Wikipedia Dataset Analysis

Revature Project 1

Christopher Chee

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

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

Data Used:

All Pageviews for October 20, 2020

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

```scala
class PageMapper extends Mapper[LongWritable, Text, Text, IntWritable] {

  override def map(key: LongWritable, value: Text, context: Mapper[LongWritable, Text, Text, IntWritable]#Context): Unit = {

    val line = value.toString
    val words = line.split("\\s+").filter(_.length > 0)

    try
    {
      if ((words(0).equalsIgnoreCase("en")) || (words(0).equalsIgnoreCase("en.m")))
        context.write(new Text(words(1)), new IntWritable(words(2).toInt))
    }
  }
}
```

```
!              72
!!             25
!!!            148
!!!!!!!!  19
!!!F____You!!!    23
!!!F____You!!!_And_Then_Some        19
!!!F____You!!!_and_Then_Some        14
!!!_(!!!_album)  16
!!!_(American_band)       14
!!!_(Chk_Chk_Chk)        14
```

Format of MapReduce output

```scala
class PageReducer extends Reducer[Text, IntWritable, Text, IntWritable] {

  override def reduce(key: Text, values: lang.Iterable[IntWritable], context: Reducer[Text, IntWritable, Text, IntWritable]#Context): Unit = {

    var count = 0

    values.forEach(count += _.get())

    context.write(key, new IntWritable(count))
  }
}
```

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

```scala
job.setInputFormatClass(classOf[KeyValueTextInputFormat])
FileInputFormat.setInputPaths(job, new Path(args(0)))
FileOutputFormat.setOutputPath(job, new Path(args(1)))

job.setMapperClass(classOf[InverseMapper[Text, Text]])
job.setReducerClass(classOf[SortReducer])

class TextToLongComparator extends WritableComparator(classOf[Text], true) {
  override def compare(a: WritableComparable[_], b: WritableComparable[_]): Int = {
    -a.toString.toLong.compareTo(b.toString.toLong)
  }
}
```

```
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
```

Top 10 viewed pages

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

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

Data Used:

Clickstream for the month of Sept.

All Pageviews for the month of Sept.

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

```scala
class InternalMapper extends Mapper[LongWritable, Text, Text, LongWritable] {

  override def map(key: LongWritable, value: Text, context: Mapper[LongWritable, Text, Text, LongWritable]#Context): Unit = {

    val fields = value.toString.split("\\s+").filter(_.length > 0)

    if (fields(2).equalsIgnoreCase("link"))
      context.write(new Text(fields(0)), new LongWritable(fields(3).toLong))
  }
}
```

```scala
class InternalReducer extends Reducer[Text, LongWritable, Text, LongWritable] {

  override def reduce(key: Text, values: lang.Iterable[LongWritable], context: Reducer[Text, LongWritable, Text, LongWritable]#Context): Unit = {

    var count = 0

    values.forEach(count += _.get().toInt)

    context.write(key, new LongWritable(count))
  }
}
```

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

internal_views.page	internal_views.views
!!	133
!!!	1938
!!!_(album)	121
!!!_(disambiguation)	19
!Hero	24
!Oka_Tokat	101
!T.O.O.H.!	53
!Women_Art_Revolution	18
!_(The_Dismemberment_Plan_album)	124
!_(The_Song_Formerly_Known_As)	20

Format of MapReduce output

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

internal_views.page	internal_views.views
!!	133
!!!	1938
!!!_(album)	121
!!!_(disambiguation)	19
!Hero	24
!Oka_Tokat	101
!T.O.O.H.!	53
!Women_Art_Revolution	18
!_(The_Dismemberment_Plan_album)	124
!_(The_Song_Formerly_Known_As)	20

JOIN

total_views.page	total_views.views
!	1016
!!	398
!!!	4488
!!!!!!!!	411
!!!F____You!!!	372
!!!F____You!!!_And_Then_Some	217
!!!F____You!!!_and_Then_Some	153
!!!_(!!!_album)	68
!!!_(American_band)	104
!!!_(Chk_Chk_Chk)	77

```
INSERT OVERWRITE DIRECTORY '/user/hive/internal_fraction'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
SELECT INTERNAL_VIEWS.PAGE,
       INTERNAL_VIEWS.VIEWS AS INTERNAL_VIEWS,
       TOTAL_VIEWS.VIEWS AS TOTAL_VIEWS,
       INTERNAL_VIEWS.VIEWS/TOTAL_VIEWS.VIEWS AS INTERNAL_FRACTION
FROM INTERNAL_VIEWS JOIN TOTAL_VIEWS
ON (INTERNAL_VIEWS.PAGE = TOTAL_VIEWS.PAGE)
ORDER BY INTERNAL_VIEWS.VIEWS/TOTAL_VIEWS.VIEWS DESC;
```

