

# Design Automation Renegades

---

GLOBETROTTING DIVISION

## BIB<sub>T</sub>E<sub>X</sub> Analytics: For Automating Reference Management and Recognizing Emerging Trends

Zhiyang Ong<sup>1</sup>

A DOCUMENT ON *Python*-BASED BIB<sub>T</sub>E<sub>X</sub> ANALYTICS  
For Reference Management ...  
and Emerging Trend Recognition

May 23, 2018

<sup>1</sup>Email correspondence to: ✉ [ongz@acm.org](mailto:ongz@acm.org)

## **Abstract**

This documents how the repository of the BibTeX *Analytics* project is organized, and its software architecture. It also describes the future goals of the project for using a data analytics approach to recognize emerging trends in research, especially emerging research trends in electrical and computer engineering, computer science, and other fields, such as medicine, agriculture, and environmental science.

Insert abstract here.

More stuff to be included.

# Revision History

Revision History:

1. Version 0.1, May 21, 2018. Initial copy of the report.
- 2.

# Contents

Revision History	i
1 Organization of the BIB <sub>T</sub> <sub>E</sub> X Analytics Repository	1
2 Software Architecture of the BIB <sub>T</sub> <sub>E</sub> X Analytics Project	4
3 Future Work	5
Bibliography	5

# Chapter 1

## Organization of the BIBTEX Analytics Repository

The main deliverables of the BIBTEX *Analytics* project are a *Python*-based software to perform reference management, and data analytics on BIBTEX entries to recognize emerging research trends.

The organization of the BIBTEX *Analytics* repository is described as follows:

1. analysis:
  - (a) Scripts to analyze BIBTEX databases:
    - i. To remove metadata from them.
    - ii. To determine the set of keywords/keyphrases from their BIBTEX entries.
    - iii. To determine the set of publishers from their BIBTEX entries.
    - iv. To determine the set of series from their BIBTEX entries.
    - v. To determine the set of journals from their BIBTEX entries.
    - vi. To determine the set of authors from their BIBTEX entries.
    - vii. To determine the set of year from their BIBTEX entries.
2. `automated_regression_testing.py`:
  - (a) Run as: `./automated_regression_testing.py`
  - (b) No input nor output required.
  - (c) *Python* script to automate regression testing.
  - (d) **Sort of deprecated.** Still works though.
3. `big_input`:
  - (a) Data set for stress testing the software deliverable of the BIBTEX *Analytics* project.
4. database
  - (a) `bibtex_database` (`bibtex_database.py`) class represents (each instance of) a BIBTEX database of BIBTEX entries.
  - (b) `bibtex_database_test.py` is a *Python* script to test the functionality of the `bibtex_database` class.
  - (c) `entry` (`entry.py`) class represents each (instance of a) BIBTEX entry.
  - (d) `entry_test.py` is a *Python* script to test the functionality of the `entry` class.
  - (e) `key_check.py` is a *Python* script to check if each BIBTEX entry is valid.
  - (f) `key_check_test.py` is a *Python* script to test the functionality of the *Python* script `key_check.py`.
  - (g) `key_frequency_pairs.py` is a *Python* script to demonstrate how to sort a set of 2-tuples based on its first/former field and its second/last/latter field.
5. `duplicate_BibTeX_entries.py`:

- (a) Run as: `./duplicate_BibTeX_entries.py [-h] [BibTeX file]`
  - (b) A *Python* script to determine if duplicate BIB<sub>T</sub>E<sub>X</sub> entries exist in a BIB<sub>T</sub>E<sub>X</sub> file/database. If such entries exist, warn the user that duplicate BIB<sub>T</sub>E<sub>X</sub> entries exist.
6. `editions.py`:
- (a) Run as: `./editions.py [input BibTeX file] [-h]`
  - (b) A *Python* script to display a set of editions from all the BIB<sub>T</sub>E<sub>X</sub> entries in a BIB<sub>T</sub>E<sub>X</sub> file/-database.
7. `front_end`:
- (a) Parse input BIB<sub>T</sub>E<sub>X</sub> files, create BIB<sub>T</sub>E<sub>X</sub> entries (i.e., instances of *entry*), and insert them into the BIB<sub>T</sub>E<sub>X</sub> database, an instance of *bibtex\_database.py*.
8. `incremental_test.py`:
- (a) Run as: `./incremental_test.py [input BibTeX file]`
  - (b) A *Python* script to incrementally test features for performing reference management and data analytics operations with BIB<sub>T</sub>E<sub>X</sub> files/databases.
9. `input`:
- (a) A set of BIB<sub>T</sub>E<sub>X</sub> files to test my *Python*-based BIB<sub>T</sub>E<sub>X</sub> *Analytics* software.
10. `institutions.py`:
- (a) Run as: `./institutions.py [input BibTeX file] [-h]`
  - (b) A *Python* script to display a set of institutions from BIB<sub>T</sub>E<sub>X</sub> entries in a BIB<sub>T</sub>E<sub>X</sub> file/database.
11. `journal_titles.py`:
- (a) Run as: `./journal_titles.py [input BibTeX file] [-h]`
  - (b) A *Python* script to display a set of journal titles from BIB<sub>T</sub>E<sub>X</sub> entries in a BIB<sub>T</sub>E<sub>X</sub> database.
12. `keywords_display.py`:
- (a) Run as: `./keywords_display.py [input BibTeX file] [-h]`
  - (b) A *Python* script to display a set of keywords/keyphrases from BIB<sub>T</sub>E<sub>X</sub> entries in a BIB<sub>T</sub>E<sub>X</sub> database.
13. `makefile`:
- (a) For build automation of *Python* scripts, not placed in subdirectories, in the repository.
14. `notes`:
- (a) `gpl-license.text`, `LICENSE`, and `mit-license.text` are text files of the GNU General Public License (GNU GPL) (`gpl-license.text`) and The MIT License (`LICENSE` and `mit-license.text`)
  - (b) `guidelines`:
    - i. A document containing a set of guidelines on how to collaborate with me.
  - (c) `report`:
    - i. This document that describes the organization of the BIB<sub>T</sub>E<sub>X</sub> *Analytics* repository, the software architecture of the BIB<sub>T</sub>E<sub>X</sub> *Analytics* software, and future work of the BIB<sub>T</sub>E<sub>X</sub> *Analytics* project.
15. `organizations.py`:
- (a) Run as: `./organizations.py [input BibTeX file] [-h]`
  - (b) A *Python* script to display a set of organizations from BIB<sub>T</sub>E<sub>X</sub> entries in a BIB<sub>T</sub>E<sub>X</sub> database.
16. `publishers.py`:
- (a) Run as: `./publishers.py [input BibTeX file] [-h]`

