Patterns in Popularity: A Cross-Platform Analysis of Wikipedia and Google Trends

Grace DesJardins

2024-12-10

GitHub repository link: https://github.com/gracedesj/727-Final-Project

Introduction

For my final project, I web scraped information from the "Most Popular Wikipedia Articles of the Week" webpage from four consecutive weeks (October 6 to 12, 2024; October 13 to 19, 2024; October 20 to 26, 2024; and October 27 to November 2, 2024). In gathering this information, I was curious if, over this four week period, there were any words/themes that appeared that were in common, and how they compared when looking at Google Trends data. With this question of interest, I performed a text analysis on the information that I scraped to find common words/themes, and then performed a Google Trends analysis using those common words/themes.

Wikipedia Web Scraping

During my Wikipedia web scraping, I essentially repeated the same process four times because I had four consecutive weeks to scrape. First, I would read in the given Wikipedia html page as an R object. Then, I would extract the table from the web page. This table had 25 rows because of the fact that it was the 25 most popular Wikipedia articles from a given week, and then 6 columns: Rank, Article, Class, Views, Image, and Notes/About. Once I extracted the table, I printed the result and noticed that there were NA's in the Class and Image columns, which made sense because there were images in both of those columns, which cannot be scraped and were not important to my analyses. Thus, I cleaned out those columns with NA, leaving me with the Rank, Article, Views, and Notes/About columns. As previously mentioned, I did this four times for each article: October 6 to 12, 2024; October 13 to 19, 2024; October 20 to 26, 2024; and October 27 to November 2, 2024. Below are previews of the four tables I obtained from these web scrapings.

October 6 to 12, 2024

```
## # A tibble: 10 x 4
##     Rank Article               Views     `Notes/about`
##    <int> <chr>                 <chr>     <chr>
##  1     1 Ratan Tata            5,034,616 "Ratan Tata, a Parsi Indian industria~
##  2     2 Lyle and Erik Menendez 2,686,453 "Down from the top spot after three w~
##  3     3 Tata family           1,656,736 "The prestigious family of #1, includ~
##  4     4 Joker: Folie à Deux   1,517,588 "Does the description \"jukebox music~
##  5     5 Noel Tata             1,198,976 "The half-brother of #1, and his succ~
##  6     6 Hurricane Milton      1,005,478 "Less than two weeks after Hurricane ~
##  7     7 George Baldock        931,894   "This English-born footballer who use~
##  8     8 Deaths in 2024        927,208   "Yesterday I got so oldI felt like I ~
```

```
##  9     9 Sean Combs             830,145   "As everyone discusses all the debauc~
## 10    10 Jamshedji Tata         785,225   "Relatives of #1: his great-grandfath~
```

October 13 to 19, 2024

```
## # A tibble: 10 x 4
##     Rank Article                Views     `Notes/about`
##    <int> <chr>                  <chr>     <chr>
##  1     1 Liam Payne             7,904,304 "\"It seems to me that when I dieThes~
##  2     2 Cheryl (singer)        2,011,202 "From 2016 to 2018, #1 dated this sin~
##  3     3 Lyle and Erik Menendez 1,800,441 "People who watch the Netflix show Mo~
##  4     4 One Direction          1,757,832 "In 2010, #1 had just had his second ~
##  5     5 Baba Siddique          1,747,594 "Indian politician Baba Siddique was ~
##  6     6 Lawrence Bishnoi       1,560,755 "Indian politician Baba Siddique was ~
##  7     7 Yahya Sinwar           1,239,034 "As the war in Gaza completed one yea~
##  8     8 Deaths in 2024         1,029,961 "All these places had their momentsWi~
##  9     9 Terrifier 3            917,431   "The third installment in the Terrifi~
## 10    10 Rodney Alcala          859,974   "In 1978, this guy won a date on tele~
```

October 20 to 26, 2024

```
## # A tibble: 10 x 4
##     Rank Article                                 Views     `Notes/about`
##    <int> <chr>                                   <chr>     <chr>
##  1     1 Rodney Alcala                           2,490,646 "In 1979, this repr~
##  2     2 Lyle and Erik Menendez                  1,468,336 "More murderers on ~
##  3     3 Liam Payne                              1,269,238 "The shocking death~
##  4     4 Venom: The Last Dance                   1,266,811 "Venom wasn't a goo~
##  5     5 Woman of the Hour                       1,242,489 "This American crim~
##  6     6 Deaths in 2024                          1,028,425 "Mumbling good morn~
##  7     7 Fernando Valenzuela                     882,113   "\"El Toro\", who p~
##  8     8 Fascism                                 762,276   "With election seas~
##  9     9 Kamala Harris                           722,806   "The Democratic Par~
## 10    10 2024 United States presidential election 697,072   "This election ente~
```

