CSE 201 - Advanced Programming
Project 1
DBLP Query Engine

The purpose of this project is to build a GUI search engine which retrieves required information from a given offline dataset.

Dataset: Represents information about publications by different authors along with other information like the year of publication, the number of pages etc.
Data to be used is available at http://dblp.uni-trier.de/xml/
The zipped version is about 342MB. For information about the XML data you can visit http://dblp.uni-trier.de/faq/What+do+I+find+in+dblp+xml
Since the file is pretty big after extraction you might face problems in opening the file directly in your text editors. We recommend using Vim and open the dblp.xml file directly in the terminal. You can also use sublime text.

You have to complete the following tasks:

Task1 : Parse the XML file using appropriate parsing techniques. This is a very common dataset and you'll find a lot of information on the web about the techniques you can use to parse it.

Task2 : Use collections in Java to retrieve and display the results of the given queries on this dataset.

Task3: Create a GUI using Java swing libraries to input the query parameters to your program.

Functionalities:

1. Display the result count you get along with the data.
2. Perform entity resolution. It simply means finding papers that belong to the same author (entity) where the author might have different names. For e.g. In research papers, sometimes an author uses his/her full name and sometimes his/her initials (or might have a different middle name). The task of entity resolution is to identify both to be the same author. During search, both the papers should be returned.
3. If the data entered is invalid show appropriate message on the GUI itself.
4. If no results are returned, display the message in the result box.
5. Appropriate checks for values in the text boxes should be present in your implementation and user should be prompted for such values.

6. Result box should display first 20 results and should give user an option (button) to see next 20 results and so on.
7. If returned results are less than 20, then show only the returned results.
8. Results are always sorted by date. (Latest ones first)
9. Results for Query1 should be in the following format:
   <s_no>, <authors>, <title>, <pages>, <year>, <volume>, <journal/Booktitle>, <url>
10. Results for Query2 should be just author names in a single column on a new line in the result box.

Queries:

1. **Query1**
   a. Find publications by a given author name.
   b. Find publications by title tags.

   Both the above queries have multiple options:
   a. Sorting by date (reverse)
   b. Sorting by relevance (matched words)
   c. Since some given year
   d. Inbetween two years
2. **Query2**
   a. Find names of authors with more than <k> publications.

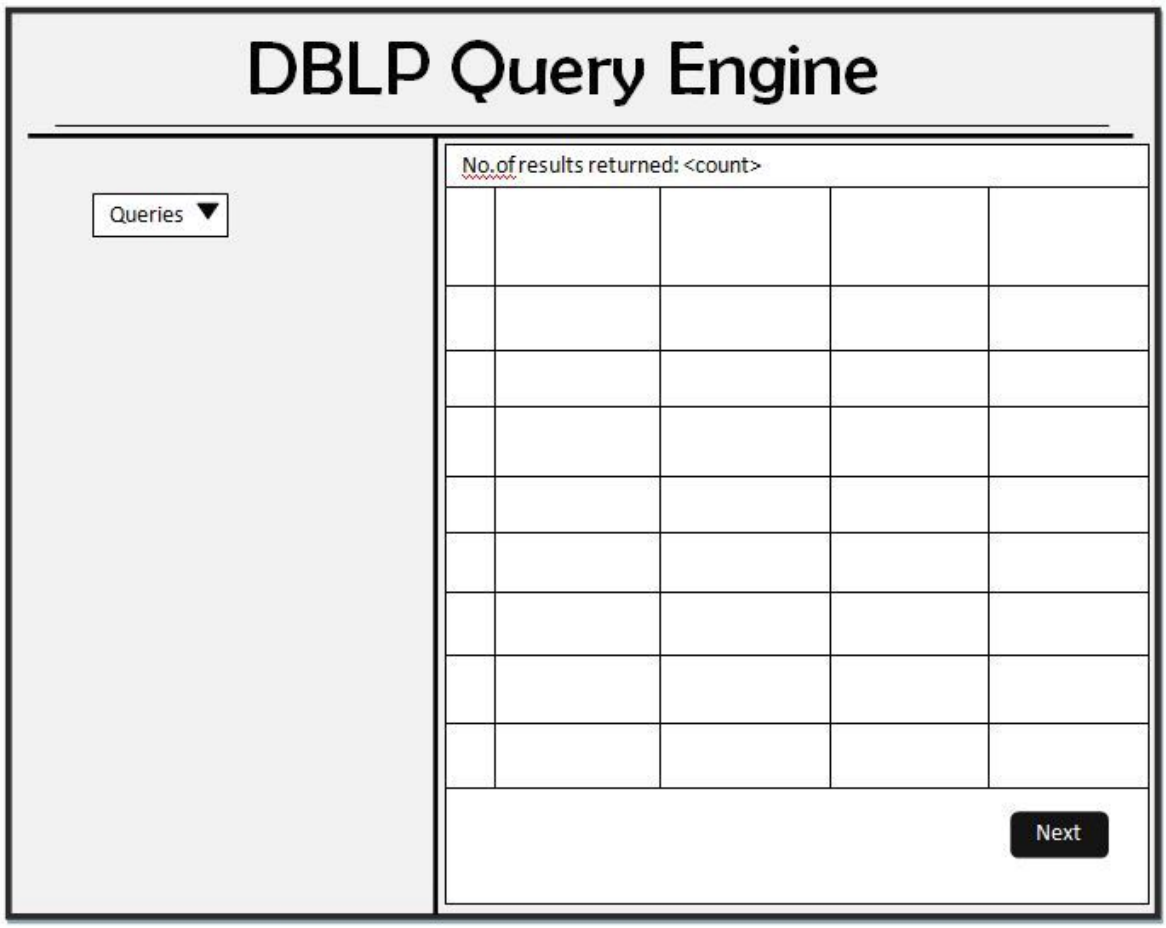
Bonus: (Max 25%)

