Project 1

Kyle M. Pacheco

November 6, 2020

Contents

Question 1

Question 2

Question 3

Question 4

Question 5

Question 6

Repository

Question 1

Question 1

Methods

Question 1

Methods

1. Get all pageview data from October 20th on Wikipedia

Question 1

Methods

1. Get all pageview data from October 20th on Wikipedia
2. Filter each entry by prefix ‘en’

Question 1

Methods

1. Get all pageview data from October 20th on Wikipedia
2. Filter each entry by prefix ‘en’
3. Consolidate entries

Question 1

Methods

1. Get all pageview data from October 20th on Wikipedia
2. Filter each entry by prefix ‘en’
3. Consolidate entries

Results:

Page Name	Page Views
Main_Page	2,726,387
Special:Search	910,309
Bible	148,726
–	124,890
Jeffrey_Toobin	116,724

Question 2

Question 2

Methods

1. Get clickstream and pageview data for all of September 2020

Question 2

Methods

1. Get clickstream and pageview data for all of September 2020
2. Filter all clickstream data where type='link'

Question 2

Methods

1. Get clickstream and pageview data for all of September 2020
2. Filter all clickstream data where type=‘link’
3. Filter and consolidate page views

Question 2

Methods

1. Get clickstream and pageview data for all of September 2020
2. Filter all clickstream data where type=‘link’
3. Filter and consolidate page views
4. Join tables on page name and calculate percentages

Question 2

Methods

1. Get clickstream and pageview data for all of September 2020
2. Filter all clickstream data where type=‘link’
3. Filter and consolidate page views
4. Join tables on page name and calculate percentages

Results:

No Filter On Page Views			
Page Name	Total Views	Links Clicked	Percentage
/r/	1	64	6400%
/\	2	56	2800%
Health//Disco	8	209	2612.5%

More Results

Page Views>10,000			
Page Name	Total Views	Links Clicked	Percentage
List_of_controversial_album_art	11271	47953	425.45%
List_of_common_World_War_II_infantry_weapons	31097	108981	350.46%
List_of_murdered_American_children	20578	71761	348.73%

More Results

Page Views>10,000			
Page Name	Total Views	Links Clicked	Percentage
List_of_controversial_album_art	11271	47953	425.45%
List_of_common_World_War_II_infantry_weapons	31097	108981	350.46%
List_of_murdered_American_children	20578	71761	348.73%

Page Views>100,000			
Page Name	Total Views	Links Clicked	Percentage
List_of_pornographic_performers_by_decade	135742	467454	344.37%
List_of_serial_killers_in_the_United_States	185479	420780	226.86%
List_of_PlayStation_5_games	100694	163494	162.37%

More Results

Page Views>10,000			
Page Name	Total Views	Links Clicked	Percentage
List_of_controversial_album_art	11271	47953	425.45%
List_of_common_World_War_II_infantry_weapons	31097	108981	350.46%
List_of_murdered_American_children	20578	71761	348.73%

Page Views>100,000			
Page Name	Total Views	Links Clicked	Percentage
List_of_pornographic_performers_by_decade	135742	467454	344.37%
List_of_serial_killers_in_the_United_States	185479	420780	226.86%
List_of_PlayStation_5_games	100694	163494	162.37%

Page Views>1,000,000			
Page Name	Total Views	Links Clicked	Percentage
Dune_(2020_film)	1278838	1201459	93.95%
Cobra_Kai	2459988	2241751	91.13%
COVID-19_pandemic_by_country_and_territory	1207880	1093321	90.52%

Question 3

Question 3

Methods

1. Use clickstream data from Question 2

Question 3

Methods

1. Use clickstream data from Question 2
2. Filter clickstream data to only show rows where referrer=‘Hotel_California’

Question 3

Methods

1. Use clickstream data from Question 2
2. Filter clickstream data to only show rows where referrer=‘Hotel_California’
3. Get the top referred by occurrences

Question 3

Methods

1. Use clickstream data from Question 2
2. Filter clickstream data to only show rows where referrer=‘Hotel_California’
3. Get the top referred by occurrences
4. Filter clickstream so that the referrer is the top referred from previous step

