

# Project 1: Wikipedia Data Analysis

Liam Hood

# Most traffic on October 20th

- Limited to English articles
- Used all hour files from October 20th
- Did not group subsections of pages together

| Site                      | Views   |
|---------------------------|---------|
| Main Page                 | 5963378 |
| Search                    | 1477350 |
| -                         | 544735  |
| Jeffrey_Toobin            | 322001  |
| C._Rajagopalachari        | 210564  |
| The_Haunting_of_bly_Manor | 185144  |

# Largest fraction of viewers referred in September

- Sorted clickstream by reference type being link and matched referrer to site from pageview from September
  - Percent was clicks from referrer over views from pageview for the month
  - Did two mapreduces on pageview month
    - 5 days at a time first then those 5 days of data together
- /r/ has 64 links followed but only 1 view giving it a 6400% click rate
- Limiting to over 100 views
  - /pol/ has 1413% click rate with 490 views
- Limiting to over 1000 views
  - NetScout\_Systems has 598% with 1708 views
- ~1000 views is where the results start to make sense
  - Most results here are lists in which people are likely to follow many links

| clickstreampercent.referrer                    | clickstreampercent.occurences | clickstreampercent.views | clickstreampercent.perc |
|--|-------------------------------|--------------------------|-------------------------|
| NetScout_Systems                               | 10222                         | 1708                     | 598.4777517564403       |
| List_of_hāfu_people                            | 18536                         | 3127                     | 592.7726255196674       |
| Boys_(disambiguation)                          | 47594                         | 8070                     | 589.7645600991326       |
| List_of_extinct_in_the_wild_animals            | 16576                         | 3248                     | 510.3448275862069       |
| List_of_prototype_World_War_II_combat_vehicles | 13633                         | 3117                     | 437.3756817452679       |
| Æscwine_of_Wessex                              | 4481                          | 1037                     | 432.1118611378978       |
| List_of_controversial_album_art                | 47953                         | 11271                    | 425.45470676958564      |

# Series of articles from Hotel California

- Searched clickstream for referrer and order by percent descending
  - Percent followed for a particular path is link followed on the path divided by total clicks on the referrer page
    - Not a true percent of clicks relative to page views but with the same referrer page each time it is proportional
  - Hotel California (16.0%)-> Hotel\_California\_(Eagles\_album) (18.4%)-> The\_Long\_Run\_(album) (24.3%)-> Eagles\_Live (54.3%)-> Eagles\_Greatest\_Hits,\_Vol.\_2 (69.3%)->The\_Very\_Best\_of\_the\_Eagles (78.9%)-> Hell\_Freezes\_Over (16.7%)->Selected\_Works:\_1972-1999
  - (81.9%)->The\_Very\_Best\_Of\_(Eagles\_album) (43.9%)->Eagles\_(box\_set)
- 
- If every user followed some path, as in leaving a page by clicking on a new link, then there would be ~1 of the original viewers left by the end
    - 2222 clicked from Hotel California to Hotel California (Eagles Album) and 0.08% followed the full path

# Popularity by country (UK, USA, AUS)

- Only looking at the first week of September 2020
- Using time to infer most likely country
- I used 9am to 5pm UTC to get 8 hours time span to limit overlap

Missouri is centroid of us population UTC -5 (14-22)

London UTC +1 (8-16) in summer (until last sunday in October)

Australia is ~UTC +10(0-7) where most people are

# Popularity by country (UK, USA, AUS)

- Results are normalized by the total views in each time category
- Then looked at percent difference
- Main\_Page more popular in UK
- US political figures are more popular in US
  - This may be most clear due to people looking at political articles during the day
- Entertainment articles tend to be more popular in Australia
  - Makes me think that this is also helped by US users looking at entertainment articles in the afternoon and evening

