History Views: Final Report

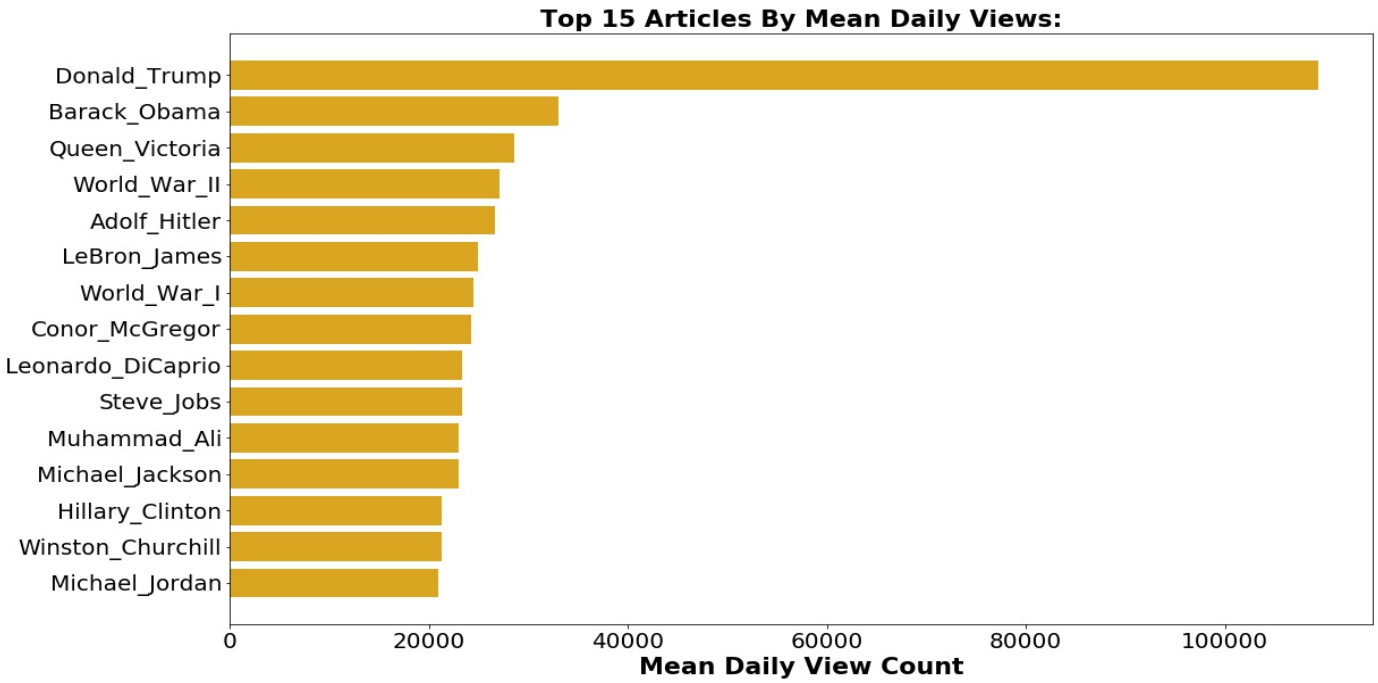
By Brian Penny

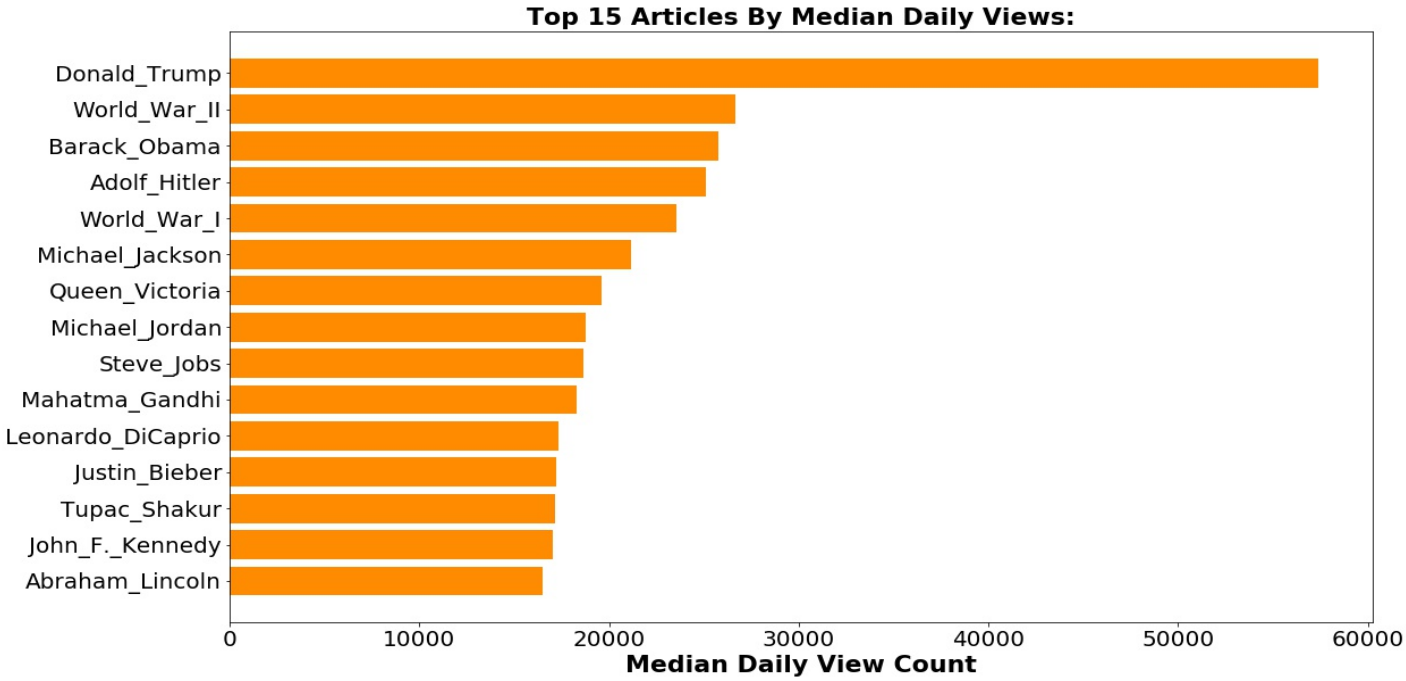
Gathering The Data:

To source this data, we used Wikipedia's Pageview API. We requested daily page view data, spanning from July 1, 2015 to September 30, 2018, for 400 articles on the English language Wikipedia: 100 on historical figures (people), 100 on historical ideologies/institutions/events (ie), 100 on world cities (cities), 100 on famous people of the modern era (moderns). The end result of this was 4 separate csv files, one for each of the four classes, each with 1188 rows (one for each day) and 101 columns (the 100 articles, plus 1 for the date column).

