File Analysis

Systems Programming | Section 5

Names:
Christopher Naporlee - cmn134
Michael Nelli - mrn73

Description

Our File Analyzer takes in a directory as an argument and walks down the path, recursively, until it reaches the bottom of the directory tree. During this walk, we take note of files and the words they contain. Each file and directory is handled on their own thread to decrease the time it takes to parse all files and scan their words. These threads share a file database to store their file entries and the files words.

The file database is then used to calculate the Jensen Shannon Distance (JSD) between each pair of files within the database. After calculations have been completed, the program prints out the JSD for each pair with colors representing how similar the files are.

Design

1. int main(int argc, char **argv)
    - Ensures that the user is passing in a valid directory path
    - Calls start_dirhandler() on the directory path to parse the directory.
    - Once returned from start_dirhandler(), begins calculation JSD of filepairs in the file database.
    - Prints JSD output.
2. void *start_dirhandler(void *thread_data)
    - Handles opening and parsing directories.
    - If a directory is found, starts a new pthread_t and will call start_dirhandler on that subdirectory.
    - If a file is found, starts a new pthread_t and will call start_filehandler on that file.

3. void *start_filehandler(void *thread_data)
    - Handles opening and parsing files, and creating entries of a file in the database.
    - For every word in the file, it adds it to a list of already found words, sorted alphabetically.
        - ➢ Words only count as alphabetical characters and '-'.
    - After every word has been parsed, passes through the word list and updates the frequency of each word
        - ➢ Frequency = (# times word seen / total # of words)
4. struct file_pair *compare_files(struct file_database *)
    - Creates a list of all possible file pairs, with each pair being stored in a file_pair node.
        - ➢ Total $\binom{n}{2}$ pairs where n = number of files.
    - Compares all of the words in both files, finding the mean probability of each word and storing it in a file_pair node. This allows us to later find the KLD.
5. void get_JSD_values(struct file_pair *list_ptr)
    - Goes through a list of file pairs and for each pair calculates JSD.

NOTE: Test cases for the program can be found in testcases.txt