Project 1: Wikipedia Data Analysis

Ernie Chu

This year, which Wikipedia article got the most traffic on October 20th?

The **main_page** of Wikipedia was the most visited site on October 20th. There were **3,234,621** hits for the mobile site, and **2,726,387** hits for the desktop site. Combined, there were **5,961,008** views on that day.

The next slide shows the top 10 viewed articles on this date. I've shown the combined views in this table.

Top 10 Articles Viewed on October 20th

title	most_popular_on_oct_20
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

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

I first compared the total number of views for a page to the entire month of September 2020 dataset to get more accurate data. I then took the number of times the article was labeled as a referrer to another as a link in the clickstream data set for September 2020. Since the clickstream data covers a month of data, the number of view totals in both files should be approximately the same.

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

To get my answer, I divided the above number by the total number of views and multiplied by 100 to get a percentage. The article with the highest proportion of links click was **Dune_(2020_film)**, at **93.95%**.

I limited page popularity to a certain threshold to get reasonable results. I ended up choosing from within the top 100 most viewed articles in September 2020.

The next slide shows the top 10 articles with the highest proportion of internal links followed.

Top 10 Articles with Highest Proportion of Links Followed

total_views_in_sept.title	total_views_in_sept.total_views	total_views_in_clickstream_sept.links_followed	percentage_links_clicked
The_Karate_Kid	804345	860567	106.99
Dune_(2020_film)	1278838	1201459	93.95
Cobra_Kai	2459988	2241751	91.13
COVID-19_pandemic_by_country_and_territory	1207880	1093321	90.52
Schitt's_Creek	1493588	1339942	89.71
Elizabeth_II	1065045	922145	86.58
Sarah_Paulson	1252257	987550	78.86
Supreme_Court_of_the_United_States	1278921	1002716	78.4
Lucifer_(TV_series)	925240	713085	77.07
One_Flew_Over_the_Cuckoo's_Nest_(film)	804015	590614	73.46

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

To find out, I query the next higher link in the chain starting with Hotel California. On this first time, ‘Hotel California’ is the URL to request. I get back the URL with the highest number of clicks from this origin. The new article is then set at the original requester, and the cycle repeats.

Here’s the table I used to generate the first link in the chain.

Top 10 Followed Articles From “Hotel_California”

clickstream_sept.previous_referrer_url	clickstream_sept.current_requester_url	clickstream_sept.type	clickstream_sept.occurrences
Hotel_California	Hotel_California_(Eagles_album)	link	2222
Hotel_California	Don_Henley	link	1537
Hotel_California	Don_Felder	link	1519
Hotel_California	Eagles_(band)	link	1335
Hotel_California	Glenn_Frey	link	1021
Hotel_California	Joe_Walsh	link	683
Hotel_California	Loree_Rodkin	link	434
Hotel_California	Coda_(music)	link	357
Hotel_California	The_Magus_(novel)	link	344
Hotel_California	Julia_Phillips	link	306

First 5 Links From Hotel California

These are the first 5 iterations of this link following “Hotel California”.

1. **Hotel_California_(Eagles_Album) (2222)**
2. **The_Long_Run_(album) (2127)**
3. **Eagles_Live (1333)**
4. **Eagles_Greatest_Hits,_Vol._2 (1136)**
5. **The_Very_Best_of_the_Eagles (996)**

An interesting future application would involve applying Spark to find the full chain. This implementation would not retain the acyclic nature of traditional Hive MapReduce.

What is a Wikipedia article that is relatively more popular in the UK, the US, and Australia?

I ended up making a lot of simplifying assumptions in this question.

I reasoned that pages with more revisions were relatively more popular during their normal business hours, which I define on the following slide.These were UTC times, and I manually counted some point times for three articles I selected purposefully.

