



# University of Victoria

**Faculty of Engineering**

**CSC 370**  
**Assignment 4**

Submitted by Team: Kaizen Blitz

Ryan Woodward - V00857268  
Brendan Ciccone - V00871008  
Sarah Sparrow - V00819941

# Table of Contents

## [1. Github Repositories Dataset](#)

- [1.1. What repos have C files?](#)
- [1.2. What repo had file with biggest size?](#)
- [1.3. What are the top licenses used?](#)
- [1.4. What are the top 10 most watched repos?](#)

## [2. Wikipedia Pageviews Dataset](#)

- [2.1. What day in the beginning of October 2018 has the highest amount of views on the page "Halloween"?](#)
- [2.2. What day in the beginning from November to December 2018 has the highest amount of views on the page "Anxiety"?](#)
- [2.3. Between Nov 1st and Nov 13th, 2018, what pages with titles like "Battle of" were viewed the most?](#)
- [2.4. What days in January 2018 did the Main Page have the most views](#)
- [2.5. What was the avg hourly views on the page "Pi Day" on March 14, 2018](#)


## [3. Google Patents Public Dataset](#)

- [3.1. How many patents have been filed?](#)
- [3.2. How many patents has each country filed? With a minimum of 100 patents filed](#)
- [3.3. How many patents did each country file in the year 2018](#)
- [3.4. What has been the number of patents Canada put out each year?](#)






# 1. Github Repositories Dataset



The following queries were done on the Github Repositories dataset which holds contents from 2.9M public, open source licensed repositories on GitHub.

## 1.1. What repos have C files?

 2019-11-21 11:48:48 Edited


```
1 SELECT repo_name, language
2 FROM `bigquery-public-data.github_repos.languages`
3 CROSS JOIN UNNEST(language) as single_language
4 WHERE single_language.name = "C"
```

 Run  Save query  Save view  Schedule query  More

Query results  SAVE RESULTS  EXPLORE WITH DATA STUDIO

Query complete (9.9 sec elapsed, 195.7 MB processed)

Job information Results JSON Execution details

 Some repeated values have been hidden to improve performance.

Row	repo_name	language.name	language.bytes
1	Siorn/school	C	9567
2	fasaxc/boost-controller	C	3280
3	macpod/lasersharklib	C	18294

## 1.2. What repo had file with biggest size?

Query editor

```
1 SELECT repo_name
2 FROM `bigquery-public-data.github_repos.files`
3 WHERE id in (SELECT R1.id
4               FROM `bigquery-public-data.github_repos.contents` R1
5               WHERE R1.size = (SELECT MAX(R2.size)
6                                FROM `bigquery-public-data.github_repos.contents` R2
7                                LIMIT 1000
8                                )
9               )
```

Run

Save query

Save view

Schedule query

More

Query results

SAVE RESULTS

EXPLORE WITH DATA STUDIO

Query complete (11.1 sec elapsed, 152.3 GB processed)

Job information

Results

JSON

Execution details

Row	repo_name
1	tamilarasi/linux-kernel-beaglebone

### 1.3. What are the top licenses used?

Query editor

```
1 SELECT license, COUNT(license) as license_num
2 FROM `bigquery-public-data.github_repos.licenses`
3 GROUP BY license
4 ORDER BY license_num DESC
```

Run

Save query

Save view

Schedule query

More

Query results

SAVE RESULTS

EXPLORE WITH DATA STUDIO

Query complete (1.6 sec elapsed, 24.8 MB processed)

Job information

Results

JSON

Execution details


Row	license	license_num
1	mit	1708987
2	apache-2.0	494089
3	gpl-2.0	345030
4	gpl-3.0	343185
5	bsd-3-clause	152379


## 1.4. What are the top 10 most watched repos?


Unsaved query


Edited


```
1 SELECT repo_name, watch_count
2 FROM `bigquery-public-data.github_repos.sample_repos`
3 ORDER BY watch_count DESC
4 LIMIT 10
```

 Run


 Save query


 Save view

 Schedule query

 More

Query results

 SAVE RESULTS

 EXPLORE WITH DATA STUDIO

Query complete (0.9 sec elapsed, 12.5 MB processed)

Job information

Results

JSON

Execution details

Row	repo_name	watch_count
1	FreeCodeCamp/FreeCodeCamp	90457
2	firehol/netdata	13208
3	joshbuckea/HEAD	13125
4	browdie/HowToBeAProgrammer	12010

## 2. Wikipedia Pageviews Dataset

The following queries were done on the Wikipedia Pageviews dataset which holds Wikipedia page views during 2018.

