Tasks by Weeks:

**Feb 1 – Feb 15 Qualitative Analysis**

1. Started Qualitative analysis on every mention of ‘Housing’ in the 53-text file dataset in the Excel file GRI\_V5

**Feb 16 – Feb 28 Completion of V5: Hard Code Phrases**

1. Completed Qualitative analysis on every mention of ‘Housing’ in the 53-text file dataset in the Excel file GRI\_V5
2. Started writing a new Python script for social\_protection\_analytics\_V5
   1. New list groups for transfer, subsidies, health, labor market, children, migration, finance, gender, and food.
   2. Old script version was only key words; therefore, the script used .replace() method which was repetitive and insufficient. New script implements new python dictionary method to captures count of key phrases with a for loop to see each and every line in the text file one by one.
   3. Polished diagram from matplotlib default version – provided percentage of Key Phrase compared relative to counterpart phrases.

**Mar 1 – Mar 15**

1. Completed Write-Up on qualitative regarding housing mentions.
2. Logged tasks for Feb1 – Mar1.

**Mar 16 – Mar 31 Completion of V6: Measures number of mentions where a hard coded word is accompanied by a fuzzywuzzy word near the hard searched word.**

1. Introduced FuzzyWuzzy into script
2. Issue with V6: A fuzzywuzzy score of 75 is high enough for cash transfer because this includes the token transferring for cash concordance; however, a score of 75 is too low for cash assistance because tokens like distance and instance are results for the token assistance. 🡪 Score of 80 seems to be optimal since it keeps ‘transferring’, ‘pretransfer’, ‘transferred’ while removing ‘instance’, ‘distance.’
   1. Version 6.5 Add-on:
3. Used an if statement to add a single word search capability for words such as loans and credit in our finance category.

**April 01 – April 15**

1. Changed width of concordance function to 60 (default 80) to decrease the amount of words being scrutinized near the ‘hard mention’ word. This attempts to lower the count of irrelevant mentions (See Examples below)
2. Altered key phrase and word categories into:
   1. Cash
   2. Transfer
   3. Pension
   4. Subsidies
   5. Labor market
   6. Health
   7. Children
   8. Other: migration; finance; gender
   9. Housing
      1. Dropped agricultural mentions besides ‘agricultural subsidies’
      2. Drop ‘child poverty’ and ‘absolute poverty line’ mentions because these are indicators rather than programs/policy.
      3. Cash now houses ‘basic income’ and similar income mentions
      4. Pension has its own category
      5. New category column to create a legend

**April 16 – April 30**

1. Completed a new categorization for the key phrases. The new categories are:
2. Cash\_transfer
3. Children
4. Finance
5. Health
6. Inkind\_transfer
7. Labor
8. Pension
9. Subsidies
10. Other
11. Housing
12. Waiver
13. Created new horizontal bar chat visualizations with respect to these new categories for key phrases with a concordance width of 60 and a fuzzywuzzy ratio of 80. This ratio is a parameter to tell python how similar the soft mention has to be in order to be counted toward the key phrase count.

**May 01- May 15**

1. Created v8\_concor which uses Tkinter library to create a Python GUI application for user-friendly analysis. The width of the concordance length and the FuzzyWuzzy match ratio is now user selected.
2. Wrote the housing\_initiatives\_documentation to illustrate the methodology and conclusions.

Examples of Concordance:

This is the first result from the script when analyzing “cash transfers.”

The hard mention is “cash.” Hard mention is when the query is directly looking for every mention of the exact word “cash,” in this example highlighted in yellow – not “money” or “dollars” etc. The rest of the words near the vicinity of the hard mention is scrutinized with fuzzywuzzy to see if the word is similar (in spelling, length, and letters, not synonyms) to the word “transfer.” Fuzzywuzzy enables us to capture similar words like “transferring, transfer, pretransfer, transferred” like in the example below in red. This mention will be counted toward the key phrase “cash transfers” because “cash” was a direct match and the transfer match was an 80% or higher match rate according to FuzzyWuzzy.

[ConcordanceLine(left=['prevent', '”', 'deprivation', 'public', 'works', 'projects', 'for', 'instance', 'aim', 'both', 'at', 'transferring', 'shortterm', 'food', 'or'], query='cash', right=['and', 'building', 'useful', 'longterm', 'infrastructure', 'figure', '21', 'illustrates', 'the', 'relationship', 'between', 'these', 'measures', 'and'], offset=5473

This is an example that would not be counted because transfer is not detected:

ConcordanceLine(left=['million', 'people', '(', '319', 'per', 'cent', 'of', 'the', 'population', ')', 'in', '200724', 'the', 'latter', 'provided'], query='cash', right=['beneﬁts', 'job', 'training', 'and', 'small', 'loans', 'to', 'unemployed', 'temporary', 'workers25', 'finally', 'universal', 'coverage', 'for'], offset=94194

Potential pitfall and solution:

In the following made up example, the script will count this example as a mention of cash transfers because the hard mention does exist while fuzzywuzzy will detect the word “transfers” as 100% compared to the searched word “transfers.” Though I have not seen an instance like this in the actual sentences in our data, I am sure that such an instance must have occurred because of the size of the dataset. A way to increase the accuracy of the script is to decrease the width of the concordance function. This width (the variable name for width in the phrase\_search function is concor\_width) is the character length of each line being scrutinized; therefore, decreasing the character length of each line can increase accuracy, especially when key phrases should have both words as near ‘neighbors.’

Example sentence: ... the program supports food transfers to 3 million people. Moreover, cash grants for the elderly have been more ...

Decreasing the width to 30:

This example would not be counted toward the key phrase mention count for ‘cash transfer.’

to 3 million people. Moreover, cash grants for the elderly have b ...