

# PA2

John Vincent  
Ryan McCullough

## MinHash

### Collecting Terms

To collect terms we get all of the file names in the given directory. We then use another class called Preprocessor to read the file line by line and match valid terms on demand. A Preprocessor is created for each file one after another so that only one file is open at any time. We collect the terms in each document by creating a list of hashmaps one for each document. When a term is found in a document it gets set in the documents hashmap. After all the documents are read the keylist of the documents hashmaps are used to get all the unique terms in each document.

### Assigning Integers to Terms

We use another hashmap that maps strings to integers to assign a value to each term. While reading the files if a read term is not a key in this hashmap then it is added to the map with a value equal to the current value of a counter. The counter is then incremented. This maps each term to a integer value from 0 to N-1 where N is the number of unique terms read.

### Permutations

We used the form recommended for our permutations  $(ax+b)\%p$ . For each permutation a different prime is used start at the next prime larger than the total number of unique terms. Each permutation has a p value that is the next prime larger than the one used in the previous permutation. The a and b values are random numbers that are at most p-1 where p is the prime corresponding to the given a and b.

## MinHashAccuracy

500500 total comparisons	400	600	800
.04	3769	435	0
.07	10	2	0
.09	0	0	0

Conclusion: Adding more permutations increases the accuracy of MinHash and predictably increasing the error margin decreases the number of inaccurate pairs found. It's clear that with a higher permutation count the amount of error decreases. The increase in accuracy is not as drastic as the increase in computational time and memory usage however, as all three were within 9% of the actual jaccard similarity.

## Screenshots from the accuracy test

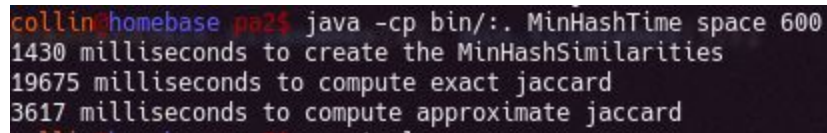
```
space-259.txt space-458.txt
  epsilon 0.0426 approx: 0.1550 exact: 0.1124
space-547.txt space-458.txt
  epsilon 0.0409 approx: 0.1800 exact: 0.1391
space-357.txt space-458.txt
  epsilon 0.0445 approx: 0.1425 exact: 0.0980
space-691.txt space-746.txt
  epsilon 0.0429 approx: 0.1600 exact: 0.1171
space-458.txt space-502.txt
  epsilon 0.0417 approx: 0.1450 exact: 0.1033
space-458.txt space-661.txt
  epsilon 0.0438 approx: 0.1475 exact: 0.1037
space-458.txt space-391.txt
  epsilon 0.0446 approx: 0.2050 exact: 0.1604
space-458.txt space-767.txt
  epsilon 0.0424 approx: 0.1825 exact: 0.1401
space-647.txt space-988.txt
  epsilon 0.0506 approx: 0.2525 exact: 0.2019
space-91.txt space-366.txt
  epsilon 0.0431 approx: 0.1125 exact: 0.1556
