Introduction to GitHub:

A Crash Course for Social Science Researchers

Brian Heseung Kim
University of Virginia

May 3, 2021

Lesson Outline

- Setting the Stage
  - Why GitHub? Why Now?
- Frameworks and Principles
  - Git, GitHub, and GitHub Desktop, Oh My
- Getting Started
  - Level 1: GitHub as Google Drive but Worse
  - Level 2: GitHub as a Dissemination Tool
  - Level 3: GitHub as a Version Control Platform
  - Level 4: GitHub as a Facilitator of Complex Collaboration
- Next Steps

Setting the Stage

Why GitHub? Why Now?

Why GitHub? Why Now?

- In today's education research space:

Why GitHub? Why Now?

- In today's education research space:
  - Collaboration across methods, languages, and disciplines is becoming more and more common

Why GitHub? Why Now?

- In today's education research space:
  - Collaboration across methods, languages, and disciplines is becoming more and more common
  - Analyses are becoming more complex and multi-stage

Why GitHub? Why Now?

- In today's education research space:
  - Collaboration across methods, languages, and disciplines is becoming more and more common
  - Analyses are becoming more complex and multi-stage
  - Entire analyses can change on the back of one line of code

Why GitHub? Why Now?

- In today's education research space:
  - Collaboration across methods, languages, and disciplines is becoming more and more common
  - Analyses are becoming more complex and multi-stage
  - Entire analyses can change on the back of one line of code
  - Research is increasingly conducted by multiple analysts building on the same codebase

Why GitHub? Why Now?

- In today's education research space:
  - Collaboration across methods, languages, and disciplines is becoming more and more common
  - Analyses are becoming more complex and multi-stage
  - Entire analyses can change on the back of one line of code
  - Research is increasingly conducted by multiple analysts building on the same codebase
  - Researchers can benefit from the existing code of other researchers, even in unrelated projects

Why GitHub? Why Now?

- In today's education research space:
  - Collaboration across methods, languages, and disciplines is becoming more and more common
  - Analyses are becoming more complex and multi-stage
  - Entire analyses can change on the back of one line of code
  - Research is increasingly conducted by multiple analysts building on the same codebase
  - Researchers can benefit from the existing code of other researchers, even in unrelated projects

=

Why GitHub? Why Now?

Frictions and dangers in analytic work abound

Frameworks and Principles

Git, GitHub, and GitHub Desktop

Git, GitHub, and GitHub Desktop

Git, GitHub, and GitHub Desktop

- For now, you can think of Git, GitHub, and GitHub Desktop roughly as:
  - GitHub = Box
  - GitHub Desktop = Box Sync
  - Git = Crazy back-end magic (JavaScript, Python, SQL)

Git, GitHub, and GitHub Desktop

- For now, you can think of Git, GitHub, and GitHub Desktop roughly as:
  - GitHub = Box
  - GitHub Desktop = Box Sync
  - Git = Crazy back-end magic (JavaScript, Python, SQL)
- And just as there are many cloud storage services…

Git, GitHub, and GitHub Desktop

- For now, you can think of Git, GitHub, and GitHub Desktop roughly as:
  - GitHub = Box
  - GitHub Desktop = Box Sync
  - Git = Crazy back-end magic (JavaScript, Python, SQL)
- And just as there are many cloud storage services…

Version Control Systems	File Repository Services	Desktop Clients
Git	GitHub	GitHub Desktop
Subversion	BitBucket	SourceTree
…	…	…

Getting Started

Level 1: GitHub as Google Drive but Worse

- GitHub can work as a file storage solution for code, documentation, and smaller datasets
- Can do this manually via GitHub
- Can do this semi-automatically via GitHub Desktop

Level 1: GitHub as Google Drive but Worse

- GitHub can work as a file storage solution for code, documentation, and smaller datasets
- Can do this manually via GitHub
- Can do this semi-automatically via GitHub Desktop

- **Quick Demo:** COVID Dashboard

Level 2: GitHub as a Dissemination Tool

- Level 1 plus:
  - Can share public “repositories” (repos) easily and freely
  - Can keep it updated at all times
  - Can write simple “read-me’s” for public use
  - Looks fancier than it is

Level 2: GitHub as a Dissemination Tool

- Level 1 plus:
  - Can share public “repositories” (repos) easily and freely
  - Can keep it updated at all times
  - Can write simple “read-me’s” for public use
  - Looks fancier than it is

