blurred and spectacle has taken over
Trump Is Vandalizing Our Wild Heritage
Time to Talk Impeachment
With Few Hurdles Left,  G.O.P. Sprints to Send  Tax Overhaul to Trump
Trump and Foreign Service
Wary Response to Trump’s Expected Recognition of Jerusalem as Capital
Manipulating the Media
The Obsession With Iran
Jerusalem, Explained: Why Trump’s Decision Matters and What’s Next
Trump’s Red Line Is Holding Up Tax Cuts
On Tax Plan,  It’s President  Vs. His Peers
What Is Driving Americans Apart?
Trump to Keep Embassy in Tel Aviv, but Weigh In on Jerusalem
Missile’s 53-Minute Test Flight  Is Longest Yet by North Korea
Without a Mandate, ‘You Open the Floodgates’ for Skimpy Health Plans
In Truth, A Bounty Of Riches For Trump
Trump, Proxy of Racism
Casting Wall Street as Victim, Trump Leads Charge on Deregulation
‘Rocket Fuel’ Plan for an Economy That Is Already Moving Fast
The G.O.P.’s  Tax Trap
What Republicans Were Born to Do
Would You Buy a Condo From the Trumps?
Trump’s Doubts on Vulgar Tape Stun Aides at Delicate Moment
Dueling Tax Plans: Here’s What the Senate and House Have to Resolve
Trump Retweets Anti-Muslim Videos
Trump Hits a Favorite Punching Bag, CNN
```



```
President’s Opioid Report Is Short on Prevention
A Lesson for Democrats
Solar Panels Loom as the Stakes in Trump’s First Trade Battle With China
Religious Right Stands to Gain In Tax Debate
Thankfully Recommitting to Resistance
President Serves Up Gratitude and Accolades, With a Side of Ridicule
Keep ICE Arrests Out of Courts
MacBook-Pro-5:final_project Hana$ █
```

A snippet of what my collected data looks like in an online JSON editor is pictured below. Every article had its own entry like what you see below, please understand that it would not be feasible for me to include a picture of every single entry.



```
{"Donald Trump": [{"web_url": "https://www.nytimes.com/2017
/11/20/us/politics/donald-trump-foundation-closing.html",
"headline": {"main": "Under Scrutiny, Trump Foundation
Raised, and Gave, More in 2016", "kicker": null,
"content_kicker": null, "print_headline": "Trump Fund
Collected And Gave More in 2016", "name": null, "seo": null
, "sub": null}, "keywords": [{"name": "subject", "value":
"Presidential Election of 2016", "rank": 1, "major": "N"},
{"name": "subject", "value": "Philanthropy", "rank": 2,
"major": "N"}, {"name": "subject", "value": "Conflicts of
Interest", "rank": 3, "major": "N"}, {"name": "subject",
"value": "United States Politics and Government", "rank": 4
, "major": "N"}, {"name": "subject", "value": "Campaign
Finance", "rank": 5, "major": "N"}, {"name": "subject",
"value": "Veterans", "rank": 6, "major": "N"}, {"name":
"organizations", "value": "Trump, Donald J, Foundation",
"rank": 7, "major": "N"}, {"name": "persons", "value":
"Schneiderman, Eric T", "rank": 8, "major": "N"}, {"name":
"persons", "value": "Trump, Donald J", "rank": 9, "major":
"N"}], "pub_date": "2017-11-21T02:58:45+0000", "new_desk":
"Washington", "byline": {"original": "By KENNETH P. VOGEL",
"person": [{"firstname": "Kenneth", "middlename": "P.",
"lastname": "VOGEL", "qualifier": null, "title": null,
"role": "reported", "organization": "", "rank": 1}],
"organization": null}, "word_count": 534, "score": 6
.577696}, {"web_url": "https://www.nytimes.com/2017/11/13
/business/trump-regulations-usda-lobbyists.html",
"headline": {"main": "An Open Door for Pesticide Lobbyists
at the U.S.D.A.", "kicker": "Trump Rules", "content_kicker"
: "Trump Rules", "print_headline": "At U.S.D.A., Pesticide
Lobby Encounters a Welcome Mat", "name": null, "seo": null,
"sub": null}, "keywords": [{"name": "persons", "value":
```