Definition of Peak Times for Question #5

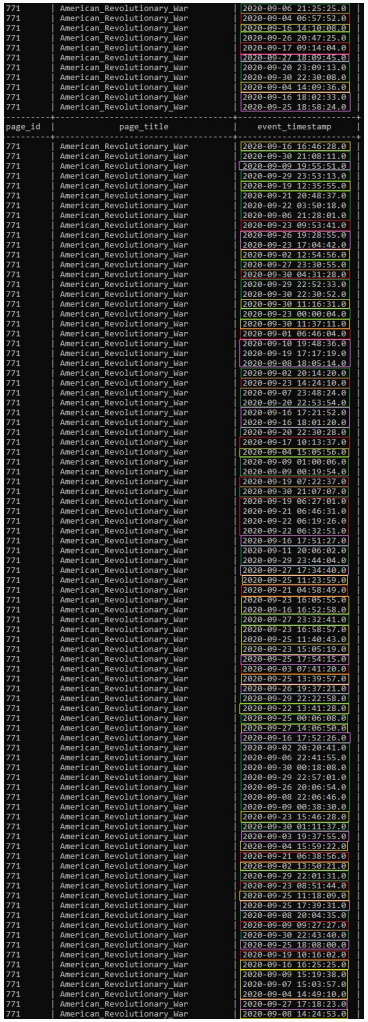
On the following slides, I present some tables with colored boxes.

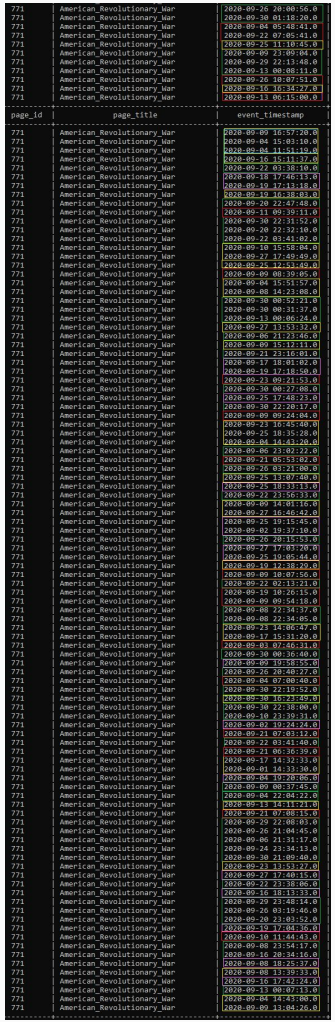
In **green** are Australian peak times, defined as **20:00 UTC to 04:00 UTC**.

In **red** are US peak times, defined as **04:00 UTC to 10:30 UTC**.

In **yellow** are UK peak times, defined as **10:30 UTC to 17:00 UTC**.

In **pink** are other times left out, defined as **17:00 UTC to 20:00 UTC**.





To the left is the table of all revision times for the article **American_Revolutionary_War** that took place in September 2019.

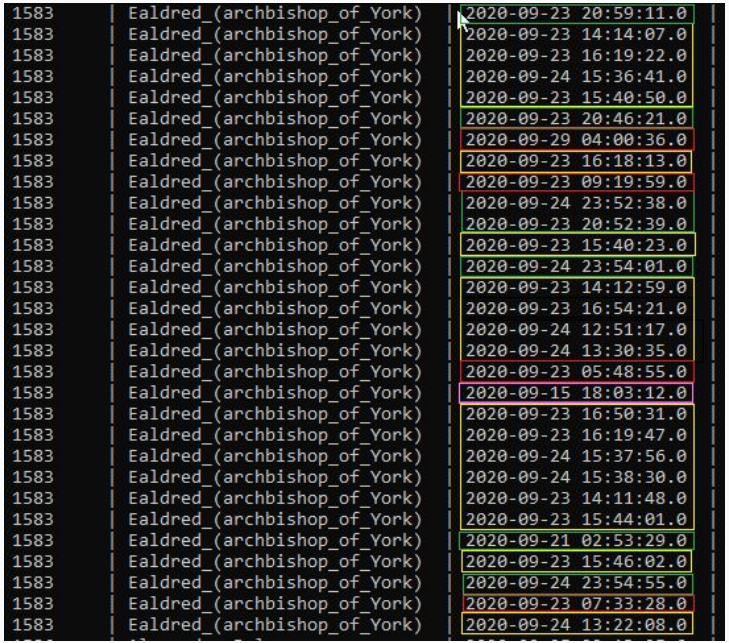
There are 80 times highlighted in green, 46 in red, 56 in yellow, and 37 in pink. As such, I concluded that the article **American_Revolutionary_War** was relatively more popular in **Australia** compared to the UK and America.