- **Quick Demo**: Education Deserts w/ Dan

Level 3: GitHub as a Version Control Platform

- Level 2 plus:
  - Can be edited simultaneously by multiple users w/o conflict
  - Easily identify who changed what, and when
  - Easily revert erroneous or catastrophic changes to code
  - Force yourself to write good notes on each update

Level 3: GitHub as a Version Control Platform

- Level 2 plus:
  - Can be edited simultaneously by multiple users w/o conflict
  - Easily identify who changed what, and when
  - Easily revert erroneous or catastrophic changes to code
  - Force yourself to write good notes on each update

- **Quick Demo:** N2FL Text Analysis

Level 4: GitHub as a Facilitator of Complex Collaboration

- Level 3 plus:
  - Can formally request changes, document bugs, and keep track of future changes needed
  - Can incorporate suggestions ***from*** total strangers, or suggest your own changes ***to*** total strangers ("open source!")
  - Can allow for the same codebase to exist in many forms at once, be many things to different people
  - Automate codebase updates

- **Quick Demo:** Parttree Package

Next Steps

Resources to Explore

- GitHub's Hello World tutorial
- Similar tutorial from HubSpot
- Similar tutorial, but using more of the back-end command line
- Just play around with it!

Any Questions?

- Setting the Stage
  - Why GitHub? Why Now?
- Frameworks and Principles
  - Git, GitHub, and GitHub Desktop, Oh My
- Getting Started
  - Level 1: GitHub as Google Drive but Worse
  - Level 2: GitHub as a Dissemination Tool
  - Level 3: GitHub as a Version Control Platform
  - Level 4: GitHub as a Facilitator of Complex Collaboration
- Next Steps

Thank you!

