Wikipedia Dataset Analysis

Eunice Lee

Link to GitHub

Q1. Which English wikipedia article got the most traffic on January 20, 2021?

- Downloaded 1/20/2021 pageviews data
  - Used 6 random files to represent the day
- Created PAGEVIEW table to load downloaded data file
- Used *domain_code* as a partition column
  - Partitioned by **'en'** and **'en.m'** and inserted to PAGEVIEW table

Q1. Result

```
SELECT page_title,
SUM(count_views) AS total_pageviews
FROM pageview_en
GROUP BY page_title
ORDER BY total_pageviews DESC
LIMIT 10;
```

page_title	total_pageviews
Main_Page	4,937,737
Joe_Biden	**2,170,628**
Amanda_Gorman	1,584,305
Kamala_Harris	1,542,483
Special:Search	1,074,654
Donal_Trump	902,545
Beau_Biden	652,683
Jill_Biden	587,221
Doug_Emhoff	533,895
President_of_the_United_States	484,480

Q2. What English wikipedia article has the largest fraction of its readers follow an internal link to another wikipedia article?

- December, 2020 pageview and clickstream data was used
  - Used 5 random files to represent the month of December
- Used ***type = 'link'*** as a partition column in clickstream table
- Fraction: link_clicked / page_views

```
SELECT pageview.page_title, linkview.linkcount,,
pageview.totalcount AS total,
ROUND(linkview.linkcount/pageview.totalcount, 5) AS
fraction
FROM q2_pageview_final AS pageview
INNER JOIN q2_link_final AS linkview
ON pageview.page_title = linkview.prev
WHERE pageview.totalcount > 100,000 //Disregarded else
ORDER BY fraction DESC;
```

Q2. Result (pageview > 100,000)

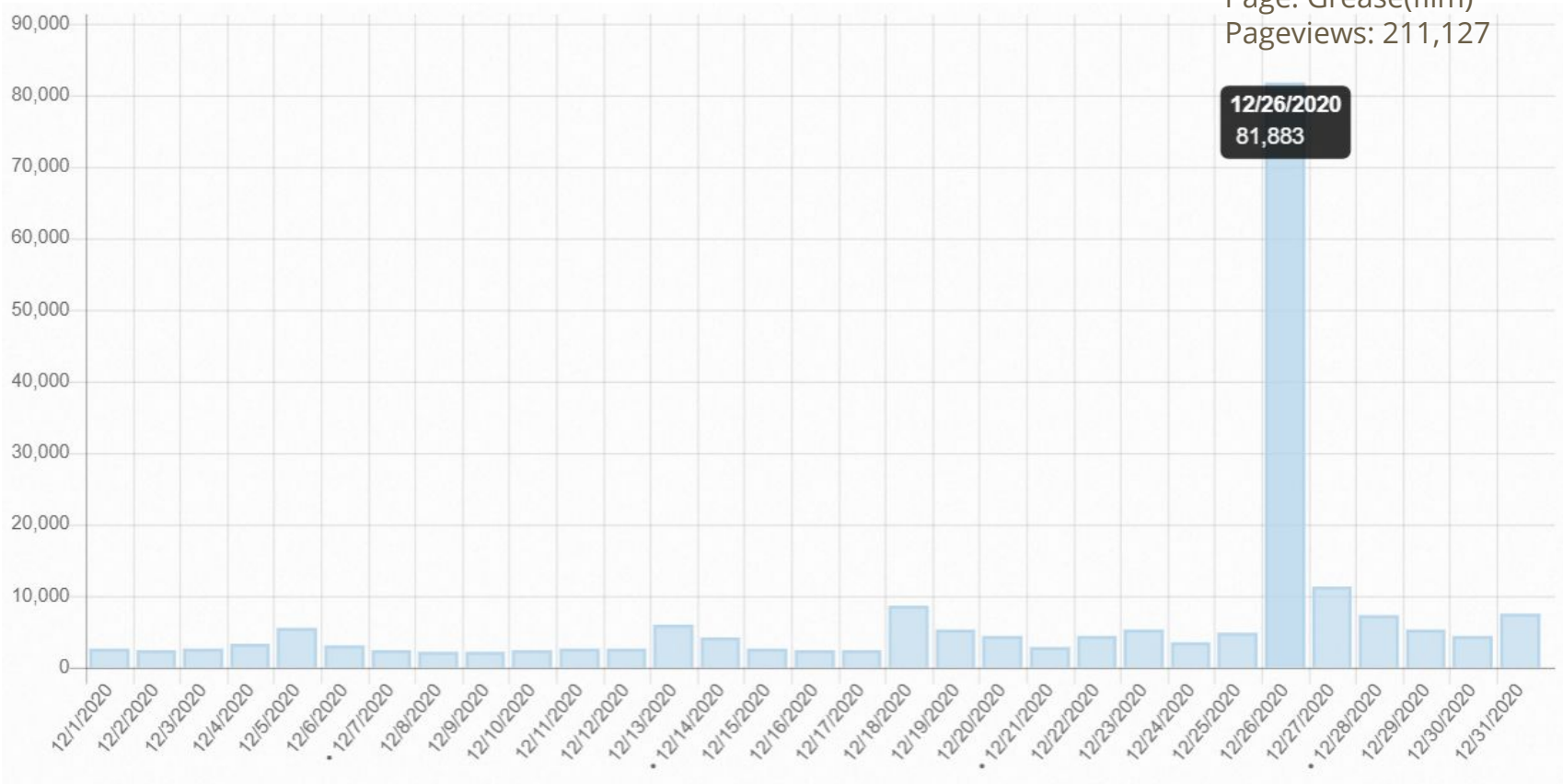
page_title	fraction
Queen_Victoria	**0.99986**
Once_Upon_a_Time_in_Hollywood	0.99805
Gilligan's_Island	0.99792
George_I_of_Great_Britain	0.99727
George_Clooney	0.99694
List_of_Netflix_original_programming	0.99509
John_Lennon	0.99215

