

SE305 DBST Final Report

Illustration of the Books

Yuqi Yi* Hongjie Chen†

December 2018

Contents

1	Project Description	1
2	Database Design	2
2.1	E-R Model	2
2.2	Relational Model	3
3	Database Construction	3
3.1	Crawl Data	3
3.2	Process Data	4
3.3	Create Database (MySQL)	4
4	User Interface	6
5	Experiments	7
6	Code	9

1 Project Description

Books are a rich source of both fine-grained information. In order to improve the readability of the book, many books have illustrations which are strictly illustrative of the text. However the E-book rarely contains illustrations. So in this work you are required to choose illustrations for each chapter in a book. This project can be seen partly as an information retrieval task.

There is a dataset called BookCorpus, but this corpus has not been maintained. Firstly, you are required to crawl all the books in BookCorpus. You can find the booklist on <https://www.smashwords.com/books/category/892> and you can download books from <https://www.gutenberg.org> or other website.

*515030910596, awonderfullife@sjtu.edu.cn

†515030910597, chen hongjie@sjtu.edu.cn

What you also need to crawl is a picture dataset about the books. The requirement is that you need to crawl at least one picture about a named entity or a concept in *each* chapter of every book in the data set.

In this project, you are required to create a relational database to store the entire data, the book text and pictures. You can store the books with separate chapters. In each chapter, you are required to extract some important information which you believe. And then, you can label for pictures and match the pictures to the chapter according to the labels you get from some tools and the important information you extract from the chapter text. For example, I can search "dodo" in Alice in Wonderland , and I will get figure 1.

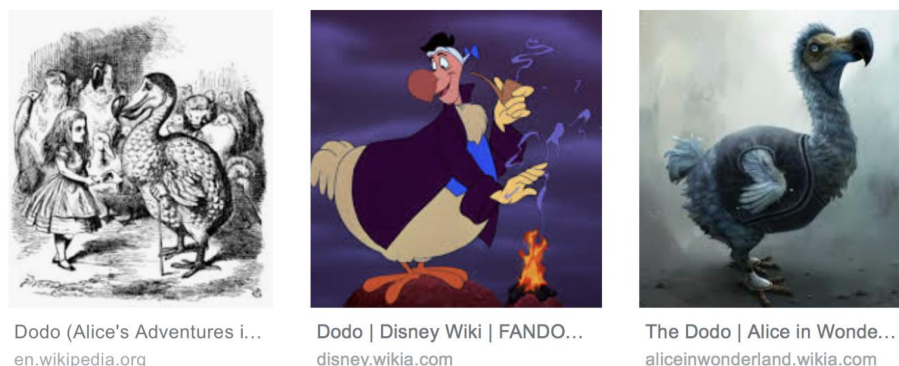


Figure 1: dodo

Required queries:

- Find the books according to the name, author.
- Given a book and chapter number, you can return some relative pictures.
- Given some words, you can find the relative pictures according to the words.

2 Database Design

In order to implement the required queries mentioned above, we first design a proper Entity-Relation (ER) model accordingly. Then we convert the E-R model to a relational model.

2.1 E-R Model

There should be two entities, "*Book*" and "*Picture*". Since a picture must belong to some book, there should be a "*belong_to*" relation between "*Book*" and "*Picture*". Moreover, a picture also appears in some chapter of a book.

Thus, we need a additional attribute for the “*belong_to*” relation. Figure 2 shows the E-R model.

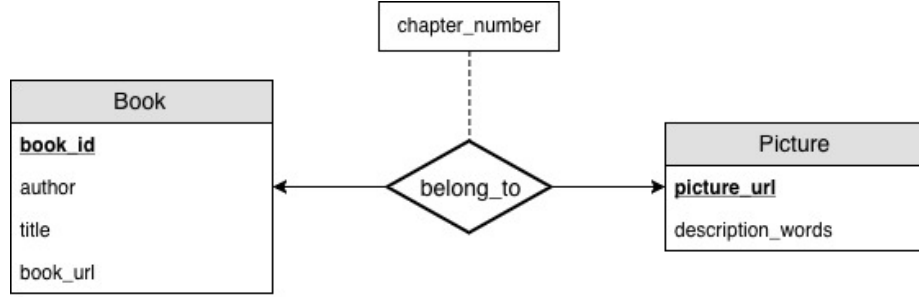


Figure 2: E-R Model

2.2 Relational Model

Based on the above E-R Model, we obtain the following 3 relation schemata naturally.