3769/500500 pairs of documents had error greater than 0.04
collin@homebase pa2$ java -cp bin/. MinHashAccuracy space 400 .07
space-627.txt space-613.txt
  epsilon 0.0705 approx: 0.5250 exact: 0.4545
space-303.txt space-253.txt
  epsilon 0.0821 approx: 0.5600 exact: 0.4779
space-36.txt space-130.txt
  epsilon 0.0707 approx: 0.3175 exact: 0.2468
space-863.txt space-317.txt
  epsilon 0.0764 approx: 0.2375 exact: 0.1611
space-130.txt space-418.txt
  epsilon 0.0827 approx: 0.2350 exact: 0.1523
space-130.txt space-702.txt
  epsilon 0.0784 approx: 0.3575 exact: 0.2791
space-99.txt space-858.txt
  epsilon 0.0721 approx: 0.3150 exact: 0.2429
space-632.txt space-253.txt
  epsilon 0.0763 approx: 0.3500 exact: 0.2737
space-561.txt space-317.txt
  epsilon 0.0716 approx: 0.2275 exact: 0.1559
space-860.txt space-659.txt
  epsilon 0.0891 approx: 0.4425 exact: 0.3534
10/500500 pairs of documents had error greater than 0.07
collin@homebase pa2$ java -cp bin/. MinHashAccuracy space 400 .09
0/500500 pairs of documents had error greater than 0.09
collin@homebase pa2$ spectacle
```

```
space-57.txt space-404.txt
  epsilon 0.0480 approx: 0.1667 exact: 0.1186
space-809.txt space-547.txt
  epsilon 0.0433 approx: 0.2717 exact: 0.2283
space-133.txt space-338.txt
  epsilon 0.0407 approx: 0.1767 exact: 0.1360
space-930.txt space-872.txt
  epsilon 0.0426 approx: 0.2733 exact: 0.2308
space-207.txt space-206.txt
  epsilon 0.0467 approx: 0.2517 exact: 0.2049
space-207.txt space-688.txt
  epsilon 0.0462 approx: 0.2133 exact: 0.1672
space-207.txt space-547.txt
  epsilon 0.0427 approx: 0.2200 exact: 0.1773
space-207.txt space-91.txt
  epsilon 0.0417 approx: 0.2083 exact: 0.1667
space-207.txt space-178.txt
  epsilon 0.0497 approx: 0.2300 exact: 0.1803
space-95.txt space-661.txt
  epsilon 0.0451 approx: 0.1733 exact: 0.1282
space-206.txt space-688.txt
  epsilon 0.0408 approx: 0.2417 exact: 0.2009
435/500500 pairs of documents had error greater than 0.04
collin@homebase pa2$ java -cp bin/. MinHashAccuracy space 600 .07
space-422.txt space-520.txt
  epsilon 0.0702 approx: 0.4017 exact: 0.3315
space-654.txt space-944.txt
  epsilon 0.0727 approx: 0.3600 exact: 0.2873
2/500500 pairs of documents had error greater than 0.07
collin@homebase pa2$ java -cp bin/. MinHashAccuracy space 600 .09
0/500500 pairs of documents had error greater than 0.09
collin@homebase pa2$ spectacle
```

```
pa2 : spectacle — Konsole
collin@homebase pa2$ java -cp bin/. MinHashAccuracy space 800 .4
0/500500 pairs of documents had error greater than 0.4
collin@homebase pa2$ java -cp bin/. MinHashAccuracy space 800 .7
0/500500 pairs of documents had error greater than 0.7
collin@homebase pa2$ java -cp bin/. MinHashAccuracy space 800 .9
0/500500 pairs of documents had error greater than 0.9
collin@homebase pa2$ spectacle
```

