CMPUT 291 - Mini Project 2 Design Document

1 Overview

The following software package is cmprosed of 3 phases. Phase 1 takes email data from an XML file and outputs the data into four files: terms.txt, emails.txt, dates.txt, and recs.txt. Phase 2 sorts these files and builds four indexes. Phase 3 is responsible for data retrival, and queries the data.

1.1 User Guide

1.1.1 Phase 1: Preparing Data Files

The software for phase 1 is a python program that reads email records in an XML file and prodcues four output files: terms.txt, emails.txt, dates.txt, and recs.txt. To run phase 1, execute the program in the terminal

```
$ python3 phase1.db
```

The program will propmt the user to enter an XML file.

```
Enter .xml file: datafile.xml
```

If the provided file is not found, or an incorrect file extension is given, an error message will display, and the user can re-enter a filename. If no extension is specified, the program will assume it is .xml.

When a valid file is entered, phase 1 parses the XML file into the four text files mentioned above. If those files already exist in the directory, the program will overwrite the data currently on the files.