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

```
SELECT PAGE, INTERNAL_VIEWS, TOTAL_VIEWS, ROUND(INTERNAL_FRACTION * 100, 2) AS INTERNAL_PERCENTAGE
FROM INTERNAL_FRACTION
LIMIT 10;
```

page	internal_views	total_views	internal_percentage
/r/	64	1	6400.0
/\	56	2	2800.0
Health//Disco	209	8	2612.5
Strange_haircuts_//_cardboard_guitars_//_and_computer_samples	26	1	2600.0
List_of_listed_buildings_in_Musselburgh,_East_Lothian	662	28	2364.29
Flourish_//_Perish	19	1	1900.0
Lost_Forever_//_Lost_Together	463	29	1596.55
2006_Chicago_Rush_season	185	12	1541.67
Baeolidia_gracilis	121	8	1512.5
Finally_//_Beautiful_Stranger	282	19	1484.21

Invalid results

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

```sql
SELECT PAGE, INTERNAL_VIEWS, TOTAL_VIEWS, ROUND(INTERNAL_FRACTION * 100, 2) AS INTERNAL_PERCENTAGE
FROM INTERNAL_FRACTION
WHERE INTERNAL_FRACTION < 1 AND TOTAL_VIEWS > 500000
LIMIT 10;
```

page	internal_views	total_views	internal_percentage
Dune_(2020_film)	1201459	1278838	93.95
Cobra_Kai	2241751	2459988	91.13
COVID-19_pandemic_by_country_and_territory	1093321	1207880	90.52
Christopher_Nolan	612734	680233	90.08
Schitt's_Creek	1339942	1493588	89.71
Bill_&_Ted_Face_the_Music	458119	518671	88.33
Elizabeth_II	922145	1065045	86.58
The_Babysitter:_Killer_Queen	663863	767589	86.49
72nd_Primetime_Emmy_Awards	465626	539024	86.38
2020_US_Open_(tennis)	436071	536585	81.27

Filtered results

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

3. What series of Wikipedia articles, starting with [Hotel California](#),
   keeps the largest fraction of its readers clicking on internal links?

Data Used:

   Clickstream for the month of Sept.

   All Pageviews for the month of Sept.

3. What series of Wikipedia articles, starting with [Hotel California](), keeps the largest fraction of its readers clicking on internal links?

referrals.referrer	referrals.referred	referrals.clicks
Bathtubs_Over_Broadway	Industrial_musical	97
Industrial_music	Industrial_musical	67
Skittles_Commercial:_The_Broadway_Musical	Industrial_musical	14
Yvonne_Craig	Industrial_musical	23
Kander_and_Ebb	Industrial_musical	15
Descendants_of_Ibn_Saud	Saud_bin_Abdulaziz_bin_Nasser_Al_Saud	20
Nasser_bin_Abdulaziz_Al_Saud	Saud_bin_Abdulaziz_bin_Nasser_Al_Saud	70
Abdullah_of_Saudi_Arabia	Saud_bin_Abdulaziz_bin_Nasser_Al_Saud	59
LGBT_rights_in_Saudi_Arabia	Saud_bin_Abdulaziz_bin_Nasser_Al_Saud	142
List_of_gay,_lesbian_or_bisexual_people:_A	Saud_bin_Abdulaziz_bin_Nasser_Al_Saud	66

JOIN

page	internal_views	total_views	internal_percentage
Dune_(2020_film)	1201459	1278838	93.95
Cobra_Kai	2241751	2459988	91.13
COVID-19_pandemic_by_country_and_territory	1093321	1207880	90.52
Christopher_Nolan	612734	680233	90.08
Schitt's_Creek	1339942	1493588	89.71
Bill_&_Ted_Face_the_Music	458119	518671	88.33
Elizabeth_II	922145	1065045	86.58
The_Babysitter:_Killer_Queen	663863	767589	86.49
72nd_Primetime_Emmy_Awards	465626	539024	86.38
2020_US_Open_(tennis)	436071	536585	81.27

```
INSERT OVERWRITE DIRECTORY '/user/hive/referral_fractions'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
SELECT REFERRALS.REFERRER,
       REFERRALS.REFERRED,
       INTERNAL_FRACTION.INTERNAL_VIEWS,
       INTERNAL_FRACTION.TOTAL_VIEWS,
       INTERNAL_FRACTION.INTERNAL_FRACTION
FROM REFERRALS JOIN INTERNAL_FRACTION
ON (REFERRALS.REFERRED = INTERNAL_FRACTION.PAGE)
ORDER BY REFERRALS.REFERRER ASC;
```

3. What series of Wikipedia articles, starting with [Hotel California](), keeps the largest fraction of its readers clicking on internal links?