## MinHashTime

It took 19675 milliseconds to compute the exact jaccard and 3617 to compute the approximate. Making it 5.5 times faster to compute the approximate, especially since its very accurate as discussed above.

A terminal window with a dark background and light-colored text. The prompt is 'collin@homebase:pe2\$'. The command 'java -cp bin/:. MinHashTime space 600' has been executed. The output shows three lines of timing information: '1430 milliseconds to create the MinHashSimilarities', '19675 milliseconds to compute exact jaccard', and '3617 milliseconds to compute approximate jaccard'.

```
collin@homebase:pe2$ java -cp bin/:. MinHashTime space 600
1430 milliseconds to create the MinHashSimilarities
19675 milliseconds to compute exact jaccard
3617 milliseconds to compute approximate jaccard
```

## Near Duplicate

For this section screenshots will be used to show the duplicates for the 10 different files because there are a lot of them

1000 permutations,  $s=.9$

```
collin@homebase ps2$ java -cp bin/. NearDuplicateTest F17PA2/ 1000 .9
Command: dups space-200.txt
baseball238.txt.copy1, baseball611.txt, baseball678.txt.copy6
baseball786.txt.copy6, baseball830.txt.copy1, baseball882.txt.copy1
baseball897.txt, baseball897.txt.copy4, baseball897.txt.copy7
baseball93.txt.copy1, hockey186.txt.copy4, hockey188.txt.copy4
hockey398.txt.copy2, hockey541.txt.copy7, hockey547.txt.copy1
hockey554.txt.copy6, hockey674.txt.copy4, hockey689.txt.copy4
hockey718.txt, hockey73.txt.copy1, hockey85.txt.copy2
hockey856.txt.copy6, hockey883.txt.copy5, hockey893.txt.copy3
hockey925.txt.copy3, space-200.txt.copy1, space-200.txt.copy2
space-200.txt.copy3, space-200.txt.copy4, space-200.txt.copy5
space-200.txt.copy6, space-200.txt.copy7, space-333.txt.copy3
space-368.txt, space-368.txt.copy5, space-886.txt.copy7
space-886.txt.copy7
Command: dups baseball22.txt
baseball141.txt.copy7, baseball22.txt.copy1, baseball22.txt.copy2
baseball22.txt.copy3, baseball22.txt.copy4, baseball22.txt.copy5
baseball22.txt.copy6, baseball22.txt.copy7, baseball327.txt.copy7
baseball452.txt, baseball452.txt.copy1, baseball452.txt.copy2
baseball452.txt.copy3, baseball452.txt.copy4, baseball452.txt.copy5
baseball452.txt.copy6, baseball452.txt.copy7, baseball628.txt.copy3
baseball642.txt.copy7, baseball689.txt, baseball689.txt.copy1
baseball689.txt.copy5, baseball689.txt.copy6, baseball8.txt.copy6
baseball814.txt, baseball814.txt.copy1, baseball814.txt.copy7
baseball853.txt.copy6, hockey0.txt.copy2, hockey385.txt.copy7
hockey813.txt.copy7, hockey844.txt, hockey944.txt.copy1
hockey998.txt.copy1, space-187.txt.copy1, space-191.txt.copy1
space-529.txt.copy2, space-768.txt.copy1, space-858.txt.copy5
space-858.txt.copy5
Command: dups hockey103.txt
baseball18.txt.copy6, baseball261.txt.copy2, baseball417.txt.copy6
baseball427.txt.copy1, baseball608.txt.copy6, baseball657.txt.copy1
baseball80.txt.copy6, baseball850.txt.copy5, hockey103.txt.copy1
hockey103.txt.copy2, hockey103.txt.copy3, hockey103.txt.copy4
hockey103.txt.copy5, hockey103.txt.copy6, hockey103.txt.copy7
hockey233.txt.copy1, hockey266.txt.copy1, hockey369.txt
hockey401.txt.copy2, hockey486.txt.copy2, hockey504.txt
hockey504.txt.copy1, hockey504.txt.copy4, hockey504.txt.copy5
hockey541.txt.copy6, hockey562.txt.copy3, hockey746.txt.copy7
hockey926.txt.copy7, hockey969.txt, hockey969.txt.copy1
hockey969.txt.copy3, hockey969.txt.copy4, hockey969.txt.copy6
space-383.txt.copy4, space-42.txt.copy7, space-599.txt.copy5
space-606.txt.copy5, space-612.txt.copy3, space-640.txt.copy1
```



Command: dups baseball200.txt

baseball200.txt.copy1, baseball200.txt.copy2, baseball200.txt.copy3  
baseball200.txt.copy4, baseball200.txt.copy5, baseball200.txt.copy6  
baseball200.txt.copy7, baseball305.txt.copy2, baseball317.txt  
baseball317.txt.copy3, baseball565.txt, baseball565.txt.copy2  
baseball565.txt.copy5, baseball565.txt.copy6, baseball657.txt.copy1  
baseball797.txt.copy3, baseball982.txt.copy2, hockey140.txt.copy5  
hockey180.txt.copy7, hockey655.txt.copy7, hockey714.txt.copy1  
hockey787.txt, hockey822.txt.copy5, hockey965.txt.copy5  
space-125.txt.copy1, space-169.txt.copy3, space-175.txt.copy2  
space-20.txt.copy4, space-200.txt.copy6, space-310.txt.copy4  
space-367.txt.copy5, space-382.txt, space-382.txt.copy3  
space-382.txt.copy5, space-405.txt.copy5, space-407.txt.copy4  
space-428.txt.copy7, space-559.txt.copy3, space-60.txt.copy5  
space-626.txt.copy7, space-806.txt.copy6, space-81.txt.copy6  
space-81.txt.copy6

