WikiShark

An Online Tool for Analyzing Wikipedia Traffic and Trends

Elad, Vardi
Hebrew University of Jerusalem
elad.vardi@mail.huji.ac.il

Lev, Muchnik
Hebrew University of Jerusalem
lev.muchnik@mail.huji.ac.il

Alex, Conway
University of Cape Town
alex@numberboost.com

Micha, Y, Breakstone
Hebrew University of Jerusalem and MIT
micha.breakstone@gmail.com

ABSTRACT

Wikipedia is a major source of information utilized by internet users around the globe for fact-checking and access to general, encyclopedic information. For researchers, it offers an unprecedented opportunity to measure how societies respond to events and how our collective perception of the world evolves over time and in response to events. Wikipedia use and the reading patterns of its users reflect our collective interests and the way they are expressed in our search for information – whether as part of fleeting, zeitgeist-fed trends or long-term – on most every topic, from personal to business, through political, health-related, academic and scientific. In a very real sense, events are defined by how we interpret them and how they affect our perception of the context in which they occurred, rendering Wikipedia invaluable for understanding events and their context. This paper introduces WikiShark (www.wikishark.com) – an online tool that allows researchers to analyze Wikipedia traffic and trends quickly and effectively, by (1) instantly querying pageview traffic data; (2) comparing traffic across articles; (3) surfacing and analyzing trending topics; and (4) easily leveraging findings for use in their own research.

CCS CONCEPTS

• **Information systems** → World Wide Web; World Wide Web; Web interfaces; Wikis; World Wide Web; Web mining; Data extraction and integration; World Wide Web; Web mining; Web log analysis; World Wide Web; Web mining; Traffic analysis; World Wide Web; Web services; RESTful web services; • **Human-centered computing** → Collaborative and social computing; Collaborative and social computing systems and tools; Wikis;

KEYWORDS

WikiShark, Wikipedia Page Views, Wikipedia Trends, Data Dumps

ACM Reference Format:
Elad, Vardi, Lev, Muchnik, Alex, Conway, and Micha, Y, Breakstone. 2021. WikiShark: An Online Tool for Analyzing Wikipedia Traffic and Trends. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3442442.3452341

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia
© 2021 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3452341

1 INTRODUCTION

With this project, our goal is to build a robust analysis platform that will enable researchers to access, manipulate and analyze vast amounts of data in a straightforward manner without being experts on databases or Big Data, and without requiring any significant effort in data preprocessing, construction or validation. Additionally, our intent is to empower more tech-savvy researchers to conduct fully-programmatic parallel analyses across a wide range of queries and hypotheses. To achieve this, we have indexed and analyzed traffic data dating back to 2008 (data which we continue to analyze and update hourly) and have leveraged this data to develop a robust analysis platform, following a 2-tier design paradigm. On the first level, we created a simple web-based User Interface (UI), with intuitive, self-explanatory, multiple-querying, graph-plotting, graph-scaling and peak-flagging capabilities; it presupposes the existence of no technical skills on the part of the user. On the second level, we created a compressed and optimized database allowing programmatic access.

The web-based UI (see screen captures in the Appendix) supports a wide range of features including (i) automatic completion for Wikipedia search terms; (ii) date range selection; (iii) simultaneous plotting of several searches; (iv) comparison across different languages; (v) normalization across terms; (vi) peak flagging; (vii) exporting data in several formats; (viii) various graph plotting options; and, last but not least, (ix) trend information analysis.

In the trend information analysis section (see Appendix, images 3 and 4) researchers can view: trending pages, trending pages by category, and trending categories. For each page, we show (i) its growth related to long-term average page views; (ii) a preview chart of traffic data for the last few days; (iii) number of pageviews; (iv) a thumbnail image; and (v) category.

Additionally, we created a Chrome browser extension allowing researchers to view the pageview data from the WikiShark graph directly on the Wikipedia website (see Appendix, Image 6) and a shortcut button on the article's full traffic data on WikiShark (see Appendix, Image 7).

Finally, the trends are automatically tweeted on our new Twitter page (https://twitter.com/wikishark_) every few hours, using the Twitter API (see Appendix, Image 8) in order to bring more attention to Wikipedia's global trends, and to open a door to new discussion and research.

These provide researchers untrained in the field of Big Data a way to straightforwardly explore patterns and obtain insights into Wikipedia activity data.

2 PLATFORM ARCHITECTURE

The user interface presented (see Appendix, Image 1) runs on top of a highly optimized infrastructure providing fast access to efficiently stored original pageview data, which are not otherwise easily accessible. To explain, the original dataset [1] represents hourly snapshots of the number of views for each of the Wikipedia pages. All pageview data since December 2007 are included, and data is further updated on an hourly basis. Corresponding to the 13-year timeframe that covers the span from January 1st 2008 to January 1st, 2021, the repository comprises 113,880 hourly log files, consuming about 30 uncompressed TB (or, on average, 2.25 TB and 8760 files per year). However, for any practical purpose, the way in which the dataset is organized renders it extremely hard to take advantage of, because the values corresponding to a single article are stored across all files in different locations: zero values are omitted, and so the addresses corresponding to a single article are different in each hourly log file.

To make the data useful, we had to (i) allow for rapid extraction of large amounts of pageview data (milliseconds); and (ii) significantly reduce the size of the data so that they could be manipulated on a standard hard drive.

To accommodate fast pageview data extraction, we transposed the databases. Wikipedia data is sampled on an hourly basis, with approximately 55M Wikipedia articles across all languages, so that accessing pageview data for a single article requires some 100,000 distinct disk-access operations (8760 hourly dumps per year * 13 years), all at potentially different locations (since zero pageviews are omitted). By transposing the databases so that files now represent yearly data for articles (rather than a single sampling hour for all articles), per-year-pageviews for an article could now be efficiently extracted from a single file, where the data are contiguously stored, with a single disk-access operation. Additional optimizations are explained below, with the architecture presented in detail.

The space necessary to store the data was reduced by (a) removing all unneeded and ancillary data; and (b) moving to a binary representation. With regard to removal of redundant data, note that, with the databases transformed, article names no longer need to be listed within the files, as in the Wikipedia database. Rather, each file now aggregates yearly views per hour. Hours were pre-indexed, so that our files simply contain a contiguous count of views. With regard to binary representation, we made use of the fact that the vast majority of articles had fewer than 255 pageviews per hour throughout their entire history, and could therefore be represented by single bytes, as detailed below. The combined result was an effective compression rate of 1:8.5, yielding a repository of approximately 1.6TB (or 270GB/year). Such a size allows the database to be easily disseminated and utilized from a standard large local hard drive.

Below we describe the architecture outlined above in greater detail. As noted, the vast majority of articles have fewer than 255 pageviews per hour over their entire history, and can therefore be represented by single bytes. To accommodate all articles, we designed a two-layer database, one with a single byte recording pageviews (0-255) and one with 3 bytes (0-224-1) to accommodate the most popular articles (e.g. the page dedicated to Michael Jackson experienced approximately 1M pageviews in a single hour around the time of his death).