referral_fractions.referrer	referral_fractions.referred	referral_fractions.internal_views	referral_fractions.total_views	referral_fractions.internal_fraction
!!	Double-negation_translation	14	291	0.048109965635738834
!!	Retroflex_click	76	335	0.22686567164179106
!!	!_(disambiguation)	145	781	0.1856594110115237
!!	Chess_annotation_symbols	1262	4108	0.30720545277507305
!!	Double_factorial	404	7530	0.053652058432934926
!!	!!!_(disambiguation)	19	181	0.10497237569060773
!!!	Strange_Weather,_Isn't_It?	94	242	0.3884297520661157
!!!	!!!_(disambiguation)	19	181	0.10497237569060773
!!!	Tyler_Pope	121	765	0.15816993464052287
!!!	As_If_(album)	94	248	0.3790322580645161

Resulting table from join

3. What series of Wikipedia articles, starting with [Hotel California](#), keeps the largest fraction of its readers clicking on internal links?

```
SELECT *
FROM REFERRAL_FRACTIONS
WHERE REFERRER='Hotel_California'
ORDER BY INTERNAL_FRACTION DESC;
```

referral_fractions.referrer	referral_fractions.referred	referral_fractions.internal_views	referral_fractions.total_views	referral_fractions.internal_fraction
Hotel_California	Eagles_(band)	132994	139366	0.954278661940502
Hotel_California	Jethro_Tull_(band)	51257	61744	0.8301535371857994
Hotel_California	Steely_Dan	66279	85960	0.7710446719404375
Hotel_California	The_Twilight_Zone_(1959_TV_series)	31414	42170	0.7449371591178563
Hotel_California	American_Horror_Story:_Hotel	53807	73110	0.7359731910819314
Hotel_California	Desperado	1255	1718	0.7305005820721769
Hotel_California	Cameron_Crowe	38391	58203	0.6596051749909798
Hotel_California	Anton_LaVey	36586	56460	0.6479985830676586
Hotel_California	John_Fowles	4350	7491	0.5806968362034441
Hotel_California	Hotel_California_(disambiguation)	144	248	0.5806451612903226

Hotel_California -> Eagles_(band) (95.4%)

3. What series of Wikipedia articles, starting with [Hotel California](#), keeps the largest fraction of its readers clicking on internal links?

```sql
SELECT *
FROM REFERRAL_FRACTIONS
WHERE REFERRER='Eagles_(band)'
ORDER BY INTERNAL_FRACTION DESC
LIMIT 10;
```

referral_fractions.referrer	referral_fractions.referred	referral_fractions.internal_views	referral_fractions.total_views	referral_fractions.internal_fraction
Eagles_(band)	Eagles_discography	16815	13799	1.2185665627944053
Eagles_(band)	2008_Universal_Studios_fire	26158	22063	1.185604858813398
Eagles_(band)	Deep_Purple	106701	98931	1.0785395881978348
Eagles_(band)	Earth,_Wind_&_Fire	103052	99649	1.0341498660297646
Eagles_(band)	Led_Zeppelin	242143	239290	1.0119227715324501
Eagles_(band)	Yes_(band)	81239	80624	1.0076280015876167
Eagles_(band)	Emerson,_Lake_&_Palmer	34792	34941	0.9957356686986635
Eagles_(band)	Fleetwood_Mac	262893	274440	0.9579252295583734
Eagles_(band)	Grammy_Award_for_Album_of_the_Year	38177	40811	0.9354585773443435
Eagles_(band)	Crosby,_Stills,_Nash_&_Young	55053	58977	0.93465588280177

Hotel_California -> Eagles_(band) (95.4%) -> Emerson,_Lake_&_Palmer (99.6%)

3. What series of Wikipedia articles, starting with [Hotel California](), keeps the largest fraction of its readers clicking on internal links?

```
SELECT *
FROM REFERRAL_FRACTIONS
WHERE REFERRER='Emerson,_Lake_&_Palmer' AND INTERNAL_FRACTION < 1
ORDER BY INTERNAL_FRACTION DESC
LIMIT 10;
```

referral_fractions.referrer	referral_fractions.referred	referral_fractions.internal_views	referral_fractions.total_views	referral_fractions.internal_fraction
Emerson,_Lake_&_Palmer	Atomic_Rooster	7230	7542	0.95863166268842
Emerson,_Lake_&_Palmer	King_Crimson	54767	63808	0.8583093029087262
Emerson,_Lake_&_Palmer	Asia_(band)	22870	28514	0.8020621449112717
Emerson,_Lake_&_Palmer	The_Nice	3897	5743	0.6785652098206513
Emerson,_Lake_&_Palmer	Moog_Music	1634	2617	0.6243790599923577
Emerson,_Lake_&_Palmer	Emerson,_Lake_&_Powell	1418	2305	0.6151843817787419
Emerson,_Lake_&_Palmer	H._R._Giger	20205	36074	0.5600986860342629
Emerson,_Lake_&_Palmer	The_Crazy_World_of_Arthur_Brown	3025	5456	0.554354838709677
Emerson,_Lake_&_Palmer	Welcome_Back_My_Friends_to_the_Show_That_Never_Ends_-_Ladies_and_Gentlemen	1366	2765	0.49403254972875227
Emerson,_Lake_&_Palmer	Progressive_rock	20568	42038	0.4892716113992102