- **Book** (book_id, author, title, book_url)
- **Picture** (picture_url, description_words)
- **belong_to** (book_id, picture_url, chapter_number)

3 Database Construction

Now we already have relation the schemata we need. Next, we first crawl data online and download useful data to local hard disks as much as possible. Then we retrieve information that we need to

3.1 Crawl Data

In order to build a database, we need to get data first. Therefore, we crawled Project Gutenberg (see figure 3). For each book, we need its meta-information (such as title, author and language) and its content. After careful analysis of the structure of Gutenberg websites, we found that for each book there is a corresponding website containing information about its author, title, and etc (see figure 4). And there is also a link, i.e. “Read this book online: HTML”, on that website leading to a new website which contains book content (see figure 4). Thus, we crawled two websites and then downloaded them for each book. Finally, we crawled more than **50,000** books.



Figure 3: Project Gutenberg

3.2 Process Data

Now we have book data on local disks, but we need to process these data before creating the final database. Each book has two HTML files of which one contains its meta-information and the other contains its content. We use BeautifulSoup, a Python library for pulling data out of HTML and XML files, to extract useful data from these HTML files. It's not difficult to extract a book's author, title and URL from its HTML file. However, the case is different for pictures. To find out which chapter a picture belongs to, we need to traverse the whole HTML tree. For each picture, we get its URL from the HTML file, and then we download it to local disks. We also get rid of books whose information is not complete. After processing, we get nearly **20, 000** books and more than **80, 000** pictures.

3.3 Create Database (MySQL)

We use MySQL to create the final database using the above relational model we designed and the above data we processed. The code for creating tables is as follows:

```

1 TABLES[ 'Book' ] = (
2     """CREATE TABLE 'books' (
3         'book_id' ' varchar(10) ,
4         'author' ' varchar(100) ,
5         'book_url' ' varchar(100)
6         'book_name' ' varchar(300)
7     ) """ )
8
9 TABLES[ 'Picture' ] = (

```

Project Gutenberg offers 58,355 free ebooks to download.

[Donate](#)
[Rate this!](#)

[Search](#)
[Latest](#)
[Terms of Use](#)
[Donate?](#)
[Mobile](#)

[Search Project Gutenberg](#)
[Help](#)

Project Gutenberg • 58,355 free ebooks

The Speeches & Table-Talk of the Prophet Mohammad by Muhammad ibn 'Abd Allah

[Download](#)
[Bibrec](#)

Download This eBook

Format	Size
Read this book online: HTML	391 kB
EPUB (with images)	201 kB
EPUB (no images)	177 kB
Kindle (with images)	716 kB
Kindle (no images)	679 kB
Plain Text UTF-8	307 kB
More Files...	

INTRODUCTION.

The aim of this little volume is to present all that is most enduring and memorable in the public orations and private sayings of the prophet Mohammad in such a form that the general reader may be tempted to learn a little of what a great man was and of what made him great. At present, it must be allowed that although "Auld Mahound" is a household word, he is very little more than a word. Things are constantly being said, written, and preached about the Arab prophet and the religion he taught, of which an elementary acquaintance with him would show the absurdity. No one would dare to treat the ordinary classics of European literature in this fashion; or, if he did, his exposure would immediately ensue. What I wish to do is to enable any one, at the cost of the least possible exertion, to put himself into a position to judge of popular fallacies about Mohammad and his creed as surely and certainly as he can judge of errors in ordinary education and scholarship. I do not wish to mention the Korán by name more than can be helped, for I have observed that the word has a deterrent effect upon readers who like their literary food light and easy of digestion. It cannot, however, be disguised that a great deal of this book consists of the Korán, and it may therefore be as well to explain away as far as possible the prejudice which the ill-fated name is apt to excite. It is not easy to say for how much of this prejudice the standard English translator is responsible. The patient and meritorious George Sale put the Korán into tangled English and heavy quarto,—people read quartos then and did not call them *éditions de luxe*,—his version then appeared in a clumsy octavo, with most undesirable type and paper; finally it has come out in a cheap edition, of which it need only be said that utility rather than taste has been consulted. One can hardly blame any one for refusing to look even at the outside of these volumes. And the inside,—not the mere outward inside, if I may so say, the type and paper,—but the heart of hearts, the matter itself, is by no means calculated to tempt a reluctant reader. The Korán is there arranged according to the orthodox form, instead of in chronological order,—it must be allowed that the chronological order was not discovered in Sale's time,—and the result is that impression of chaotic indefiniteness which impressed Carlyle so strongly, and which Carlyle has impressed upon most of the present generation. A large disorderly collection of prophetic rhapsody did not prove inviting, as the state of popular knowledge about Mohammad very clearly shows.