Additionally, we used the following design to optimize access to data. For each data-year we created a single file containing all articles, each storing a vector containing per-hour data. Reading and writing data points is a time-costly operation, and since we needed to accommodate millions of such operations, we defined an area of the RAM where approximately 2,000 articles were stored contiguously (see Figure 1), the logic being that writing a single data point takes the same amount of time as writing an entire block, so contiguously storing the data on the RAM, and writing them as blocks, can dramatically reduce the writing overhead.

Each article was entered into the database with its own ID, which served as a pointer to its location within the data-year files, allowing access to the relevant data points. Each year comprises 8,760 bytes of data (24 hours x 365 days). Multiplying this by 55M articles, we have the binary data-year file containing approximately 270GB of data (see Figure 2). As mentioned earlier, some articles had more than 255 hourly pageviews (e.g. Michael Jackson). Such articles used a separate file with 3 bytes per hour.

Finally, all articles are contiguous, so pointers point to the beginning of the file, along with an offset, allowing for speedy retrieval (see Figure 3).

3 USING THE WIKISHARK SYSTEM

The primary mode of access to WikiShark is through the home page (wikishark.com) which is divided into 5 main sections (marked A,B,C,D,E on Appendix Image 1.1) – Main Search bar, Trending Pages, Popular Pages, Trending Categories and Trending Pages by Category.

Below we briefly explain how to use each section.

3.1 A – Main Search bar

The main search bar (Marked as "A" in the Appendix, Image 1.1) is the main access point for the system. In the search bar one can enter Wikipedia article names to view their associated pageview traffic data. When a user starts typing the name of a Wikipedia article, a list of entries will open with auto-completion (see Appendix, Image 9) according to the names of the existing Wikipedia article names. Users can enter one or more article names. When a user enters more than one Wikipedia article name, all associated article traffic will be displayed on the same graph, for easy comparison. In the example shown in Image 5, below, two entries were entered in the search bar: Donald Trump and Joe Biden. Clicking on the search icon, located on the right part of the search bar, activates the search query and takes users to the results page, where they can view the pageview traffic data result chart (see detailed explanation below).

B – Trending Pages

The Trending Pages section shows the current five top trending pages. The trends are calculated and updated hourly using the list of formulas below.

For each trending page, we show (i) its name; (ii) thumbnail and category; (iii) traffic increase rate; (iv) a small preview chart for traffic data; and (v) number of pageviews over the past 24 hours.

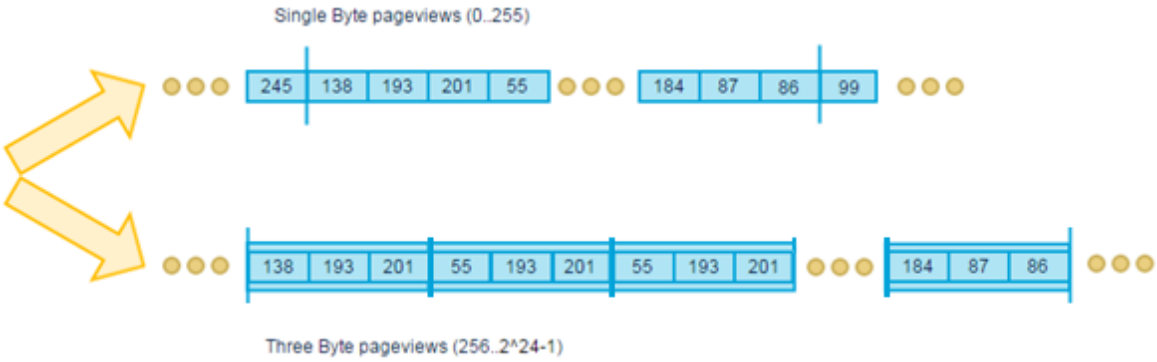

Figure 1: Single-byte and three-byte pageview representation.

Figure 2: A binary data file containing pageviews per year.

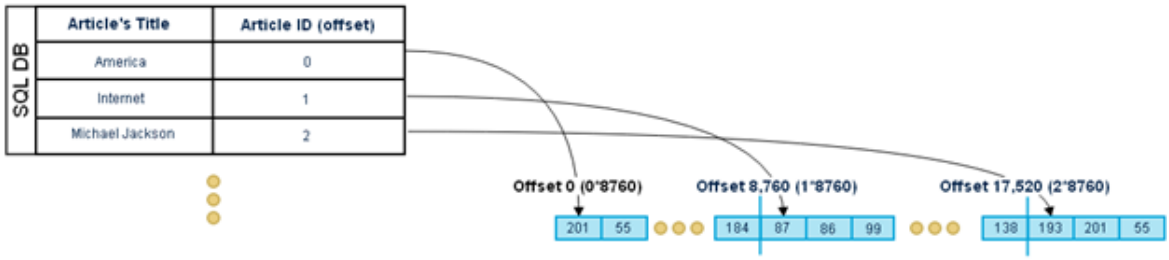

Figure 3: Pointers use offsets to facilitate quick retrieval.

Traffic increase rate.
this is the measure used to decide the top trends. It is calculated using the following formula. First, we define two time frames, a 'short-term' time frame, for which we measure the amount of traffic per page in the last 24 hours. Then we measure an 'ongoing time frame' which is calculated as the average number of daily page views over one week, starting two weeks prior. We use this as a reliable measure of ongoing traffic, as it is not immediately adjacent to a potential emerging trend on the one hand, and on the other, it is not too far in the distant past. The traffic increase rate is calculated by dividing the 'short-term' traffic by the traffic in the 'ongoing time frame'.

To further eliminate noise and ensure that trending pages only include truly interesting articles, we apply the following additional restrictions: (i) An article is considered trending only if its traffic increase rate is at least 2x; (ii) We only look at articles with more than 5000 page views over the last 24 hours, and (iii) for the 'ongoing time frame' we increase the count to 240 page views per day to avoid exaggerated traffic increase rates.

After calculating the increase rate for each article, we sort them by their associated traffic increase rate, and showcase the top 5 on the homepage.

3.1.1 The Image thumbnail. the image is taken by Wikipedia's image API [5], the API is called once for each title, and then stored and cached in our database. If we cannot retrieve an image from the API, we use the WikiShark logo.

3.1.2 A small preview chart for traffic data. The small preview chart for traffic is taken for the last 7 days using jQuery Sparklink [7] which provides a quick glimpse of the trend without the need to perform a full search and open a new page.

