Problem Set 1

Regex

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.

To solve the problem using SQL and Regex in a Google Sheets query, the general approach is as follows:

1. **Identify the structure of the URL**: You need to extract keywords from the URL, so first identify the format of the URL (e.g., the base URL, query parameters, etc.).
2. **Use Regex to extract the desired keywords**: Regex can be used within the QUERY function in Google Sheets to match patterns. For example, you may want to extract parts of the URL after the domain, such as the path or query parameters, depending on the keywords you are looking for.

Steps:

- Assume your data is in Column A (with headers) in Google Sheets, and URLs are in the format http://example.com/keyword1/keyword2?query=xyz.

Here is how you can write the query using REGEXEXTRACT or REGEXREPLACE in the QUERY function:

Example:-

```
=QUERY(A2:A, "SELECT REGEXEXTRACT(A,
'https?://(?:www\.)?([^/]+)(/[^?]+)?')", 1)
```

Explanation:

- https?://: Matches both http:// and https://.
- (?:www\.)?: Non-capturing group to optionally match www..
- ([^/]+): Captures the domain part of the URL.
- (/[^?]+)?: Captures the path after the domain (which often contains keywords), excluding the query string.

Extracting Keywords:

If you are looking for keywords in the path, you could refine the regex to extract multiple words, such as keywords from a query string, or path segments.

For example, to get the part after domain/ up to the first query parameter (if present), you could modify the query to:

=QUERY(A2:A, "SELECT REGEXEXTRACT(A, 'https?://(?:www\.)?[^/]+(/[^?]+)')", 1)

Handling Multiple Words:

If you need to extract multiple words or specific query parameters from the URL, adjust the regex to match different URL components. You can use REGEXREPLACE to remove unwanted characters and isolate keywords.