DETAILED REPORT ON DATASET PREPARATION **START TO END(YEAST)** :

Step 1: Go on site

http://arep.med.harvard.edu/biclustering/
It has Microarray data and list of gene names for the yeast.

Microarray data name	Yeast expression matrix
Corresponding Gene name	Genes associated with rows of the yeast data

There are total 2884 genes of Yeast.
I have stored it by name yeast_name.txt for list of yeast gene name
Yeast_dataset for microarray data

Step 2: Go on site

http://www.geneontology.org/
It gene ontology consortium site
(GO Consortium is up-to-date, comprehensive, computational model of biological systems, from the molecular level to larger pathways, cellular and organism-level systems.)

Step 3: Ontology -> Download Ontology -> go-basic.obo
Download the file I have saved it by name gene_ontology_dataset.txt

Step 4 Go on: Go on site
GENERIC GENE ONTOLOGY (GO) TERM MAPPER

https://go.princeton.edu/cgi-bin/GOTermMapper
Give the yeast_name.txt file or any gene list file and the process u want(biological, cellular, molecular) and organism type and download the goterm-gene mapping.

I have 3 folder named biological, cellular, molecular and each have their name as

Biological Process

complete_file_biological_process.txt	The file downloaded from Go term mapper
go_term_gene.txt	Applied some preprocessing on complete_file_biological to remove useless information it just have the goterm_id and all gene mapped to it

matrix_process_gene_go_term.txt	It is matrix for biological process where we have genes in x-axis and goterm in y-axis and each value indicate if a goterm present or not
process_mapped_genes.txt	It mean all genes mapped in biological process
process_unique_goterm.txt	It mean all unique go term of biological process mapped
process_unmapped_genes.txt	Out of the 2884 gene there were few not mapped in Biological process it is list of that
yeast_name.txt	It is the total list of 2884 gene that was already mentioned before

Cellular Component:

complete_file_cellular_component.txt	The file downloaded from Go term mapper
go_term_gene.txt	Applied some preprocessing on complete_file_cellular_component.txt to remove useless information it just have the goterm_id and all gene mapped to it
matrix_process_gene_go_term.txt	It is matrix for biological process where we have genes in x-axis and goterm in y-axis and each value indicate if a goterm present or not
process_mapped_genes.txt	It mean all genes mapped in biological process
process_unique_goterm.txt	It mean all unique go term of biological process mapped
process_unmapped_genes.txt	Out of the 2884 gene there were few not mapped in Biological process it is list of that
yeast_name.txt	It is the total list of 2884 gene that was already mentioned before

Same goes for Molecular Process too.

Combined Data

Here we got the seperate gene mapping for biological,cellular and molecular what we do is we find the genes common in all 3 and goterm union of all three. And have a matrix of gene goterm.

The files are as followed

check_if_go_term_common.py	It checks if goterm is common and saveit in seperate list (goterm list is disjoint)
create_combined_data_matrix.py	It takes input all anotation from (BP, MF, CC) and make a final resultant matrix of gene common in all and goterm unique
create_combined_data_matrix.py	Microarray data preprocessing helps to extract the colu

	Column corresponding to goterm present in all.
find_common_gene_in_all.py	Helps to find gene common in all(BP,MF,CC)
phase2_extract_data.py	Helps to make individual goterm gene matrix and relevant files

OTHER IMPORTANT RESULTS AND ANALYSIS:

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

Combined Analysis:

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
****
unique go term : there are 160 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

PHASE 2 of the WORK

The second phase of the work is we now have the goterms list union of all three. There is A gene ontology that specifies the relationship between the goterms and parent child relationship in a DAG format. From that we can retrive various information like parent, all parents , child, child of child, depth, level, Information Content, etc...
Our main task is to find:
Lins similarity between all go-terms pair.
Shen Similarity between all go-terms pairs.
Depth based Similarity between all go-term pairs.
Multifactored similarity i.e arctan(lin+shen+depth)*2/pi
Overlap similarity i.e for gene : all overlapping goterms/min(goterm annotation)

For relevant information read the following paper:
Multi-factored gene-gene proximity measures exploiting biological knowledge extracted from Gene Ontology : application in gene clustering.

The Tool I have used that gave me support of the Ontology so that I could extract the necessary relavant information are as followed.
GOATOOLS: A Python library for Gene Ontology analyses

Github Repisotory:
https://github.com/tanghaibao/goatools#available-statistical-tests-for-calculating-uncorrected-P-values

U need to setup the environment before using this tool. Check the readme file for that.
The file name and relevant functionality is given here.

lins_similarity_matrix.ipynb	It is jupiter notebook code for finding goterm-goterm lins similarity only thing input is the gotrem list

Extract_dep_lev_ic_parent.ipynb	Gives the following file as output: Go_term info_remove_obsolete.txt Which has depth level information content and immediate parents.
Go_find_LCA.py	Helps to find LCA Lowest common ancestor of the goterm pair It takes the relavant files as input: Go_term info_remove_obsolete.txt Goterm_list_without_obs.txt Outputs: Go_term_all_parents_info.txt (which has all parent info) LCA_and_Depth_Info_unique.txt(goterm pair lca and depth)
Depth_based_similarity.py	outputs : Depth_Based_Matrix.txt where goterm-pair and their corresponding depth based similarity is stored
Shen_similarity.py	It takes the following files as input: Go_term info_remove_obsolete.txt LCA_and_Depth_Info_unique.txt Goterm_list_without_obs.txt Outputs: Shen_similarity_matrix.txt (goterm pair wise Shen Similarity)

Inside LIN_SHEN_DEPTH folder is the final multifactored similarity code
Multi_factored_similarity_goterm.py file that does the operation
arctan(Lins+Shen+Depth)*2/pi
It take input as :
Shen_similarity_matrix.txt, Lins_similarity_matrix.txt, Depth_Based_Matrix.txt, Goterm_list.txt
Outputs:
Multifactored_Similarity_Matrix.txt

Merged_result.py that does the following operation (Overlap_ratio+ Gene-gene based (goterm pair multifactored similarity))/2
And outputs the Multifactored_TermOverlap_Matrix.txt