3.1.3 Category Classification. We performed the category classification by analyzing the content of each article. In order to classify the data into categories, we used a tool called Textrazor API, which was simple to use and gave us fast, solid results without the need

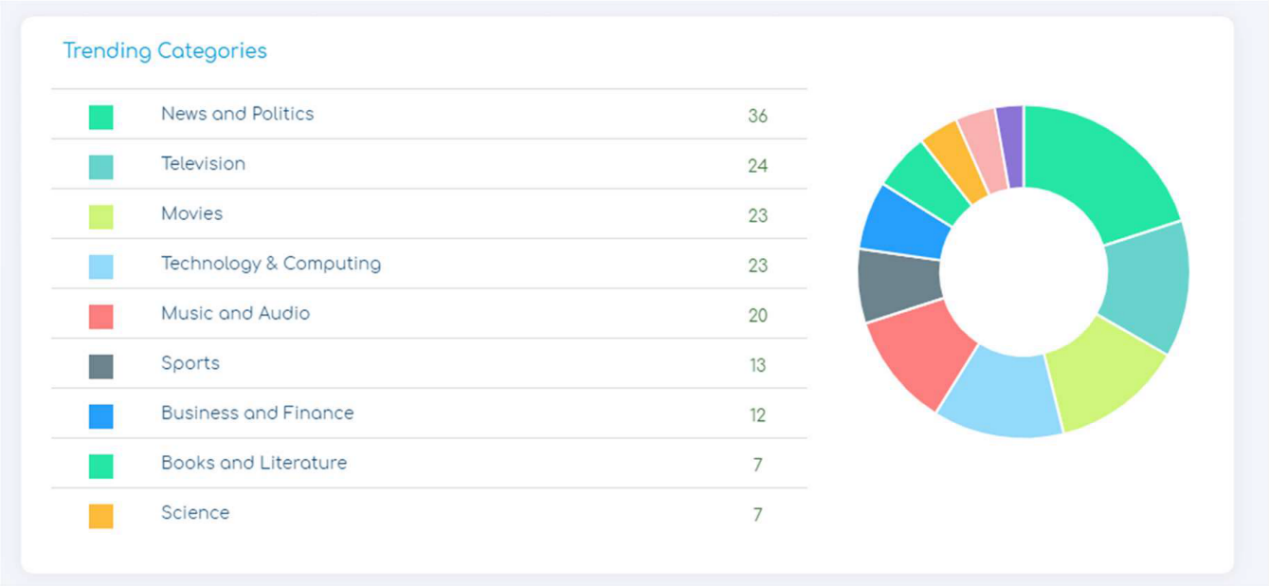

Image 4: – Current top trending categories.

to build our own classifier. We chose to use the IAB Content Taxonomy's categories [9] which was one of the classification options Textrazor offers. Originally used in advertising, it has a simple and comprehensive list of categories and so was a good fit for us. Some of the main categories of the IAB Content Taxonomy are: Technology & Computing, News and Politics, Sports, Business and Finance, Movies, Books and Literature, Music and Audio, Religion & Spirituality. Using trial and error, we found that the best results were achieved by classifying only the first paragraph of the article text (using Wikipedia API [5]), which also saved us time and resources.

3.2 C- Popular Pages

Popular pages are the Wikipedia articles with the most absolute-count views over the last 24 consecutive hours. Note that a large absolute number of views does not necessarily mean the page is trending, as many articles regularly receive high amounts of interest and views, such as the 'bible' article, which consistently has more than 100,000 views per day, and so many times appears as 'popular' even though it is not trending.

3.3 D – Trending Categories

We look at all trending pages per category (recall that a page must have at least a 2x traffic increase rate to be considered trending). We then sum the trending pages by category (in English). We then display the trending categories, defined as all categories which include at least 5 trending articles. See Image 4 below.

3.4 E – Top Trending Pages by Category

(Marked as "E" on Appendix, Image 1.1) As the name suggests, this section displays the trends as described above, presenting the top 5 for each category. This allows for a quick understanding of the current trends within a category.

3.5 Results page – Pageview Traffic Data

After a search for one or more Wikipedia titles is performed, the search results page is displayed. On this page we see a chart with the pageviews of the selected Wikipedia entries plotted on the same graph. The default view will show pageview information for the last few days, but we can easily change the date range using the bar at the bottom, below the graph, or the date inputs at the top right (see Appendix, Image 10). Note that we can select a specific date range by typing a date in the top corner, while the bottom time-range bar enables faster albeit less accurate time selection. If the user wants to drill down and zoom in on a part of the chart, they can click and drag their mouse over the chart as it selects that section and zooms in. As noted above, the information is continuously updated, on an hourly basis, and recorded going back to 2008, so that over 13 years of data are currently accessible via the results page.

Our chart's interface is generated and plotted via a javascript library called HighStock [3] which is a subversion of HighCharts [4]. We used Highcharts, as it is very stable and popular, and is also used by Facebook and Twitter. We specifically used HighStock, as it (1) provides a dedicated solution for TimeSeries charts and it is able to handle the loading of many data points, which is important for us because we have information covering thirteen-plus years. We also want to display multiple entries at once, (2) and, as it also supports asynchronous loading, it is possible to load the data without letting the user see a blank page until the data is available. Also, (3) the HighStock/HighCharts license is free to use for academic purposes.

3.6 Exporting the Data

To allow users to use the data they see on the WikiShark results page, we provide a number of ways to export the data via in several standard formats. The simplest way to export is directly from the Pageview traffic data results page. At the bottom of each traffic data results page there is a 'Download' link that allows users to save

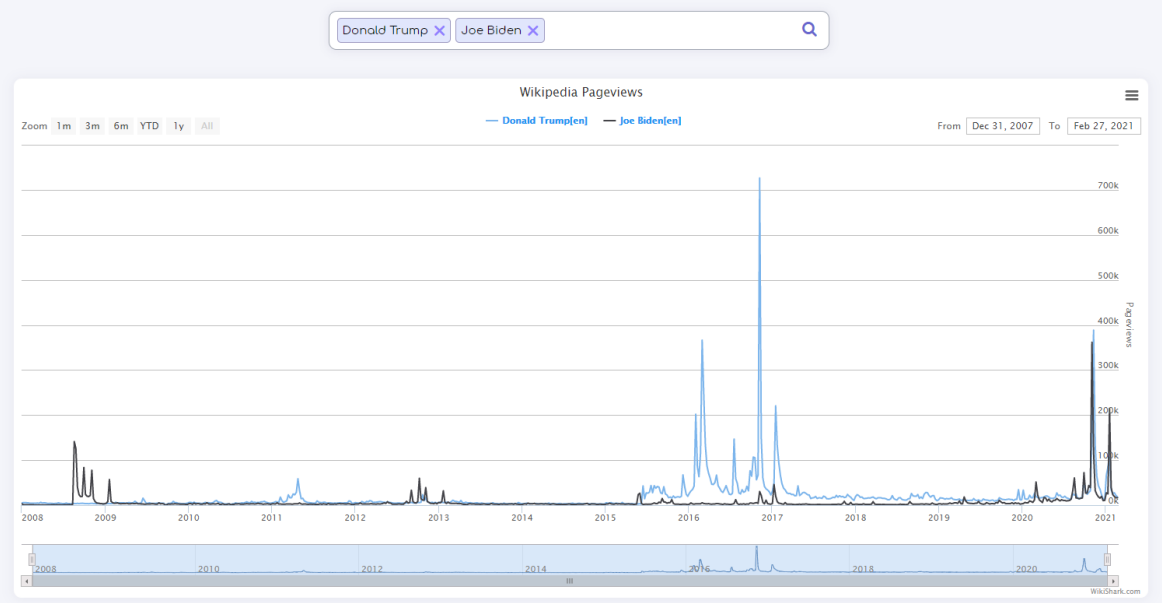

Image 5: Results page – with Pageview traffic data. In this example we compare Donald Trump and Joe Biden.

the data in popular formats such as JSON and CSV (see Appendix, Image 11). The data can be exported for one or more search entries.

