BIS 420 PROGRAMMING FOR DATA SCIENCE

PRAJAKTA POHARE

CHAPTER 14 EXERCISE 14.4

ILLINOIS STATE UNIVERSITY

In a large collection of MP3 files, there may be more than one copy of the same song, stored in different directories or with different file names. The goal of this exercise is to search for duplicates.

1. Write a program that searches a directory and all of its subdirectories, recursively, and returns a list of complete paths for all files with a given suffix (like .mp3). Hint: os.path provides several useful functions for manipulating file and path names.

2. To recognize duplicates, you can use md5sum to compute a “checksum” for each files. If two files have the same checksum, they probably have the same contents.

3. To double-check, you can use the Unix command diff.