Command: dups baseball308.txt

baseball104.txt.copy5, baseball139.txt.copy7, baseball308.txt.copy1  
baseball308.txt.copy2, baseball308.txt.copy3, baseball308.txt.copy4  
baseball308.txt.copy5, baseball308.txt.copy6, baseball308.txt.copy7  
baseball410.txt.copy7, baseball468.txt.copy6, baseball476.txt.copy1  
baseball563.txt.copy7, baseball600.txt.copy5, baseball830.txt.copy3  
baseball955.txt.copy3, hockey160.txt.copy2, hockey447.txt.copy2  
hockey606.txt.copy4, hockey630.txt.copy3, space-170.txt.copy2  
space-284.txt.copy2, space-317.txt.copy5, space-366.txt  
space-366.txt.copy1, space-366.txt.copy6, space-366.txt.copy7  
space-375.txt.copy5, space-40.txt.copy4, space-461.txt.copy5  
space-483.txt, space-483.txt.copy3, space-550.txt  
space-550.txt.copy1, space-550.txt.copy4, space-550.txt.copy5  
space-550.txt.copy7, space-569.txt.copy7, space-670.txt.copy1

Command: dups hockey4.txt

baseball15.txt.copy1, baseball241.txt.copy6, baseball415.txt.copy5  
baseball43.txt, baseball43.txt.copy1, baseball43.txt.copy2  
baseball43.txt.copy3, baseball43.txt.copy4, baseball43.txt.copy5  
baseball43.txt.copy6, baseball43.txt.copy7, baseball459.txt  
baseball459.txt.copy3, hockey188.txt.copy5, hockey4.txt.copy1  
hockey4.txt.copy2, hockey4.txt.copy3, hockey4.txt.copy4  
hockey4.txt.copy5, hockey4.txt.copy6, hockey4.txt.copy7  
hockey626.txt.copy4, space-120.txt.copy6, space-121.txt.copy6  
space-308.txt.copy1, space-448.txt.copy4, space-481.txt.copy2  
space-492.txt.copy2, space-628.txt.copy1, space-668.txt.copy1  
space-706.txt.copy2, space-71.txt.copy6, space-749.txt.copy5  
space-9.txt.copy6, space-9.txt.copy6

Command: dups hockey8.txt

baseball546.txt.copy7, baseball556.txt.copy1, baseball614.txt  
baseball674.txt.copy2, baseball740.txt.copy3, baseball89.txt.copy7  
baseball920.txt.copy4, hockey135.txt.copy5, hockey14.txt.copy3  
hockey555.txt.copy7, hockey609.txt.copy2, hockey8.txt.copy1  
hockey8.txt.copy2, hockey8.txt.copy3, hockey8.txt.copy4  
hockey8.txt.copy5, hockey8.txt.copy6, hockey8.txt.copy7  
space-361.txt.copy1, space-38.txt.copy4, space-551.txt.copy2  
space-653.txt, space-653.txt.copy1, space-653.txt.copy2  
space-653.txt.copy4, space-653.txt.copy7, space-774.txt.copy2

Command: dups hockey20.txt

baseball1239.txt, baseball1239.txt.copy4, baseball1239.txt.copy5  
baseball1239.txt.copy7, baseball1550.txt.copy2, baseball607.txt.copy4  
baseball613.txt.copy4, baseball812.txt.copy7, baseball863.txt.copy3  
hockey20.txt.copy1, hockey20.txt.copy2, hockey20.txt.copy3  
hockey20.txt.copy4, hockey20.txt.copy5, hockey20.txt.copy6  
hockey20.txt.copy7, hockey449.txt.copy5, hockey457.txt.copy5  
hockey584.txt.copy6, hockey716.txt.copy2, space-247.txt.copy4  
space-250.txt.copy1, space-362.txt, space-362.txt.copy2  
space-362.txt.copy3, space-362.txt.copy4, space-362.txt.copy7  
space-386.txt.copy6, space-604.txt.copy5, space-652.txt  
space-652.txt.copy5, space-652.txt.copy6, space-660.txt.copy7  
space-666.txt.copy4, space-682.txt.copy7, space-69.txt.copy7  
space-69.txt.copy7

Command: dups baseball23.txt

baseball167.txt, baseball167.txt.copy2, baseball167.txt.copy4  
baseball167.txt.copy5, baseball23.txt.copy1, baseball23.txt.copy2  
baseball23.txt.copy3, baseball23.txt.copy4, baseball23.txt.copy5  
baseball23.txt.copy6, baseball23.txt.copy7, baseball371.txt.copy4  
baseball390.txt.copy4, baseball567.txt.copy1, baseball618.txt.copy3  
baseball8.txt.copy3, baseball926.txt.copy3, hockey156.txt.copy6  
hockey296.txt.copy1, hockey378.txt.copy6, hockey429.txt.copy1  
hockey58.txt.copy2, hockey733.txt.copy3, hockey757.txt.copy1  
hockey997.txt.copy5, space-228.txt.copy7, space-335.txt  
space-335.txt.copy5, space-335.txt.copy7, space-367.txt.copy6  
space-532.txt.copy2, space-946.txt, space-946.txt.copy1  
space-946.txt.copy2, space-946.txt.copy3, space-946.txt.copy4

