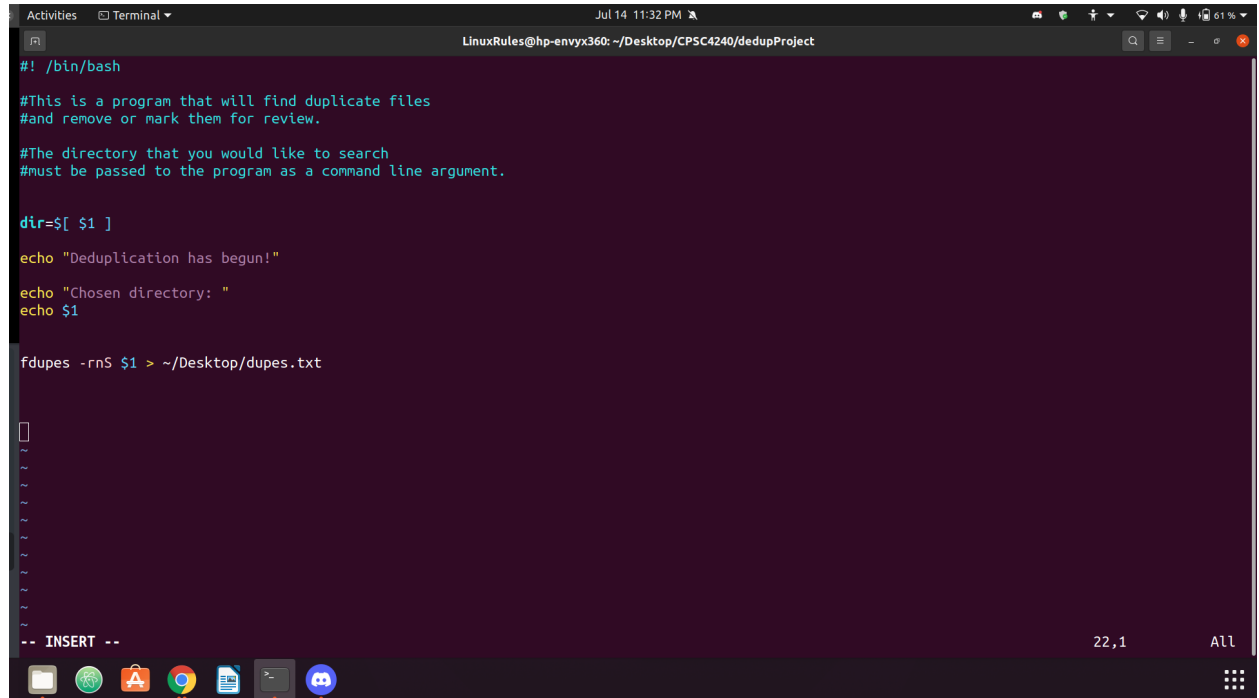


## CPSC 4240 Milestone 2

We have found our four articles to back why what we are doing is so important. One of them was mentioned in milestone one and will be explained more in depth now. While reading this article "<https://techchannel.com/SMB/10/2018/Human-Error-Top-Cause-of-Downtime>," we found the following quote, "Information Technology Intelligence Corp.'s (ITIC) 2018 Global Server Hardware, Server OS Reliability Survey ([ibm.co/2LCX6gz](http://ibm.co/2LCX6gz)), which surveyed more than 800 customers worldwide, found that 59 percent of respondents cited human error as the No. 1 cause of unplanned downtime." This article shows the prevalence and reason that human error is cited and believed to cause. Taking that it is not hard to assume that human error from accidental duplication can easily add up. Next is an article that goes into many different reasons as to why data duplication is bad for a company one of the main ones it talks about is the amount to store the same data twice sure if it's one or two files, but the bigger the company the bigger that begins to add up in unnecessary storage space and storage cost. All of these reasons are in "<https://www.dataaxlegenie.com/blog/this-is-why-duplicate-data-is-bad-for-you/>." We found an article as well that lists a good way to take care of the duplicate data problem and then gives it a name for what we are doing. It is called data deduplication and it is introduced as a solution to data duplication caused by human error in the following quote, "To counteract the problems outlined, the solution is a process called, unsurprisingly, data deduplication. This is a blend of human insight, data processing and algorithms to help identify potential duplicates based on likelihood scores and common sense to identify where records look like a close match. This process involves analysing your database to identify potential duplicate records, and unravelling these to identify definite duplicates." This suggests the exact method that we are using and goes even further to give us the same insight to the need to think through all the possible duplicates that can come about, "Deduplication rules also need to be implemented, based on your own unique data issues, in order to create a bespoke deduplication strategy. The rules should take into account your decisions about how 'strict' you want to be with your deduplication, in terms of maintaining the balance between losing valuable customer data and having a clean database." The last two quotes are both pulled from the following article, "<https://www.mycustomer.com/marketing/data/why-duplicate-records-are-costly-for-your-company-and-what-can-be-done-about-it>." Our last article also has to do with future advancement of this project. It has a good suggestion to take into consideration when data deduplication is occurring and you are deleting all the duplicates through an algorithm. "...before you go the automated route, consider a safer option: Set your duplicate file finder to ignore files smaller than 20MB. That way, you'll have far fewer files to worry about, yet still free up a lot of space." This is definitely something we will look into doing in the future with this where we set the search to only worry about the big storage takers. This would make more sense seeing as they cause the biggest problem in the first place (pun intended). The previous suggestion was made from the last of our four articles, "<https://www.pcworld.com/article/2039794/automatically-delete-a-huge-amount-of-duplicate-files.html>."

