SELECT

directory,

COUNT(\*) AS count

FROM (

SELECT

REGEXP\_EXTRACT(files.path, r"([^/]\*)/[^/]\*$") AS directory,

files.path

FROM

`bigquery-public-data.github\_repos.sample\_files` AS files

WHERE

path LIKE '%/robots.txt')

GROUP BY

directory

ORDER BY

count DESC

LIMIT

10

This gives the following list, which seems to match our expectations:

1. public
2. web
3. static
4. app
5. assets
6. templates
7. docs
8. dist
9. www
10. src

Next, we’ll query common file names in these directories:

SELECT

filename,

COUNT(\*) AS count

FROM (

SELECT

REGEXP\_EXTRACT(files.path, r"/([^/]\*)$") AS filename

FROM

`bigquery-public-data.github\_repos.sample\_files` AS files

WHERE

REGEXP\_CONTAINS(files.path, r"/(public|web|static|app|assets|templates|docs|dist|www|src)/[^/]\*$"))

GROUP BY

filename

ORDER BY

count DESC

LIMIT

1000;

And this gives our list of files, starting with index.html