20210113\_weekly workflow for mutual fund lab

## Summary

1. Weekly workflow posted to teams.
2. Lab slides posted to teams (may be a good reminder for folks about larger project goals, team members, etc.)
3. All work due posted to teams by **7:00 am on Wednesdays**—please save files and code in teams folder and include a weekly summary of your work, questions, etc. The summaries can be very informal (unless requested otherwise).
4. All referenced filed on teams
5. Email [amtucker@gsu.edu](mailto:amtucker@gsu.edu) with any questions.

## Individual assignments/projects

### Public health key word counts and text extraction 2019-2020 data (Chris, Hiwot, and Mohit)

* Using the master data file on teams, and the starter code from Dr. Xia, please extract sentences with public health and natural disaster keywords. We want an excel file for each with each row a unique sentence. Retain accession numbers (filing id assigned by the SEC that will act as our key). Drop duplicate sentences for public health and natural disasters (I said that we may not need to worry about duplicates with this approach, but that isn’t true you may have more than one keyword in a sentence so the same sentence may be puled more than once.) Please save code and excel files to teams, and provide a weekly summary before lab next week. In your summary please provide some summary stats on # of keywords identified, counts of key words, number of sentences, etc. Please share in your summary and/or put in slides. Dr. Xia also posted some additional instructions (see post-lab emails)
  + Hiwot please work with Chris and Mohit to provide the latest keywords for public health and natural disaster keywords. Please run the code on the new 2019/2020 data now available on teams and provide key word count analysis.

### Public health and natural disaster visualizations- Xiaofie

* When new data is available, we will ask you to update your visualization scripts

### SEC-preferred language current market conditions indicator variable.—Hiwot start on this. Xiaofie can maybe assist with questions

* Take existing approach//research on current market conditions. Dr. Xia may provide additional starter code. I am happy to talk with you about this more too.
* Build binary classification code on principal risk text in master data file. We are creating a yes(1)/no(0) indicator variable where 1 is conditional on the presence of any of the following phrases: not yet known”; “persist or worsen”; “continues to face”; “continues to experience”; “current market”; “recent events”; or “ongoing.”
* Run code on 2010-2020 data.

### Hand coding public health disclosure text as boilerplate vs. covid-specific disclosures. Shane and Nyssa

* See file 20110113 in the public health disclosures\text analysis folder. The first tab has the instructions/code book for how to label your entries. There is a list of sample substantive topics in a tab, and my sample coding in another. You are to review the tab labeled “to be labeled”. There are roughly ~1050 rows in the sample I reviewed. In your tab there are ~350 rows of unique sentences (I removed duplicates). Work through these labels. The most important part is the disclosure type (boilerplate versus covid). The secondary task is analyzing the substantive text—which will be harder because there is great er variation. I provided you with a list of possible substantive labels that I observed if this is helpful. Try your best, and we can spend more time discussing this next week as needed. Focus on getting the first label done and recording your questions. While you are reading for the first label, try your hand at the substantive labels (each sentence can have up to 5 but most have 1-2). If you have additional time, go back and spend more time with the substantive labels. Write down your questions and patterns. Include those in your summaries. Save your labels in a new excel file with your name and date to teams.

### 

### Public Health topic modeling (Achal)

-We want to topic model the extracted public health and natural disaster text. Ronald was working on getting the sentence level data, but we are bypassing that intermediate step with the work assigned to Chris and Mohit. Instead read the memo written by Dr. Smelcer on selecting the optimal K and the recommended documentation for Structural Topic Modeling in R. Also, review Ronald code and initial output in topic modeling folder and AT notes on initial output (file labeled with date 20201119…)

### Fund family language similarity (Achal)

-Please pick up where you left off last semester in comparing the mean cosine for top and bottom quartiles of funds (all, and then broken down by CRSP categories). Sample results below. Complete the t-tests to see if the size of the companies differs by top and bottom quartile of cosine and inversely if cosine scores vary by size of the company.  Please put your results in a few slides to be shared with the group in lab. Please update the similarity code with any additional work from this week.