However, there is a high degree of uncertainty due to the times that fell outside of peak hours!

1495	Australian_Labor_Party	2020-09-02 01:09:53.0
1495	Australian_Labor_Party	2020-09-14 01:10:59.0
1495	Australian_Labor_Party	2020-09-21 04:21:37.0
1495	Australian_Labor_Party	2020-09-02 10:42:18.0
1495	Australian_Labor_Party	2020-09-13 04:04:55.0
1495	Australian_Labor_Party	2020-09-02 02:07:23.0
1495	Australian_Labor_Party	2020-09-02 01:14:10.0
1495	Australian_Labor_Party	2020-09-14 00:29:47.0
1495	Australian_Labor_Party	2020-09-13 04:15:15.0
1495	Australian_Labor_Party	2020-09-14 02:14:29.0
1495	Australian_Labor_Party	2020-09-16 07:14:42.0
1495	Australian_Labor_Party	2020-09-02 02:06:57.0
1495	Australian_Labor_Party	2020-09-13 04:09:13.0
1495	Australian_Labor_Party	2020-09-13 04:23:53.0
1495	Australian_Labor_Party	2020-09-21 04:17:55.0
1495	Australian_Labor_Party	2020-09-13 04:13:45.0

To the left is the table of all revision times for the article **Australian_Labor_Party** that took place in September 2019.

There are 7 times highlighted in green, 8 in red, 1 in yellow, and none in pink. As such, I concluded that the article **Australian_Labor_Party** was relatively more popular in **America** compared to the UK and Australia.



1583	Ealdred_(archbishop_of_York)	2020-09-23 20:59:11.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 14:14:07.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 16:19:22.0
1583	Ealdred_(archbishop_of_York)	2020-09-24 15:36:41.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 15:40:50.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 20:46:21.0
1583	Ealdred_(archbishop_of_York)	2020-09-29 04:00:36.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 16:18:13.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 09:19:59.0
1583	Ealdred_(archbishop_of_York)	2020-09-24 23:52:38.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 20:52:39.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 15:40:23.0
1583	Ealdred_(archbishop_of_York)	2020-09-24 23:54:01.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 14:12:59.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 16:54:21.0
1583	Ealdred_(archbishop_of_York)	2020-09-24 12:51:17.0
1583	Ealdred_(archbishop_of_York)	2020-09-24 13:30:35.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 05:48:55.0
1583	Ealdred_(archbishop_of_York)	2020-09-15 18:03:12.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 16:50:31.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 16:19:47.0
1583	Ealdred_(archbishop_of_York)	2020-09-24 15:37:56.0
1583	Ealdred_(archbishop_of_York)	2020-09-24 15:38:30.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 14:11:48.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 15:44:01.0
1583	Ealdred_(archbishop_of_York)	2020-09-21 02:53:29.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 15:46:02.0
1583	Ealdred_(archbishop_of_York)	2020-09-24 23:54:55.0
1583	Ealdred_(archbishop_of_York)	2020-09-23 07:33:28.0
1583	Ealdred_(archbishop_of_York)	2020-09-24 13:22:08.0

To the left is the table of all revision times for the article **Ealdred_(archbishop_of_York)** that took place in September 2019.