| pageviewcountry.site                  | pageviewcountry.views | pageviewcountry.gbr_rel2_usa | pageviewcountry.aus_rel2_usa |
|---------------------------------------|-----------------------|------------------------------|------------------------------|
| Main_Page                             | 15789153              | 10.566777074846291           | -7.829884334439259           |
| Special:Search                        | 4063522               | 8.780721560505352            | -22.968604010378847          |
| -                                     | 1695349               | 12.973514357634357           | -13.752376374894233          |
| Tenet_(film)                          | 642594                | -6.223729394066322           | 13.618372132044257           |
| Chadwick_Boseman                      | 578915                | 2.003013674810487            | 29.2101649814229             |
| Mulan_(2020_film)                     | 539431                | -8.683189393918157           | 14.9239822035991             |
| Afghan_Girl                           | 427888                | -1463.112622514609           | -366.13940178872997          |
| Cobra_Kai                             | 403248                | -17.23512830900356           | 27.016431149298803           |
| The_Boys_(2019_TV_series)             | 395666                | -14.099647821136898          | 3.25065388434368             |
| Deaths_in_2020                        | 349344                | 0.242857989774829            | -10.924428600901734          |
| Bible                                 | 323103                | -6.379190791118862           | 20.30794578864198            |
| Robert_F._Kennedy_Jr                  | 318871                | -20.99961665874185           | -19.81482808430435           |
| I'm_Thinking_of_Ending_Things_(film)  | 263022                | -27.226634229209264          | -2.190727305431809           |
| The_1619_Project                      | 236800                | -51.14563575473039           | -112.88187993900938          |
| Labor_Day                             | 216587                | -4.964719557919316           | -115.85942410604353          |
| William_Zabka                         | 205698                | -25.317227718996513          | 34.93613258030882            |
| XXXX                                  | 205420                | 0.5477066340282362           | -41.80408947755288           |
| Avengers_(2020_video_game)            | 204266                | -12.850029227032358          | 2.820113059544867            |
| QAnon                                 | 198932                | -11.262155025224638          | 2.0015198429608554           |
| Ralph_Macchio                         | 197256                | -24.998532211579086          | 29.884799025899866           |
| Joe_Biden                             | 195856                | -29.480579213028488          | -2.9130572743843524          |
| Novichok_agent                        | 195399                | -9.382358993201525           | -142.15537635910394          |
| Raised_by_Wolves_(American_TV_series) | 187137                | -8.984259715287301           | 8.524745468647197            |

# How many Users see a vandalized page

- Filtered history by comments containing '%vandal%' and positive seconds since revision
- Match views to revisions
  - Sum of time since last revision is taken as time the vandalism was visible
    - Could be wrong if vandalism wasn't last revision
  - Assume that the users are evenly distributed in the month
  - Number of users on the page in the time in the time span is approximate number that see vandalism of a page
- $1005 \approx \text{avg}(\text{number\_views} * (\text{sum}(\text{seconds\_since\_last\_revision}) / 2592000))$



# How People use Wikipedia

- Grouped clickstream by reference type and sorted out Main\_Page as the reference
- $3.25E7$  enter through the main page then go to an article
- $4.62E9$  enter through a direct search for a page
- $1.98E9$  follow an internal link to a page
- $2.88E7$  do something else that gets an 'other' in clickstream

| wikiuse.reftype | wikiuse.totaloccurences |
|-----------------|-------------------------|
| other           | 58927330                |
| external        | 4621586140              |
| link            | 1979605165              |

| wikiusemain.reftype | wikiusemain.totaloccurences |
|---------------------|-----------------------------|
| other               | 30137193                    |
| link                | 2379287                     |

# GitHub

[https://github.com/LiamAHood/Project\\_1.git](https://github.com/LiamAHood/Project_1.git)

<https://www.hallaminternet.com/google-analytics-hour-of-day-day-of-week-reports/>

[https://en.wikipedia.org/wiki/Mean\\_center\\_of\\_the\\_United\\_States\\_population](https://en.wikipedia.org/wiki/Mean_center_of_the_United_States_population)

[https://en.wikipedia.org/wiki/Time\\_in\\_the\\_United\\_Kingdom#:~:text=The%20United%20Kingdom%20uses%20Greenwich,UTC%2B01%3A00](https://en.wikipedia.org/wiki/Time_in_the_United_Kingdom#:~:text=The%20United%20Kingdom%20uses%20Greenwich,UTC%2B01%3A00)

[https://en.wikipedia.org/wiki/Time\\_in\\_Australia](https://en.wikipedia.org/wiki/Time_in_Australia)