Fraction < 1
(6629 Counts)

page_title	fraction
List_of_pornographic_performers_by_decade	**3.49496**
Grease_(film)	2.5374
USP_Florence_ADMAX	2.50508
The_Vicar_of_Dibley	2.47327
The_Sound_of_Music_(film)	2.26818
Carry_On_(franchise)	2.08784
List_of_Saturday_Night_Live_cast_members	2.02468

Fraction > 1
(217 Counts)

Q2. Continued ...

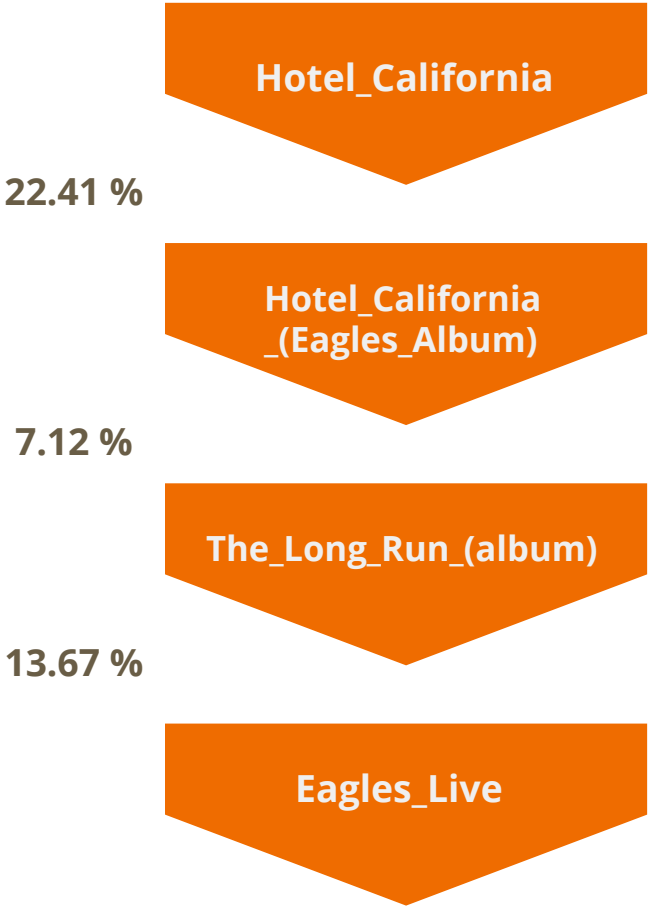


Q3. What series of wikipedia articles, starting with Hotel California, keeps the largest fraction of its readers clicking on internal links?