Image export – Users can also export the visual image of the graph itself in common image formats (SVG, PNG, PDF, JPG). This can be done using the menu button on the upper right corner of the graph and then clicking on 'Download Image' (see Appendix, Image 12).

4 EXTERNAL TOOLS

4.1 WikiShark's Chrome extension

WikiShark's Chrome extension [10] creates a button on any Wikipedia article, allowing users to directly view WikiShark stats on the WikiShark site (See Image 6 below). Moreover, it also allows a quick traffic preview by pressing the extension icon (see Appendix, Image 6).

4.2 WikiShark's Twitter page – auto-tweeting Trends page

Trends are automatically tweeted on our new Twitter page (https://twitter.com/wikishark_) every few hours using the Twitter Developer API [7] (see Appendix, Image 8) in order to bring more attention to Wikipedia's global trends and to open the door for new discussion and research.

5 RELATED WORK

When compared with other tools (e.g. Wikitrends [12], Wikipulse [14], Pageviews-Analysis [14]), WikiShark manifests three main advantages. The first is the historical perspective: WikiShark collects, aggregates and analyzes information dating back to 2008.

The second advantage has to do with the interaction of traffic and trends. Rather than displaying trends and pageviews separately, WikiShark is the only tool to combine the two, presenting both traffic and trends in the same system. For example, adjacent to each trend, we display the traffic graph over the trending period, which supplements data for each trend with an important layer of information. Clicking on the trend opens the full entry graph – in the same interface and without having to switch between systems.

The third advantage is WikiShark's ease of use. Much of the information (such as trends) appears directly on the home page, without one having to perform a search or, indeed, interact with the system in any way. Additionally, searches can be performed in any language.

A direct comparison with this popular tool Pageviews-Analysis [14] (which seems to be one the most professional pageviews tools) reveals that it includes data dating back to just 2015, and requires the pre-selection of a specific language for the Wikipedia page in question, such a English Wikipedia. This places restrictions on the search and makes it difficult to compare entries in different languages (which are on different Wikipedias) on the same graph. In WikiShark, comparing such articles is straightforward, and all data analyzed date back to 2008.

Comparing WikiShark to tools such as Wikitrends [12] or Wikipulse [13] reveals that these tools are non-intuitive and require multiple interactions to arrive at results. For example, Wikitrends requires a visual scan across the network graph and does not directly show results, whereas Wikipulse, while admirably designed, does not display the trends in a user-friendly fashion and requires searches to be performed via the search bar.

6 FUTURE PLANS

Our future plans include:

1. Full support for all the languages (currently most of the features, such as Trends, are available only in English).

Image 6: WikiShark's Chrome extension creates a button on any Wikipedia article that allows users to directly view WikiShark stats on the WikiShark site.

2. History of trends. Currently, the trends change daily and we would like to offer 'Search trends' by date and category, and also to view more than five per day or per category.
3. Better API – currently we offer a good but basic API. We plan to offer 'Search by dates', enabling some basic manipulation, such as normalization directly from the API.
4. Classification – Instead of using Textrazor, we will build and use our own classifier.

7 CONCLUSION

In this article we introduce WikiShark, an online tool that allows researchers to analyze Wikipedia traffic, and to query and view public interest data since 2008 instantly and simply, without setting up expensive systems. This opens up an exciting new opportunity for sharing trends, which can in turn provide the basis for new traffic-based research (Lampos et al.,2021; Yasseri et al. 2016). Our hope is that WikiShark can be leveraged to expose novel insights with respect to Wikipedia and how it is used and searched worldwide.

ACKNOWLEDGMENTS

We wish to thank Amazon, which granted WikiShark $20,000 worth of AWS hosting.

REFERENCES

[1] Wikipedia's Pageview dumps, https://dumps.wikimedia.org/other/pageview_complete/
[2] Yasseri, T. and Bright, J., 2016. Wikipedia traffic data and electoral prediction: towards theoretically informed models. EPJ Data Science, 5 (1), pp.1-15. https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0083-3
[3] HighCharts – Highcharts: Interactive JavaScript charts for web pages, https://www.highcharts.com
[4] High-Stock – Time Series charts for webpages, https://www.highcharts.com/demo/stock
[5] Wikipedia API., https://www.mediawiki.org/wiki/API:Main_page
[6] jQuery Sparklink. A web plugin which generates small inline charts directly in the browser, using data supplied either inline in the HTML or via javascript, https://omnipotent.net/jquery.sparkline,
[7] Twitter Developer API, https://developer.twitter.com/en/docs/twitter-api
[8] TextRazor – Natural Language Processing and Artificial Intelligence techniques to parse, analyze and extract semantic metadata from your content, https://www.textrazor.com/
[9] Internet Advertising Bureau Content Taxonomy v2, https://www.iab.com/guidelines/content-taxonomy/
[10] WikiShark Chrome Extension https://chrome.google.com/webstore/detail/wikishark-Wikipedia-stati/jmbdjjmajaloijoimjbheaohdjfednge
[11] Lampos, Vasileios, Maimuna S. Majumder, Elad Yom-Tov, Michael Edelstein, Simon Moura, Yohhei Hamada, Molebogeng X. Rangaka, Rachel A. McKendry, and Ingemar J. Cox. "Tracking COVID-19 using online search." NPJ digital medicine 4, no. 1 (2021): 1-11. https://www.nature.com/articles/s41746-021-00384-w
[12] Wikitrends - Graph Visualization of Wikipedia, https://wiki-insights.epfl.ch/wikitrends/
[13] Wikipulse, - Wikipedia popularity trends, https://wikipulse.com/
[14] Pageviews Analysis, https://pageviews.toolforge.org/Adya, Paramvir Bahl, Jitendra Padhye, Alec Wolman, and Lidong Zhou. 2004. A multi-radio unification protocol for IEEE

A APPENDICES

Image 1: WikiShark analysis platform's homepage (www.wikishark.com).