Brian Heseung Kim

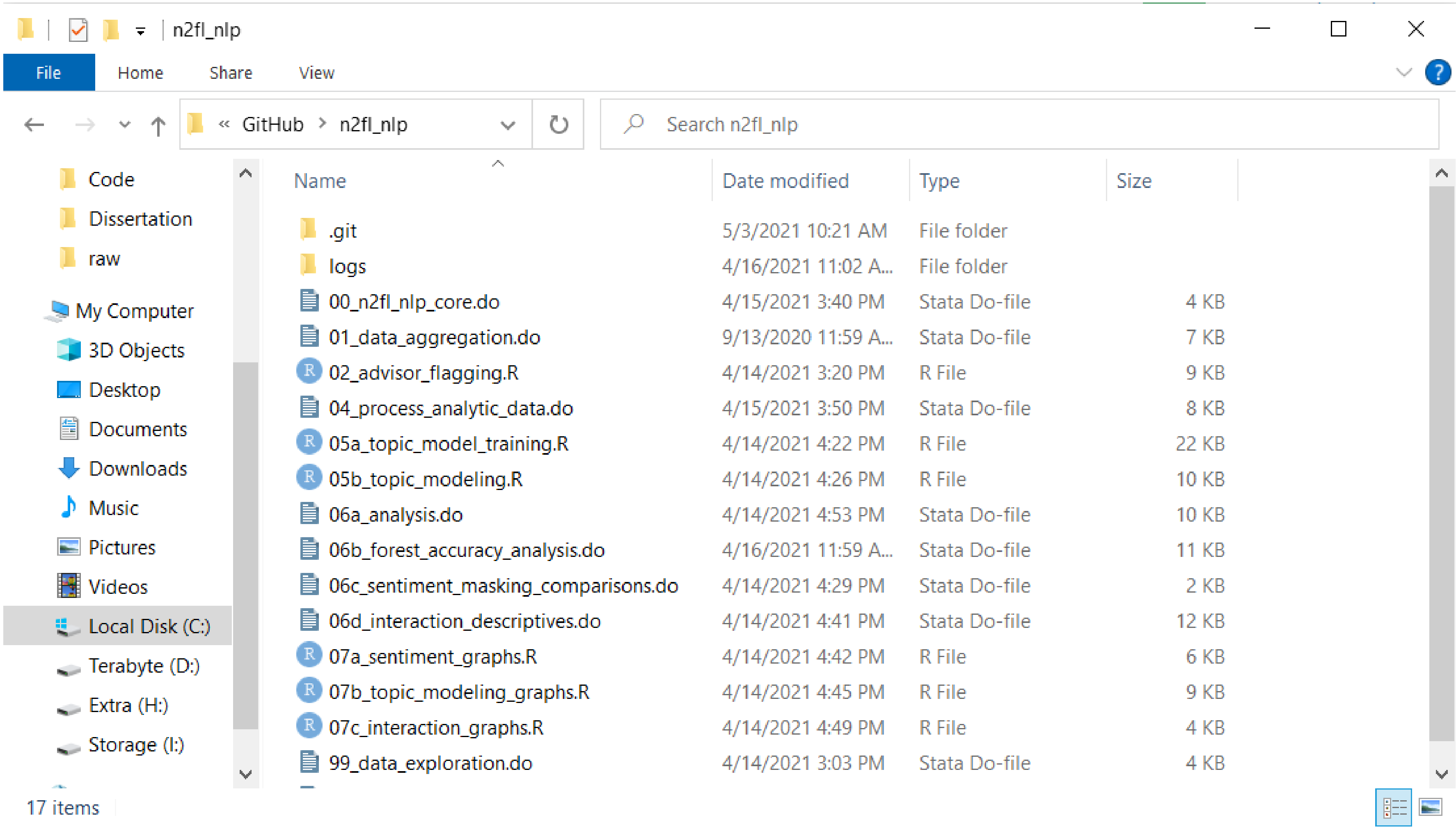
@brhkim

brian.kim
@virginia.edu

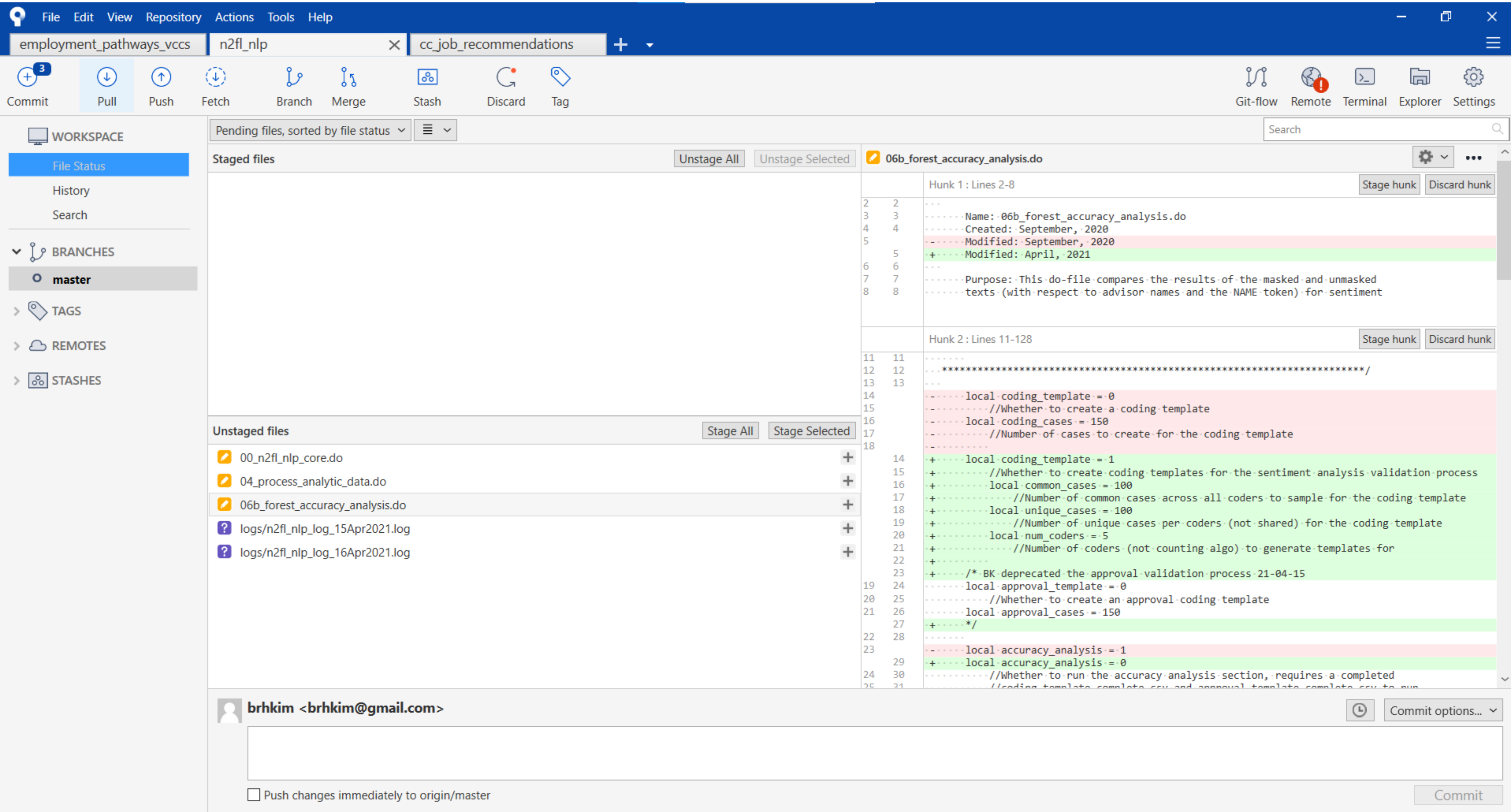
brhkim.com

@brhkim

For the most up-to-date information from EdPolicyWorks,
please contact us at:
EdPolicyWorks@virgina.edu