2.1. What day in the beginning of October 2018 has the highest amount of views on the page “Halloween”?

Query editor

```
1 SELECT *
2 FROM `bigquery-public-data.wikipedia.pageviews_2018`
3 WHERE DATE(datehour) >= "2018-10-01" AND DATE(datehour) <= "2018-10-10" AND title = "Halloween" AND views > 100
4 ORDER BY views DESC
5 LIMIT 10
```

Run

Save query

Save view

Schedule query

More

Query results

SAVE RESULTS

EXPLORE WITH DATA STUDIO

Query complete (0.3 sec elapsed, cached)

Job information

Results

JSON




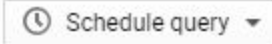

Execution details



Row	datehour	wiki	title	views
1	2018-10-08 00:00:00 UTC	en.m	Halloween	1044
2	2018-10-07 21:00:00 UTC	en.m	Halloween	916
3	2018-10-02 02:00:00 UTC	en.m	Halloween	887
4	2018-10-07 18:00:00 UTC	en.m	Halloween	886
5	2018-10-06 21:00:00 UTC	en.m	Halloween	866
6	2018-10-07 19:00:00 UTC	en.m	Halloween	857
7	2018-10-07 17:00:00 UTC	en.m	Halloween	856
8	2018-10-07 20:00:00 UTC	en.m	Halloween	846
9	2018-10-01 18:00:00 UTC	en	Halloween	836
10	2018-10-08 19:00:00 UTC	en.m	Halloween	832



2.2. What day in the beginning from November to December 2018 has the highest amount of views on the page “Anxiety”?

```
1 SELECT views, datehour
2 FROM `bigquery-public-data.wikipedia.pageviews_2018`
3 WHERE title = "Anxiety" AND DATE(datehour) >= "2018-11-01" AND views > 100
4 ORDER BY views DESC
5 LIMIT 10;
```

**Query results**  



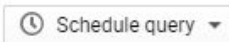

Query complete (5.2 sec elapsed, 183.1 GB processed)

[Job information](#) [Results](#) [JSON](#) [Execution details](#)

Row	views	datehour
1	531	2018-12-08 05:00:00 UTC
2	525	2018-12-09 10:00:00 UTC
3	279	2018-12-06 02:00:00 UTC
4	211	2018-12-04 23:00:00 UTC
5	184	2018-11-18 22:00:00 UTC

2.3. Between Nov 1st and Nov 13th, 2018, what pages with titles like “Battle of” were viewed the most?

```
1 SELECT DISTINCT(A.title) as title, SUM(A.views) as views
2 FROM `bigquery-public-data.wikipedia.pageviews_2018` A
3 WHERE A.title LIKE 'Battle_of_%' AND DATE(A.datehour) >= "2018-11-01" AND DATE(A.datehour) <= "2018-11-13"
4 GROUP BY title
5 ORDER BY views DESC
6 LIMIT 10;
```

 Run  Save query  Save view  Schedule query  More

Query results

 SAVE RESULTS

 EXPLORE WITH DATA STUDIO




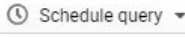

Query complete (2.7 sec elapsed, 47.7 GB processed)

Job information [Results](#) JSON Execution details

Row	title	views			
1	Battle_of_Loudoun_Hill	143120	6	Battle_of_Stalingrad	60889
2	Battle_of_the_Somme	133369	7	Battle_of_Passchendaele	60404
3	Battle_of_Bannockburn	121154	8	Battle_of_Verdun	59597
4	Battle_of_France	102585	9	Battle_of_the_Bulge	58346
5	Battle_of_Belleau_Wood	61108	10	Battle_of_Waterloo	56070

## 2.4. What days in January 2018 did the Main Page have the most views

```
1 SELECT DISTINCT(DATE(datehour)) as date, SUM(views) as views
2 FROM `bigquery-public-data.wikipedia.pageviews_2018`
3 WHERE title LIKE 'Main_Page' AND DATE(datehour) >= "2018-01-01" AND DATE(datehour) <= "2018-01-31" AND views > 10000
4 GROUP BY date
5 ORDER BY views DESC
6 LIMIT 10;
```

