Duplicate Analysis:Quantify the files that exist in more than one directory

As part of Candidate Interview

Mir AlMasud

November 2, 2021

Introduction

Context

- There is a set of files in NAS to be loaded into a PLM system
- Only single version can be loaded even if multiple copies exist in the disk
- file existing in multiple directories make which version to load ambiguous

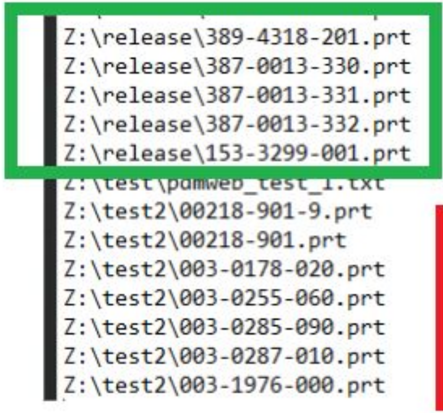
Problem statement

We need to quantify the files that exist in more than one directory

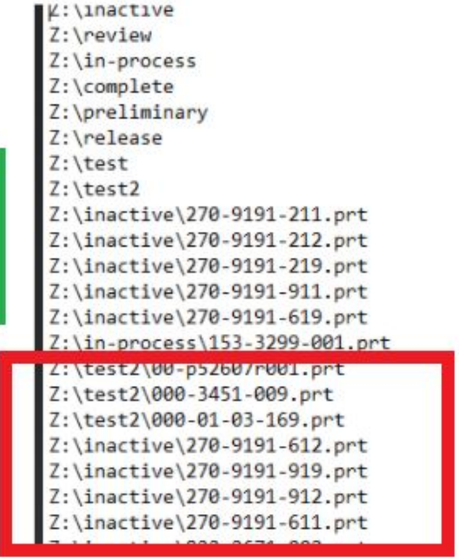
Points to consider

1: Extra Directories

- 173609 line entries in the nas_top_level_dir.txt
- Directories exist that is not based on lifecycle state
- [“Review”, “release”, “in-process”] Matters



```
Z:\release\389-4318-201.prt
Z:\release\387-0013-330.prt
Z:\release\387-0013-331.prt
Z:\release\387-0013-332.prt
Z:\release\153-3299-001.prt
Z:\test\pdmweb_test_1.txt
Z:\test2\00218-901-9.prt
Z:\test2\00218-901.prt
Z:\test2\003-0178-020.prt
Z:\test2\003-0255-060.prt
Z:\test2\003-0285-090.prt
Z:\test2\003-0287-010.prt
Z:\test2\003-1976-000.prt
```



```
Z:\inactive
Z:\review
Z:\in-process
Z:\complete
Z:\preliminary
Z:\release
Z:\test
Z:\test2
Z:\inactive\270-9191-211.prt
Z:\inactive\270-9191-212.prt
Z:\inactive\270-9191-219.prt
Z:\inactive\270-9191-911.prt
Z:\inactive\270-9191-619.prt
Z:\in-process\153-3299-001.prt
Z:\test2\00-p52607r001.prt
Z:\test2\000-3451-009.prt
Z:\test2\000-01-03-169.prt
Z:\inactive\270-9191-612.prt
Z:\inactive\270-9191-919.prt
Z:\inactive\270-9191-912.prt
Z:\inactive\270-9191-611.prt
```

Points to consider

2: Case Insensitive

- This is a windows NAS
- Filenames are not case sensitive
- Detection algorithm should ignore case

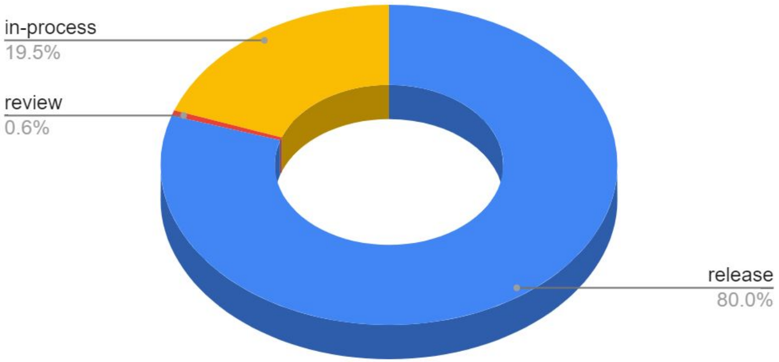
Points to consider

3: unsorted

- Filenames are not in any particular order
- An alphabetical listing will require sorting

Total Files Distribution

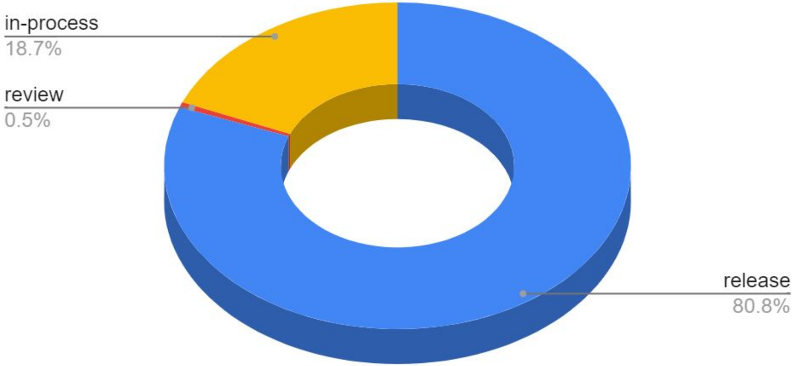
Number of files in release	:	138753
Number of files in review	:	969
Number of files in in-process	:	33766
===================================		
Total:		173488



Unique Files Distribution:

Files that has no duplicate

Number of unique files in release	:	136319
Number of unique files in review	:	841
Number of unique files in in-process	:	31451
======================================		
Total:		168611



Duplicate Files

Duplicates	**Number**
Number of files in both release and review:	123
Number of files in both release and in-process:	2310
Number of files in both review and in-process:	4
Number of files existing in all 3 directory	1
Total	**2438**

Observations:

- We are unable to check if file with same name are identical
- We can’t compare the checksum
- We can only assume they are same unless more information is provided about these files (i.e., metadata, timestamp,history, version number)
- All files are .prt files

Programming Solution

script :   duplicate_analysis.py

Input: nas_top_level_dir.txt

Output: duplicate_analysis.txt

Data structure

```
prt_dict={   'file1': {
                       ‘count’: 1,
                       'in-process': 0,
                       'release': 1,
                       'review': 0
                       },
             'file2': {
                       ‘count’: 2,
                       'in-process': 0,
                       'release': 1,
                       'review': 1
                       },
             .
             .
             .
             }
```

Summary: Alphabetical listing

```
no.   |filenames            : review   in-process release
===============================================================
     1|000588d-00.prt       :        0        1        1
     2|003-0287-020.prt     :        0        1        1
     3|003-1424-000.prt     :        0        1        1
     4|003-1424-070.prt     :        0        1        1
     5|01-p40203k.prt       :        0        1        1
     6|01-p40397k001.prt    :        0        1        1
     7|01-p56123u.prt       :        0        1        1
     8|015-4101-020.prt     :        0        1        1
     9|015-4101-040.prt     :        0        1        1
    10|015-4101-050.prt     :        0        1        1
```

https://github.com/almasudme/duplicate_analysis/blob/main/duplicate_analysis.py

Questions?