Name	Date modified	Type	Size
.git	5/3/2021 10:21 AM	File folder	
logs	4/16/2021 11:02 A...	File folder	
00_n2fl_nlp_core.do	4/15/2021 3:40 PM	Stata Do-file	4 KB
01_data_aggregation.do	9/13/2020 11:59 A...	Stata Do-file	7 KB
02_advisor_flagging.R	4/14/2021 3:20 PM	R File	9 KB
04_process_analytic_data.do	4/15/2021 3:50 PM	Stata Do-file	8 KB
05a_topic_model_training.R	4/14/2021 4:22 PM	R File	22 KB
05b_topic_modeling.R	4/14/2021 4:26 PM	R File	10 KB
06a_analysis.do	4/14/2021 4:53 PM	Stata Do-file	10 KB
06b_forest_accuracy_analysis.do	4/16/2021 11:59 A...	Stata Do-file	11 KB
06c_sentiment_masking_comparisons.do	4/14/2021 4:29 PM	Stata Do-file	2 KB
06d_interaction_descriptives.do	4/14/2021 4:41 PM	Stata Do-file	12 KB
07a_sentiment_graphs.R	4/14/2021 4:42 PM	R File	6 KB
07b_topic_modeling_graphs.R	4/14/2021 4:45 PM	R File	9 KB
07c_interaction_graphs.R	4/14/2021 4:49 PM	R File	4 KB
99_data_exploration.do	4/14/2021 3:03 PM	Stata Do-file	4 KB



Unstaged files

- 00_n2fl_nlp_core.do
- 04_process_analytic_data.do
- 06b_forest_accuracy_analysis.do
- logs/n2fl_nlp_log_15Apr2021.log
- logs/n2fl_nlp_log_16Apr2021.log

06b_forest_accuracy_analysis.do

Hunk 1 : Lines 2-8

```
       Name: 06b_forest_accuracy_analysis.do
       Created: September, 2020
-      Modified: September, 2020
+      Modified: April, 2021

       Purpose: This do-file compares the results of the masked and unmasked
       texts (with respect to advisor names and the NAME token) for sentiment
```