Hotel_California -> Eagles_(band) (95.4%) -> Emerson,_Lake_&_Palmer (99.6%) -> Atomic_Rooster (95.9%)

3. What series of Wikipedia articles, starting with [Hotel California](#), keeps the largest fraction of its readers clicking on internal links?

```
SELECT *
FROM REFERRAL_FRACTIONS
WHERE REFERRER='Atomic_Rooster' AND INTERNAL_FRACTION < 1
ORDER BY INTERNAL_FRACTION DESC
LIMIT 10;
```

referral_fractions.referrer	referral_fractions.referred	referral_fractions.internal_views	referral_fractions.total_views	referral_fractions.internal_fraction
Atomic_Rooster	Emerson,_Lake_&_Palmer	34792	34941	0.9957356686986635
Atomic_Rooster	Colosseum_(band)	4822	5122	0.9414291292463881
Atomic_Rooster	Cactus_(American_band)	3699	5200	0.7113461538461539
Atomic_Rooster	The_Crazy_World_of_Arthur_Brown	3025	5456	0.5544354838709677
Atomic_Rooster	Hard_Stuff	331	599	0.5525876460767947
Atomic_Rooster	Dexys_Midnight_Runners	13884	25479	0.5449193453432238
Atomic_Rooster	Homework_(Atomic_Rooster_album)	119	221	0.5384615384615384
Atomic_Rooster	List_of_Atomic_Rooster_members	306	609	0.5024630541871922
Atomic_Rooster	Made_in_England_(Atomic_Rooster_album)	514	1031	0.498545101842871
Atomic_Rooster	Death_Walks_Behind_You	1011	2126	0.4755409219190969

Hotel_California -> Eagles_(band) (95.4%) -> Emerson,_Lake_&_Palmer (99.6%) -> Atomic_Rooster (95.9%) -> Emerson,_Lake_&_Palmer

4. Find an example of an English Wikipedia article that is relatively more popular in the UK. Find the same for the US and Australia.

4. Find an example of an English Wikipedia article that is relatively more popular in the UK. Find the same for the US and Australia.

Data Used:

Pageviews during peak internet traffic hours (7 – 9 PM)

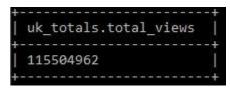Used 1 week of pageview data from Sept. 21 - 25 (Mon – Fri)

Assumptions:

Peak internet traffic hours is similar in these countries

Traffic during peak hours is representative of traffic all the time

Wikipedia traffic during peak hours is representative as well

4. Find an example of an English Wikipedia article that is relatively more popular in the UK. Find the same for the US and Australia.

uk_views.page	uk_views.views
!	18
!!	7
!!!	92
!!!!!!!!	4
!!!F██████_You!!!	5
!!!F██████_You!!!_And_Then_Some	2
!!!_(!!!_album)	1
!!!_(American_band)	1
!!!_(Chk_Chk_Chk)	9
!!!_(album)	12

```
INSERT OVERWRITE DIRECTORY 'user/hive/uk_total'
SELECT SUM(VIEWS) FROM UK_VIEWS;
```

uk_totals.total_views
115504962

Sum of views during UK peak hours

4. Find an example of an English Wikipedia article that is relatively more popular in the UK. Find the same for the US and Australia.

```
INSERT OVERWRITE DIRECTORY 'user/hive/uk_percentage'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
SELECT PAGE, VIEWS, VIEWS / TOTAL_VIEWS * 100
FROM UK_VIEWS, UK_TOTALS
ORDER BY VIEWS / TOTAL_VIEWS DESC;
```

```
+----------------------------+---------------------+--------------------------------+
|     uk_percentage.page     | uk_percentage.views | uk_percentage.view_percentage  |
+----------------------------+---------------------+--------------------------------+
| Shooting_of_Breonna_Taylor | 251305              | 0.2175707395150695             |
| Amy_Coney_Barrett          | 173273              | 0.15001346868544055            |
| Ruth_Bader_Ginsburg        | 154539              | 0.13379425205992448            |
| Ratched_(TV_series)        | 124418              | 0.10771658450482846            |
| Dennis_Nilsen              | 119860              | 0.10377043368924704            |
| Enola_Holmes_(film)        | 115567              | 0.10005371024666455            |
| Rosemary_West              | 107911              | 0.09342542357617502            |
| Fred_West                  | 104913              | 0.0908298640884363             |
| Millie_Bobby_Brown         | 78851               | 0.06826633127674636            |
| S._P._Balasubrahmanyam     | 66924               | 0.05794036796445161            |
+----------------------------+---------------------+--------------------------------+
```

