**Question 1**

**Write a query to extract the keywords from the url column. This query should be run on a Google Sheet. This should be done only via SQL (Query Function) and Regex and not via any formulas. Also, please note that keywords - with multiple words - should be separated by space.**

Assuming your data is in a Google Sheet in the following structure:

| **URL** | **Count\_of\_search\_results** |
| --- | --- |
| <https://play.google.com/store/search?q=messenger%20app&c=apps> | 50 |
| <https://play.google.com/store/search?q=torch&c=apps> | 20 |
| <https://play.google.com/store/search?q=personal%20loan%20app&c=apps> | 10 |
| <https://play.google.com/store/search?q=pnr%20status%20in%20india&c=apps> | 10 |

Assuming the URL column is column A and the Count\_of\_search\_results column is column B, here is how you can set up your QUERY function in a new column to extract and format the keywords:

1. **Create a new column header for the keywords.** Let's assume you put "Keywords" in cell C1.
2. **Use the following QUERY function:**

=QUERY(A2:B, "SELECT A, B, REGEXREPLACE(REGEXEXTRACT(A, 'q=([^&]+)'), '%20', ' ') AS keywords")

**Explanation:**

* **A2**

: The range of your data, assuming your data starts from row 2.

* **SELECT A, B, REGEXREPLACE(REGEXEXTRACT(A, 'q=([^&]+)'), '%20', ' ') AS keywords**:
  + REGEXEXTRACT(A, 'q=([^&]+)'): Extracts the part of the URL after q= and before & (if present), capturing the keywords.
  + REGEXREPLACE(REGEXEXTRACT(A, 'q=([^&]+)'), '%20', ' '): Replaces %20 with a space in the extracted keywords.
  + AS keywords: Labels the new column as "keywords".

**Step-by-Step:**

1. **Insert the query function in cell C2** (or any cell where you want the results to start):

=QUERY(A2:B, "SELECT A, B, REGEXREPLACE(REGEXEXTRACT(A, 'q=([^&]+)'), '%20', ' ') AS keywords", 1)

This function will process the URLs in column A, extract the keywords, replace %20 with spaces, and display the results along with the original URL and count of search results in the columns.

**Resulting Data in Google Sheets:**

| **URL** | **Count\_of\_search\_results** | **Keywords** |
| --- | --- | --- |
| <https://play.google.com/store/search?q=messenger%20app&c=apps> | 50 | messenger app |
| <https://play.google.com/store/search?q=torch&c=apps> | 20 | torch |
| <https://play.google.com/store/search?q=personal%20loan%20app&c=apps> | 10 | personal loan app |
| <https://play.google.com/store/search?q=pnr%20status%20in%20india&c=apps> | 10 | pnr status in india |

**Problem Set 2 - Working with BigQuery**

Suppose you get a raw data of your website visitors and you are asked:

1. Identify how many users came via a certain referrer url.
2. Also, convert the ***visitStartTime*** property to IST

let's write a query to count the number of users grouped by the referrer URL:

SELECT

IFNULL(h.page.referralPath, 'null') AS referral\_path,

COUNT(DISTINCT fullVisitorId) AS num\_users

FROM

`bigquery-public-data.google\_analytics\_sample.ga\_sessions\_\*`,

UNNEST(hits) AS h

GROUP BY

referral\_path

ORDER BY

num\_users DESC;

**Explanation:**

* IFNULL(h.page.referralPath, 'null') AS referral\_path: This ensures that if the referralPath is not present, it will be considered as 'null'.
* COUNT(DISTINCT fullVisitorId) AS num\_users: Counts the distinct users based on the fullVisitorId.
* UNNEST(hits) AS h: Unnests the hits array to access individual hits.
* GROUP BY referral\_path: Groups the results by the referralPath.
* ORDER BY num\_users DESC: Orders the results by the number of users in descending order.

**Query to Convert visitStartTime to IST**

The visitStartTime is in Unix timestamp format, which we need to convert to IST. IST is UTC+5:30.

SELECT

fullVisitorId,

visitId,

visitStartTime,

TIMESTAMP\_SECONDS(visitStartTime) AS visit\_start\_time\_utc,

TIMESTAMP\_SECONDS(visitStartTime) + INTERVAL 5 HOUR + INTERVAL 30 MINUTE AS visit\_start\_time\_ist

FROM

`bigquery-public-data.google\_analytics\_sample.ga\_sessions\_\*`

LIMIT 100;

**Explanation:**

* visitStartTime: The original visit start time in Unix timestamp.
* TIMESTAMP\_SECONDS(visitStartTime) AS visit\_start\_time\_utc: Converts the Unix timestamp to a UTC timestamp.
* TIMESTAMP\_SECONDS(visitStartTime) + INTERVAL 5 HOUR + INTERVAL 30 MINUTE AS visit\_start\_time\_ist: Converts the UTC timestamp to IST by adding 5 hours and 30 minutes.

**Combined Query**

Combining both requirements into one query:

WITH session\_data AS (

SELECT

fullVisitorId,

visitId,

visitStartTime,

TIMESTAMP\_SECONDS(visitStartTime) AS visit\_start\_time\_utc,

TIMESTAMP\_SECONDS(visitStartTime) + INTERVAL 5 HOUR + INTERVAL 30 MINUTE AS visit\_start\_time\_ist,

IFNULL(h.page.referralPath, 'null') AS referral\_path

FROM

`bigquery-public-data.google\_analytics\_sample.ga\_sessions\_\*`,

UNNEST(hits) AS h

)

SELECT

referral\_path,

COUNT(DISTINCT fullVisitorId) AS num\_users,

MIN(visit\_start\_time\_ist) AS first\_visit\_ist,

MAX(visit\_start\_time\_ist) AS last\_visit\_ist

FROM

session\_data

GROUP BY

referral\_path

ORDER BY

num\_users DESC;

**Explanation:**

* session\_data CTE: Extracts and transforms the necessary fields, including converting visitStartTime to UTC and IST, and handling referralPath.
* referral\_path: Groups by the referral path.
* COUNT(DISTINCT fullVisitorId) AS num\_users: Counts the distinct users.
* MIN(visit\_start\_time\_ist) AS first\_visit\_ist: Gets the first visit time in IST.
* MAX(visit\_start\_time\_ist) AS last\_visit\_ist: Gets the last visit time in IST.
* ORDER BY num\_users DESC: Orders the results by the number of users in descending order.

This combined query will provide the number of users per referrer URL along with the first and last visit times in IST.