Initial Statistics:

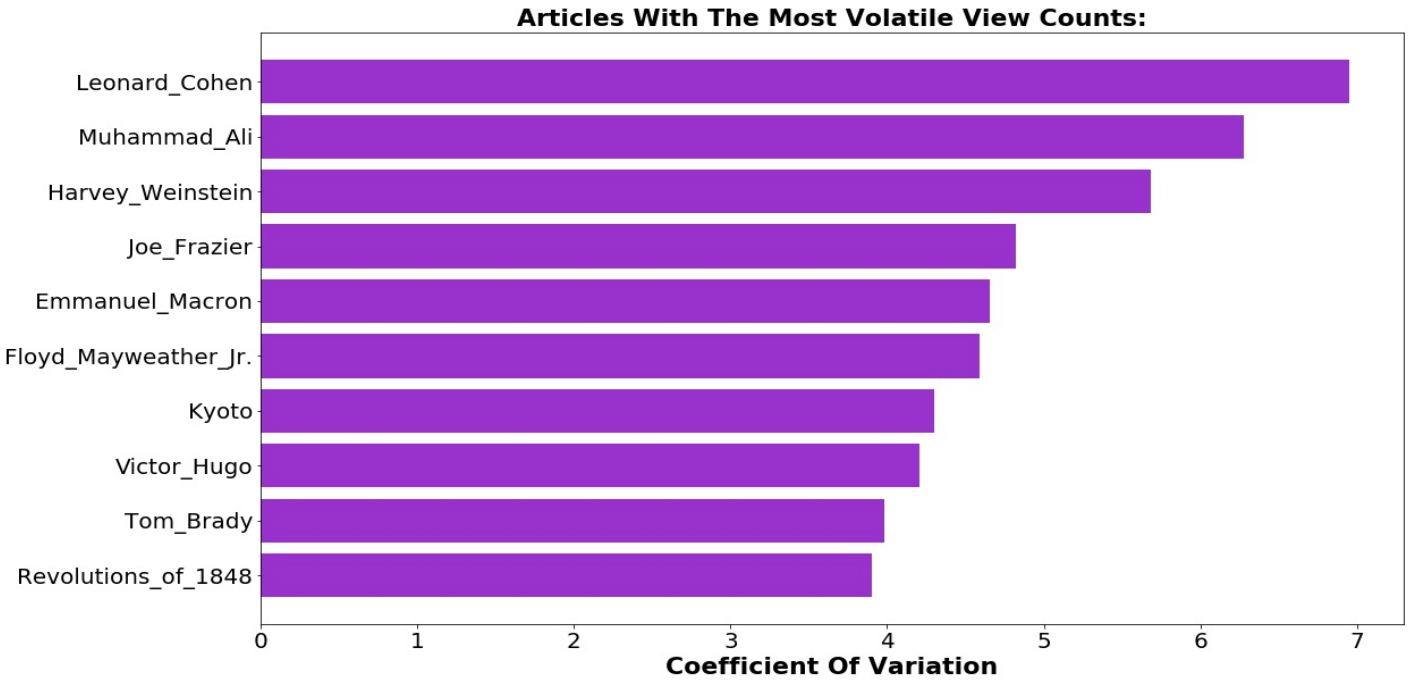
We divided our analysis work into five separate notebooks: one for analyzing each of the four classes of articles individually, and one for analyzing the combined dataset of all 400 articles. We started off by calculating some interesting summary statistics and seeing which articles had the highest values in these. We did this for each of the four classes, and also for the combination. For the purposes of this report, we will show the visualizations for the combined one only:

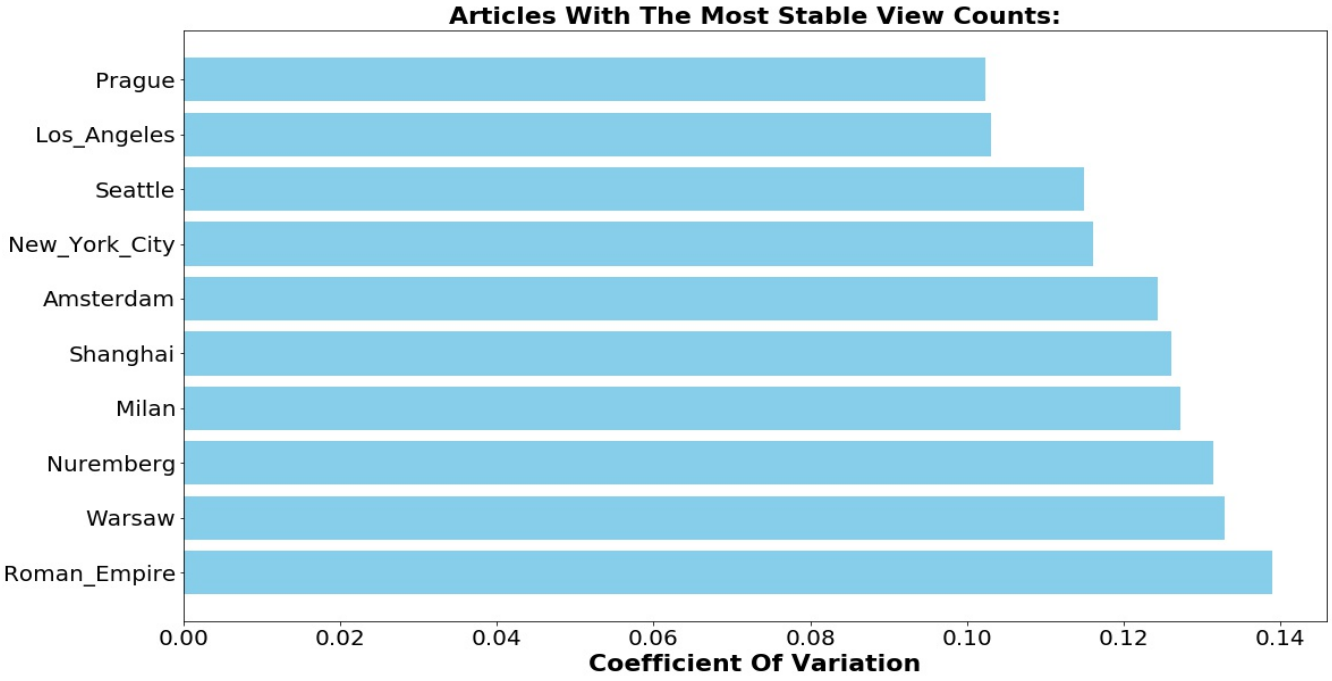




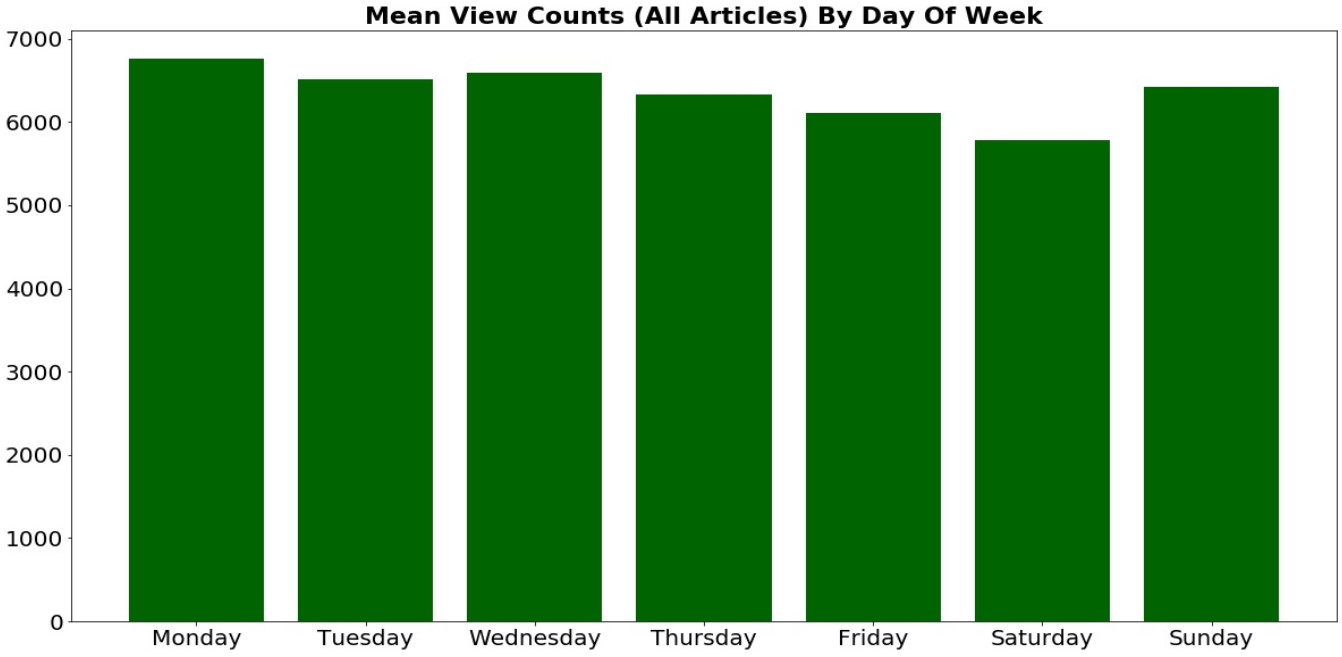
As we see, Donald Trump is by far the most popular, both by mean and by median. World War II and Adolf Hitler have more steadily high view counts than the more volatile Queen Victoria and LeBron

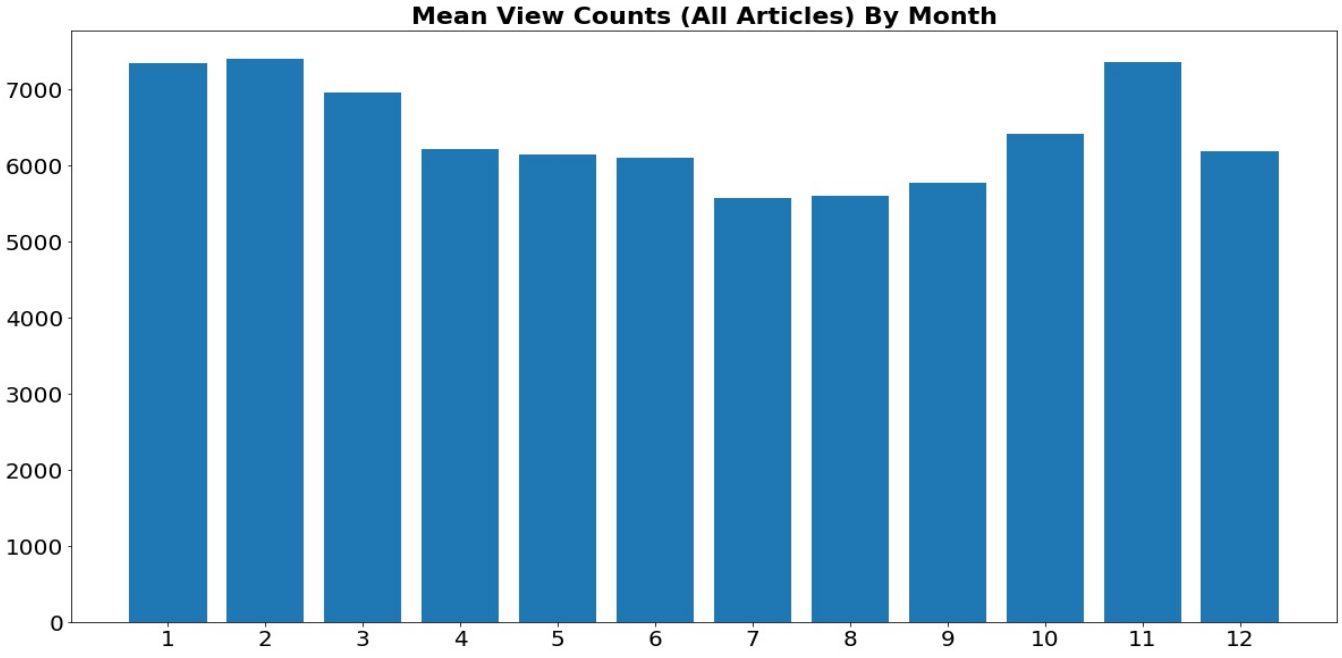
James, which is why the former two have a higher position on the median chart, whereas the latter two have a higher position on the mean chart.





As we see with the two charts above, the articles with the most volatile view counts, defined by those with the highest coefficients of variation (standard deviation divided by the mean), are primarily from the modern people set, while those with the most stable view counts (lowest coefficients of variation), tend to be from the city set. This makes sense, as the view counts of modern people tend to spike based on current events that they are involved with, whereas there is more consistent interest in cities, based on social study or touristic considerations.