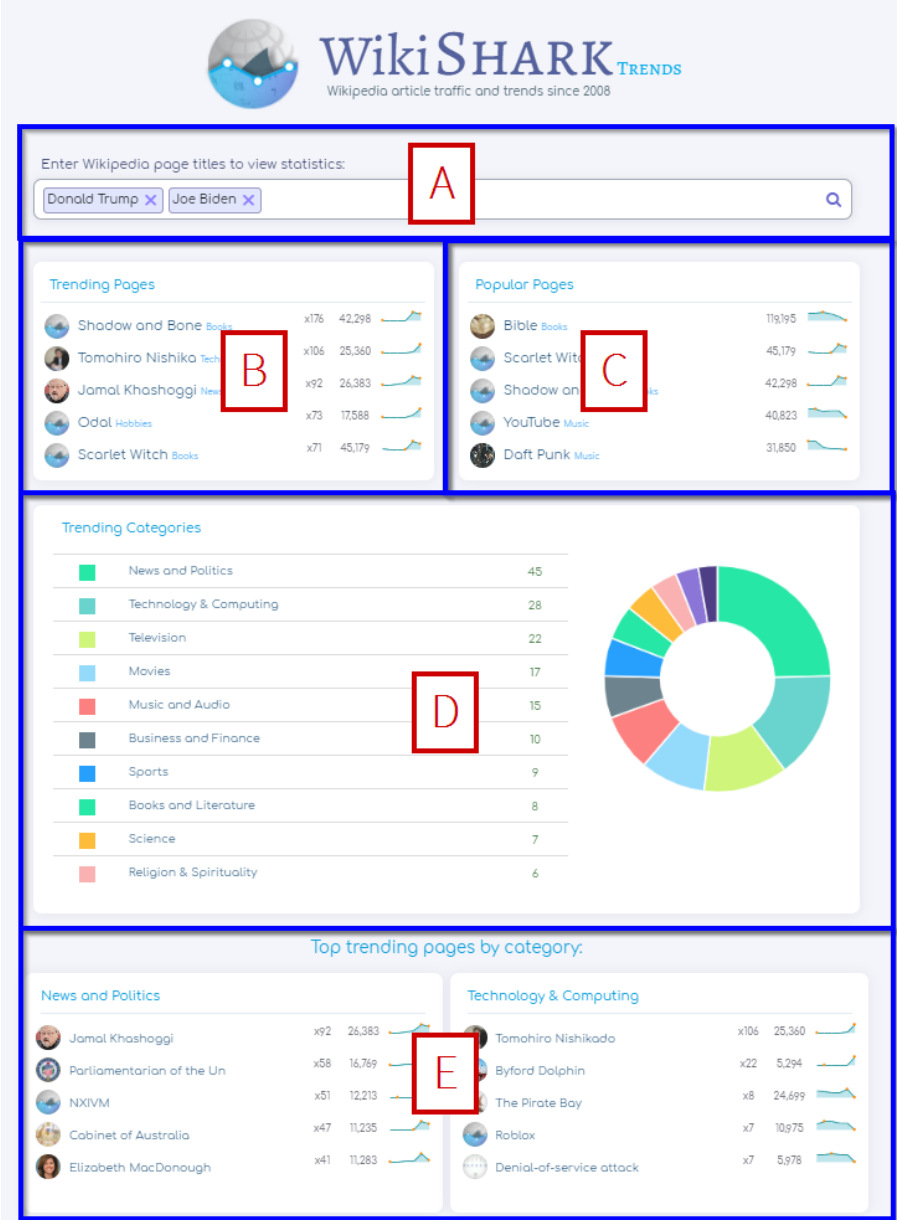

Image 1.1: WikiShark's homepage – main sections.

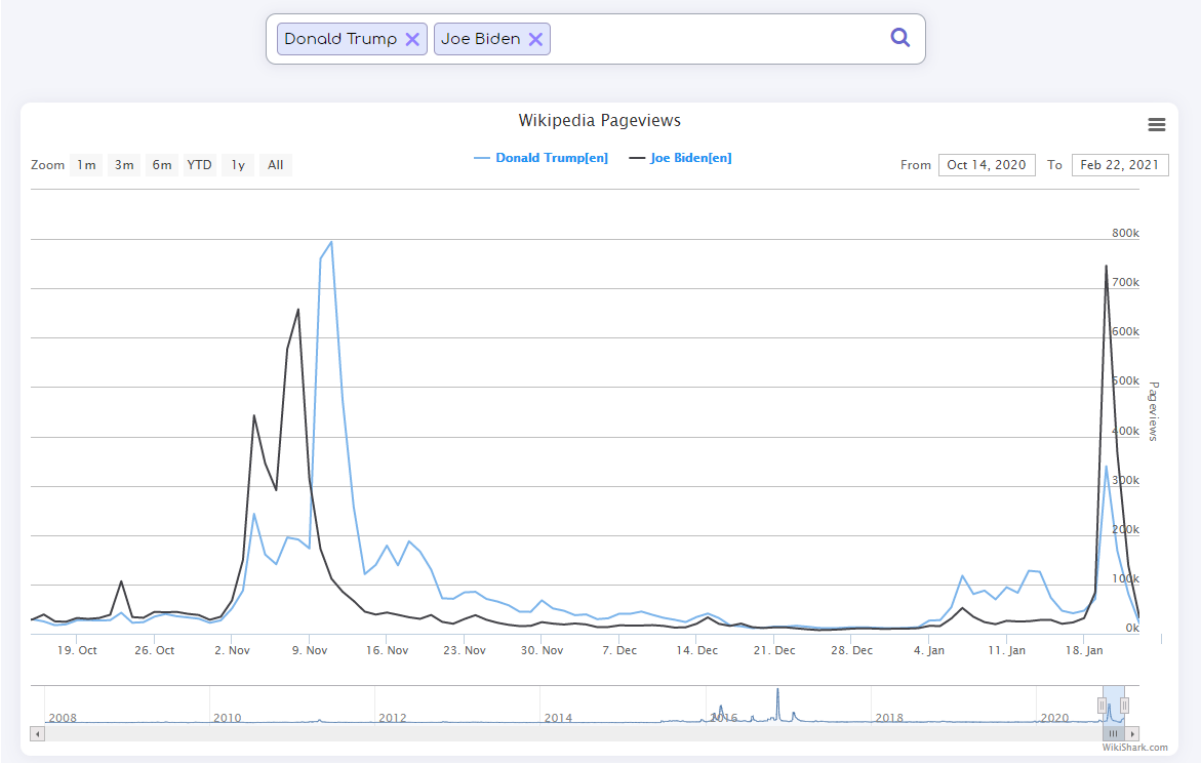

Image 2: User Interface for comparing two articles. In this example we compare Donald Trump and Joe Biden.

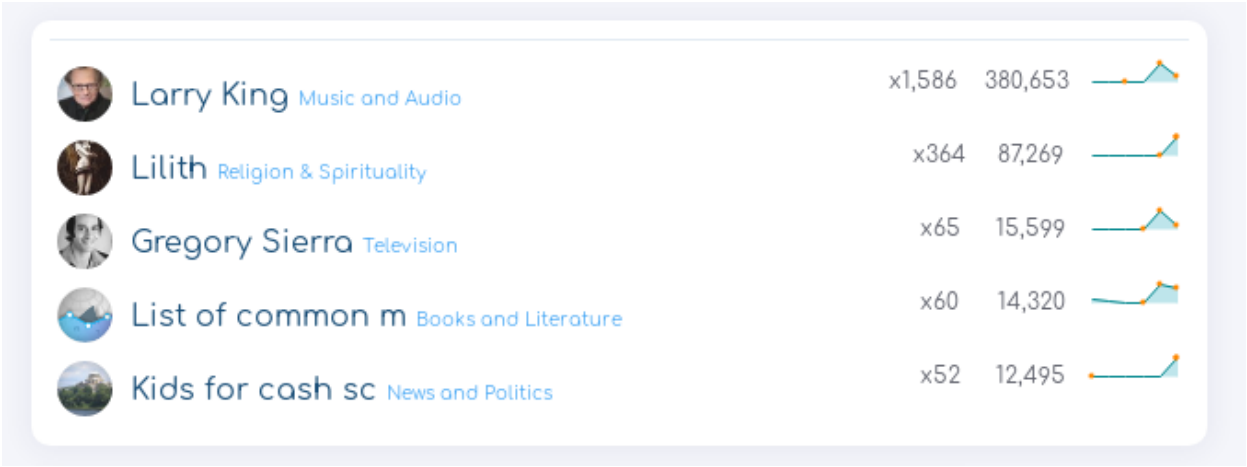

Image 3: Current trending pages. For each trending page, we show (i) its name; (ii) thumbnail and category; (iii) traffic increase rate; (iv) a small preview chart for traffic data; and (v) number of pageviews over the past 24 hours.

