Boise State University CS 535 Fall 2024
Project 1: Mutual Links
Due Date: October 18, 2024 at 11:59 PM

Introduction

Among other things, Wikipedia is a directed graph. Pages can link to other pages. We want to find all the pairs of pages that link to each other. You can use these AWS S3 URIs as data sources:

```
s3://bsu-c535-fall2024-commons/arjun-workspace/articles/
s3://bsu-c535-fall2024-commons/arjun-workspace/linktarget/
s3://bsu-c535-fall2024-commons/arjun-workspace/page/
s3://bsu-c535-fall2024-commons/arjun-workspace/pagelinks/
s3://bsu-c535-fall2024-commons/arjun-workspace/redirect/
```

The last part of each path corresponds to a dump of Wikipedia's SQL databases, except for the articles. You won't need articles for this assignment, but you'll need the other four.

Assignment Details

We want pairs of articles that link to each other. For example, if "Arctic" links to "polar bear" and "polar bear" links to "Arctic", the pair would appear in the result data set. As a counterexample, if "Lord Kelvin" links to "Compass" but "Compass" does not link to "Lord Kelvin", that pair would not appear in the result set. Your result should be a table with two columns: `page_a` and `page_b`. Column values should be the IDs of the pages that appear in the pair.

You should not have repeat pairs in your dataset. If `page_a` is "Arctic" and `page_b` is "polar bear", there should not be any other pair in your results with "Arctic", "polar bear" or "polar bear", "Arctic".

You also need to account for page redirects in your pairs. If "England" links to "Great Britain", "Great Britain" redirects to "United Kingdom", and "United Kingdom" links to "England", the pair in your result set should only contain "England", "Great Britain".

Rubric

40 points - code & submission quality
50 points - basic pair construction
30 points - redirects are accounted for
20 points - pairs are unique
40 points - code runs in under 75 minutes using a single r6gd.2xlarge
Total points: 180

Submission

Create a private Git repository to host your project. It must be private. Invite me as a collaborator to the project so I can view it. If I see commits past the due date, I will roll back the project to the latest commit before the due date.

You may not submit a Jupyter Notebook. It should be a normal Python project. Tell me what the entrypoint file is, and provide a script to generate the zip file with the rest of your dependencies. I will supply the environment variable PAGE_PAIRS_OUTPUT containing an S3 URI for your code to put the expected result table in. Make sure your code writes parquet files to this path when it's done.