As of right now we currently are able to get all the duplicate files of a chosen directory and their file size and we then print them in a file we created on the desktop. For future plans we plan to take that file and sort the file sizes to then delete them based on file size. For testing currently, first we are not having the script search the entire filesystem instead we are starting simple with the duplicates being in the same directory. For the duplicates we created duplicate .txt files to make sure we have control of the testing parameters. In the future the first feature for duplicates to test is using two duplicate .txt files one that has words and the other being blank. This way the one with words will be a larger file than the empty one. Our bash script deletes the one that is a smaller file. We will then still use .txt files but we will methodically test each different problem the script could run into making sure our test cases cover what we are looking

for. All of our tests are coming out positive and we have started from the ground and will be slowly adding more and more features to the script. We believe by presentation time that we will have the script looking through the entire file system and deleting the duplicates with most of the problems we expect that could arise in the search. Below you will see our code for the bash script and the text file that the script is piping the output into.



```
LinuxRules@hp-envyx360: ~/Desktop/CPSC4240/dedupProject
#!/bin/bash

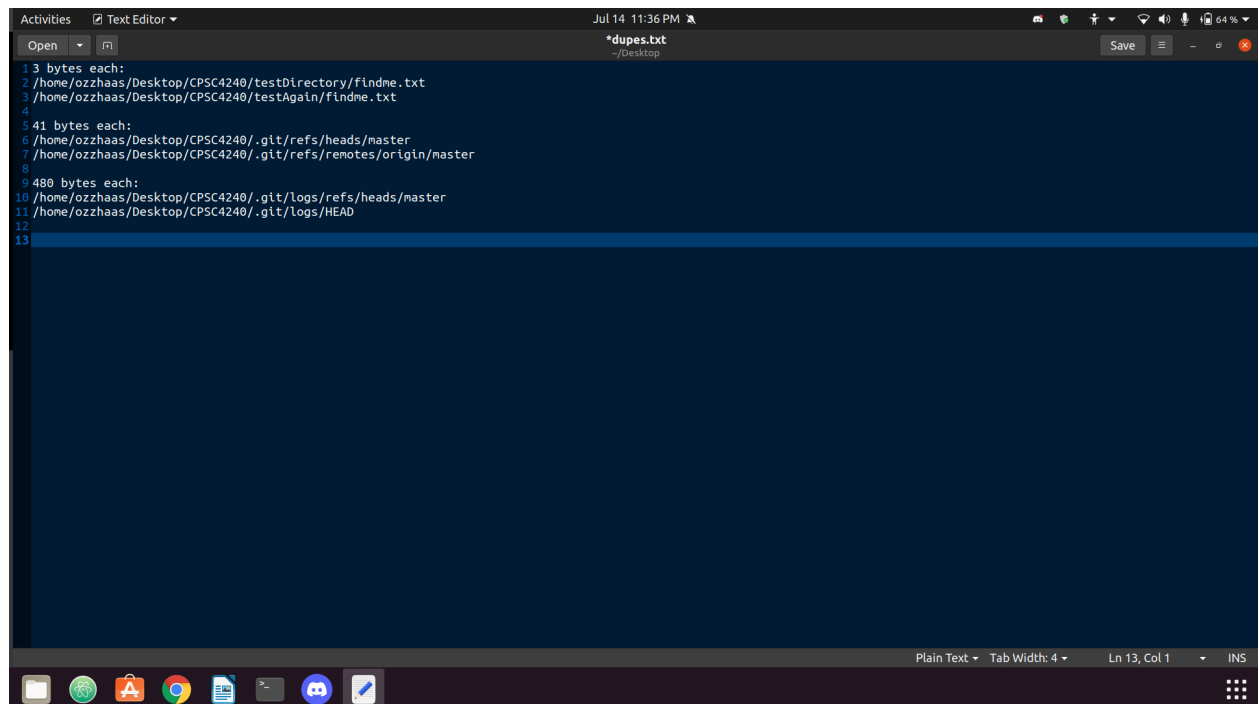
#This is a program that will find duplicate files
#and remove or mark them for review.

#The directory that you would like to search
#must be passed to the program as a command line argument.

dir=${1}

echo "Deduplication has begun!"
echo "Chosen directory: "
echo $1

fdupes -rnS $1 > ~/Desktop/dupes.txt
```



```
*dupes.txt
~/Desktop
Save

1 3 bytes each:
2 /home/ozzhaas/Desktop/CPSC4240/testDirectory/findme.txt
3 /home/ozzhaas/Desktop/CPSC4240/testAgain/findme.txt
4
5 41 bytes each:
6 /home/ozzhaas/Desktop/CPSC4240/.git/refs/heads/master
7 /home/ozzhaas/Desktop/CPSC4240/.git/refs/remotes/origin/master
8
9 480 bytes each:
10 /home/ozzhaas/Desktop/CPSC4240/.git/logs/refs/heads/master
11 /home/ozzhaas/Desktop/CPSC4240/.git/logs/HEAD
12
13
```