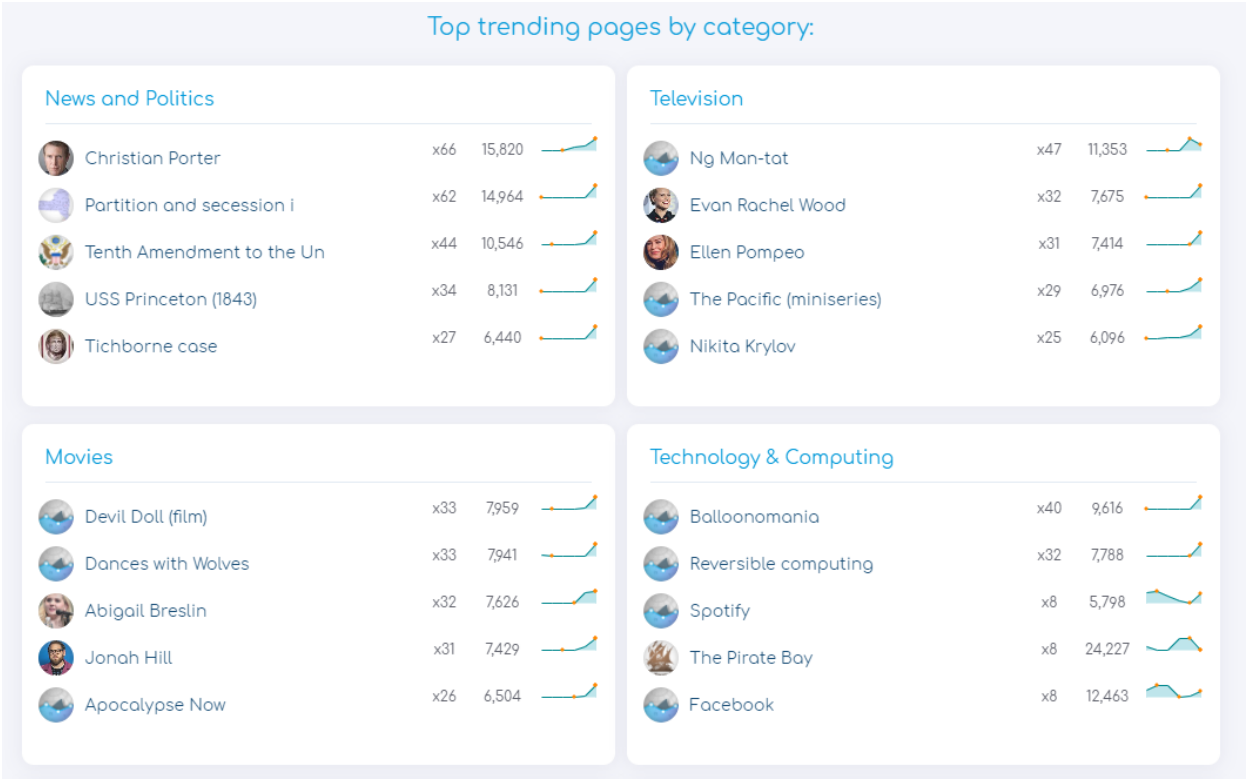

Image 4: Current trending pages by category.

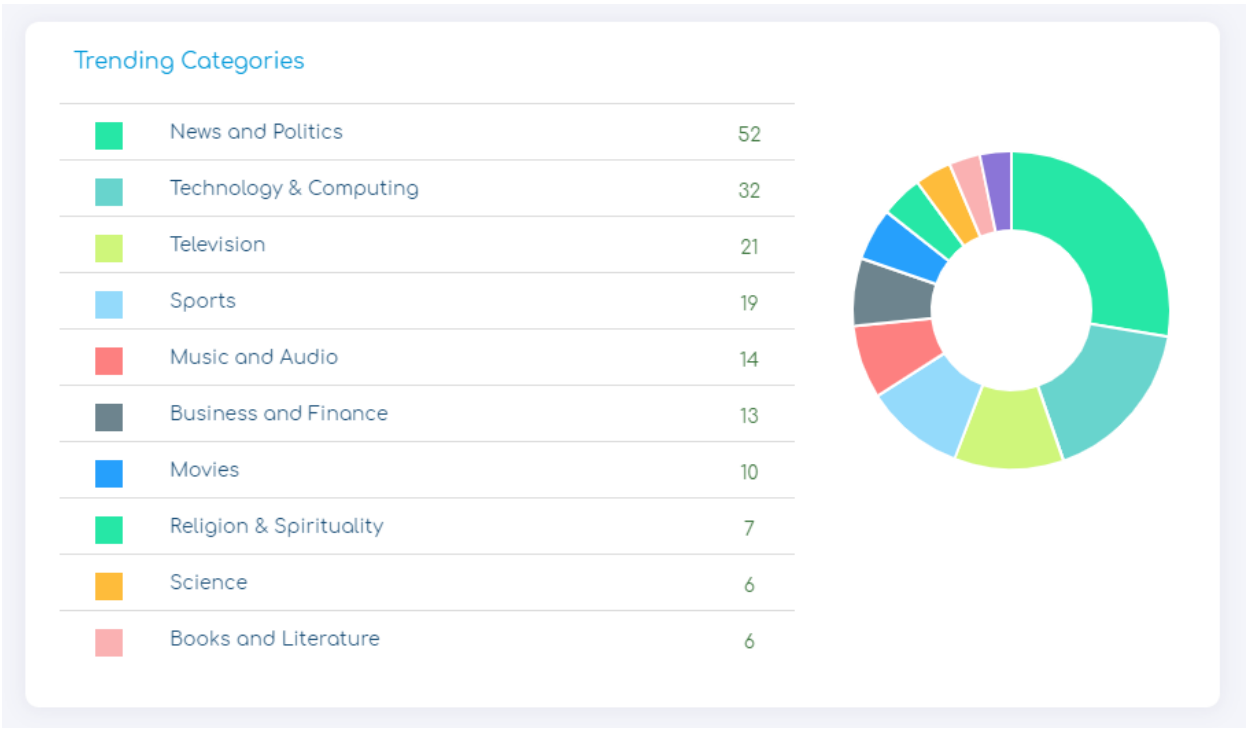

Image 5: Current top trending categories.