The attitude of the multitude towards Sale's Korán was on the whole reasonable. But if the faults that were found there are shown to belong to Sale and not to the Korán, or only partly to it, the attitude should change. In the first place, the Korán is not a large book, and in the second, it is by no means so disorderly and anarchic as is commonly supposed. Reckoned by the number of verses, the Korán is only two-thirds of the length of the New Testament, or, if the wearisome stories of the Jewish patriarchs which Mohammad told and retold are omitted, it is no more than the Gospels and Acts. It has been remarked that the Sunday edition of the *New York Herald* is three times as long. But the real permanent contents of the Korán may be taken at far less even than this estimate. The book is full—I will not say of vain repetitions, for in teaching and preaching repetition is necessary—but of reiterations of certain cardinal articles of faith, and certain standard demonstrations of these articles by the analogy of nature. Like the numerous stories borrowed by Mohammad from the Talmud, which have little but an antiquarian interest, many of these reiterated arguments and illustrations may with advantage be passed over. There is also a considerable portion of the Korán which is devoted to the exposure and confutation of those who, from political, commercial, or religious motives, made it their business to thwart Mohammad in his efforts to reform his people. These personal, one might say party, speeches are valuable only to the biographer and historian of the times. They throw but little light on the character of the man Mohammad himself. They show him, indeed, to be,—what we knew him before—a sensitive, irritable man, keenly alive to ridicule and scorn. But for this purpose one instance is sufficient. We do not form our estimate of a great statesman from his moments of irritation, but from those larger utterances which reveal the results of a life's study of men and government. So with Mohammad, we may abandon the personal and temporary element in the Korán, and base our judgment upon those utterances which stand for all time, and deal not with individuals or classes, but with man as he is, in Arabia or England, or where we will. This position is not taken with the object of saving Mohammad from himself. His attacks upon his opponents will bear comparison with those of other statesmen. They are doubtless couched in more barbaric language than we are accustomed to, and where we insinuate, Mohammad curses outright. But in the face of a treacherous and malignant opposition, the Arabian prophet comported himself with singular self-restraint. He only threatened hell-fire, and people of all denominations are still threatened with that every Sunday, to say nothing of Lent. Leaving out the Jewish stories, needless repetitions, and temporary exhortations or personal vindications, the speeches of Mohammad may be set forth in very moderate compass. One speech—*sura*, or chapter, as it is generally called—follows another so much to the same effect, that a limited number will be found to contain all the ideas which a minute study of the whole Korán could collect. I believe there is nothing important, either in doctrine or style, which is not contained in the twenty-eight speeches which fill the first hundred and thirty pages of this small volume. If I were a Mohammadan, I think I could accept the present collection as a sufficient representation of what the Korán teaches.

The obscurity of the Korán is largely due to its ordinary arrangement. This consists merely in putting the longest chapters first and the shortest last. The Mohammadans appear to be contented with this curious order, which after all is not more remarkable than that of some other sacred books. German criticism, however, has discovered the method of arranging the Korán in approximately chronological sequence. To explain how this is established would carry me too far, but the results are certain. We can state positively that the chapters of the Korán—or, as I prefer to call them, the speeches of Mohammad—fall into certain definite chronological groups, and if we cannot arrange each individual speech in its precise place, we can at least tell to which group, extending over but few years, it belongs. The effect of this critical arrangement is to throw a perfectly clear light on the development of Mohammad's teaching, and the changes in his style and method. When the Korán is thus arranged—as it is in Mr. Rodwell's charming version, which deserves to be better read than it is—the impression of anarchy disappears, and we see only the growth of a remarkable mind, the alternations of weakness and strength in a gifted soul, the inevitable inconsistencies of a great man. I do not believe any one who reads the speeches of Mohammad as I have arranged them in Professor Nöldeke's chronological order will say that they have no definite aim or coherence. They may be monotonous, and often they are rambling, but their intention and sequence of thought are to me clear as noonday.

It is something more, however, than any supposed length or obscurity that has hitherto scared people from the Korán. The truth is that the atmosphere of our Arabian prophet's thoughts is so different from what we breathe ourselves, that it needs a certain effort to transplant ourselves into it. That it can be done, and done triumphantly, may be proved by Mr. Browning's *Saul*, as Semitic a poem as ever came from the desert itself. We see the whole life and character of the Bedawy in these lines:—

Figure 4: Up: Book information. Down: Book content

```

10      """CREATE TABLE 'Picture' (
11          'picture_url' varchar(100),
12          'description_word' varchar(30)
13      ) """
14
15  TABLES['belong-to'] = (
16      """CREATE TABLE 'belong-to' (
17          'picture_url' varchar(100),
18          'book_id' varchar(10),

```

5

```

19         'chapter_number' : int(10)
20     )"""

```

We also implemented other basic database operations, such as *insert* and *delete*.