Percentage of total views per article

4. Find an example of an English Wikipedia article that is relatively more popular in the UK. Find the same for the US and Australia.

```
INSERT OVERWRITE DIRECTORY 'user/hive/uk_us'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
SELECT UK_PERCENTAGE.PAGE,
       UK_PERCENTAGE.VIEW_PERCENTAGE,
       US_PERCENTAGE.VIEW_PERCENTAGE,
       (UK_PERCENTAGE.VIEW_PERCENTAGE - US_PERCENTAGE.VIEW_PERCENTAGE)
FROM UK_PERCENTAGE JOIN US_PERCENTAGE
ON UK_PERCENTAGE.PAGE = US_PERCENTAGE.PAGE
ORDER BY (UK_PERCENTAGE.VIEW_PERCENTAGE - US_PERCENTAGE.VIEW_PERCENTAGE) DESC;
```

uk_us.page	uk_us.uk_view_percent	uk_us.us_view_percent	uk_us.difference
Dennis_Nilsen	0.10377043368924704	0.006216127892937832	0.09755430579630921
Rosemary_West	0.09342542357617502	0.001891453646464705	0.09153396992971032
Fred_West	0.0908298640884363	0.0030371035823942836	0.08779276050604201
The_7.39	0.043744441039684515	5.866056763502978E-4	0.043157835363334215
Moors_murders	0.03501321441065017	0.0017240283236915133	0.033289186086958654
Janet_Leach_(appropriate_adult)	0.031143250798177834	7.45392346105708E-4	0.030397858452072126
2020-21_UEFA_Europa_League	0.033051393930591484	0.0056718927528176056	0.02737950117777388
Appropriate_Adult	0.02712870465253259	4.619622542365949E-4	0.026666742398295997
David_Morrissey	0.021270081886179053	7.820037803239243E-4	0.02048807810585513
Rory_Delap	0.02311242697954396	0.002925212457592544	0.020187214521951418

UK Page Views vs US Page Views

Dennis_Nilsen comparison

$$\frac{0.10377}{0.006216} = 16.694$$

4. Find an example of an English Wikipedia article that is relatively more popular in the UK. Find the same for the US and Australia.

```
INSERT OVERWRITE DIRECTORY 'user/hive/aus_us'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
SELECT AUS_PERCENTAGE.PAGE,
       AUS_PERCENTAGE.VIEW_PERCENTAGE,
       US_PERCENTAGE.VIEW_PERCENTAGE,
       (AUS_PERCENTAGE.VIEW_PERCENTAGE - US_PERCENTAGE.VIEW_PERCENTAGE)
FROM AUS_PERCENTAGE JOIN US_PERCENTAGE
ON AUS_PERCENTAGE.PAGE = US_PERCENTAGE.PAGE
ORDER BY (AUS_PERCENTAGE.VIEW_PERCENTAGE - US_PERCENTAGE.VIEW_PERCENTAGE) DESC;
```

aus_us.page	aus_us.aus_view_percent	aus_us.us_view_percent	aus_us.difference
S._P._Balasubrahmanyam	0.32380010645457113	0.075400220361343	0.24839988609322813
Dean_Jones_(cricketer)	0.07238597143434429	0.014293679828992843	0.05809229160535145
F5_Networks	0.06548221326297168	0.007738999010643865	0.05774321425232782
List_of_Bollywood_actresses	0.03801161897189277	0.0018223444223224537	0.03618927454957032
Liu_Chuyu	0.03252593809548109	9.021221937140273E-4	0.03162381590176706
List_of_Indian_film_actresses	0.030716735526306237	0.001958917412889283	0.028757818113416954
Dennis_Nilsen	0.03471286443835748	0.006216127892937832	0.028496736545419645
S._P._Charan	0.03450588570822143	0.010560136267593613	0.023945749440662782
Agha_Mohammad_Khan_Qajar	0.021654343068694546	0.001404398162033602	0.020249944906660943
Anurag_Kashyap	0.025761158447940854	0.007182422937641094	0.01857873551029976

Australia Page Views vs US Page Views

Dean_Jones_(cricketer)