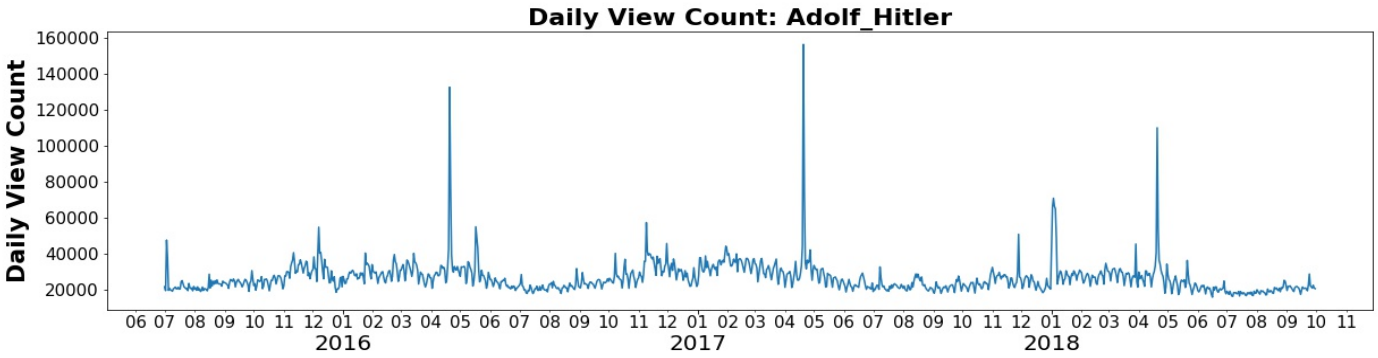




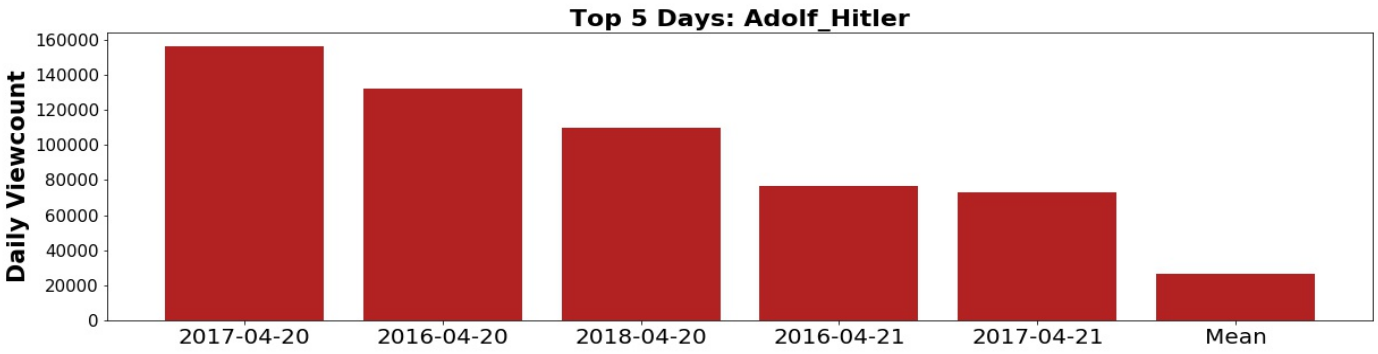
As we see with these charts, there is not a hugely significant variation in article views based on the day of the week, although Friday/Saturday (the party days) do have noticeably lower average views than Sunday-Wednesday. The monthly variation is a bit more prominent, with the highest average views in January, February, March, and November (cold months that are a part of the main grade school and university school years) , and the lowest average views in the summer months (July, August, September).

View Spike Analysis

One of the more interesting areas of analysis with this dataset involved looking at when the view counts of certain articles spiked, and then determining the probable reason for this spike. As a first example of this, let's take a look at the view counts for the highly viewed Adolf Hitler article:

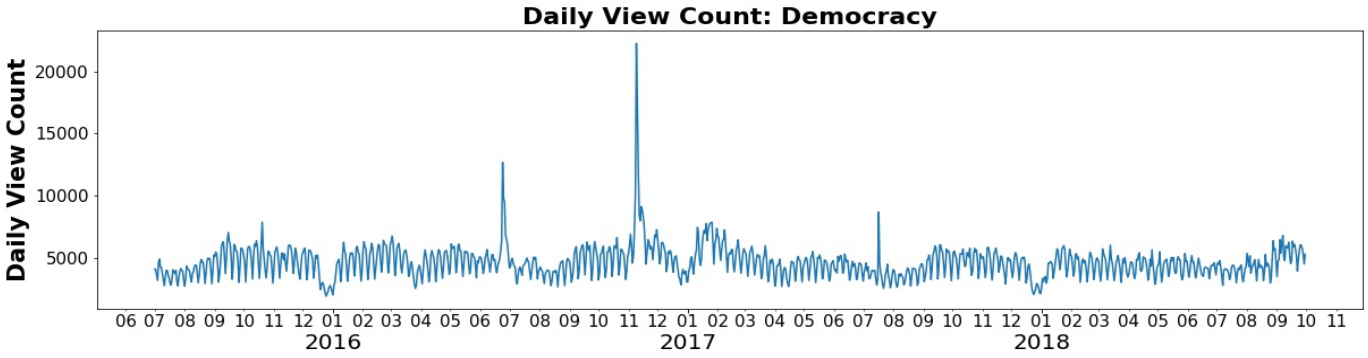


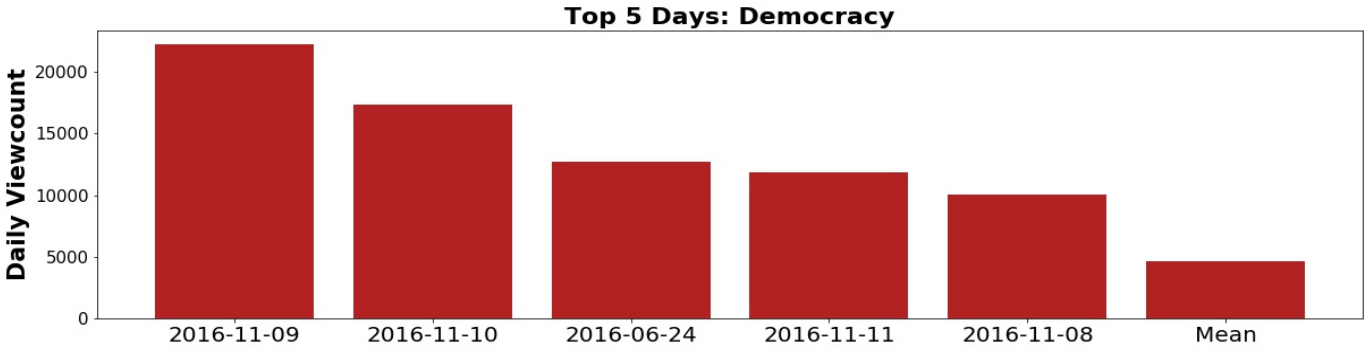
We see definite view count spikes near the middle of the month of April of all three years covered. To find out what specific days these are, we take a look at the top 5 days by view count:



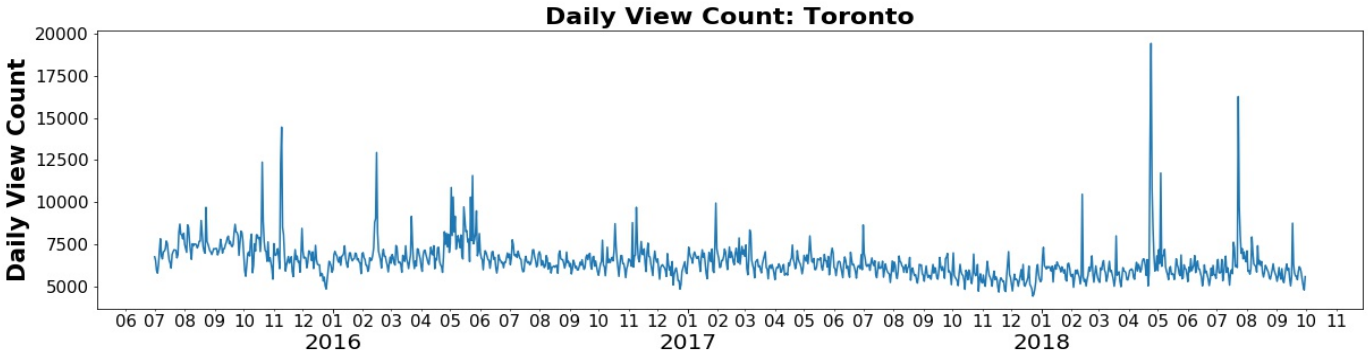
The April 20^{ths} in all three years are the top 3 days by view count. What happened on April 20 involving Adolf Hitler? It turns out to be: his birthday. For some reason, probably because of some sort of 'famous events that happened on this day' media, significantly more people are interested in Hitler's Wikipedia article on his birthday than they are on other days.