1. **(15%)** Given a data upto a particular year, predict the number of papers an author publishes in the next year.
   a. This can be a **Query 3** in your application where we will give you year (upto which you can use the data) and 5 random authors for which you have to predict the number of publications.
   b. If the prediction comes ±20% of the actual number of papers for at-least 3 authors, you get the bonus marks.

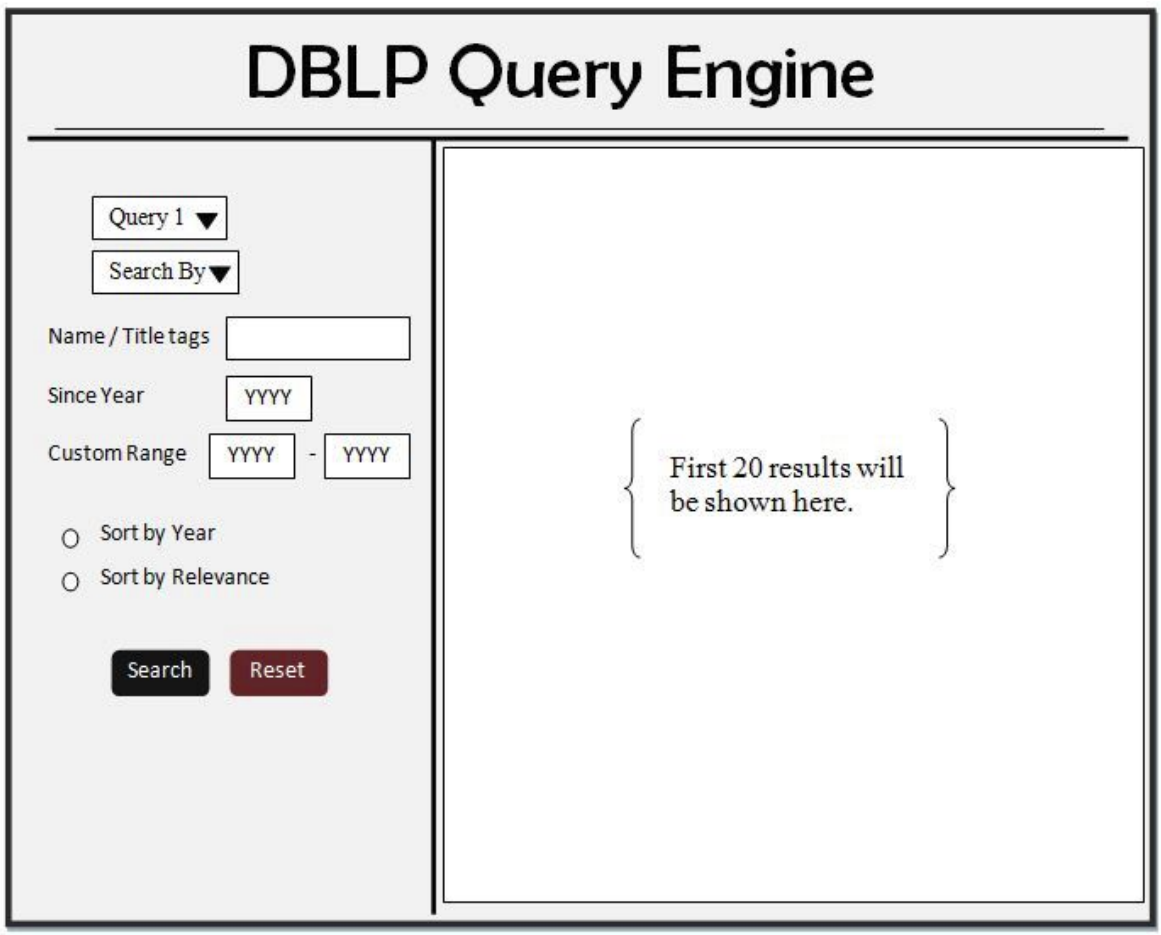
2. **(1%)** For every design pattern that you use in your implementation. **(max 5%)**
3. **(2%)** If you generate doxygen comments. Refer: (http://www.stack.nl/~dimitri/doxygen/manual/docblocks.html)
4. **(1%)** If all the classes in your system are less than 200 lines.
5. **(1%)** If every function in your system is less than 50 lines.
6. **(1%)** If you use CVS/SVN/GIT with more than 50 commits and not more than 3 commits in any day.

A reference layout is given below:

1. The left panel shows the combobox for selecting the query. Right panel is just for reference to show how the results should look like. Initially, this area should be empty.



2. The left panel on selecting Query 1 in the combobox.



3. The left panel on selecting Query 2 in the combobox