

# DBMS II – CS3563

## Assignment 2 Report– Group 19

### Team Members:

- Tanmay Garg CS20BTECH11063
- Aayush Patel CS20BTECH11001
- Tanmay Goyal AI20BTECH11021
- Tanay Yadav AI20BTECH11026

### Obtaining the database:

Please follow the necessary steps in the README.md to dump the data into the database.

### Modifications to our ER Diagram:

#### 1. Publication Venue Entity:

1. The entire entity has been changed, so now there is no inheritance of conference and journal anymore. It has now been made part of Paper Entity itself

#### 2. Author Entity:

1. Added extra Similarity\_ID, to identify authors who have slight variations in their names.
2. For example, there are names in the source.txt file like **A. Bertossi and Alan A. Bertossi**. Such names can mean different people so the *attribute author\_id* is **different** for them but due to the constraint specification in assignment 1, their *similarity\_id* attribute is **same**. Thus names having such similar formats(like the two above) will have same *similarity\_id*. We achieved this using **Levenstein distance** calculation using the [RapidFuzz](#) library.

### 3. Authored Relation:

1. The relation between author and paper has been made mandatory on both sides
2. This was done according to the dataset provided as there are no authors who don't have a research paper

### **Libraries and Methodology Used:**

1. We have used python to parse through the entire source.txt file and generate 4 tsv files which are tab separated value files.
2. The tsv\_generator.py will generate the required tsv files.
3. The source.txt file was very unsanitised. It had a lot of **html and latex encodings**. For example: `&#365;`; To convert this to required unicode utf-8 character, we use the [html.parser](#) library.
4. The pg\_loader.py will dump the data from the generated tsv's into the postgres database. The library used here is [psycopg 2](#)
5. A fuzzer.py takes all the names of authors and generates a data/name\_similarity.txt file which contains the `author_name` and the `similarity_id` assigned to it based on fuzzy logic and Levenshtein Distance over a threshold of 90% as discussed in the point 2.2 on page 1. This has already been generated and is available in the data folder else everything is generated from source.txt itself.

### **Assumptions:**

1. Names which match character by character(after html and latex sanitisation), have been assigned the same *author\_id*.
2. Names which don't match exactly(like Ken Thompson and K. Thompson) might not be the same person, that's why they have **different** *author\_id* but **same** *similarity\_id*.

\*\*\*\*\*