Query results [SAVE RESULTS](#) [EXPLORE WITH DATA STUDIO](#)

Query complete (6.5 sec elapsed, 81 GB processed)

Job information **Results** JSON Execution details


Row	date	views
1	2018-01-22	20814005
2	2018-01-15	20548520
3	2018-01-23	19688903
4	2018-01-16	19030603
5	2018-01-29	18944432

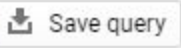
6	2018-01-31	18516841
7	2018-01-24	18490016
8	2018-01-30	18412500
9	2018-01-08	18240783
10	2018-01-27	18221161

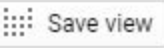
2.5. What was the avg hourly views on the page “Pi Day” on March 14, 2018

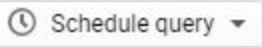
Query editor

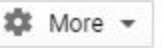
```
1 SELECT AVG(views) AS avg_views
2 FROM `bigquery-public-data.wikipedia.pageviews_2018`
3 WHERE DATE(datehour) = "2018-03-14" AND title = "Pi_Day"
```

 Run


 Save query


 Save view

 Schedule query

 More

Query results

 [SAVE RESULTS](#)

 [EXPLORE WITH DATA STUDIO](#)

Query complete (0.0 sec elapsed, cached)

Job information

[Results](#)

JSON

Execution details

Row	avg_views
1	1837.7697841726622

### 3. Google Patents Public Dataset

The following queries were executed using the Google Patents Public Dataset. The overview of this set describes it as containing data based on section 337. Section 337 declares the infringement of certain statutory intellectual property rights and other forms of unfair competition in import trade to be unlawful practices. Most Section 337 investigations involve allegations of patent or registered trademark infringement.

#### 3.1. How many patents have been filed?

Query editor

HIDE EDITOR

FULL SCREEN

```
1 SELECT COUNT(publication_number)
2 FROM `patents-public-data.patents.publications`
3 LIMIT 1000
```

Run

Save query

Save view

Schedule query

More

This query will process 1.7 GB when run.

Query results

SAVE RESULTS

EXPLORE WITH DATA STUDIO

Query complete (2.9 sec elapsed, 1.7 GB processed)

Job information

Results

JSON

Execution details

Row	f0_
1	119041383

### 3.2. How many patents has each country filed? With a minimum of 100 patents filed

```
SELECT COUNT(*) AS cnt, country_code
FROM `patents-public-data.patents.publications

GROUP BY country_code
HAVING cnt > 100
ORDER BY country_code
```

9	3241713	CA	
10	726798	CH	
11	31897	CL	
12	22734436	CN	
13	27816	CO	
14	9565	CR	
15	174257	CS	

Job information		Results	JSON	Execution details
87	1772123	TW		
88	184974	UA		
89	16802823	US		
90	13668	UY		
91	217	VN		
92	4296882	WO		
93	57072	YU		
...	.....	..		

### 3.3. How many patents did each country file in the year 2018

#### Query editor

```
1 SELECT COUNT(*) AS cnt, country_code
2 FROM `patents-public-data.patents.publications`
3 WHERE publication_date >= 20180101 AND publication_date <= 20181231
4 GROUP BY country_code
5 HAVING cnt > 100
6 ORDER BY country_code;
```

Job information

**Results**

JSON

Execution details

6	52352	BR	
7	52409	CA	
8	1791	CH	
9	4320	CL	
10	3486501	CN	
11	2628	CO	

### 3.4. What has been the number of patents Canada put out each year?

#### Query editor

```
1 SELECT COUNT(*) AS cnt, country_code, publication_date AS Year
2 FROM `patents-public-data.patents.publications`
3 WHERE country_code = 'CA'
4 GROUP BY country_code, publication_date
5 HAVING cnt > 100
6 ORDER BY publication_date;
```

Row	cnt	country_code	Year
7601	608	CA	20180308
7602	523	CA	20180313
7603	565	CA	20180315
7604	555	CA	20180320
7605	575	CA	20180322

Rows per page: 100 ▾ 7601 - 7700 of 7701 First page |<