Command: dups space-33.txt

baseball156.txt, baseball156.txt.copy5, baseball156.txt.copy7  
baseball264.txt, baseball264.txt.copy5, baseball320.txt  
baseball320.txt.copy7, baseball385.txt.copy3, baseball449.txt.copy2  
baseball489.txt.copy2, baseball493.txt.copy1, baseball502.txt.copy5  
baseball540.txt.copy3, baseball555.txt.copy5, baseball556.txt.copy7  
baseball745.txt.copy3, baseball818.txt, baseball818.txt.copy5  
baseball818.txt.copy7, baseball903.txt.copy2, hockey143.txt.copy4  
hockey398.txt.copy4, hockey401.txt, hockey401.txt.copy5  
hockey433.txt.copy6, hockey486.txt.copy6, hockey505.txt  
space-33.txt.copy1, space-33.txt.copy2, space-33.txt.copy3  
space-33.txt.copy4, space-33.txt.copy5, space-33.txt.copy6  
space-33.txt.copy7, space-374.txt.copy2, space-461.txt.copy7  
space-813.txt.copy7, space-813.txt.copy7

Command: exit



1000 permutations,  $s=.95$

```
collin@homebase pa2$ java -cp bin/:. NearDuplicateTest F17PA2/ 1000 .  
Command: dups space-200.txt  
baseball101.txt, baseball101.txt.copy2, baseball422.txt.copy1  
hockey411.txt, hockey450.txt.copy2, hockey478.txt  
hockey478.txt.copy4, hockey699.txt.copy3, hockey844.txt.copy6  
space-113.txt, space-113.txt.copy2, space-113.txt.copy3  
space-200.txt.copy1, space-200.txt.copy2, space-200.txt.copy3  
space-200.txt.copy4, space-200.txt.copy5, space-200.txt.copy6  
space-200.txt.copy7, space-265.txt.copy7, space-412.txt.copy2  
space-425.txt.copy1, space-439.txt.copy2, space-537.txt  
space-811.txt.copy1, space-811.txt.copy1  
Command: dups baseball22.txt  
baseball172.txt.copy4, baseball22.txt.copy1, baseball22.txt.copy2  
baseball22.txt.copy3, baseball22.txt.copy5, baseball22.txt.copy6  
baseball22.txt.copy7, baseball271.txt, baseball317.txt.copy5  
baseball4.txt.copy6, baseball667.txt, baseball667.txt.copy4  
baseball724.txt.copy3, baseball793.txt.copy6, baseball8.txt.copy3  
baseball896.txt.copy1, baseball981.txt, hockey165.txt.copy4  
hockey319.txt.copy7, hockey347.txt.copy1, hockey565.txt.copy3  
hockey642.txt.copy1, hockey748.txt.copy5, hockey93.txt.copy5  
space-382.txt.copy2, space-382.txt.copy2  
Command: dups hockey103.txt.copy3  
baseball216.txt.copy3, baseball268.txt.copy3, baseball373.txt.copy2  
baseball75.txt, baseball75.txt.copy3, baseball864.txt.copy5  
baseball983.txt.copy5, hockey103.txt, hockey103.txt.copy4  
hockey103.txt.copy5, hockey103.txt.copy7, hockey142.txt.copy3  
hockey274.txt.copy3, hockey345.txt.copy6, hockey438.txt.copy2  
hockey451.txt.copy1, hockey502.txt, hockey502.txt.copy2  
hockey887.txt.copy7, hockey961.txt.copy1, space-121.txt  
space-223.txt.copy4, space-329.txt.copy4, space-655.txt.copy5  
space-655.txt.copy5
```



```
Command: dups baseball362.txt
baseball233.txt.copy2, baseball362.txt.copy1, baseball362.txt.copy2
baseball362.txt.copy3, baseball362.txt.copy4, baseball362.txt.copy5
baseball362.txt.copy6, baseball362.txt.copy7, baseball714.txt.copy3
baseball862.txt.copy1, hockey207.txt.copy6, hockey258.txt.copy5
hockey311.txt.copy3, hockey373.txt.copy6, hockey374.txt.copy2