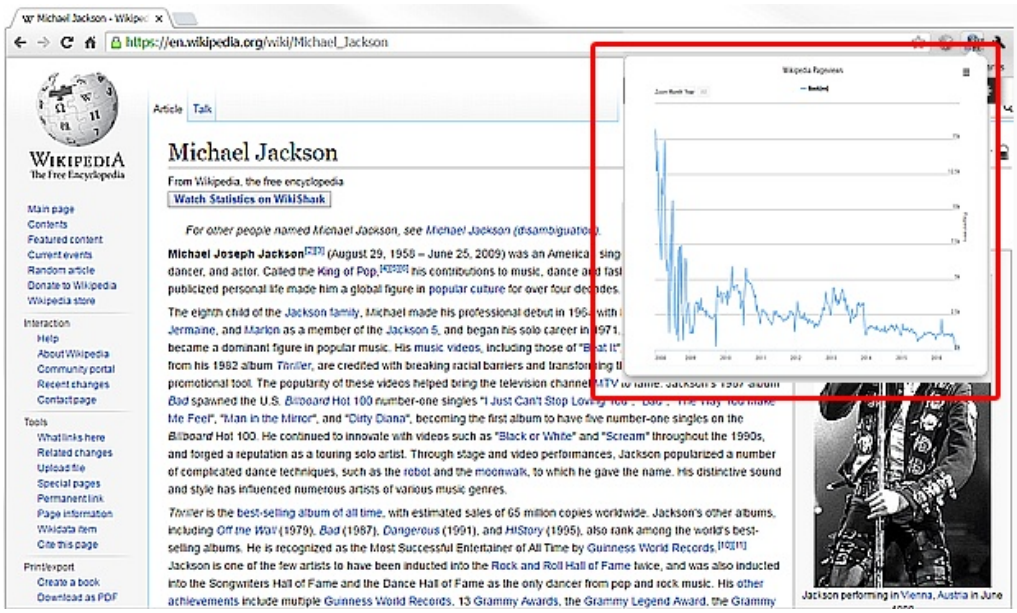

Image 6: WikiShark's Chrome extension allows quick traffic previews by pressing the extension icon.

Image 7: WikiShark's Chrome extension creates a button on any Wikipedia article, which allows users to directly view WikiShark stats on the WikiShark site.

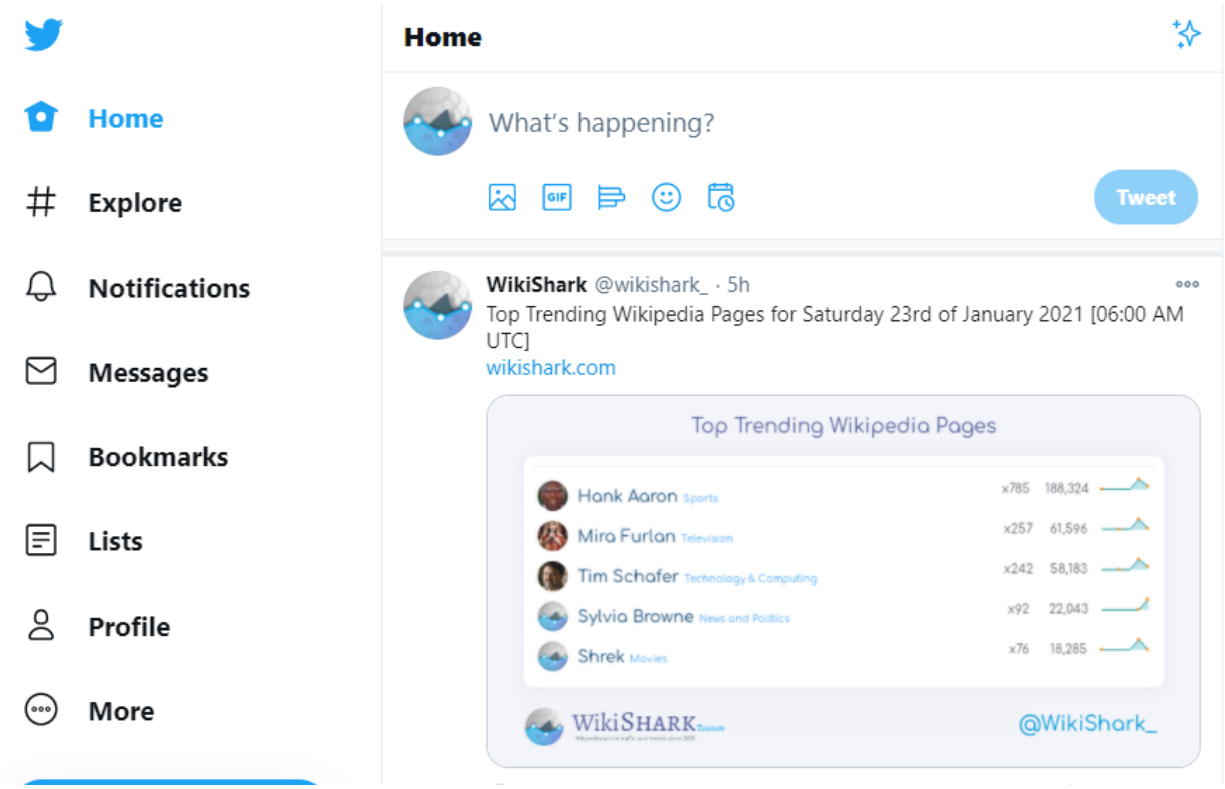

Image 8: WikiShark's Twitter page, auto-tweeting trending pages.

Image 9: Auto-complete. In this image, we pressed "A" and Wikipedia titles starting with 'a' appeared.