$$\frac{0.07238}{0.01429} = 5.065$$

List of Bollywood actresses

$$\frac{0.03801}{0.001822} = 20.86$$

4. Find an example of an English Wikipedia article that is relatively more popular in the UK. Find the same for the US and Australia.

```
+------------------------------+--------------------------+--------------------------+--------------------------+
|         us_uk.page           |   us_uk.us_view_percent  |   us_uk.uk_view_percent  |     us_uk.difference     |
+------------------------------+--------------------------+--------------------------+--------------------------+
| Tyler_Herro                  | 0.12242205419484475      | 0.0230630784502574       | 0.09935897574458735      |
| Amy_Coney_Barrett            | 0.2351535965258859       | 0.15001346868544055      | 0.08514012784044536      |
| Shooting_of_Breonna_Taylor   | 0.2894347784453299       | 0.2175707395150695       | 0.07186403893026039      |
| Ryan_Fitzpatrick             | 0.04377534575653362      | 0.006620494797444286     | 0.03715485095908933      |
| Jamal_Murray                 | 0.041187452131063965     | 0.006094976248726007     | 0.03509247588233796      |
| Chrishell_Stause             | 0.041239284049170655     | 0.007493184578511873     | 0.033746099470658784     |
| Anne_Heche                   | 0.03684673467184353      | 0.004292456284259026     | 0.032554278387584505     |
| Darren_Waller                | 0.03639793607934831      | 0.0040786126573506       | 0.03231932342199771      |
| Allegiant_Stadium            | 0.03312059567017382      | 0.0053755266375482644    | 0.02774506903262556      |
| The_Killers                  | 0.02973424368720352      | 0.0021375705054125726    | 0.02759667318179095      |
+------------------------------+--------------------------+--------------------------+--------------------------+
```

US Page Views vs UK Page Views

```
+------------------------------+--------------------------+--------------------------+--------------------------+
|         us_aus.page          |  us_aus.us_view_percent  |  us_aus.aus_view_percent |     us_aus.difference    |
+------------------------------+--------------------------+--------------------------+--------------------------+
| Amy_Coney_Barrett            | 0.2351535965258859       | 0.060953995145390814     | 0.1741996013804951       |
| Shooting_of_Breonna_Taylor   | 0.2894347784453299       | 0.14866383788786555      | 0.14077094055746434      |
| Tyler_Herro                  | 0.12242205419484475      | 0.03383829244195526      | 0.08858376175288948      |
| Ratched_(TV_series)          | 0.10891737134147594      | 0.06547178987368425      | 0.04344558146779169      |
| Ryan_Fitzpatrick             | 0.04377534575653362      | 0.0070129555829551715    | 0.03676239017357845      |
| Chrishell_Stause             | 0.041239284049170655     | 0.004769941478675008     | 0.036469342570495646     |
| Anne_Heche                   | 0.03684673467184353      | 0.0028078625332845497    | 0.03403887213855898      |
| Darren_Waller                | 0.03639793607934831      | 0.004595721972013726     | 0.0318022141073459       |
| Jamal_Murray                 | 0.041187452131063965     | 0.010188118500653822     | 0.030999333630410145     |
| Allegiant_Stadium            | 0.03312059567017382      | 0.004713357365400404     | 0.028407238304773416     |
+------------------------------+--------------------------+--------------------------+--------------------------+
```

US Page Views vs Australia Page Views

vs UK

vs Australia (+ India)

Tyler Herro

$$\frac{0.1224}{0.02306} = 5.307 \qquad \frac{0.1224}{0.03383} = 3.618$$

Amy Coney Barrett

$$\frac{0.2351}{0.1500} = 1.567 \qquad \frac{0.2351}{0.06095} = 3.857$$

Shooting of Breonna Taylor

$$\frac{0.2894}{0.2175} = 1.33 \qquad \frac{0.2894}{0.1486} = 1.947$$

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

Data used:

MediaWiki enwiki history up to October 2020

Pageviews for October 20, 2020

Assumptions:

All revisions that were reverted were vandalizations

Total views over a day are representative of normal traffic

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

```
CREATE EXTERNAL TABLE WIKIHISTORY
(WIKI_DB STRING,
    EVENT_ENTITY STRING,
    EVENT_TYPE STRING,
    EVENT_TIMESTAMP STRING,
    EVENT_COMMENT STRING,
    EVENT_USER_ID BIGINT,
    EVENT_USER_TEXT_HISTORICAL STRING,
    EVENT_USER_TEXT STRING,
    EVENT_USER_BLOCKS_HISTORICAL ARRAY<STRING>,
    EVENT_USER_BLOCKS ARRAY<STRING>,
    EVENT_USER_GROUPS_HISTORICAL ARRAY<STRING>,
    EVENT_USER_GROUPS ARRAY<STRING>,
    EVENT_USER_IS_BOT_BY_HISTORICAL ARRAY<STRING>,
    EVENT_USER_IS_BOT_BY ARRAY<STRING>,
    EVENT_USER_IS_CREATED_BY_SELF BOOLEAN,
    EVENT_USER_IS_CREATED_BY_SYSTEM BOOLEAN,
    EVENT_USER_IS_CREATED_BY_PEER BOOLEAN,
    EVENT_USER_IS_ANONYMOUS BOOLEAN,
    EVENT_USER_REGISTRATION_TIMESTAMP STRING,
    EVENT_USER_CREATION_TIMESTAMP STRING,
    EVENT_USER_FIRST_EDIT_TIMESTAMP STRING,
    EVENT_USER_REVISION_COUNT BIGINT,
    EVENT_USER_SECONDS_SINCE_PREVIOUS_REVISION BIGINT,
    PAGE_ID BIGINT,
    PAGE_TITLE_HISTORICAL STRING,
    PAGE_TITLE STRING,
    PAGE_NAMESPACE_HISTORICAL INT,
    PAGE_NAMESPACE_IS_CONTENT_HISTORICAL BOOLEAN,
    PAGE_NAMESPACE INT,
    PAGE_NAMESPACE_IS_CONTENT BOOLEAN,
    PAGE_IS_REDIRECT BOOLEAN,
    PAGE_IS_DELETED BOOLEAN,
    PAGE_CREATION_TIMESTAMP STRING,
    PAGE_FIRST_EDIT_TIMESTAMP STRING,
    PAGE_REVISION_COUNT BIGINT,
    PAGE_SECONDS_SINCE_PREVIOUS_REVISION BIGINT,
    USER_ID BIGINT,
    USER_TEXT_HISTORICAL STRING,
    USER_TEXT STRING,
    USER_BLOCKS_HISTORICAL ARRAY<STRING>,
    USER_BLOCKS ARRAY<STRING>,
    USER_GROUPS_HISTORICAL ARRAY<STRING>,
    USER_GROUPS ARRAY<STRING>,
    USER_IS_BOT_BY_HISTORICAL ARRAY<STRING>,
    USER_IS_BOT_BY ARRAY<STRING>,
    USER_IS_CREATED_BY_SELF BOOLEAN,
    USER_IS_CREATED_BY_SYSTEM BOOLEAN,
    USER_IS_CREATED_BY_PEER BOOLEAN,
    USER_IS_ANONYMOUS BOOLEAN,
    USER_REGISTRATION_TIMESTAMP STRING,
    USER_CREATION_TIMESTAMP STRING,
    USER_FIRST_EDIT_TIMESTAMP STRING,
    REVISION_ID BIGINT,
    REVISION_PARENT_ID BIGINT,
    REVISION_MINOR_EDIT BOOLEAN,
    REVISION_DELETED_PARTS ARRAY<STRING>,
    REVISION_DELETED_PARTS_ARE_SUPPRESSED BOOLEAN,
    REVISION_TEXT_BYTES BIGINT,
    REVISION_TEXT_BYTES_DIFF BIGINT,
    REVISION_TEXT_SHA1 STRING,
    REVISION_CONTENT_MODEL STRING,
    REVISION_CONTENT_FORMAT STRING,
    REVISION_IS_DELETED_BY_PAGE_DELETION BOOLEAN,
    REVISION_DELETED_BY_PAGE_DELETION_TIMESTAMP STRING,
    REVISION_IS_IDENTITY_REVERTED BOOLEAN,
    REVISION_FIRST_IDENTITY_REVERTING_REVISION_ID BIGINT,
    REVISION_SECONDS_TO_IDENTITY_REVERT BIGINT,
    REVISION_IS_IDENTITY_REVERT BOOLEAN,
    REVISION_IS_FROM_BEFORE_PAGE_CREATION BOOLEAN,
    REVISION_TAGS ARRAY<STRING>)
```

REVISION_SECONDS_TO_IDENTITY_REVERT BIGINT

5. Analyze how many users will see the average vandalized Wikipedia page before the offending edit is reversed.
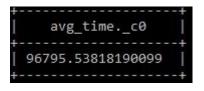
```
CREATE TABLE REVISIONS_SECONDS AS
SELECT PAGE_TITLE, REVISION_SECONDS_TO_IDENTITY_REVERT
WHERE REVISION_SECONDS_TO_IDENTITY_REVERT > 0
FROM WIKIHISTORY;
```

revisions_seconds.page_title	revisions_seconds.revision_seconds_to_identity_revert
Tityus_serrulatus	591
Sandbox	1168
Michael_Matricciani	2552
Werner_Fischer_(sailor)	35326
The_Little_Match_Girl	97
Eurodog/sandbox266	51
Xa	194
DannyS712_bot_III/Redirects.json	874
Main_Page	17
Maxie_the_Fox/sandbox	12

Time (in seconds) before reverting change on page

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

```
CREATE TABLE AVG_TIME AS
SELECT AVG(REVISION_SECONDS_TO_IDENTITY_REVERT)
FROM REVISIONS_SECONDS
WHERE REVISION_SECONDS_TO_IDENTITY_REVERT > 0;
```
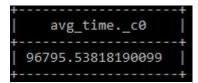
```
+---------------------+
|    avg_time._c0     |
+---------------------+
| 96795.53818190099   |
+---------------------+
```