hockey395.txt.copy1, hockey422.txt.copy7, hockey895.txt.copy7
space-108.txt, space-183.txt, space-183.txt.copy2
space-405.txt.copy1, space-519.txt.copy7, space-630.txt.copy2

Command: dups hockey528.txt
baseball155.txt.copy3, baseball47.txt.copy3, baseball497.txt.copy4
baseball76.txt.copy7, hockey12.txt.copy6, hockey35.txt.copy5
hockey457.txt.copy1, hockey528.txt.copy1, hockey528.txt.copy3
hockey528.txt.copy4, hockey528.txt.copy5, hockey528.txt.copy6
hockey528.txt.copy7, hockey758.txt.copy1, hockey922.txt.copy3
space-102.txt.copy5, space-22.txt.copy3, space-386.txt.copy5
space-531.txt.copy2, space-641.txt.copy7, space-695.txt
space-695.txt.copy3, space-695.txt.copy3

Command: dups hockey3.txt.copy5
baseball341.txt.copy7, baseball471.txt.copy4, baseball916.txt.copy2
baseball981.txt.copy6, hockey3.txt, hockey3.txt.copy4
hockey408.txt.copy1, hockey490.txt.copy1, hockey582.txt.copy3
hockey623.txt.copy4, hockey913.txt.copy6, space-238.txt.copy2
space-346.txt.copy2, space-42.txt.copy3, space-420.txt.copy6
space-420.txt.copy6

Command: dups space-4.txt.copy1
baseball110.txt, baseball459.txt.copy4, hockey312.txt
hockey490.txt.copy1, hockey590.txt, hockey590.txt.copy1
hockey590.txt.copy2, hockey590.txt.copy3, hockey590.txt.copy4
hockey590.txt.copy5, hockey590.txt.copy6, hockey590.txt.copy7
hockey86.txt.copy5, space-120.txt.copy1, space-456.txt.copy5
```

```
Command: dups hockey75.txt.copy7
baseball735.txt.copy3, baseball756.txt.copy6, baseball899.txt.copy7
hockey130.txt.copy6, hockey169.txt.copy3, hockey175.txt.copy6
hockey282.txt.copy5, hockey358.txt.copy2, hockey37.txt.copy3
hockey467.txt, hockey467.txt.copy7, hockey652.txt.copy6
hockey656.txt, hockey75.txt, hockey75.txt.copy1
hockey75.txt.copy2, hockey75.txt.copy3, hockey75.txt.copy4
hockey75.txt.copy6, hockey766.txt.copy2, space-241.txt.copy4
space-261.txt.copy2, space-470.txt.copy7, space-759.txt.copy2

Command: dups hockey33.txt
baseball117.txt, baseball117.txt.copy3, baseball295.txt.copy7
baseball405.txt.copy2, baseball59.txt.copy1, baseball62.txt.copy5
baseball620.txt.copy7, baseball639.txt, baseball639.txt.copy4
baseball657.txt.copy1, baseball809.txt.copy7, hockey33.txt.copy1
hockey33.txt.copy2, hockey33.txt.copy3, hockey33.txt.copy5
hockey33.txt.copy6, hockey33.txt.copy7, space-747.txt
space-747.txt.copy1, space-747.txt.copy1

Command: dups baseball88.txt
baseball88.txt.copy2, baseball88.txt.copy3, baseball88.txt.copy4
baseball88.txt.copy5, hockey315.txt.copy4, hockey369.txt.copy6
hockey479.txt.copy4, space-287.txt.copy6, space-449.txt.copy5
space-567.txt, space-567.txt.copy1, space-691.txt.copy4
space-729.txt, space-831.txt, space-901.txt.copy1
space-901.txt.copy1

Command: findb 1000 .95
18
Command: exit
```

It's clear that .95 is far past the sweet spot of balancing false positives with false negatives and we are getting far too many false negatives because they copies no longer consistently show for every document. 18 buckets just isn't enough to capture the appropriate about of files.

## Sources

We only use the lecture slides for this assignment