

A decorative graphic on the left side of the slide consisting of two overlapping parallelograms. The front one is blue and the back one is a light green. They are positioned diagonally, with the blue one partially covering the green one.

Project 1: Hive on WikiMedia

Samuel L Owens




Which English Wikipedia
article got the most
traffic on October 20?

<https://dumps.wikimedia.org/other/pageviews/2020/2020-10/pageviews-20201020-{00..23}0000.gz>

- A table is created containing the number of views each article gets on each wikipedia article on 10/20
 - Each page receives 24 records, one for each hour in the day
- An intermediary table is created containing only English page records
- Another intermediary table is created combining the hourly records into one daily record


pageviews_en_total.title	pageviews_en_total.views
Main_Page	5961008
Special:Search	1476831
-	544714
Jeffrey_Toobin	321459
C._Rajagopalachari	210558
The_Haunting_of_Bly_Manor	185139
Robert_Redford	178779
Jeff_Bridges	159163
Bible	151484
Chicago_Seven	149966
Harshad_Mehta	116907
The_Trial_of_the_Chicago_7	115626
Deaths_in_2020	113965
Kyler_Murray	113410
Abbie_Hoffman	107768
Murder_of_Robert_McCartney	104409
QAnon	103285
Sisters_at_Heart	99831
Dancing_with_the_Stars_(American_season_29)	99750
2016_United_States_presidential_election	91744




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

<https://dumps.wikimedia.org/other/pageviews/2020/2020-10/pageviews-202009{01..30}-{00..23}0000.gz>

<https://dumps.wikimedia.org/other/clickstream/2020-09/clickstream-enwiki-2020-09.tsv.gz>

- 
- A table is created containing the number of viewers of each article from September 2020 that follow an internal link to another wikipedia page
 - Each page receives a record for each page that it links to
 - An intermediary table is created that sums up the number of times a viewer of the article follows any internal link
 - Another table is created containing the pageview records from September 2020
 - An intermediary table is created that sums up all of the activity for each page for the entire month
 - These two tables are joined together to provide the monthly readers of a page, the number of those readers that follow internal wikipedia links, and that ratio represented as a fraction

q2_answer.title	q2_answer.views	q2_answer.links	q2_answer.link_percent
Dune_(2020_film)	1278838	1201459	0.94
COVID-19_pandemic_by_country_and_territory	1207880	1093321	0.91
Cobra_Kai	2464628	2241751	0.91
Schitt's_Creek	1493588	1339942	0.9
Elizabeth_II	1065054	922145	0.87
Sarah_Paulson	1252257	987550	0.79
Supreme_Court_of_the_United_States	1279104	1002716	0.78
2016_United_States_presidential_election	1052252	768124	0.73
Enola_Holmes_(film)	1980047	1356311	0.68
2020_United_States_presidential_election	1151217	749205	0.65

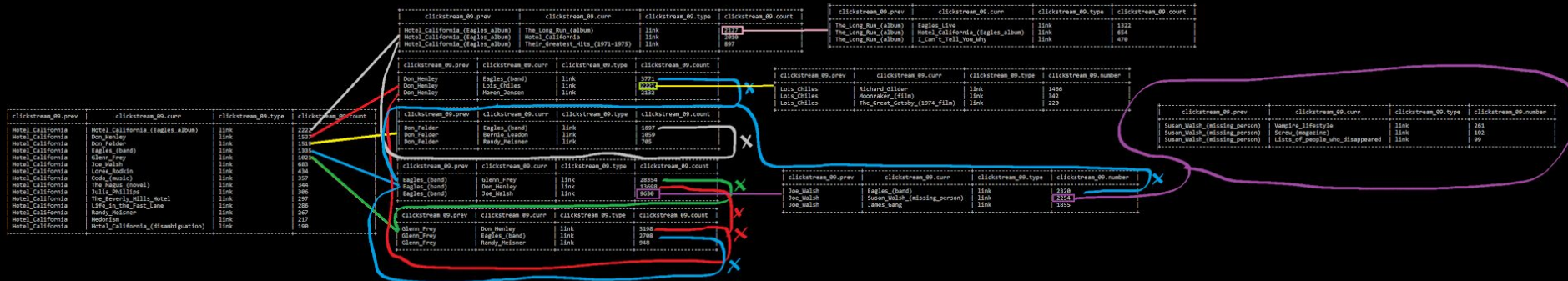



What series of wikipedia
articles, starting with
Hotel California, keeps the
largest fraction of its
reader through internal
links

<https://dumps.wikimedia.org/other/pageviews/2020/2020-10/pageviews-202009{01..30}-{00..23}0000.gz>

<https://dumps.wikimedia.org/other/clickstream/2020-09/clickstream-enwiki-2020-09.tsv.gz>

- Any series that loops back on itself will be disregarded
- When two series converge at one link, the series with more present viewers continues
- Hotel_California links out a total of 13,779 of its readers to internal Wikipedia pages
- From Hotel_California the top 5 internal Wikipedia pages linked to are found
- From that top 5, their top 3 linked to pages are found
- The series with the highest retention becomes clear after 4 levels
- Hotel_California -> Eagles_(band) -> Joe_Walsh -> Susan_Walsh_(missing_person) -> Vampire_Lifestyle