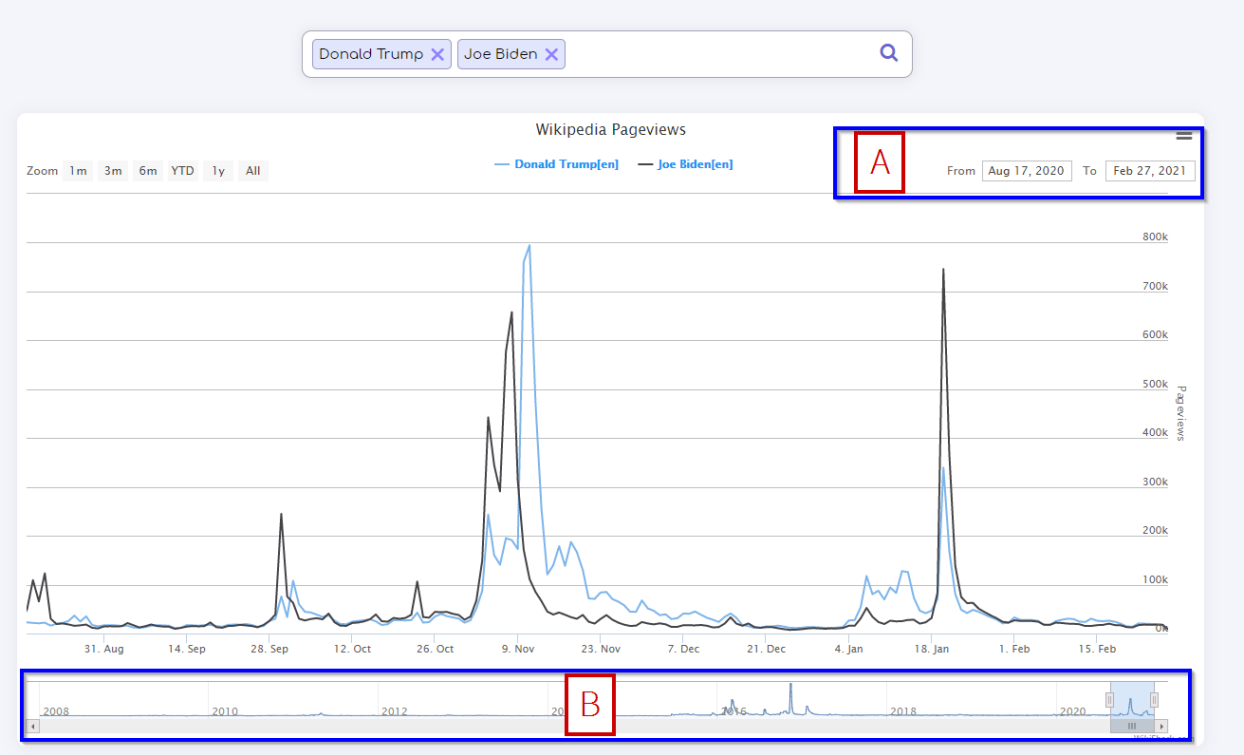

Image 10: Changing the time frame on the results page.

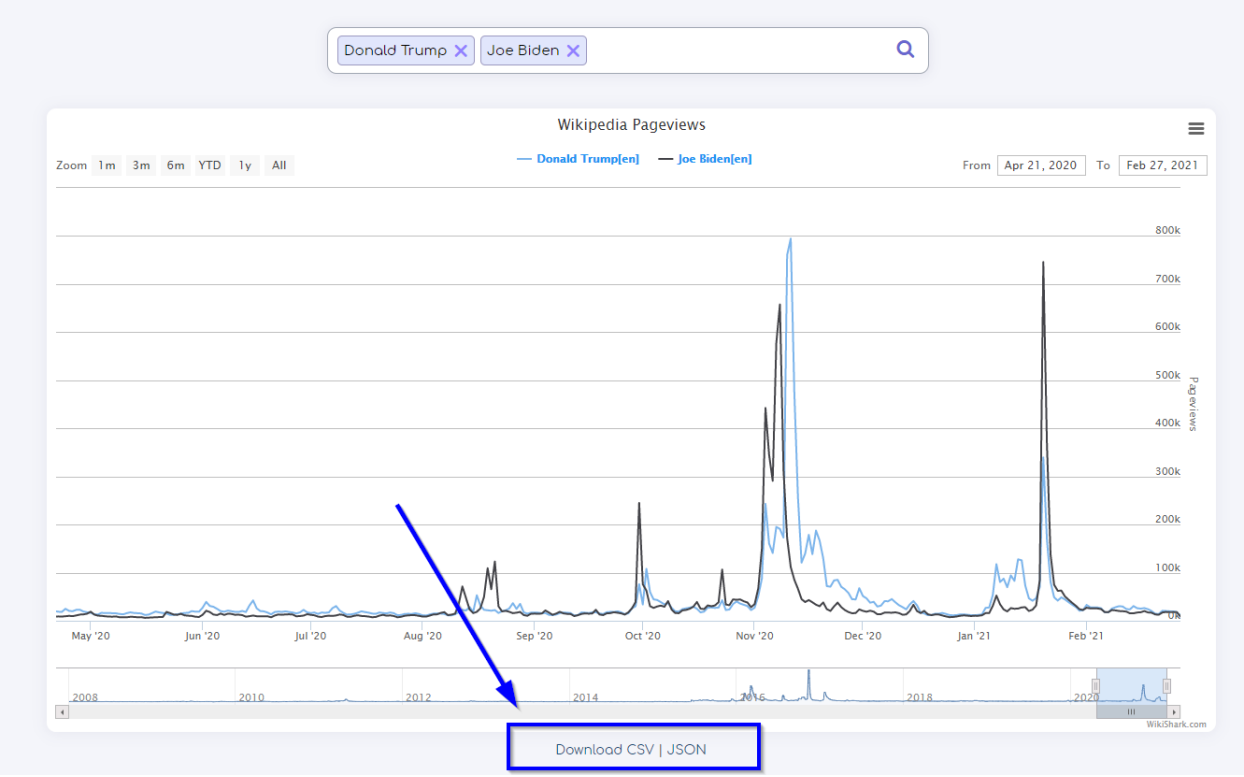

Image 11: Download data in CSV or JSON formats.

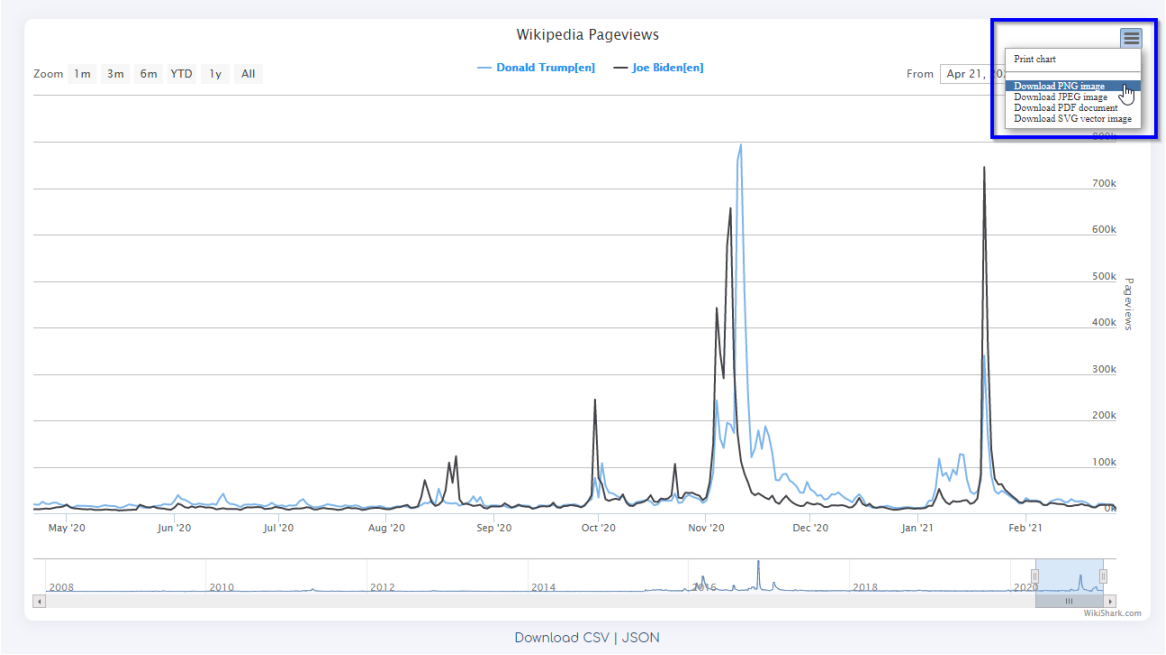

Image 12: Saving the chart as an image (JPG, PNG, PDF, SVG).