Summary of Main Processing Pipeline (20211130)

	Task	Codes	Input	Output
1	**Download Raw Data**			
1.1	Download Thomson One's Conference Calls	automatic_download .py	-	O1: [yyyymmdd-yyyymmdd].pdf O2: [yyyymmdd-yyyymmdd].xls
2	**PDF Processing**			
2.1	Convert Conference Calls from .pdf to .txt	pdftransfer.sh	O1	O3: [yyyymmdd-yyyymmdd].txt
2.2	Split Conference Call .txt files to separate out individual conference calls, and combine with report information from .xls files	ParseCCpdf.jl	O3, O2	O4: [yyyymmdd-yyyymmdd].csv
3	**Firm Identification (Firm Name Matching)**			
3.1	Download Compustat datasets	-	-	O5: ciqcompany. sas7bdat O6: ciqcountrygeo. sas7bdat O7: wrds_gvkey. sas7bdat
3.2	Download Hassan dataset	-	-	O8: Hassanfile_raw_ updated20219030.csv
3.3	Process Compustat and Hassan datasets into usable and truncated .csv files.	convert_sas7bdattocsv.py hassanfilecsv_viewable_ truncate.py	O5, O6, O7, O8	O9: ciqcompany.csv O10: ciqcountrygeo.csv O11: wrds_gvkey.csv O12: Hassanfile_raw_ updated20219030 _truncated.csv
3.4	Match titles in conference calls with firm names in Hassan and Compustat datasets, with both exact and fuzzy matching	Based on the current flow of work, stage 3 is technically done after stage 4 since it used entryfiles_combined, though the original codes were written as stage 4 after stage 3.	O9, O10, O11, O12	O13: CC_List[yyyy].csv O14: CC_List_2001-2021.csv
4	**Keyword Identification**			
4.1	Make a folder structure with x groups (default: x = 50)	mkdir.py dividefilesequallyinto folders.py	-	-
4.2	Make a list of keywords and template entry file	-	-	O15: keyterms.txt O16: Entry mask.xlsx

4.3	Identify keywords in each conference call	keyword_ident_1.py keyword_ident_1.sh	O15, O13, O4	O17: FR5.csv
4.4	Extract all paragraphs from each conference call that contains a specific keyword	keyword_ident_2.py keyword_ident_2.sh	O17	O18: TotalCircnew.xlsx
4.5	Cleans the identified matches and merges with gvkey dataset	mergeclean.do	O18, O14	O19: cric1_newtotal.xlsx
4.6	Make a paragraph record file that splits the number of entries into groups of 500	make_paragraphrecord.py	O19	O20: paragraphrecord.xlsx
4.7	Bold the keywords and separate file into "entryfiles", each containing 500 entries.	makebold.py makebold.sh	O19, O20, O15, O16	O21: [i].xlsx
4.7	Combine entry files	combine_entryfilesjason.py combine_entryfilessixun.py combine_sixunand jasonentryfiles.py	O21	O22: entryfiles_ combined.xlsx
5	**Get Front Page Descriptions**			
5.1	Extract front page descriptions from conference calls	extractdescriptioninfront page.py extractdescriptioninfront page.sh copyfiles.py copyfiles.sh	O21, O2, O3	O23: [yyyymmdd-yyyymmdd]_withfront pagedesc.xlsx
5.2	Combine xls files	combine_xlsfiles_with description.py	O23	O24: xlscombined_with frontpagedescription .xlsx
5.3	Match and add front page descriptions to combined entry files	-	O24, O22	O25: entryfiles_ combined.xlsx (updated)