There are 7 times highlighted in green, 4 in red, 18 in yellow, and 1 in pink. As such, I concluded that the article **Ealdred_(archbishop_of_York)** was relatively more popular in **the UK** compared to America and Australia.

How many users will see the average vandalized Wikipedia page before the offending edit is reversed?

I assumed that all revision events in the edits data was activity towards reverting vandalism. In reality, there could be many reasons why articles are revised.

I then counted the total number of revision events (**330,496**) and the total number of views across all pages (**6,660,118,635**) in September of 2019. By dividing the total number of views by the total number of revision events, I found that roughly **20,152 users** on average saw an article before every revision.

The next slide shows the HQL code that I wrote for this query.

HQL: Average Views Before Vandalism Reverted

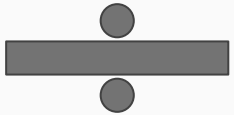
```
# Get the number of times revisions were made to articles in September 2019.
CREATE TABLE TOTAL_NUMBER_OF_REVISIONS AS
SELECT COUNT(REVISION_SECONDS_TO_IDENTITY_REVERT) AS TOTAL_NUMBER_OF_REVISIONS_IN_SEPT
FROM RELATIVE_POPULARITY
WHERE REVISION_SECONDS_TO_IDENTITY_REVERT > 0;

# Get the total number of views across all articles that were present in September 2019.
CREATE TABLE TOTAL_NUMBER_OF_VIEWS AS
SELECT SUM(OCCURRENCES) AS TOTAL_NUMBER_OF_VIEWS_IN_SEPT
FROM CLICKSTREAM_SEPT;

# Divide the number of views by the number of revisions to see the number of times an article was viewed per revision.
# 5. Analyze how many users will see the average vandalized wikipedia page before the offending edit is reversed.
SELECT ROUND(TOTAL_NUMBER_OF_VIEWS.TOTAL_NUMBER_OF_VIEWS_IN_SEPT / TOTAL_NUMBER_OF_REVISIONS.TOTAL_NUMBER_OF_REVISIONS_IN_SEPT, 2) AS AVERAGE_NUMBER_OF_VIEWS_BEFORE_REVISION
FROM TOTAL_NUMBER_OF_REVISIONS, TOTAL_NUMBER_OF_VIEWS;
```

Tables and Internal Representation

total_number_of_views_in_sept
6660118635



total_number_of_revisions_in_sept
330496

average_number_of_views_before_revision
20151.89

Significant Limitations to Consider

The given average number is likely inaccurate, due to 2 major factors - activity from users and accuracy of data.

Many articles on Wikipedia are likely never vandalized and/or never revised. Likewise, many articles are frequently viewed, revised, and have high activity. Regarding data, I had very low confidence in the data that Wikipedia itself supplies - many “revision_seconds” records are in the positive and negative **billions**. Any analysis on this data is likely inaccurate and can’t really be taken too seriously.

Given better data, an interesting future experiment could be to see the bottom quartile (25th), median, and top quartile (75th) percentile as they relate to page views and revision activity.

What were the most popular articles in September 2019?

I was interested in finding out the most popular 10 articles on Wikipedia (en and en.m) during September 2019. Curiously, it ended up being **- (dash)**, which we made a point of discussing in class. It got **7,171,434,364 views**!

That's 7 billion, 171 million, 434 thousand, 364 views - a lot!

The following table shows the top 10 most viewed pages in September 2019.

Top 10 Followed Viewed Articles in September 2019

title	total
-	7171434364
Main_Page	165044119
Special:Search	41915305
Ruth_Bader_Ginsburg	7605356
Amy_Coney_Barrett	5924508
Tenet_(film)	3877047
Shooting_of_Breonna_Taylor	3850524
Dennis_Nilsen	3564441
Deaths_in_2020	3316200
Mulan_(2020_film)	3239724

10 rows selected (2515.92 seconds)

Thanks for watching! Do you have any questions?