4 User Interface

For realize these needs we are required to implement, we design a simple GUI interface with wxpython. You can see our user interface below (Figure 5).

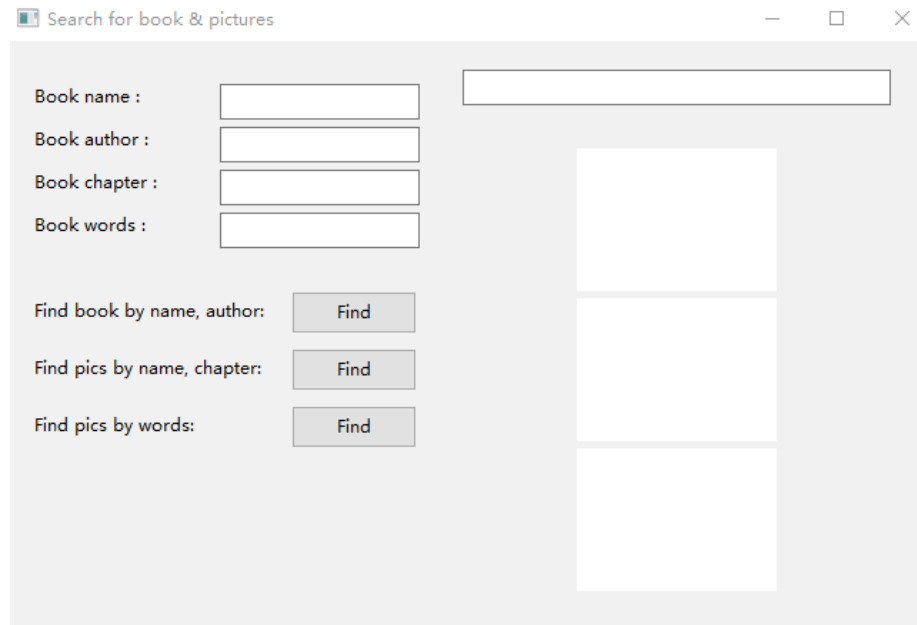


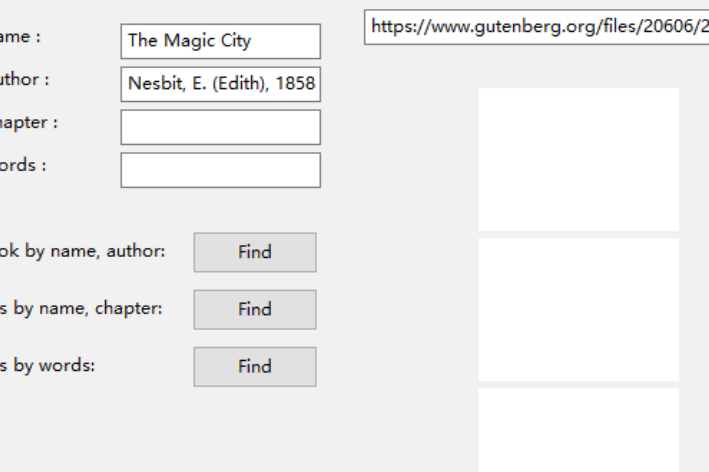
Figure 5: **Interface:** interface with three basic function buttons

there are four input text on the interface: book name, book author, chapter number, and the key words. After you input the corresponding feature that is needed for a specific search. you could click the corresponding button. For finding book through it's name and author, we would return a link for it's contents on the top of the right column. For finding pictures through some book name, chapter number or through key word. we would show the result pictures on the right column too and click same button for mutilate times, you can get for other result pictures for that instead of just three of them. If you change the corresponding input features, you just need to re-click the find button to update the result.

5 Experiments

We build a database with over 100000+ information about books and pictures in it, for simplify here we just show one example that we use the system to find the related book or pictures we wanted.

Figure 6 shows the experiment result we done for finding the corresponding book through book name and author. the result we return as a link for that book on the right top of our interface.



Search for book & pictures

Book name :

Book author :

Book chapter :

Book words :

Find book by name, author:

Find pics by name, chapter:

Find pics by words:

[https://www.gutenberg.org/files/20606/20606-h/20606-h/20606-h.htm](https://www.gutenberg.org/files/20606/20606-h/20606-h.htm)

Figure 6: **Interface:** find book through book name and author

Figure 7 shows the experiment result we done for finding pictures through book name and chapter number. We finding there are three related picture and we present then on the right column of our interface.