Question 3

Methods

1. Use clickstream data from Question 2
2. Filter clickstream data to only show rows where referrer='Hotel_California'
3. Get the top referred by occurrences
4. Filter clickstream so that the referrer is the top referred from previous step
5. Repeat from step 3 until satisfied

Question 3

Methods

1. Use clickstream data from Question 2
2. Filter clickstream data to only show rows where referrer='Hotel_California'
3. Get the top referred by occurrences
4. Filter clickstream so that the referrer is the top referred from previous step
5. Repeat from step 3 until satisfied

Chain:
Hotel_California

Question 3

Methods

1. Use clickstream data from Question 2
2. Filter clickstream data to only show rows where referrer='Hotel_California'
3. Get the top referred by occurrences
4. Filter clickstream so that the referrer is the top referred from previous step
5. Repeat from step 3 until satisfied

Chain:
Hotel_California → Hotel_California_(Eagles_album)

Question 3

Methods

1. Use clickstream data from Question 2
2. Filter clickstream data to only show rows where referrer=‘Hotel_California’
3. Get the top referred by occurrences
4. Filter clickstream so that the referrer is the top referred from previous step
5. Repeat from step 3 until satisfied

Chain:
Hotel_California → Hotel_California_(Eagles_album) →
The_Long_Run_(album)

Question 3

Methods

1. Use clickstream data from Question 2
2. Filter clickstream data to only show rows where referrer=‘Hotel_California’
3. Get the top referred by occurrences
4. Filter clickstream so that the referrer is the top referred from previous step
5. Repeat from step 3 until satisfied

Chain:
Hotel_California → Hotel_California_(Eagles_album) →
The_Long_Run_(album) → Eagles_Live

Question 3

Methods

1. Use clickstream data from Question 2
2. Filter clickstream data to only show rows where referrer='Hotel_California'
3. Get the top referred by occurrences
4. Filter clickstream so that the referrer is the top referred from previous step
5. Repeat from step 3 until satisfied

Chain:
Hotel_California → Hotel_California_(Eagles_album) →
The_Long_Run_(album) → Eagles_Live →
Eagles_Greatest_Hits,_Vol._2

Question 3

Methods

1. Use clickstream data from Question 2
2. Filter clickstream data to only show rows where referrer=‘Hotel_California’
3. Get the top referred by occurrences
4. Filter clickstream so that the referrer is the top referred from previous step
5. Repeat from step 3 until satisfied

Chain:
Hotel_California → Hotel_California_(Eagles_album) →
The_Long_Run_(album) → Eagles_Live →
Eagles_Greatest_Hits,_Vol._2 → The_Very_Best_of_the_Eagles

Question 3

Methods

1. Use clickstream data from Question 2
2. Filter clickstream data to only show rows where referrer=‘Hotel_California’
3. Get the top referred by occurrences
4. Filter clickstream so that the referrer is the top referred from previous step
5. Repeat from step 3 until satisfied

Chain:
Hotel_California → Hotel_California_(Eagles_album) → The_Long_Run_(album) → Eagles_Live → Eagles_Greatest_Hits,_Vol._2 → The_Very_Best_of_the_Eagles → Hell_Freezes_Over

Question 4

Assumptions

1. Peak internet usage occurs between 7pm-11pm in each region

Question 4

Assumptions

1. Peak internet usage occurs between 7pm-11pm in each region
2. Using a sub-interval of those hours will result in similar usage across regions

Question 4

Assumptions

1. Peak internet usage occurs between 7pm-11pm in each region
2. Using a sub-interval of those hours will result in similar usage across regions
3. 90%,96%, and 88% of the population have a broadband internet connection in the US, UK, and Australia respectively

Question 4

Assumptions

1. Peak internet usage occurs between 7pm-11pm in each region
2. Using a sub-interval of those hours will result in similar usage across regions
3. 90%,96%, and 88% of the population have a broadband internet connection in the US, UK, and Australia respectively
4. 91% of Australian population lives on either East or West coast

