BUSINESS ANALYST - ONLINE OPERATIONS

COPYRIGHT REMOVAL REQUESTS

Introduction

Report on Copyright Removal Requests focuses on visualizing the removal requests data and deriving meaningful insights from this data. Google receives requests from copyright owner to takedown content from Google's search results that may infringe on copyright. A request can be reported by the same copyright owner or by a reporting organisation who works on behalf of the copyright owner. If a pirated copy of a video is published on a any webpage without the consent of the copyright owner then the owner reports a takedown requests with Google to remove the content from Google search.

Dataset downloaded contains requests from March 10th 2011 to December 5th 2018. Data contains 3 spreadsheets naming requests.csv, domains.csv and urls-no-action-taken.csv. Detailing the columns in all the spreadsheets one by one:

1. Requests.csv - Contains 6,990,660 rows * 12 columns. Details are as follows:
    a. Request ID - Primary key for the requests table.
    b. Date - Timestamp when the request was reported.
    c. Lumen URL - Structured detailed information of the request. Gives information about the Copyright takedown request. Contains information whether the URLs reported are original or infringed.
    d. Copyright Owner ID - References the id unique value to every owner/organization. There are some organizations which have multiple owner ids assigned to them. Ex. Multi Media, LLC, Chaturbate LLC.
    e. Copyright Owner Name - Contains name of the organization owning the copyright for the content.
    f. Reporting Organization ID - Reporting Organization are identified as the one who report the takedown requests to Google on behalf of the Copyright Owner. Copyright organization can also be reporting organization.

g. Reporting Organization Name - States the name of the reporting organization.

h. URLs removed - It is continuous variable.Contains the number of URLs removed from Google search pertaining to the request. Before taking down the url's, a dedicated team checks the authentication of the request and validating the infringement. If a reporting organization reports to remove 100 URLs which they believe violate the copyright law then Google's team validates the request and takes down the content from Google's search. Also, from the request only if 90 URLs violate the law then those URLs are removed and the rest are treated appropriately.

i. URLs that were not in Google's search index - Contains the number of URLs which were not found in Google's search index. Continuing the above scenario, URLs that were not listed in Google's search index are classified in this category. It is continuous variable.

j. URLs which we took no action - Indicates the number of URLs from the request which were untreated for specific reason. I assume that this URL were original and thus no action of taking down was taken.

k. URLs pending review - Denotes number of URLs which are pending for review with the team. It is continuous variable.

l. From Abuser - It is boolean variable. Determines whether the request reported is from an abuser. Anyone can take disadvantage of the Digital Millennium Copyright Act(DMCA) and so it necessary to authenticate the request.

2. Domains.csv - Contains domain information about the URLs reported in the requests. 233,274,101 rows * 7 columns are present in the spreadsheet and their details are as follows

a. Request ID - Foreign key referencing the request id column from Requests.csv table. Each request id can have one or many URLs requested for take down.

b. Domain - Contains domain information for the URL.

c. URLs removed - Contains the number of URLs removed from Google search pertaining to the request. It is continuous variable.

d. URLs that were not in Google’s search index - Contains the number of URLs which were not found in Google’s search index.
e. URLs which we took no action - Indicates the number of URLs from the request which were untreated for specific reason.
f. URLs pending review - Denotes number of URLs which are pending for review with the team. It is continuous variable.
g. From Abuser - Determines whether the request reported is from an abuser.

3. URLs no action taken.csv - This spreadsheet contains information about the URLs on which no action was taken. Comprises of 128,941,708 rows * 4 columns which are detailed as follows:

a. Request ID - This column references the primary key request id from requests.csv.
b. Domain - States the domain information of the URL.
c. URL - Contains web address of the page which was requested to be taken down.
d. From Abuser - Boolean value stating whether the DMCA utility was a genuine practice or misused.

