DBMS 2

Assignment 2 Report

Team Members:

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

Modifications to our ER Diagram:

- **Publication Venue Entity**:
  - 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
- **Author Entity:**
  - Added extra Similarity_ID, to identify authors who have slight variations in their names
  - Such as in the dataset provided some of names of the authors have characters such as ä ç etc and the exact same names are there without those characters
- **Authored Relation:**
  - The relation between author and paper has been made mandatory on both sides
  - This was done according to the dataset provided as there are no authors who don't have a research paper

Libraries and Methodology Used:

- We have used ***python*** to parse through the entire source text file and generate 4 tsv files which are tab separated value files
- The 4 files are:
  - Authors.tsv: Has details of author
  - Author_paper.tsv: Has details of the relationship between research paper and author
  - Paper.tsv: Has details of the research paper
  - Citations.tsv: Has details of the citations
- We have another python file which creates all the tables in the database and loads all the data into the database
  - The library used here is Psycopg 2
- A python file then takes all the names of authors and assigns the similarity_id based on fuzzy logic and Levenshtein Distance over a threshold of 95% and then saves this tsv file
  - The library used for applying fuzzy is
    - thefuzz
    - RapidFuzz
- We have also sanitized the author names using libraries:
  - html.parser
- This file saves the author names tsv file along with the similarity IDs and this file should we used to load up into the database (it has already been generated by us and is available in the folder

- A readme file is also present to guide the user on how to run the program and the files present in the directory
- Please read through it carefully to execute the entire program

Assumptions Made:

- Names such as, Alan Turing and A. Turing may or may not have been considered as different depending upon the fuzzy logic program which calculates the similarity between the names
- Even though, A. Turing could also be Adam Turing or Alexander Turing, as regarding an author we just have a name and no other information to confirm the person's identity