Question 4

Assumptions

1. Peak internet usage occurs between 7pm-11pm in each region
2. Using a sub-interval of those hours will result in similar usage across regions
3. 90%,96%, and 88% of the population have a broadband internet connection in the US, UK, and Australia respectively
4. 91% of Australian population lives on either East or West coast

Methods

1. Convert 7pm-9pm for each region into UTC and get corresponding pageview data from Wikipedia

Question 4

Assumptions

1. Peak internet usage occurs between 7pm-11pm in each region
2. Using a sub-interval of those hours will result in similar usage across regions
3. 90%,96%, and 88% of the population have a broadband internet connection in the US, UK, and Australia respectively
4. 91% of Australian population lives on either East or West coast

Methods

1. Convert 7pm-9pm for each region into UTC and get corresponding pageview data from Wikipedia
2. Consolidate, filter by prefix ‘en’, and normalize by population (per million) for each region

Results

Results

Claim 1

Taskmaster (TV Series) is relatively more popular in the UK than in the US.

Results

Claim 1

Taskmaster (TV Series) is relatively more popular in the UK than in the US.

	US Page Views	UK Page Views
Before Normalizing	11916	9198
After Normalizing	19.97	143.75

Results

Claim 1

Taskmaster (TV Series) is relatively more popular in the UK than in the US.

	US Page Views	UK Page Views
Before Normalizing	11916	9198
After Normalizing	19.97	143.75

Claim 2

Marmite is relatively more popular in Australia than in the US.

Results

Claim 1

Taskmaster (TV Series) is relatively more popular in the UK than in the US.

	US Page Views	UK Page Views
Before Normalizing	11916	9198
After Normalizing	19.97	143.75

Claim 2

Marmite is relatively more popular in Australia than in the US.

	US Page Views	AUS Page Views
Before Normalizing	3552	3858
After Normalizing	5.95	96.39

Question 5

Methods

1. Get revisions and pageviews history from Wikipedia

Question 5

Methods

1. Get revisions and pageviews history from Wikipedia
2. Filter revisions so that revision_seconds_to_identity_revert is a positive integer

Question 5

Methods

1. Get revisions and pageviews history from Wikipedia
2. Filter revisions so that revision_seconds_to_identity_revert is a positive integer
3. Join with pageviews on page_title

Question 5

Methods

1. Get revisions and pageviews history from Wikipedia
2. Filter revisions so that revision_seconds_to_identity_revert is a positive integer
3. Join with pageviews on page_title
4. Get average pageviews and average revision_seconds_to_identity_revert

Question 5

Methods

1. Get revisions and pageviews history from Wikipedia
2. Filter revisions so that revision_seconds_to_identity_revert is a positive integer
3. Join with pageviews on page_title
4. Get average pageviews and average revision_seconds_to_identity_revert

Average seconds to revert a revision: 81687.76

Average pageviews for vandalized pages in September 2020: 38499.08

Question 5

Methods

1. Get revisions and pageviews history from Wikipedia
2. Filter revisions so that revision_seconds_to_identity_revert is a positive integer
3. Join with pageviews on page_title
4. Get average pageviews and average revision_seconds_to_identity_revert

Average seconds to revert a revision: 81687.76

Average pageviews for vandalized pages in September 2020: 38499.08

Result: A vandalized page gets 1213.31 views on average before being reverted.

Question 6: Most Popular English Lists On Wikipedia In September

Methods

1. Filter and consolidate pageviews from English Wikipedia

Question 6: Most Popular English Lists On Wikipedia In September

Methods

1. Filter and consolidate pageviews from English Wikipedia
2. Find pages whose page names begin with 'List_of'

Question 6: Most Popular English Lists On Wikipedia In September

Methods

1. Filter and consolidate pageviews from English Wikipedia
2. Find pages whose page names begin with ‘List_of’

Results

Page Name	Page Views
List_of_Marvel_Cinematic_Universe_films	852,758
List_of_presidents_of_the_United_States	756,816
List_of_James_Bond_films	650,084
List_of_justices_of_the_Supreme_Court_of_the_United_States	623,624
List_of_The_Boys_characters	574,700

https://github.com/KylePacheco1021/PJ1_Pacheco