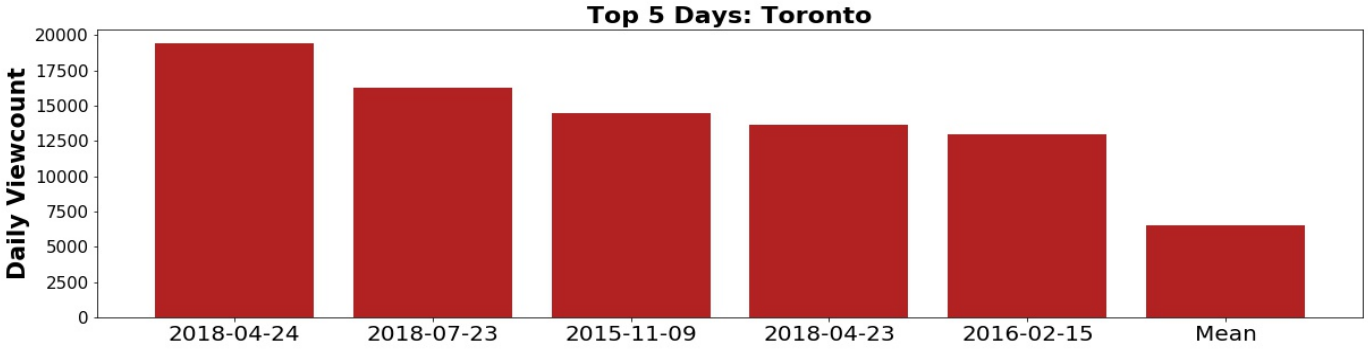
Let's take an in-depth look at some of the other articles:



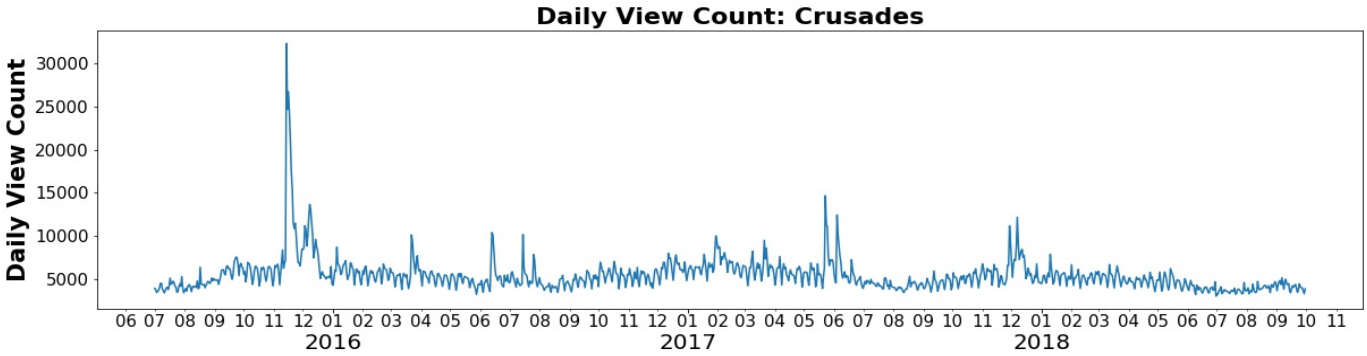


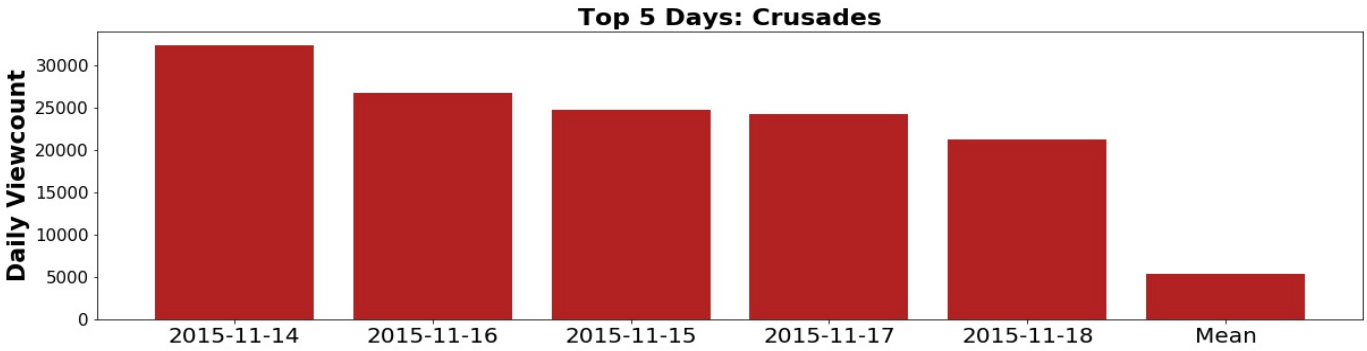
Here, we see that for the article on 'Democracy', view counts spike around the U.S. Presidential election, held on November 8, 2016, that brought Donald Trump to power. We also see a spike on June 24, 2016, the day after the U.K. referendum on 'Brexit', a vote that those in favour of the U.K. leaving the European Union won.





The largest spike for the Toronto article occurred around April 24, 2018, the day after the deadly Toronto ‘Van Attack’. Another major spike occurred on July 23, 2018, the day after the ‘Danforth Shooting’. The reason for the November 9, 2015 spike isn’t as obvious, but it is probably related to significant developments on that day in two Toronto child murder cases (involving the murders of Melonie Biddersingh and Katelynn Sampson). February 15, 2016 was the day after the 2016 NBA All-Star Game, held in Toronto at the Air Canada Centre.





For the article on the Crusades, we see the biggest spike on November 14, 2015, with the four subsequent days also being in the top 5 days by view count. What happened on this date? A large-scale terrorist attack on Paris, France occurred on November 13, 2015. On the following day, November 14, ISIS claimed responsibility for the attack, celebrating their blow against the so-called '**Crusader** Nation of France'. This appeared to get people interested in the Crusades.