October 27 to November 2, 2024

```
## # A tibble: 10 x 4
##     Rank Article                                 Views     `Notes/about`
##    <int> <chr>                                   <chr>     <chr>
##  1     1 Teri Garr                               1,355,055 "This American actr~
##  2     2 2024 Ballon d'Or                        1,273,764 "European champion ~
##  3     3 Rodney Alcala                           1,258,084 "Netflix brought at~
##  4     4 2024 United States presidential election 1,234,532 "At least it's over~
##  5     5 Tony Hinchcliffe                        1,121,021 "The 2024 Trump ral~
##  6     6 Rúben Amorim                            1,110,284 "Manchester United ~
##  7     7 Liam Payne                              1,069,395 "Two weeks after th~
##  8     8 Diwali                                  1,053,976 "The Hindu festival~
##  9     9 Deaths in 2024                          1,005,464 "\"From that fatefu~
## 10    10 Freddie Freeman                         988,883   "As the Los Angeles~
```

Text Analyses

From the Wikipedia web scrapings, I wanted to analyze the text data to see if there were any prominent words or themes across the four weeks. In order to ease this process, I combined the four tables above into one table. I also narrowed down the columns, leaving me with Article and Notes/about, as those columns had the textual data that I needed. In the end, this combined table had 100 rows (25 rows per week) and 2 columns.

Next, I cleaned the data in this table utilizing the tidytext package. This allowed me to break the text in the Notes/about column into individual words (tokens), so each word became a row in a new data set. Cleaning the data also removed stop words (such as "and", "the"), which do not add much meaning for the text analysis. With this cleaned data, I was ready to move onto performing my text analyses.

My first text analysis involved finding the common words used overall. In other words, across all of the 100 rows of articles, the following were the top ten words that appeared in the Notes/about column:

```
## # A tibble: 10 x 2
##    word         n
##    <chr>    <int>
##  1 million     22
##  2 1           14
##  3 released    13
##  4 film        12
##  5 movie       12
##  6 netflix     12
##  7 series      11
##  8 tata        11
##  9 time        11
## 10 week        10
```

This particular analysis was not the most helpful with regards to my question of interest because the common words used overall were not very specific or thematic words. For example, the top three words, "million", "1", and "released", would not help me in my Google Trends analyses, as these words can be used to describe a plethora of events.

I then conducted another exploratory textual analysis to see if I could get words with more substance. In this analysis, I grouped the data by the article name, and then pulled the top ten words from each article in descending order. This analysis was also not helpful with regards to my question of interest because a majority of the "common" words in a given article only appeared once (n=1).

After these exploratory textual analyses, I decided to take another approach to find the common words/themes across the four weeks. Namely, I sorted the articles by how often they appeared in the "Most Popular Wikipedia Articles of the Week" across the four weeks. Shown below are my results, which show three articles that appeared in all four weeks: Deaths in 2024, Elon Musk, and Lyle and Erik Menendez. Because of the fact that these articles showed up in all four weeks, I concluded that these were popular themes for the four week period, and could be used as keywords for my Google Trends analysis.

```
## # A tibble: 10 x 2
##    Article                                n
##    <chr>                              <int>
##  1 Deaths in 2024                         4
##  2 Elon Musk                              4
##  3 Lyle and Erik Menendez                 4
##  4 Agatha All Along (miniseries)          3
##  5 Kamala Harris                          3
```

```
##  6 Liam Payne                                3
##  7 Rodney Alcala                             3
##  8 Sean Combs                                3
##  9 Terrifier 3                               3
## 10 2024 United States presidential election  2
```

