

A decorative graphic on the left side of the slide, consisting of a network of white lines and small circles on a dark blue background, resembling a circuit board or a data network.

P1 – WIKIPEDIA DATASET ANALYSIS

MICHAEL SPLAVER

HIGHEST TRAFFICKED ARTICLE ON OCTOBER 20


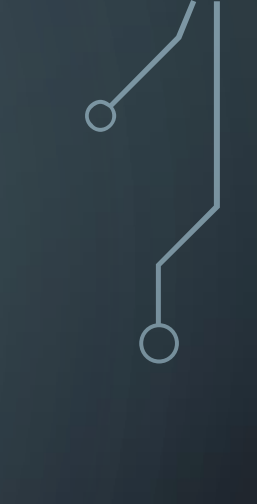
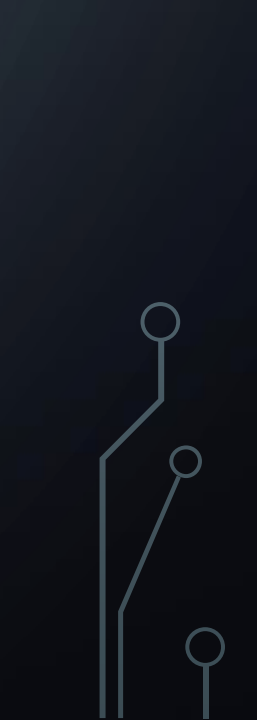
Page Title	Page Views
Jeffrey Toobin	321,459
C. Rajagopalachari	210,558
The Haunting of Bly Manor	185,139
Robert Redford	178,779
Jeff Bridges	159,163

- Jeffery Toobin's article was the most trafficked on Oct. 20th
- Page was viewed 321,459 times
- Excludes insignificant articles like the Main Page



STEPS TAKEN

HIGHEST TRAFFICKED ARTICLE ON OCTOBER 20

- Utilized Pageviews hourly data from Oct 20th, 2020
 - Grouped all the pages with English domain codes together
 - Totaled all the page views up
 - Sorted by the most views pages
- 
- 
- 