- (b) A *Python* script to display a set of publishers from BIB<sub>T</sub>E<sub>X</sub> entries in a BIB<sub>T</sub>E<sub>X</sub> database.
- 17. `readme.md`:
  - (a) A *Markdown*-based `readme` document briefly describing this project.
- 18. `rm_bibtex_metadata.py`:
  - (a) Run as: `./rm_bibtex_metadata.py [input BibTeX file] [output BibTeX file] [-h]`
  - (b) `[output BibTeX file]` is an optional parameter.
  - (c) A *Python* script to delint/remove BIB<sub>T</sub>E<sub>X</sub> metadata from a BIB<sub>T</sub>E<sub>X</sub> database/file.
- 19. `sandbox`:
  - (a) A set of *Python* scripts to test different concepts in *Python*.
- 20. `statistics`:
  - (a) `test_statistics_tester.py` is a *Python* script to test the functionality of the `test_statistics` class.
  - (b) `test_statistics` (`test_statistics.py`) class to perform statistical analysis on results of automated testing of a *Python* script.
- 21. `tutti_series.py`:
  - (a) Run as: `./tutti_series.py [input BibTeX file] [-h]`
  - (b) A *Python* script to display series from BIB<sub>T</sub>E<sub>X</sub> entries in a BIB<sub>T</sub>E<sub>X</sub> database.
- 22. `utilities`:
  - (a) `file_io.py` is a *Python* script to perform input/output (I/O) operations on files, such as BIB<sub>T</sub>E<sub>X</sub> databases/files and L<sup>A</sup>T<sub>E</sub>X documents.
  - (b) `queue_ip_arguments.py` is a *Python* script to process input arguments for a script to clean BIB<sub>T</sub>E<sub>X</sub> databases/files.
- 23. `validate_url.py`:
  - (a) Run as: `./validate_url.py [input BibTeX file] [output BibTeX file] [-h]`
  - (b) `[output BibTeX file]` is an optional parameter.
  - (c) A *Python* script to check each BibTeX entry in a BibTeX database if it has the non-standard BIB<sub>T</sub>E<sub>X</sub> field(s) “Bdsk-Url-1” (and “Bdsk-Url-2”), and if the “Url” (and “Doi”) field(s) is(/are) missing; if these conditions are true, copy their values to the “Url” BibTeX field (and “Doi” field, if it is a DOI).
- 24. `z_booktitles.py`:
  - (a) Run as: `./z_booktitles.py [input BibTeX file] [-h]`
  - (b) A *Python* script to display booktitles from all BIB<sub>T</sub>E<sub>X</sub> entries in a BIB<sub>T</sub>E<sub>X</sub> database.

## Chapter 2

# Software Architecture of the BIB<sub>T</sub>E<sub>X</sub> Analytics Project

Software Architecture of the BIB<sub>T</sub>E<sub>X</sub> *Analytics* Project



# Chapter 3

## Future Work

Future work of the  $\text{BIB}\text{T}_{\text{E}}\text{X}$  *Analytics* project is described as follows:

1. Clustering of keywords/keyphrases:
  - (a) **Problem statements:**
    - i. For an author, find clusters of keyphrases, publishers, journal titles, conferences, ...
    - ii. For each keyphrase, determine the cluster of publishers, years, journal titles, conferences, ...
  - (b) Build dictionary of (*keyphrase*, *frequency*) two-tuples (or pairs).
  - (c) Sort the dictionary based on the frequency term/element, *frequency*, in these two-tuples.
  - (d) Alternate solution:
    - i. Build a set of (*keyphrase*, [*list of years*]) lists.
    - ii. [*list of years*] is a list of years of publications; or it is a set of years for publications that include the keyphrase *keyphrase*.
    - iii. Sort the set based on length of the list of years, [*list of years*].
- 2.
- 3.
- 4.
- 5.
- 6.
- 7.
8. Find emerging research trends to consider pivoting towards, or to get involved in side projects
  - (a) E.g., benchmark adiabatic quantum computers with topological computers and universal quantum computers [?].