Average time before
vandalization is reverted

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

```
CREATE TABLE AVG_TIME AS
SELECT AVG(REVISION_SECONDS_TO_IDENTITY_REVERT)
FROM REVISIONS_SECONDS
WHERE REVISION_SECONDS_TO_IDENTITY_REVERT > 0;
```

```
SELECT AVG(VIEWS) AS AVG_VIEWS
FROM PAGEVIEWS_20_10_20;
```

```
+---------------------+
|    avg_time._c0     |
+---------------------+
|  96795.53818190099  |
+---------------------+
```

Average time before vandalization is reverted

```
+---------------------+
|     avg_views       |
+---------------------+
|  38.663661475679675 |
+---------------------+
```

Average views an article receives per day

5. Analyze how many users will see the average vandalized Wikipedia page before the offending edit is reversed.
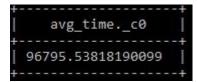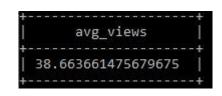
```
CREATE TABLE AVG_TIME AS
SELECT AVG(REVISION_SECONDS_TO_IDENTITY_REVERT)
FROM REVISIONS_SECONDS
WHERE REVISION_SECONDS_TO_IDENTITY_REVERT > 0;
```

```
SELECT AVG(VIEWS) AS AVG_VIEWS
FROM PAGEVIEWS_20_10_20;
```

```
+---------------------+
|    avg_time._c0     |
+---------------------+
| 96795.53818190099   |
+---------------------+
```
Average time before
vandalization is reverted

```
+---------------------+
|     avg_views       |
+---------------------+
| 38.663661475679675  |
+---------------------+
```
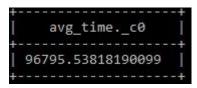Average views an article
receives per day

$$\frac{96795.538\ seconds}{revert}\ x\ \frac{1\ minute}{60\ seconds}\ x\ \frac{1\ hour}{60\ minutes}\ x\ \frac{1\ day}{24\ hours}\ =\ \frac{1.12\ days}{revert}$$

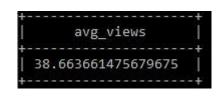# 5. Analyze how many users will see the average vandalized Wikipedia page before the offending edit is reversed.

```
CREATE TABLE AVG_TIME AS
SELECT AVG(REVISION_SECONDS_TO_IDENTITY_REVERT)
FROM REVISIONS_SECONDS
WHERE REVISION_SECONDS_TO_IDENTITY_REVERT > 0;
```

```
SELECT AVG(VIEWS) AS AVG_VIEWS
FROM PAGEVIEWS_20_10_20;
```

```
+---------------------+
|    avg_time._c0     |
+---------------------+
| 96795.53818190099   |
+---------------------+
```
Average time before
vandalization is reverted

```
+---------------------+
|     avg_views       |
+---------------------+
| 38.663661475679675  |
+---------------------+
```
Average views an article
receives per day

$$\frac{96795.538 \; seconds}{revert} \; x \; \frac{1 \; minute}{60 \; seconds} \; x \; \frac{1 \; hour}{60 \; minutes} \; x \; \frac{1 \; day}{24 \; hours} = \frac{1.12 \; days}{revert}$$

$$\frac{1.12 \; days}{revert} \; x \; \frac{38.663 \; views}{day} = \frac{43.303 \; views}{revert}$$

6. Which popular English Wikipedia pages have the lowest percentage of people click an internal link?

6. Which popular English Wikipedia pages have the lowest percentage of people click an internal link?

Data Used:

Clickstream for the month of Sept.

All Pageviews for the month of Sept.

6. Which popular English Wikipedia pages have the lowest percentage of people click an internal link?

```
SELECT PAGE, INTERNAL_VIEWS, TOTAL_VIEWS, ROUND(INTERNAL_FRACTION * 100, 2) AS INTERNAL_PERCENTAGE
FROM INTERNAL_FRACTION
WHERE INTERNAL_FRACTION < 1 AND TOTAL_VIEWS > 500000
ORDER BY INTERNAL_PERCENTAGE ASC
LIMIT 10;
```

```
+------------------+---------------+--------------+---------------------+
|       page       | internal_views | total_views  | internal_percentage |
+------------------+---------------+--------------+---------------------+
| F5_Networks      | 513           | 1487955      | 0.03                |
| Flag_of_Scotland | 3815          | 527267       | 0.72                |
| Bible            | 32110         | 3170711      | 1.01                |
| Microsoft_Office | 23379         | 2136261      | 1.09                |
| Gmail            | 6667          | 508260       | 1.31                |
| Google_Classroom | 11300         | 790056       | 1.43                |
| 123Movies        | 8201          | 573309       | 1.43                |
| Main_Page        | 2379287       | 165044119    | 1.44                |
| XXXX             | 49387         | 2056847      | 2.4                 |
| The_Pirate_Bay   | 14763         | 570331       | 2.59                |
+------------------+---------------+--------------+---------------------+
```

10 popular pages with lowest percentage

Questions?

GitHub: https://github.com/ChrisChee1/RevatureProject1