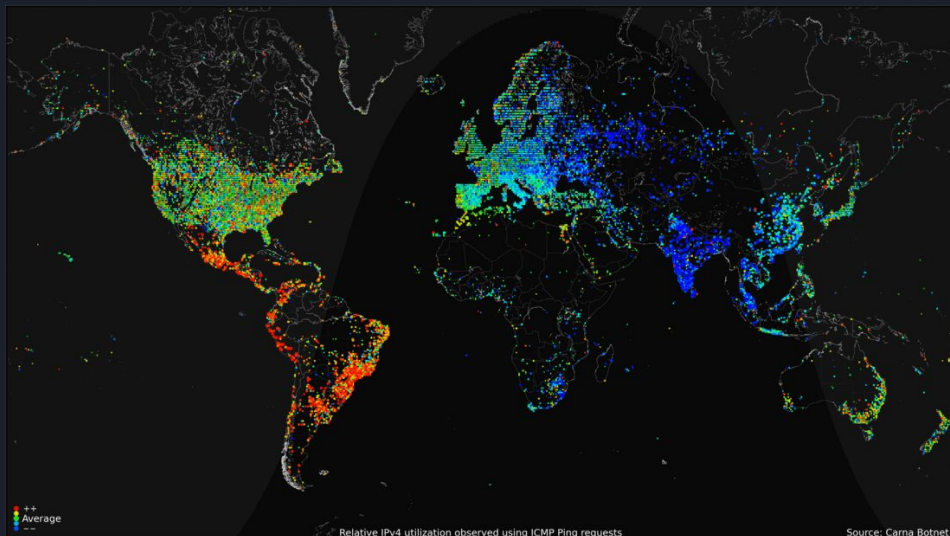


Find an example of an
English Wikipedia article
that is relatively more
popular in the UK, US,
and Australia


<https://dumps.wikimedia.org/other/pageviews/2020/2020-10/pageviews-20201020-{00..23}0000.gz>

<https://darknetdiaries.com/episode/13/>

- It would appear the hours of highest activity in the US are bracketed about Business hours
 - 0900-1800 converted forward 5 hours for UTC is 1400-2300, this time frame is attributed to the US
- The UK's hours of highest activity span approximately from midday to sunset
 - 1300 - 1800 UK time is the same as UTC and need not be adjusted
- Australia appears to have some hours of activity in the early morning, followed by a lull, and more Activity towards the end of business hours
 - 0700 - 1000 and 1400 - 1800 ACT converts to 1800 - 2100 and 0100 - 0500 UTC
- A table is loaded for from the files representing the respective time spans
- We repeat the process of the first question on each of these tables




q4_usa_answer.title	q4_usa_answer.views
Main_Page	78773541
Special:Search	21294630
-	8254202
Amy_Coney_Barrett	2957345
Ruth_Bader_Ginsburg	2711613
Dennis_Nilsen	2620063
Shooting_of_Breonna_Taylor	1960292
Tenet_(film)	1731853
q4_uk_answer.title	q4_uk_answer.views
Main_Page	57771634
Special:Search	15519336
-	6055166
Ruth_Bader_Ginsburg	2039267
Amy_Coney_Barrett	1579484
Shooting_of_Breonna_Taylor	1330111
Deaths_in_2020	1141635
Tenet_(film)	1128742
q4_aus_answer.title	q4_aus_answer.views
Main_Page	58408304
Special:Search	15612897
-	6269236
Ruth_Bader_Ginsburg	1815496
Amy_Coney_Barrett	1267117
S._P._Balasubrahmanyam	1222834
Tenet_(film)	1130105
Deaths_in_2020	1067903
Shooting_of_Breonna_Taylor	1065502

A decorative graphic on the left side of the slide consisting of two overlapping parallelograms. The front one is blue and the back one is a light greenish-blue. They are positioned diagonally, with the blue one in front of the green one.


Analyze how many users
will see the average
vandalized Wikipedia
page before the
offending edit is reversed

https://dumps.wikimedia.org/other/mediawiki_history/2020-10/enwiki/2020-10.enwiki.2020-09.tsv.bz2

- 
- A table is created containing the 70 columns provided in the Wiki_History data set for September 2020
 - From the 70 only 2 are checked, one to verify that the record represents an edit being reverted and one to show the time in seconds before that edit was eventually reverted.
 - From this the average time a reverted edit is visible is calculated and a table is saved
 - The table has seconds, minutes, hours, and days
 - The result is an average visibility of just over 3 days
 - The average views a page gets in a month is then calculated from our previously created table of monthly view totals per page
 - This average is divided by 10 due to the average visibility time being roughly 10% of 30 days


q5_answer.average_views
36.3

reverted_avg_time.seconds	reverted_avg_time.minutes	reverted_avg_time.hours	reverted_avg_time.days
282766.49	4712.77	78.55	3.27



By the same metric as
the previous question,
what is the most
vandalized Article?

https://dumps.wikimedia.org/other/mediawiki_history/2020-10/enwiki/2020-10.enwiki.2020-09.tsv.bz2

- 
- A table is created containing the 70 columns provided in the Wiki_History data set for September 2020
 - An intermediary table is created containing the number of times a revision is reverted on each page
 - Many of the top records appear to be internal pages for logging and administration but the top appears to be “Teahouse”

q6_answer.title	q6_answer.reversions
Administrator_intervention_against_vandalism	14990
Requests_for_page_protection	7986
Administrators'_noticeboard/Incidents	7784
Teahouse	7782
Username_for_administrator_attention	6476
Sandbox	6174
Administrator_intervention_against_vandalism/TB2	5004
In_the_news/Candidates	4686
Help_desk	4246
Administrators'_noticeboard	3786