step 1: we will get the yeast data from
****
http://arep.med.harvard.edu/biclustering/
****
step2: here we have to take the gene list of 2884 gene and also the gene expression matrix

step3: next go on site
****
https://go.princeton.edu/cgi-bin/GOTermMapper
****
go term mapper and select SGD slim and corresponding data
a) biological process
b) cellular component
c) molecular function
in our case we considered all three of them.
select plain text and html both and copy paste the data generated after submit

---

step 4: statistics:
biological process: total gene : 2884
mapped gene : 2264
unmapped :620
(1 identified ambigious, 224 unannotated, 116 not annotated in slim, 292 had no root annotation)
unique go term:100
(2 go term has no membership in any gene)
7730 gene-goterm pairs
****
molecular function: total gene: 2884
mapped gene: 1978
unmapped gene: 906
(1 found ambigious, 224 unannotated, 77 not annotated in slim, 593 no-root annotation)
unique go term: 43
(3 go term has no membership in any gene)
4595 gene-goterm pair
****
cellular component: total gene: 2884
mapped gene: 2466
unmapped gene: 418
(1 ambigious, 20 not annotated in slim, 168 has no root annotation)
unique go term:23
(1 go term has no membership in any gene)
7389 unique gene-go term pair

---

step 5: now find gene common in all three of the senario we are considering i.e biological process,
molecular function, cellular component
results are as followed:
process "intersect" function = 1884
process " intersect" component = 2200
function "intersect" component = 1914
****
function "intersect" component "intersect" process = 1842
****
function "union" component "union" process =2552
****

---

step 6: unique go term : there are 116 unique go term as go term are unique for molecular function, biological process and cellular component.
****
so the final matrix is in combined data folder with name = "matrix_combined_all_gene.txt"
****
gene_common_in_all3.txt has gene common in all three dataset
****
individual folder has its own dataset
a) complete file from go term mapper
b) mapped gene
c) unmapped gene
d) unique go term
e) preprocessed goterm gene list from complete file
f) matrix of gene on x-axis and go term on y-axis