Besides analyzing these and other articles' view count spikes individually, we also did some calculations to determine which article, in the combined dataset, had the most dramatic relative view count spike in each of the months covered in the dataset. To do this, we calculated the proportion of each article's total view count covered by each month, and then outputted the name of the article with the highest proportion in that month. We then discovered the most likely reason for this result.

Articles With The Most Dramatic Relative View Count Spike That Month

Year	Month	Article	Reason
2015	7	Amy_Schumer	Trainwreck, a movie she wrote and starred in, was released.
	8	Ronda_Rousey	She won another UFC fight, and got profiled in a number of articles.
	9	Stephen_Colbert	His debut as host of The Late Show.
	10	Stephen_Harper	He went up against Justin Trudeau in the Canadian Federal Election.
	11	Ronda_Rousey	She was defeated, for the first time, by Holly Holm.
	12	Ludwig_van_Beethoven	Featured in a Google Doodle, celebrating 245 years since his baptism.
2016	1	Celine_Dion	Her husband died, and she cancelled the rest of her Vegas performances this month.
	2	Ted_Cruz	Won the Iowa Republican Primary. Serious contender against Trump at this point.
	3	Brussels	ISIS terrorist attack.
	4	Andrew_Jackson	Announcement that he will be replaced on the front of the $20 bill by Harriet Tubman.
	5	Elizabeth_Warren	Engaged in a Twitter war with Donald Trump.
	6	Muhammad_Ali	Died, and was remembered fondly.
	7	Theresa_May	Became Prime Minister of the U.K.
	8	Rio_de_Janeiro	Hosted the Olympic Games.
	9	Ann_Coulter	Became the primary target for comedic insults at Comedy Central's The Roast of Rob Lowe.
	10	Joe_Frazier	41 year anniversary of his epic third battle with Muhammed Ali: The Thrilla In Manila.
	11	Leonard_Cohen	Died, and was remembered fondly.
	12	Aleppo	Assad government retakes control of the important city after years of fighting.
2017	1	Richard_B._Spencer	Video goes viral of him getting punched in the face by a protester on Trump's Inauguration Day.
	2	Tom_Brady	Quarterbacked his victorious team in the Super Bowl game.
	3	Paul_Ryan	Speaker of the House as Republicans failed to pass a 'repeal and replace' bill for 'Obamacare'.
	4	Marine_Le_Pen	Placed 2nd in the first round of the French Presidential Election.
	5	Emmanuel_Macron	Won the French Presidency by defeating Le Pen in the second round.
	6	Victor_Hugo	Featured in a Google Doodle, celebrating 145 years since he published the final chapter of Les Miserables.
	7	Floyd_Mayweather_Jr.	Lead-up to a giant boxing match between him and UFC champion Conor McGregor.
	8	Floyd_Mayweather_Jr.	He defeated Conor McGregor.
	9	Ben_Shapiro	He delivered a speech at U.C. Berkeley, prompting large-scale protests.
	10	Harvey_Weinstein	Numerous accusations of sexual misconduct levelled against him, kicking off the #MeToo movement.
	11	Robert_Mugabe	Military coup removes him from power in Zimbabwe, which he had ruled for 30 years.
	12	Jerusalem	Trump administration recognized Jerusalem as the capital of Israel, a very controversial move.
2018	1	Martin_Luther_King_Jr.	Martin Luther King Day.
	2	Jacob_Zuma	Announced his resignation as President of South Africa.
	3	Revolutions_of_1848	170th anniversary of its breakout.
	4	George_H._W._Bush	His wife, Barbara Bush, died.
	5	Copenhagen	Copenhagen Fashion Summit 2018.
	6	Copenhagen	Copenhagen Democracy Summit 2018.
	7	Zagreb	Croatia was in the World Cup 2018 Final, and Zagreb is its capital.
	8	Militarism	Russia announces largest war exercises since 1981, Vostok 2018, to take place.
	9	Jeff_Bezos	Forbes did an exclusive interview with him. Earlier that year, he surpassed Bill Gates as the richest person alive.

As we can see, the majority of the articles in this list are from the modern people set, as these tend to get the largest relative view count spikes from current events in the month in which they were involved. There are also some cities which normally do not get a lot of interest, but which current events that month forced into the limelight. There are also a few historical people/events, as a result of some kind of anniversary commemoration that massively boosted relative interest in them.

In general, our view spike analysis reveals that people's interest in Wikipedia articles tends to be very tied either to current events, or to anniversaries. For most articles, there is some base rate of interest in the topic, which keeps the view count fairly steady for the majority of the time, but with a current event or anniversary prompt, the view count skyrockets in relative terms, only to fall back to the base rate again in relatively short order.

Article Clustering

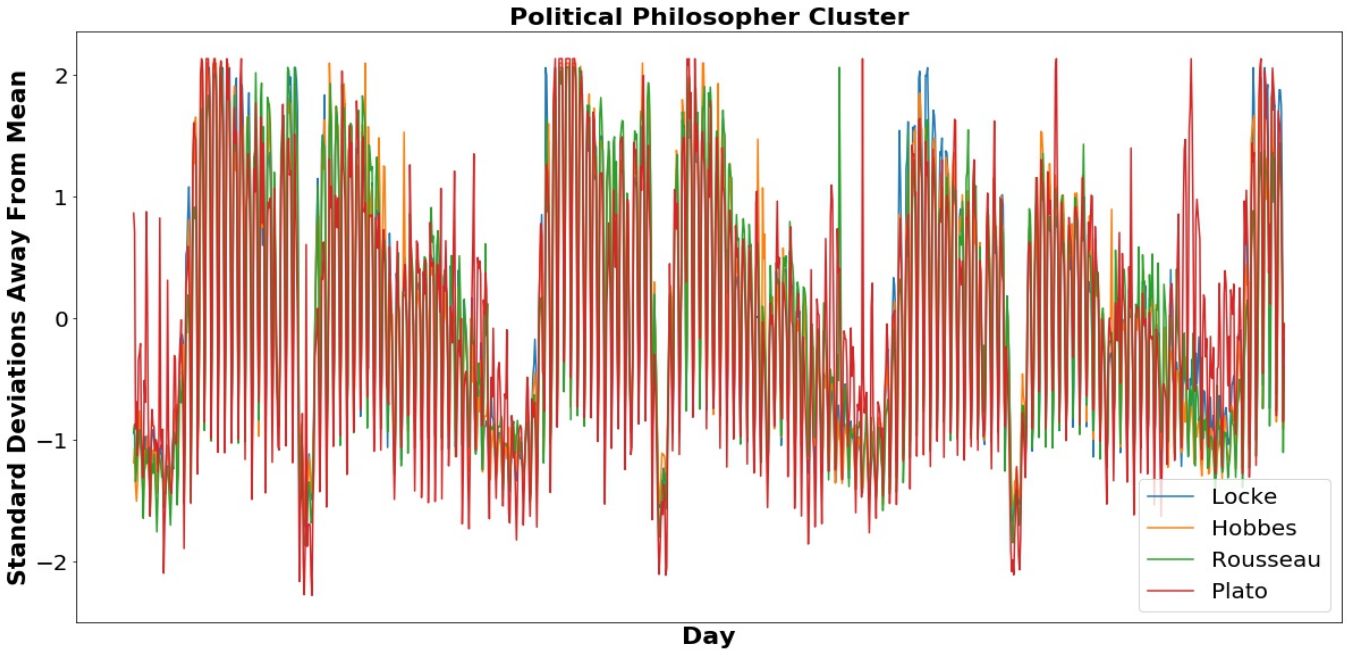
Another area of analysis that we were interested in exploring with this dataset was article clustering. Would the daily view counts of articles about, for instance, a similar kind of person, go up and down together? Would they be correlated enough to be clustered together via an unsupervised learning algorithm?