Tools Used:

1. MySQL Workbench
2. SQLLite Studio
3. Tableau
4. Jupyter Notebook

Approach:

1. After downloading the data, README.txt file helped me understand how the data would be structured.
2. Data being very large it needed third party application to see the actual data.
3. MS - Excel was in capable to handle more than 1 million rows and so data couldn't be visualized in MS-Excel.
4. Delimit is a third party application used to check and visualize how the data is structured and understand relationship amongst the spreadsheets.
5. After understanding the relationship, Tableau was used to visualize and infer some stats from data. All the 3 spreadsheets naming requests.csv, domains.csv and urls-no-action-taken.csv were visualized individually. Also, requests.csv and domains.csv were joined on request.request_id=domains.request_id to find some meaningful observations.
6. Also, all the 3 spreadsheets were Bulk Inserted individually into MySQL database using python. A connection with MySQL was created from Python. SQL is my comfort zone and allowed me to derive more inferences from the data.

Stats:

Supporting screenshots are provided in Google slides. Also, Sql Scripts and Tableau file is forwarded to support the stats.

1. Top 5 Copyright Owner based on Sum of Requests:

Sr No	Copyright Owner ID	Copyright Owner Name	Sum of Requests
1	39071	BPI LTD MEMBER COMPANIES	259,061
2	22818	BPI(British Recorded Music Industry) Ltd	141,950
3	76274	NUCLEAR BLAST RECORDS	98,101
4	73592	Universal Music GmbH	54,634
5	34071	IFPI	52,973

2. Top 5 Reporting Organizations based on Sum of Requests:

Sr No	Reporting Organization Name	Sum of Requests
1	AudioLock.NET	1,855,946
2	MUSO.com Anti-Piracy	545,788
3	BPI(British Recorded Music Industry) Ltd	437,996
4	proMedia	394,543
5	Digimarc	251,575

3. Top 5 Reporting Organization who are Abuser’s of the utility:

Sr No	Reporting Organization Name	# of Requests which were flagged True
1	Multi Media, LLC	7,858
2	Web Kontrol Ltd	4,071
3	ProtectYourContent LLC	4,020
4	Anti-piracy net	3,629
5	ChaturbateLLC	3,220

4. Top 5 domains from which the URLs were removed:

Sr No	Domain	# of URLs removed
1	4shared.com	62,474,414
2	mp3toys.xyz	51,260,635
3	rapidgator.net	24,301,504
4	chomikuj.pl	22,037,725
5	uploaded.net	20,091,560

5. Top 5 domains from which no action taken against URLs:

Sr No	Domain	# of URLs
1	iplusfree.com	6,274,698
2	rapidgator.net	4,829,957
3	mangapark.me	4,572,362
4	genteflowmp3.org	3,023,888
5	uploaded.net	2,937,647

6. Top 5 domains for which URLs not found in Google search Index.

Sr No	Domain	# of URLs
1	mangapark.me	6,394,505
2	mangapark.com	6,172,600
3	ninemanga.com	6,021,379
4	unblocksites.co	5,340,071
5	remusicos.com	4,409,291

Insights:

1. Reporting Organizations have multiple IDs for same name
    a. Those could be same companies and have multiple IDs. Need to update such IDs to provide better results.
    b. Also, they could be different companies with same name, so for such cases, there is no need to update the records.
2. Also, there have been cases where the reporting utility is abused but still the URLs are removed. We need to recheck such records.
3. Considering a threshold value for the URLs removed, domains can be delisted from search engine. Example, domain 4shared.com had 68,019,982 URLs reported and of those 62,474,414 were removed, which is 91.84% then 4shared.com can be delisted from the search engine. Generalising, if a threshold value of 90% then based on the data available the domain can be delisted from search engine
4. Considering a threshold value for the Abuser column. If a reporting organization is abusing the utility then the requests from such organization can be ignored.
5. Another insight is of owner's abusing the takedown request utility. If a domain falls under true abuser filter then considering a threshold such domains can be ignored from removal. Example, domain hqcollect.tv had a total of 330,159 URLs which were reported by Abuser's but were still taken down. Such, URLs need revalidation and reconsideration.