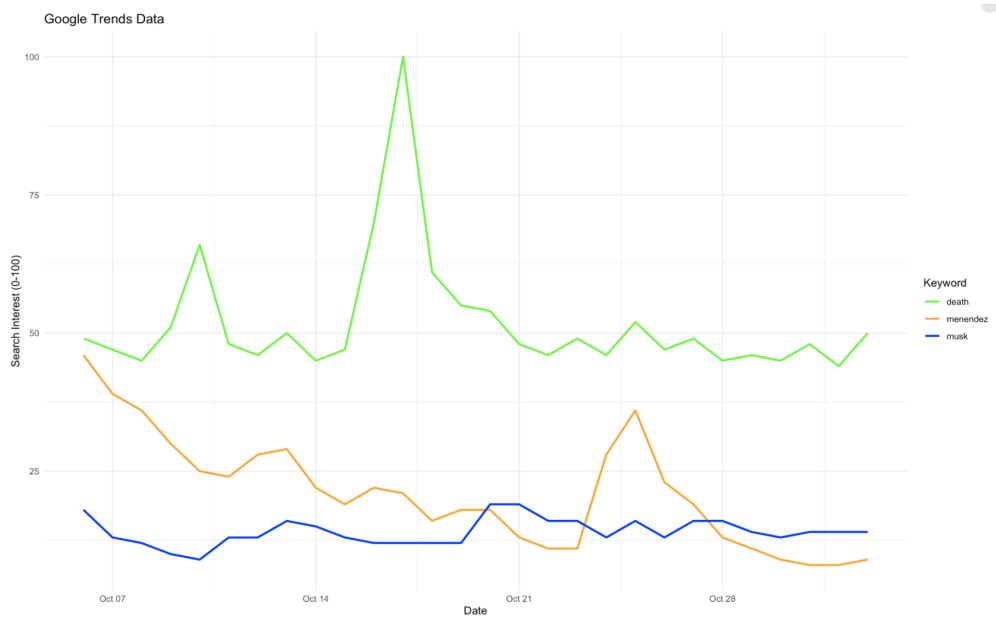
Before moving onto my Google Trends analysis, I wanted to do one more text analysis to see if it would give me any other informative results. In particular, I found what the common words were only for the articles that appeared in all four weeks (Deaths in 2024, Elon Musk, and Lyle and Erik Menendez). Like the analysis where I grouped the data by the article name and then pulled the top ten words from each article, this analysis was also not very helpful because the majority of the "common" words in these three articles only appeared, at most, four times. With that, I concluded my mostly exploratory text analyses. Though they weren't the most helpful with regards to my question of interest, the experimentation helped lead me to the popular themes for the four week period: Deaths in 2024, Elon Musk, and Lyle and Erik Menendez.

Google Trends API Analyses

In my Google Trends analyses, I wanted to see how common themes from my Wikipedia analyses compared to Google searching. In doing these analyses, I used the keywords "Death", "Musk", and "Menendez". I decided to have my keywords be one word, rather than what they originally were (Deaths in 2024, Elon Musk, and Lyle and Erik Menendez) because searching for the original article titles would return data specific to those exact phrases. Making my keywords one word helped make my queries a bit broader, as I was nervous that using the original article titles wouldn't give me much data.

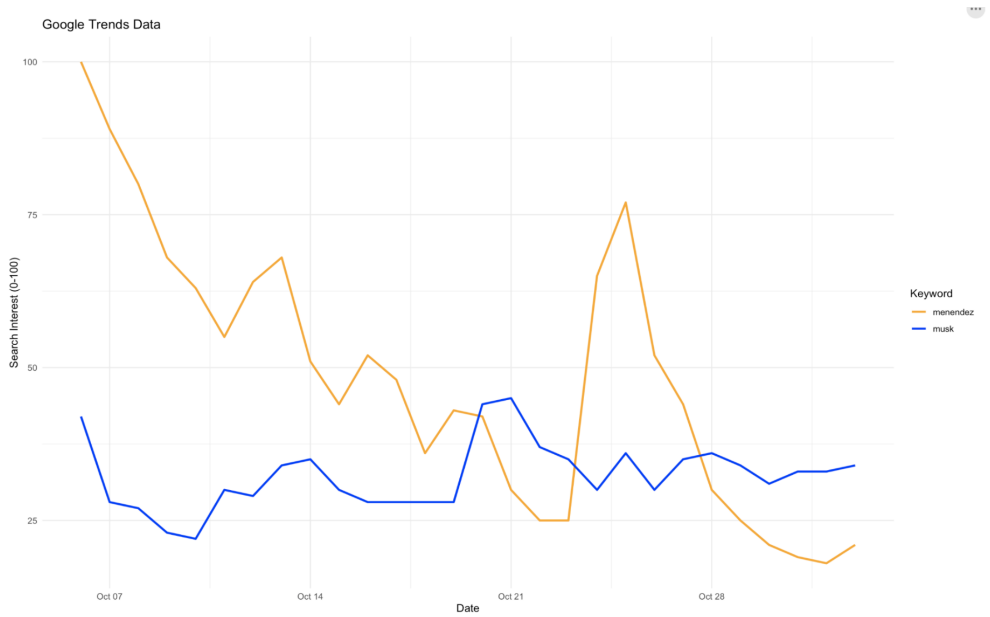
Global Analysis

First, I decided to run my Google Trends query globally, while using the time frame of October 6th, 2024 to November 2nd, 2024. I decided to do this because Wikipedia is a globally used website, so I wanted to see how the Google Trends compared on that level. The following graph shows the results:



As shown in the graph, the green line represents the "Death" keyword, the orange line represents the "Menendez" keyword, and the blue line represents the "Musk" keyword. From this, we see that all of the

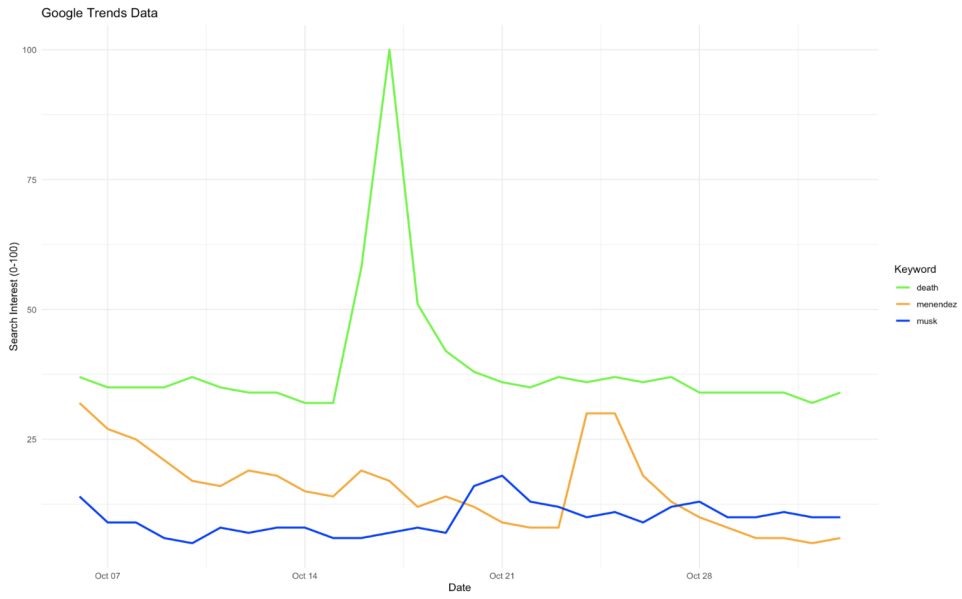
keywords have search interest, however, the "Death" keyword dominates. The "Death" keyword has the most search interest, whereas the "Menendez" and "Musk" keyword search interests have varying, but quite similar levels. Because of the fact that the search interests are relative to each other, I wanted to take a narrower look at the "Musk" and "Menendez" keywords. This is because of the fact that "Death" is a broader word than the last names "Musk" and "Menendez", and because I wanted to see how the data looked without the "Death" line that has a lot of search interest. After running the query without the "Death" keyword, these were the results that I obtained:



The graph above shows a similar pattern between the "Menendez" and "Musk" lines that were plotted with the "Death" keyword, but now the search interest is scaled differently due to the fact that the "Death" keyword was removed. This graph was useful as it helped me zoom further into the patterns between these two keywords, as it previously was not as pronounced.

