CMPT 291 Mini Project 2

Nebye Berhe (berhe) | Tristan Carlson(tcarlson) | Victor Nguyen(victor3)

# Overview of Project

The goal of this project was to work with data at the physical layer. We built an information retrieval system via Berkeley DB library to query a database of email records. It efficiently evaluates various database queries using algorithms and index files.

This project consists of three phases, each of which contributes to the creation of a DBMS. In Phase 1, the read.py takes an XML data file and converts it into four output data text files required to make index files. In Phase 2, we take the output data text files and load it into an index database file using the index.sh shell script to make hash/B+ indexes. Finally, in Phase 3, the query.py with the help of query\_parser.py allows a user to make queries. These user input conditions will access the index files with the help of query.py and display the results.

The query.py uses the Berkeley DB Python API in order to reference any data via key-value pairs via. Which allows the program to either get an exact or partial match from a single value or a range of values from an index key.

# User Guide

* To build the output files from an XML datafile:
  + python read.py <XML file>
* To build the indexes:
  + ./index.sh
* To make queries:
  + python

# Algorithm to evaluate queries

# Testing Strategy

The testing for this project was done separately for each phase. For Phase 1, Victor had a **diff** command built in with his IDE ‘PyCharm’. He used this to compare the output of read.py with the test results given on eClass. The program would highlight the line that is different, which helped Victor pinpoint the error. For Phase 2, to test our shell script **index.sh** result and view our index files. We used the command **db\_dump -p** to read andwrite the results of the index files to a directory. We can then view each index file in a readable text file. Victor also used a shell script checker (reference in README.txt) to double check if the code was in the style shell script wanted.

For Phase 3,

# Group work breakdown

* Phase 1: Victor Nguyen
* Phase 2: Nebye Berhe / Victor Nguyen
* Phase 3: Tistan Carlson / Nebye Berhe / Victor Nguyen
* Project Document: Victor Nguyen
* Testing: Tistan Carlson / Nebye Berhe / Victor Nguyen