**Database folder that store all the following table:** **main.db**

*Database tables related to Indexer (5 tables):*

Table name: **body\_inverted\_index**

Columns:

* wordId (format: uuid4)
* value (format: uuid4;frequency;position1,position2,… uuid4;frequency;position1,position2,…)

Example:

|  |  |
| --- | --- |
| wordId | value |
| 11 | 34;1;456 275;3;75,80,221 |
| 23 | 97;2;45,992 |

Description: Given a word id, return the url id, word frequency, positions

Table name: **title\_inverted\_index**

Columns:

* wordId (format: uuid4)
* value (format: uuid4;frequency;position1,position2,… uuid4;frequency;position1,position2,…)

Example:

|  |  |
| --- | --- |
| wordId | value |
| 11 | 34;1;456 275;3;75,80,221 |
| 23 | 97;2;45,992 |

Description: Given a word id, return the url id, word frequency, positions

Table name: **forward\_index**

Columns:

* urlId (format: uuid4)
* value (format: uuid4 uuid4 …)

Example:

|  |  |
| --- | --- |
| urlId | value |
| 101 | 4 70 91 301 |
| 200 | 2 67 9 |

Description: Given a url id, return the word id(s)

Table name: **word\_to\_id**

Columns:

* word (format: string)
* wordId (format: uuid4)

Example:

|  |  |
| --- | --- |
| word | wordId |
| fruit | 7 |
| read | 99 |

Description: Given a word, return the word id

Table name: **id\_to\_word**

Columns:

* wordId (format: uuid4)
* word (format: string)

Example:

|  |  |
| --- | --- |
| wordId | word |
| fruit | 7 |
| read | 99 |

Description: Given a word id, return the word

*Database tables related to Spider (8 tables):*

Table name: **url\_to\_id**

Columns:

* url (format: string)
* urlId (format: uuid4)

Example:

|  |  |
| --- | --- |
| url | urlId |
| https://www.cse.ust.hk/~kwtleung/COMP4321/testpage.htm | 75 |
| https://www.cse.ust.hk/~kwtleung/COMP4321/ust\_cse.htm | 1001 |

Description: Given a url, return the url id

Table name: **id\_to\_url**

Columns:

* urlId (format: uuid4)
* url (format: string)
* Example:

|  |  |
| --- | --- |
| url | urlId |
| https://www.cse.ust.hk/~kwtleung/COMP4321/testpage.htm | 75 |
| https://www.cse.ust.hk/~kwtleung/COMP4321/ust\_cse.htm | 1001 |

* Description: Given a url id, return the url

Table name: **crawled\_page\_to\_id**

Columns:

* url (format: uuid4)
* urlId (format: uuid4)

Example:

|  |  |
| --- | --- |
| url | urlId |
| https://www.cse.ust.hk/~kwtleung/COMP4321/testpage.htm | 75 |
| https://www.cse.ust.hk/~kwtleung/COMP4321/ust\_cse.htm | 1001 |

Description: Given a url, return the url id

Table name: **id\_to\_page\_title**

Columns:

* urlId (format: uuid4)
* pageTitle (format: string)

Example:

|  |  |
| --- | --- |
| urlId | pageTitle |
| 75 | Test page |
| 1001 | CSE department of HKUST |

Description: Given a url id, return the page title

Table name: **id\_to\_last\_modification\_date**

Columns:

* urlId (format: uuid4)
* lastModificationDate (format: date)

Example:

|  |  |
| --- | --- |
| urlId | lastModificationDate |
| 75 | 17/12/2024 21:09:37 |
| 1001 | 3/3/2025 08:02:58 |

Description: Given a url id, return the last modification date

Table name: **id\_to\_page\_size**

Columns:

* urlId (format: uuid4)
* pageSize (format: string)

Example:

|  |  |
| --- | --- |
| urlId | pageSize |
| 75 | 603 |
| 1001 | 108 |

Description: Given a url id, return the page size

Table name: **id\_to\_children\_url\_id**

Columns:

* urlId (format: uuid64)
* childrenUrlId (format: uuid64 uuid64 …)
* Example:

|  |  |
| --- | --- |
| urlId | childrenUrlId |
| 75 | 7 71 546 7839 |
| 1001 | 938 |

* Description: Given a url id, return all the children url id. Note that if the given url has no more children (excluding cyclic that go back to any parents), they should not be shown in the table.

Table name: **id\_to\_parents\_url\_id**

Columns:

* urlId (format: uuid64)
* parentsUrlId (format: uuid64 uuid64 …)
* Example:

|  |  |
| --- | --- |
| urlId | parentsUrlId |
| 57 | 1 551 |
| 1003 | 2 |

* Description: Given a url id, return all the parents url id. Since the TA may modify the testing webpage (i.e., hidden test cases), we CANNOT assume there is only ONE parent for each urlId. Note that if the given url has no more parent (excluding cyclic that go back to any children), they should not be shown in the table.