**Data Collection Method**

**Introduction**

To uncover how the United States Census Bureau (USCB) data is being utilized post-publication, a comprehensive data collection strategy leveraging the Common Crawl dataset was employed. This paper outlines the methodology used to gather, process, and analyze web data to identify references to USCB data.

**Data Acquisition**

To begin, 11 WET files were downloaded from the Common Crawl dataset, focusing on English language sites. This initial dataset comprised approximately 136,000 website links and their corresponding content.

**Web Scraping**

To identify websites referencing USCB data, a web scraper was developed using the Beautiful Soup package. The scraper performed the following steps:

1. Made HTML requests to each site.
2. Downloaded the HTML content into text documents.
3. Extracted all anchor (<a>) tags.
4. Returned links containing keywords such as “census”, “census bureau”, “censusbureau”, and “census.gov”.

If a site did not reference the Census Bureau, the output was ‘None’. If the request failed, the output was ‘Error’ along with the error code. 403 error codes indicated forbidden access to the site which could indicate sites that require user logins or paywalls. The errored category was retained to identify potential insights from sites behind paywalls.

**Initial Findings**

Upon scraping the 136,000 sites, 60 sites with references to the Census Bureau were identified. Recognizing the time and resource intensity of this process, the team sought to optimize the approach to handle larger datasets efficiently.

**Optimization Strategy**

To enhance efficiency, a keyword search method was tested on the content of the 136,000 sites that were scraped. This involved:

1. Cleaning the content by removing special characters, newline indicators (\n), and converting text to lowercase.
2. Implementing a keyword search using terms like ‘census’, ‘uscb’, and ‘american community survey’.

The dataset was then filtered to isolate rows without these keywords and had links to the USCB found with the web scraper, identifying 14 additional sites that would have been missed by keyword search alone. To enhance our key word search, a named entity recognition (NER) was performed on these 14 sites to identify additional keywords.

**Expanded Keyword Search**

From the NER analysis, keywords such as ‘american’, ‘nih’, ‘government’, ‘survey’, and ‘population’ were added to the keyword search. This expanded search successfully reduced the number of links to be scraped from 136,000 to 29,000 while capturing all relevant links.

**Increased Sample Size**

To ensure robust pattern identification, over 100 additional WET files were downloaded to increase our sample size. These additional links were narrowed down through our keyword search and then scraped for any links to the USCB. This expanded dataset included nearly 500,000 links, from which over 700 additional sites were found with references to the Census Bureau.

**Data Storage and Analysis**

The final dataset was saved into two files:

1. Sites with 403 errors for potential insights into paywalled content.
2. Sites with found links for further parsing to distinguish between repackaging and citing of Census Bureau data.

**Geographic Analysis**

To visualize the geographic distribution of Census Bureau data usage, a second web scraper was developed to extract IP addresses and locations of base sites. This data was used to create a map highlighting areas with significant Census Bureau data activity.

**Conclusion**

With the final datasets and initial exploration complete, the team proceeded to build models and gather insights on the most mentioned topics and accessed Census Bureau data. This methodology provides a scalable approach to understanding the dissemination and utilization of USCB data across the web.