Figure 8 shows the experiment result we done for finding pictures through book name and chapter number. the result of these also shown as above.

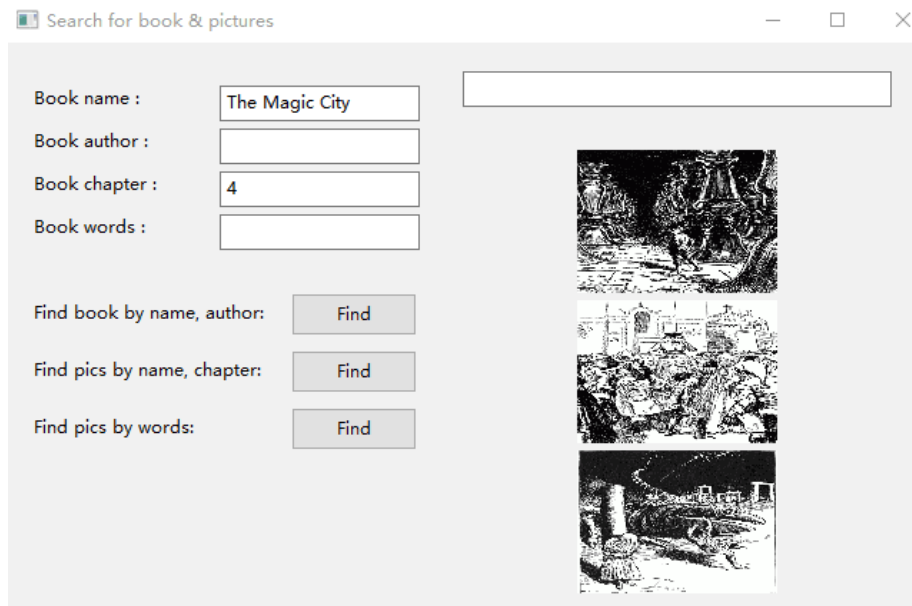


Figure 7: **Interface:** find pictures through book name and chapter number

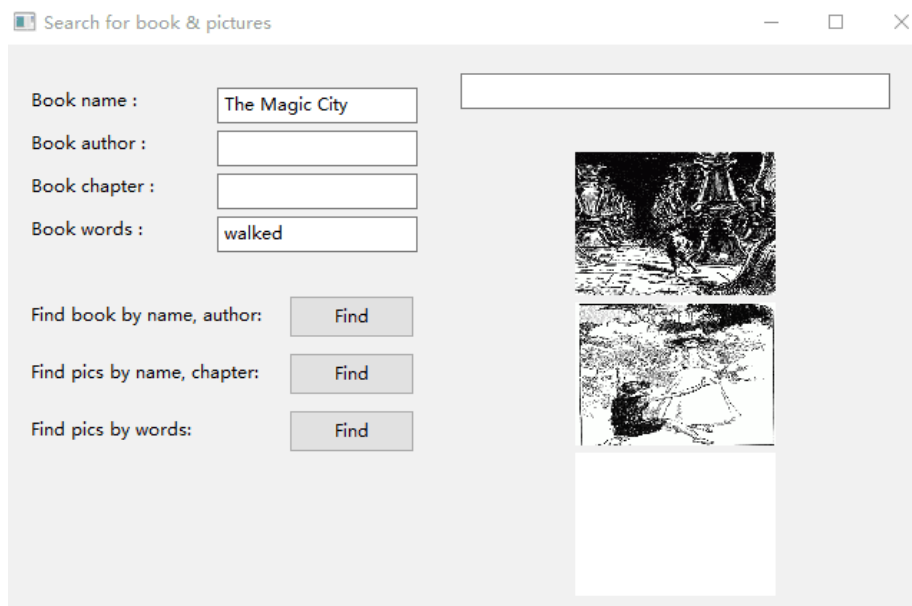


Figure 8: **Interface:** find picture through book name and key words

6 Code

All the code is available online¹. We also give the structure of our code here.

```
code
├── crawler
│   ├── crawl_book_info.py
│   └── crawl_book_content.py
├── process
│   ├── process.py
│   └── download_picture.py
├── buildMySQL.database
│   ├── create_table.py
│   ├── delete_table.py
│   └── insert_item.py
└── interface
    └── interface.py
```

¹https://github.com/CoolPhilChen/SE305_DBST_Project