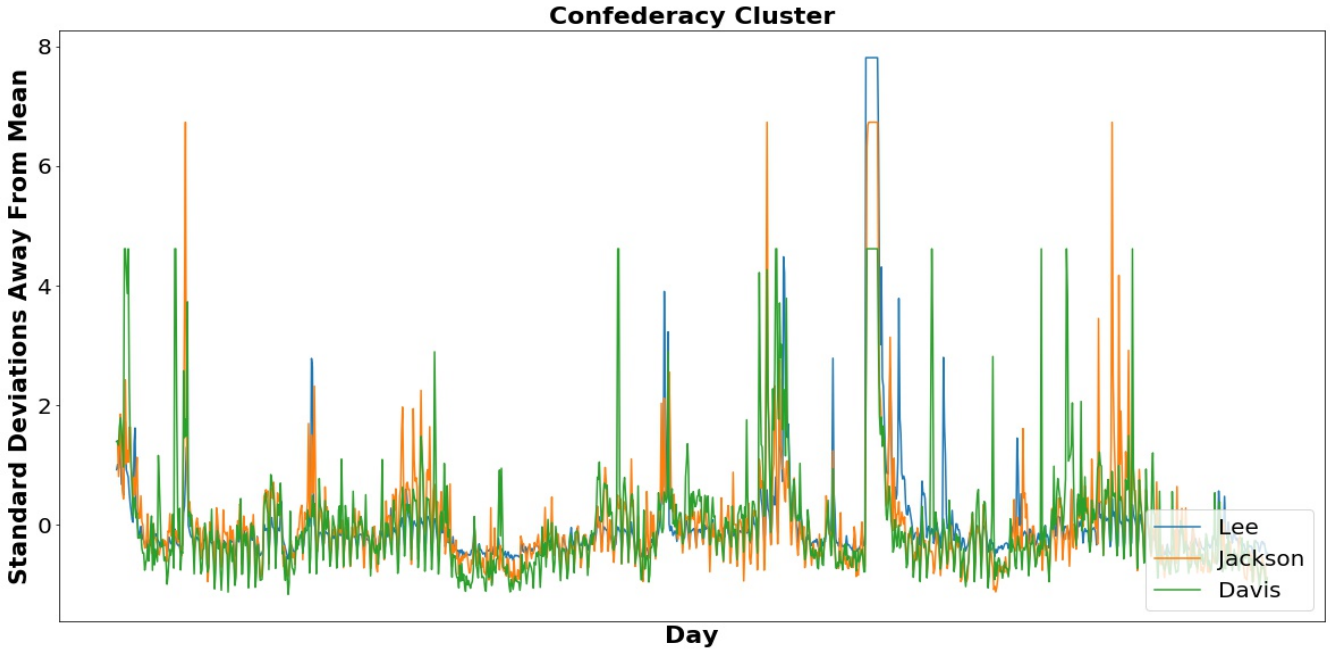
Before finding out the answer, we had to apply a few transformations on the data first. Firstly, we had to smooth out the kinds of big view count spikes that we were looking at in the section above. Although interesting and informative in their own right, for the purpose of clustering, these massive spikes in view counts (major outliers) would distract from the overall correlational trends in the data. Therefore, we smoothed them out a bit by setting all view count values greater than the mean of that article plus two of its standard deviations, to be equal to the mean of that article plus two of its standard deviations.

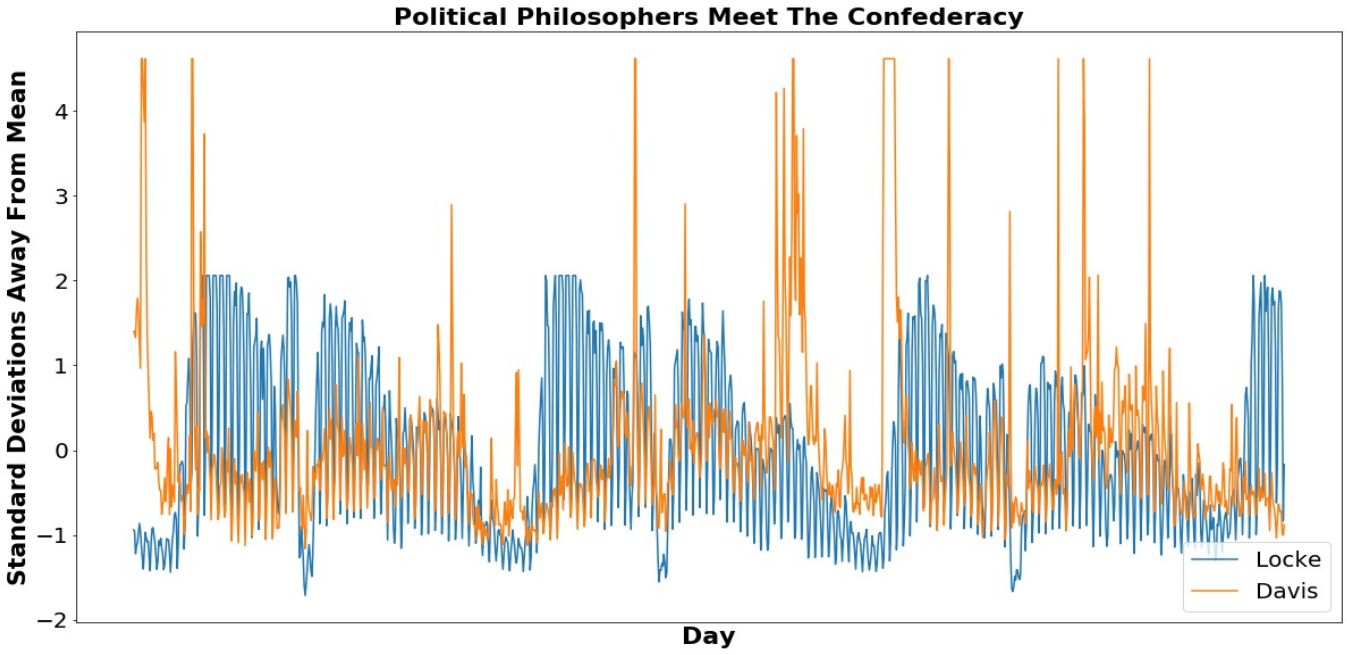
We then applied a Standard Scaler transform (turning the values into the number of standard deviations away from the mean, rather than actual view counts) in order to make the view count data of articles with different overall popularity levels comparable to one another, and useful in a distance-based clustering algorithm such as K Means. This done, we ran the transformed data, for each of the four classes of articles individually, through a K Means clustering algorithm, setting the number of clusters to find at 60 (a number reached through a process of trial and error aimed at finding the best clusters, based on our subject matter knowledge). We ran this clustering process on each of the four classes individually because the output was better and more manageable this way, as opposed to running a massive clustering on the combined dataset.