```
▼ Donald Trump [130]
   ▼ 0 {8}
        web_url : https://www.nytimes.com/2017/11/20/us
                  /politics/donald-trump-foundation-
                  closing.html
      ▼ headline {7}
           main : Under Scrutiny, Trump Foundation
                  Raised, and Gave, More in 2016
           kicker : null
           content_kicker : null
           print_headline : Trump Fund Collected And
                            Gave More in 2016
           name : null
           seo : null
           sub : null
      ▶ keywords [9]
        pub_date : 2017-11-21T02:58:45+0000
        new_desk : Washington
      ▼ byline {3}
           original : By KENNETH P. VOGEL
         ▶ person [1]
```

Instructions for running my code

To run my code, the following import statements are located at the top of my Sublime file:

```
import requests
import json
import sqlite3
import time
import collections
import random
from os import path
from PIL import Image
import numpy as np
import matplotlib.pyplot as plt
from wordcloud import WordCloud, ImageColorGenerator, STOPWORDS
```

All of the modules will need to be installed using pip install in the terminal window. If you have trouble installing the modules I would suggest using the sudo pip install commands and typing in your password.

On line 39 of my Sublime file, at the end of the line the code reads ‘api-key’: ‘enter-api-key’}). If you do not enter a key and are using the cache data, then the program will run just fine. If you want to enter your own access token, go to this website https://developer.nytimes.com and fill out the necessary information to obtain an access token. Then enter that access token where it says enter-api-key on line 39 of the code.

In terms of running the code after everything is installed, certain files need to be in the same directory. My Sublime file with all of the commented code is named Hana_Bezark_Final.py. If you wish to use the cache data, that file is named 206final_project.json. The SQL database that is created is named final_project_database.sqlite. The text file named trump needs to be in the same directory in order for the Word Cloud to be formed, because the Word Cloud code takes this text file as input to generate the Word Cloud. The picture file named trump_pic.png also needs to be in the same directory because the Word Cloud code uses the colors that are found in this image to mirror the color of the words that are generated in the Word Cloud. Once the Word Cloud code is run, it will generated the completed Word Cloud and save it to the directory as trump.png.

Documentation for each function I wrote

I commented each line of code in my Sublime file, so line by line descriptions of what my functions do should be clearly accessed within the file. Line 29 of my code in the Sublime file starts my first function, def newyorktimes_data(search_term). This function states to use the cache if it exists, or to make a request if a cache file does not exist. It sets up how to make the request to the New York Times API and loop through the pages to get all of the requested information. It then writes this information to the cache and returns the list of dictionaries received.

Line 50 of my code in the Sublime file starts my second function, newyorktimes_info(search_term). This function creates the variable New_York_Times_articles that contains all of the data gathered above in the previous function. It then loops through every article and checks to see if all the information is present. If the information is present then it is stored, and if it is not present then it is set to EMPTY. Once the empty dictionary (named info) is filled, it is then returned.

Document all resources

Date	**Issue Description**	**Location of Resource**	**Result (did it solve the issue?)**
11/28	High level research of New York Times API documentation	https://developer.nytimes.com/article_search_v2.json	Yes
11/28	High level research of Yelp API	https://www.yelp.com/developers/docume	Yes, but ended up not using Yelp

	documentation	ntation/v3	
12/4	How many articles are on each page on New York Times	https://developer.nytimes.com/article_search_v2.json#/Documentation/GET/articlesearch.json	Yes
12/5	How to put program to sleep	https://www.pythoncentral.io/pythons-time-sleep-pause-wait-sleep-stop-your-code/	Yes
12/6	How to resolve the issue of having duplicate headlines in New York Times data	https://data-gov.tw.rpi.edu/wiki/How_to_use_New_York_Times_Article_Search_API	Yes
12/6	How the keywords are set up within New York Times API	https://developer.nytimes.com/article_search_v2.json#/Documentation/GET/articlesearch.json	Yes
12/7	Learn how to code a Word Cloud in Python	https://github.com/amueller/word_cloud	Yes
12/7	Import Word Cloud module not working in terminal	https://github.com/amueller/word_cloud/issues/181	Yes
12/8	How to use Image Color Generator to change colors of Word Cloud	https://amueller.github.io/word_cloud/auto_examples/colored.html	Yes