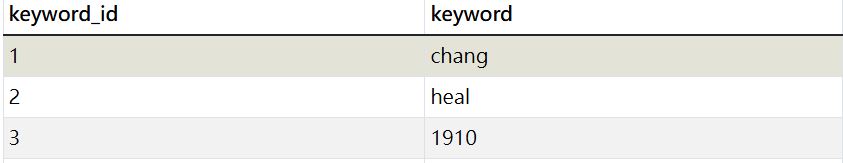
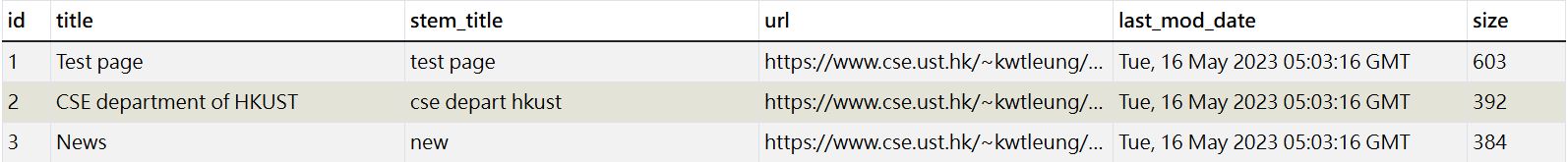
## Table Design

The schema consists of five primary tables:

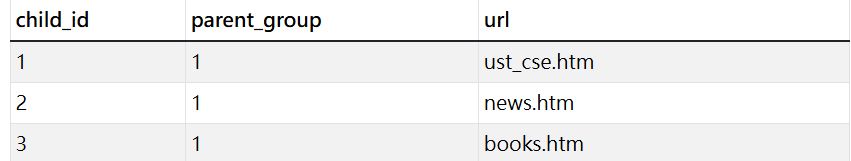
1. Keywords
   * Purpose: Maps words to unique Word IDs for efficient storage and lookup.
   * Attribute:[**keyword\_id, keyword**]
     1. **keyword\_id** (Primary Key): A unique identifier for each keyword
     2. **keyword**: The actual filtered and stemmed word



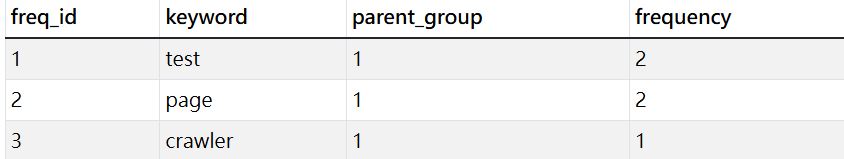
1. links
   * Purpose: Stores metadata for each crawled URL, including its unique Page ID.
   * Attribute: [**id, title, stem\_title, url, last\_mod\_date, size**]
     1. **id** (Primary Key): A unique identifier for each page (Page ID)
     2. **title:** The title of the page
     3. **stem\_title:** The stemmed title of the page
     4. **url**: The actual URL of the page
     5. **last\_mod\_date**: The last modification date of the page
     6. **size**: The size of the page in bytes



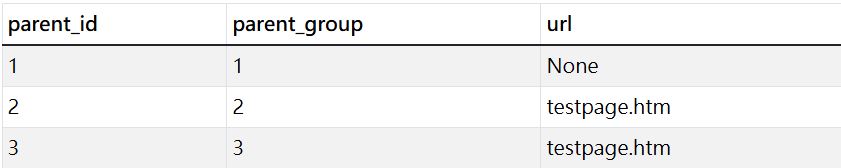
1. child\_links
   * Purpose: Tracks the child links (URLs) for each parent link.
   * Attribute: [**child\_id, parent\_group, url**]
     1. **child\_id**(Primary Key): A unique identifier for each child link
     2. **parent\_group**: Represents which parent link the child URL originates from
     3. **url:** The URL of the child link.



1. keywords\_freq
   * Purpose: Tracks the frequency of each keyword for a given parent link (Page ID).
   * Attribute: [**freq\_id, keyword, parent\_group,frequency**]
     1. **freq\_id**(Primary Key): A unique identifier for the keyword frequency entry
     2. **keyword:**The keyword (linked to the Keywords table).
     3. **parent\_group**:The Page ID of the parent link where the keyword appears.
     4. **frequency:**The frequency of the keyword on the page.

****

1. parent\_links
   * Purpose: Tracks parent links for each group of child links.
   * Attribute: [**parent\_id, parent\_group, url**]
     1. **parent\_id**(Primary Key): A unique identifier for each parent link
     2. **parent\_group**: Identifies the group of child links originating from this parent link
     3. **url:**The URL of the parent link



## ER diagram

## Database ER diagram (crow's foot).png

## Explanation of the Design

Forward Index:

* The forward index is implemented using the Links and Keywords Frequency tables:
  + The Links table stores metadata about each page (Page ID, title, URL, etc.).
  + The Keywords Frequency table records the keywords and their frequencies for each page (linked by parent\_group to the Page ID in the Links table).

Inverted Index:

* The inverted index is implemented using the Keywords and Keywords Frequency tables:
  + The Keywords table maps each word to a unique Word ID.
  + The Keywords Frequency table links keywords to the Page IDs where they appear, along with their frequencies.

Mapping Tables:

* URL ↔ Page ID:
  + The Links table serves as the mapping table between URLs and Page IDs.
* Word ↔ Word ID:
  + The Keywords table serves as the mapping table between words and Word IDs.

Child and Parent Links:

* The Child Links and Parent Links tables track the relationships between parent and child links:
  + The Parent Links table identifies the parent link for each group of child links.
  + The Child Links table stores the child URLs for each parent link.