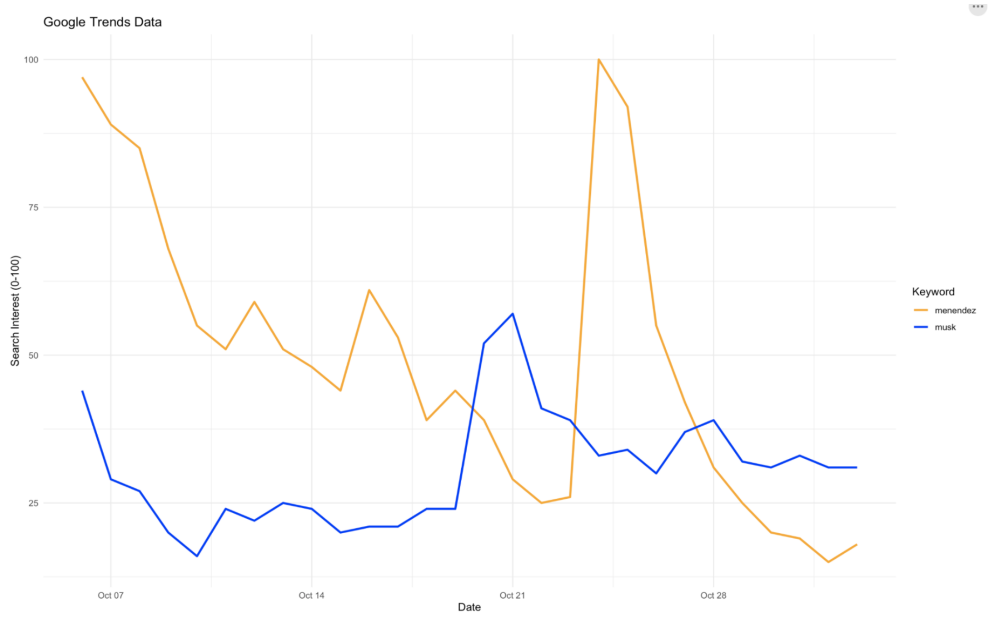
United States Analysis

I then decided to run my Google Trends query, but just within the United States, while using the time frame of October 6th, 2024 to November 2nd, 2024. I decided to do this because of the fact that Elon Musk and the Menendez brothers live in the United States. I was thinking that these people would concern a majority United States-based audience, so I wanted to see how this query compared to the global query. The following graph shows the results:



From a United States standpoint, I noticed pretty similar patterns to the global analysis. This likely means that the United States was a major contributor towards the global trends during my specified time frame, which makes sense due to the fact that Musk and the Menendez brothers live in the United States. Another interesting aspect of these graphs that I noticed was how the "Death" line had a massive increase around October 16th, which I knew was when English singer-songwriter, Liam Payne, died. Given that he was well known in the United States, as well as around the world because he was from the United Kingdom, it made sense that this was likely a major reason why the "Death" line spiked like it did globally and in the United States. However, there is another increase in death searches that is seen in the global graph around October 9th, but is not very prominent at all in the United States graph. This made me think that this was likely due to a death of someone that was better known in a different part of the world other than the United States. In fact, I looked at the early October dates in the "Deaths in 2024" Wikipedia article, as well as the tables from my Wikipedia web scrapings, and realized that this increase may have been due to the death of an Indian industrialist and philanthropist, Ratan Tata. Not only did his name appear on October 9th in the "Deaths in 2024" Wikipedia article, but his article title: "Ratan Tata" appeared twice across all four weeks. This would make sense because he was well-known in India, which has a very large population, and likely contributed to the global "Death" searches around the day he died.

Next, I wanted to run the same United States Google Trends query, but with only the "Menendez" and "Musk" keywords, just as I did previously. After running the query without the "Death" keyword, these are the results that I obtained:



This United States-standpoint graph looks pretty similar to the graph from the global-standpoint. However, I see a bit more pronunciation in this graph. Namely, the stark increase in the "Menendez" keyword in this graph reaches its peak at a search interest of 100, whereas in the global graph, the "Menendez" keyword reaches its peak at a search interest of about 75. Additionally, the "Musk" keyword in this graph reaches its peak at a search interest of about 60, whereas the "Musk" keyword in the global graph reaches its peak at a search interest of about 45. These results make sense because, as mentioned before, Musk and the Menendez brothers are based in the United States, so people likely have more amplified search interest for them in the United States than internationally.

Conclusions, Limitations, and Future Directions

Overall, in my Wikipedia analyses, I was able to infer that the articles searched across all four weeks (Deaths in 2024, Elon Musk, and Lyle and Erik Menendez) were popular themes during my specified time period. With that, I was able to see interest in my Google Trends analyses via the search interest for the keywords "Death", "Musk", and "Menendez". Though I see general interest in these topics during my specified time frame, it is hard to completely compare search interest between Wikipedia and Google Trends. One limitation is that Wikipedia and Google are fundamentally different websites/search engines, so it is hard to compare the two when different processes are involved within each website. Another important aspect to consider is the context that is being looked at. Especially for the Google Trends analyses, it was important for me to remember what time frame and geographical area I was looking at, as changing those up could have had major implications on the conclusions I made. In completing this project, I also considered some future directions I would take. First, I would want to dig deeper to see if my text analyses could give more specific, substantial common words. Taking the route that I took with the articles that appeared across all four weeks was helpful in determining the common themes, but I would also like to find a way to find 'better' common words. Additionally, I would like to look into articles that were not in all four weeks (e.g., 3 out of the 4 weeks) to see those trends, and how they compare to the articles that were in all four weeks. In doing this project, I was able to gain some insight into people's interests throughout October and a bit of November, and how those interests vary over time.