- December, 2020 pageview and clickstream data from Q2 was used
- Used Subquery and Inner Join to get the results for each series
- Limitation: Used the single highest percentage path on each article

Q3. Result

```
SELECT pageview.page_title AS referrer,
link.curr AS requested,
link.linkcount AS link_click,
pageview.pageviewcount AS page_views,
ROUND(link.linkcount/pageview.pageviewcount, 4) AS fraction
FROM
(SELECT page_title, ((SUM(count_views)*5)*31) AS
pageviewcount
FROM q3_pageview
GROUP BY page_title) AS pageview
JOIN
(SELECT prev, curr, SUM(numpair) AS linkcount
FROM internal_view
GROUP BY prev, curr) AS link
ON pageview.page_title = link.prev
WHERE pageview.page_title = 'PAGE TITLE GOES HERE'
ORDER BY fraction DESC
LIMIT 10;
```



Q4. Find an example of an English wikipedia article that is relatively more popular in the Americas than elsewhere.

- 1/20/2021 pageview data was used

Time Zone used for Analysis

PST	7:00 a.m - 3:30 p.m
EST	10:00 a.m - 6:30 p.m
UTC	3:00 p.m - 11:30 p.m
AWST	11:00 p.m - 12:00 a.m

```
SELECT page_title,
SUM(count_views) AS page_views
FROM q4_pageview_en
GROUP BY page_title
ORDER BY page_views DESC
LIMIT 15;
```

page_title	page_view
Main_Page	1,570,601
Joe_Biden	**710,870**
Kamala_Harris	660,925
Special:Search	348,458
Donald_Trump	289,365

Q5. Analyze how many users will see the average vandalized wikipedia page before the offending edit is reversed.

- December, 2020 Page Revision and User History data
- December 20, 2020 Pageview data (English Articles)
- Loaded data to VANDAL table
- Searched for '%VANDAL%' from event_comment column

6. Interesting Findings

```
SELECT wiki_db, country,
activity_level, SUM(upper_bound)
AS upper
FROM geoeditor
GROUP BY wiki_db, country,
activity_level
ORDER BY upper DESC;
```

enwiki	United States	23300
jawiki	Japan	8430
enwiki	United Kingdom	6660
itwiki	Italy	6620
frwiki	France	6130
enwiki	India	5200
dewiki	Germany	5650
eswiki	Spain	3080
kowiki	South Korea	2470
zhwiki	Taiwan	2320
ptwiki	Brazil	2000

Time Zone

USA	4:00 p.m ~
Brazil	6:00 p.m ~
UK	9:00 p.m ~
Spain	10:00 p.m ~
Taiwan	5:00 a.m ~
Japan	6:00 a.m ~
South Korea	6:00 a.m ~

Q6. Result

Country	Page_Title	Views	Total Views
USA	Zodiac_Killer	38,535	884,302
Italy	Isola_delle_Rose _(micronazione) = Rose_Island_(micronation)	14,259	2,084,147
Korea	조두순 = Cho Doo-Soon	11,910 - higher than main page	322,960
Spain	Nuestra_Señora_de_Guadalupe_(México) = Our_Lady_of_Guadalupe_ (Mexico)	10,621	2,608,318
Japan	今際の国のアリス = Alice In Borderland	7,755	3,633,919
Taiwan	S風暴 = S Storm	6,366	1,679,254
Germany	Zahlennamen = Names of large numbers	6,226	2,931,904
Brazil	Cleópatra	3,347	910,682
France	Dominique_Strauss-Kahn	2,762	2,647,120

Country Views: 15,319,606
Total Views: 50,710,482

30% of Total Views

Thank you for listening :)