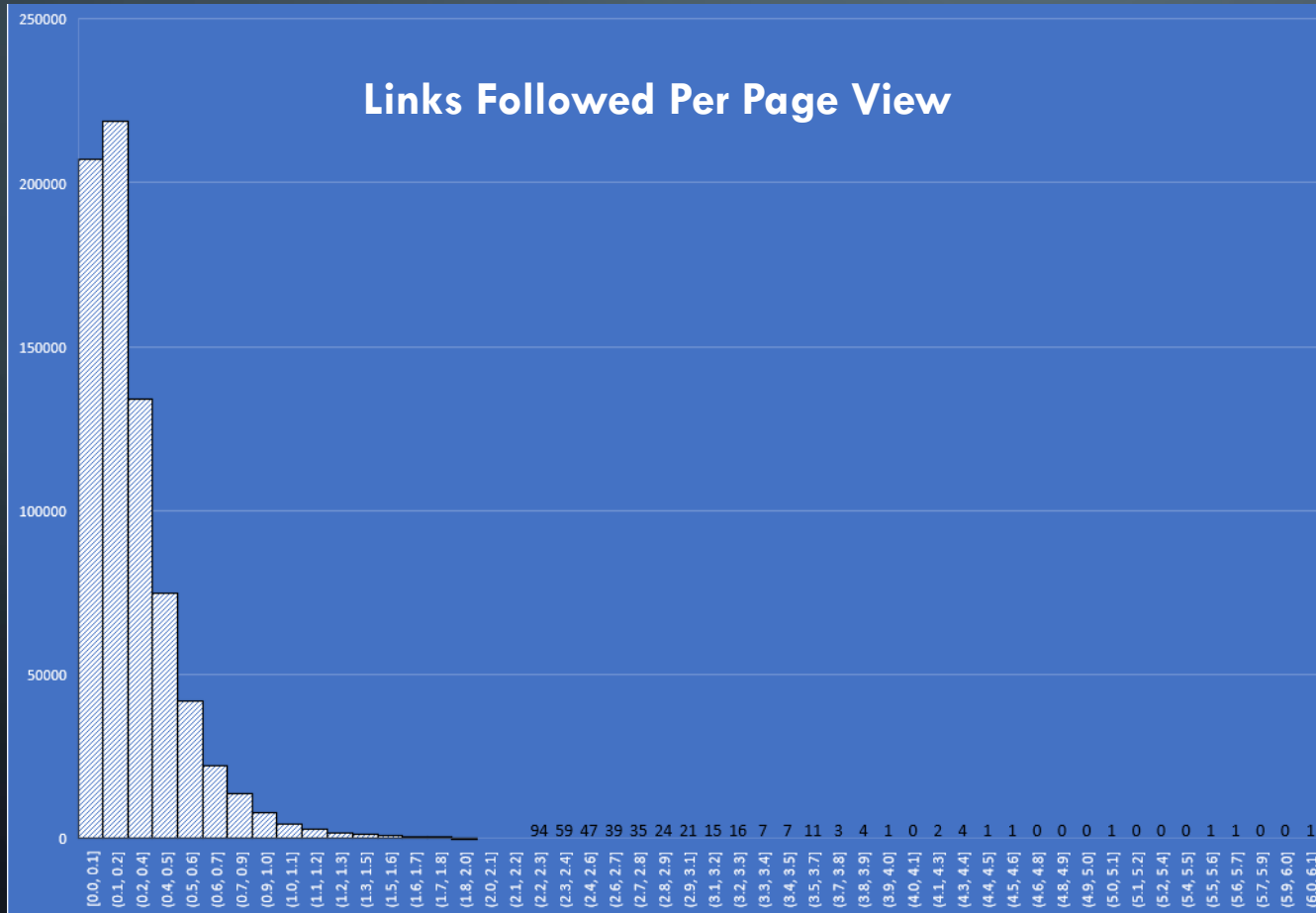
HIGHEST CLICKTHROUGH TO ANOTHER WIKI

Page Title	Links Followed Per Page View
NetScout Systems	6.1
List of extinct in the wild animals	5.65
List of Panathinaikos_F.C. players	5.61
Boys (disambiguation)	5.08

- This data is for the month of August
- Only includes Pages with more than 1000 page views in August
- The NetScout Systems article had the highest average links followed per page view with 6.1

RESULTS EXPLANATION

HIGHEST CLICKTHROUGH TO ANOTHER WIKI



- Values can be greater than 1 because multiple links can be followed in a single page view
- The highest links/page view are extremely rare
- Most pages are between 0.0 and 0.9

STEPS TAKEN

HIGHEST CLICKTHROUGH TO ANOTHER WIKI

- Utilized Pageviews and Clickstream monthly data from August
- Replicated previous steps to produce page views for the month of August
- Grouped all the clickstream pages on the referrer with type “link”
- Totaled all the occurrences of clickstream referrer following a link
- Calculated average links followed per page view on each page
- Sorted by highest links followed per page view

HIGHEST PERCENT CLICKTHROUGH SEQUENCE



- This data is for the month of August
- Only searched down the single highest percentage path on each article
- Possible another path results in a higher overall percentage (unlikely though)

STEPS TAKEN

HIGHEST PERCENT CLICKTHROUGH SEQUENCE

- Utilized Pageviews and Clickstream monthly data from August
- Replicated previous steps to produce average links followed per page view
- Sorted by highest links followed per page view
- Repeated for the highest percentage links followed on each page

MORE POPULAR ARTICLES BY REGION

PAGE: Nate Bjorkgren	
Region	Views during region hours of Total Views
US	99.94%
UK	69.57%
AUS	21.05%

PAGE: Frankenstein_Castle	
Region	Views during region hours of Total Views
UK	98.87%
US	19.98%
AUS	Less than 1000 page views.

- Data is from Oct 20th
- Assumes most active internet hours are 8am-7pm local time
- Percentages not calculated on pages with less than 1000 views
- Tried my best to find representative articles out of the top 20 most popular

MORE POPULAR ARTICLES BY REGION CONTINUED

PAGE: Malissa Stribling	
Region	Views during region hours of Total Views
AUS	91.69%
UK	Less than 1000 page views.
US	Less than 1000 page views.

- Really difficult to find popular Australian articles because of significant overlap with late-night US internet hours
- Data can be skewed by big event taking place causing an influx of article traffic on specific hours

STEPS TAKEN

MORE POPULAR ARTICLES BY REGION

- Utilized Pageviews hourly data from Oct 20th, 2020
- Replicated previous steps to produce page views for each selected time zone (US – 14:00 -1:00 UTC, UK – 8:00 – 19:00 UTC, AUS – 21:00 – 8:00 UTC)
- Calculated page views for each time zone selection over the total page views for that day
- Sorted by highest percentage of views of total page views
- Repeated for each time zone selection (US, UK, AUS)

AVERAGE VIEWS ON A VANDALIZED PAGE

Average days before page is reverted

5.24 days

Average views per day on each page

14.78 views/day

**Average views
before vandalized edit is reversed**

77.45 views

- Data is from August
- Assuming all reverts are from vandalized pages
- Average views per day includes all articles (including some with no page views)

STEPS TAKEN

AVERAGE VIEWS ON A VANDALIZED PAGE

- Utilized wiki-history data up to Oct 2020
- Averaged the total seconds between the articles being reverted
- Averaged the number of people who will view an average page on a day
- Calculated Average number of views before article is reverted

USER WITH THE MOST REVISIONS

Username	Revisions
WP 1.0 bot	7,379,526
ClueBot NG	5,545,179
AnomieBOT	4,470,964
InternetArchiveBot	4,404,937
Ser Amantio di Nicolao	3,723,296

- Uses data up to October 2020
- Bot accounts currently take up the top 4 spots for revisions made
- “Ser Amantio di Nicolao” famously Steven Pruitt is the only human in the top 5

STEPS TAKEN

USER WITH THE MOST REVISIONS

- Utilized wiki-history data up to Oct 2020
- Found the records with the maximum number of revisions for each user
- Sorted by users with most revisions



THANK YOU!

All slides and hive queries are available at my github repo:

<https://github.com/Michael-Splaver/p1-wikipedia-analysis>