We found that this process did indeed result in some meaningful clustering of articles. For instance, for the historical people dataset, a cluster for famous political philosophers was formed, including Aristotle, Plato, John Locke, Niccolo Machiavelli, Voltaire, Jean-Jacques Rousseau, Thomas Hobbes, and Pericles (technically a political leader, but one who valued, and is associated with, ancient greek philosophy). Another cluster, this time for leading figures of the Confederate States Of America, was also formed, including Robert E. Lee, Stonewall Jackson (both famous Confederate generals), and Jefferson Davis (the Confederate President).

To demonstrate the kind of pattern that the clustering algorithm was able to pick up on, we display three line charts: one with the smoothed and scaled view counts of four members of the political philosopher cluster, one with those of the three members of the Confederacy cluster, and one with one member of one cluster and one of the other:


Political Philosopher Cluster
Standard Deviations Away From Mean
2
1
0
−1
−2
Day
Locke
Hobbes
Rousseau
Plato


Confederacy Cluster
Standard Deviations Away From Mean
8
6
4
2
0
Day
Lee
Jackson
Davis



We see here that for both clusters, there is a distinctive view count pattern, and that individual articles within these clusters largely conform to the overall pattern of the cluster. But when an article from one cluster is superimposed on an article from another, then we see two distinct patterns.

In addition to these two clusters, the clustering for the historical people dataset also created clusters for classical composers, Mongol Great Khans, heads of the Medici family, two clusters of U.S. Presidents, and others. Interesting clusters were also created for the other three datasets. To explore these clusters in more detail, open the html file (in the project file) corresponding to the article class that you are interested in to get an interactive view of the different clusters (with more than one article in them) created.

In general, the results of the clustering process suggest that, outside of the major view spike events, people tend to look at multiple related Wikipedia articles in one sitting, rather than just one. People

might get to these related articles based on embedded links to one article in another, or based on an interest in a category of articles, rather than just a single one.

Exploratory Linear Modelling

Although this project is focused on exploratory data analysis for general insights and understanding, rather than on predictive modelling of a designated target variable, we decided that it would be interesting to run some simple linear models, just to see how accurately we were able to predict the daily view counts of certain articles, given those of other, related articles. In order to avoid our model getting ruined by outliers, we applied the same smoothing transformation in this case as we described above, but this time on the combined dataset of all four classes of articles.

The 10 models run are summarized below:

Target	Features	Model Specifications	Training R Squared Score	4-Fold Cross Validation R Squared Score
'John_Locke'	'Thomas_Hobbes', 'Mercantilism', 'Monarchy'	LinearSVR (C=0.0001)	0.952	**0.945**
'Adolf_Hitler'	'World_War_II', 'Imperialism', 'Authoritarianism', 'Benito_Mussolini', 'Edmund_Burke', 'Johann_Wolfgang_von_Goethe', 'Abraham_Lincoln'	LinearSVR (C=0.00001)	0.597	**0.531**
'Democracy'	'Republicanism', 'Individualism', 'Dictatorship', 'Monarchy', 'Aristocracy', 'Jean-Jacques_Rousseau’, 'Edmund_Burke'	LinearSVR (C=0.00001)	0.693	**0.625**
'Wolfgang_ Amadeus_Mozart'	'Franz_Schubert', 'Romanticism', 'Age_of_Enlightenment', 'Leonardo_da_Vinci', 'Frédéric_Chopin', 'Johann_Sebastian_Bach', 'Niccolò_Machiavelli'	LinearSVR (C=0.0001)	0.541	**0.415**

'Communism'	'Capitalism', 'Industrial_Revolution', 'Karl_Marx', 'Russian_Revolution', 'Totalitarianism', 'Social_democracy', 'Cicero'	LinearSVR (C=0.00001)	0.736	**0.557**
'Islam'	'Muhammad', 'Crusades', 'Judaism', 'Dubai', 'Mecca'	LinearSVR (C=0.00001)	0.751	**0.513**
'London'	'Rome', 'Islam', 'Boston', 'Vienna', 'Paris', 'Middle_Ages', 'Romanticism'	Linear Regression()	0.609	**0.442**
'Toronto'	'San_Francisco', 'Boston', 'Seattle', 'Chicago', 'Vienna', 'Vancouver', 'Sydney'	LinearSVR (C=0.00001)	0.543	**0.402**
'Ronald_Reagan'	'Franklin_D._Roosevelt', 'George_W._Bush', 'Bill_Clinton', 'Washington,_D.C.', 'Donald_Trump', 'Bernie_Sanders', 'Authoritarianism'	LinearSVR (C=0.00001)	0.747	**0.619**
'Justin_Trudeau'	'Pierre_Trudeau', 'Stephen_Harper', 'John_A._Macdonald'	LinearSVR (C=0.0001)	0.775	**0.751**

For feature selection on these models, I outputted the 20 articles that had the highest Pearson correlation coefficients for the target article, then selected, through a process of trial and error with the intent of optimizing the cross validation R squared score, at most 7 articles from this list as features. Given the exploratory nature of these models, where totally optimizing the R squared scores wasn't incredibly important, we determined that a more rigorous process of feature selection was not required. We stuck to linear models so that the coefficients for each model would be easily interpretable.

In general, the pretty respectable cross validation R squared scores achieved by these models further illustrates the point made in the section above, which is that there are indeed significant and non-spurious correlations between the daily view counts of related articles.

Conclusions

With that, we now have a better understanding of how the public interacts with online, history-related information sources. Through our view spike analysis, we found that sudden, major interest can be generated in articles based on current events or significant anniversaries. Through clustering and exploratory linear modelling, we found that there are significant correlations between the daily view counts of articles about related subjects, suggesting that people tend to ‘chain surf’ through related articles.

Although for this project, we decided to focus on articles related to history, a very similar process could be applied to other classes of Wikipedia articles in order to get more insight on how the public interacts with them. For instance, we could look at the articles of companies in industries of interest to get a better understanding of what significant P.R. or marketing events generate view count spikes, and which companies’ interest levels (view counts) are correlated with one another. This could help the analyst get a better understanding of the industry landscape in general, and about how certain events affect interest in specific companies. These insights could then be used to make more informed business decisions.