Hunk 2 : Lines 11-128

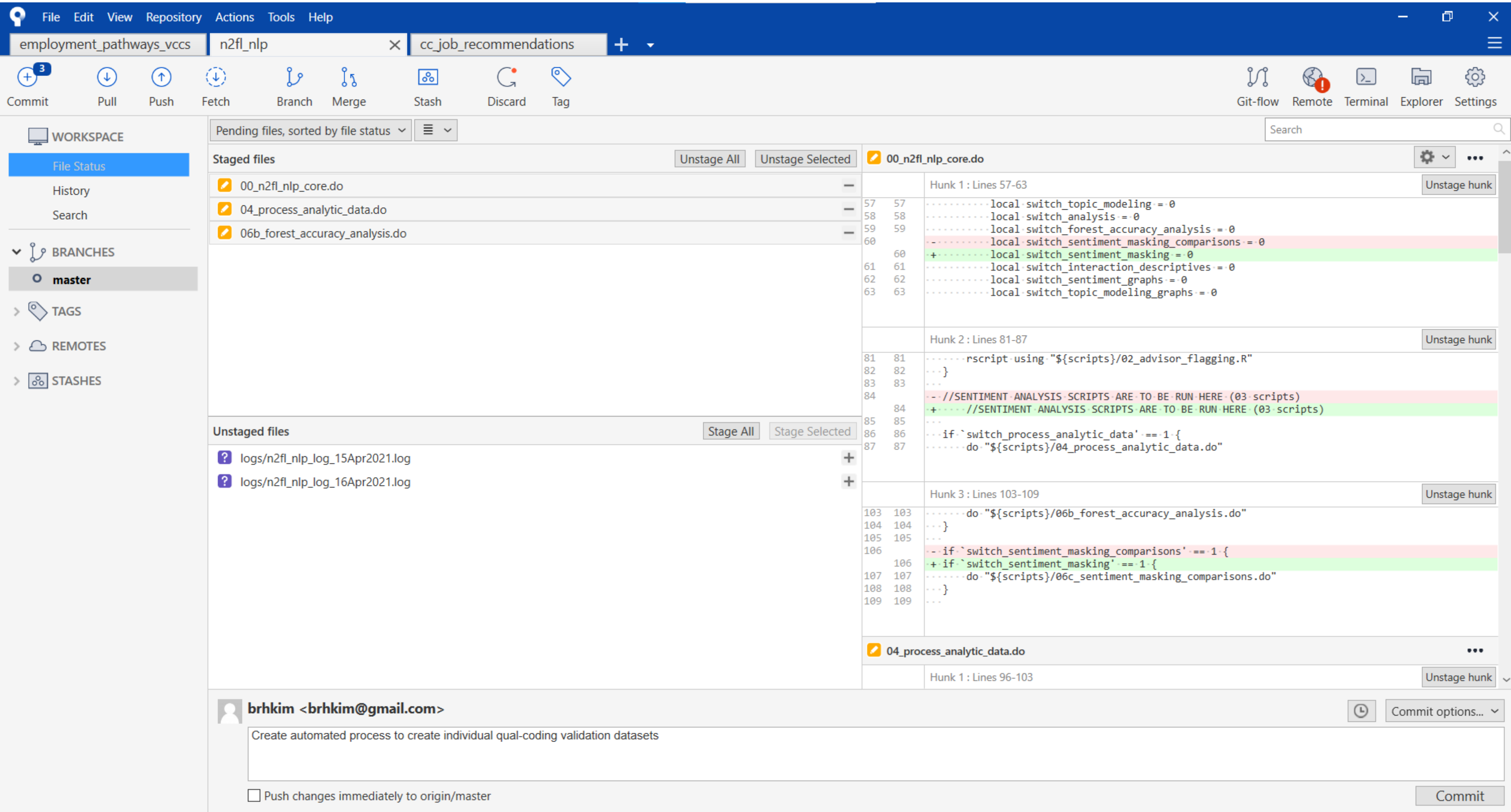
```
   ***********************************************************************/

-      local coding_template = 0
-          //Whether to create a coding template
-      local coding_cases = 150
-          //Number of cases to create for the coding template
-
+      local coding_template = 1
+          //Whether to create coding templates for the sentiment analysis validation process
+          local common_cases = 100
+              //Number of common cases across all coders to sample for the coding template
+          local unique_cases = 100
+              //Number of unique cases per coders (not shared) for the coding template
+          local num_coders = 5
+              //Number of coders (not counting algo) to generate templates for
+
+      /* BK deprecated the approval validation process 21-04-15
       local approval_template = 0
           //Whether to create an approval coding template
       local approval_cases = 150
+      */

-      local accuracy_analysis = 1
+      local accuracy_analysis = 0
           //Whether to run the accuracy analysis section, requires a completed
```

brhkim <brhkim@gmail.com>

☐ Push changes immediately to origin/master



Staged files

- 00_n2fl_nlp_core.do
- 04_process_analytic_data.do
- 06b_forest_accuracy_analysis.do

Unstaged files

- logs/n2fl_nlp_log_15Apr2021.log
- logs/n2fl_nlp_log_16Apr2021.log

00_n2fl_nlp_core.do

Hunk 1 : Lines 57-63

```
          local switch_topic_modeling = 0
          local switch_analysis = 0
          local switch_forest_accuracy_analysis = 0
-         local switch_sentiment_masking_comparisons = 0
+         local switch_sentiment_masking = 0
          local switch_interaction_descriptives = 0
          local switch_sentiment_graphs = 0
          local switch_topic_modeling_graphs = 0
```

Hunk 2 : Lines 81-87

```
      rscript using "${scripts}/02_advisor_flagging.R"
  }

- //SENTIMENT ANALYSIS SCRIPTS ARE TO BE RUN HERE (03 scripts)
+     //SENTIMENT ANALYSIS SCRIPTS ARE TO BE RUN HERE (03 scripts)

  if `switch_process_analytic_data' == 1 {
      do "${scripts}/04_process_analytic_data.do"
```

Hunk 3 : Lines 103-109

```
      do "${scripts}/06b_forest_accuracy_analysis.do"
  }

- if `switch_sentiment_masking_comparisons' == 1 {
+ if `switch_sentiment_masking' == 1 {
      do "${scripts}/06c_sentiment_masking_comparisons.do"
  }

```

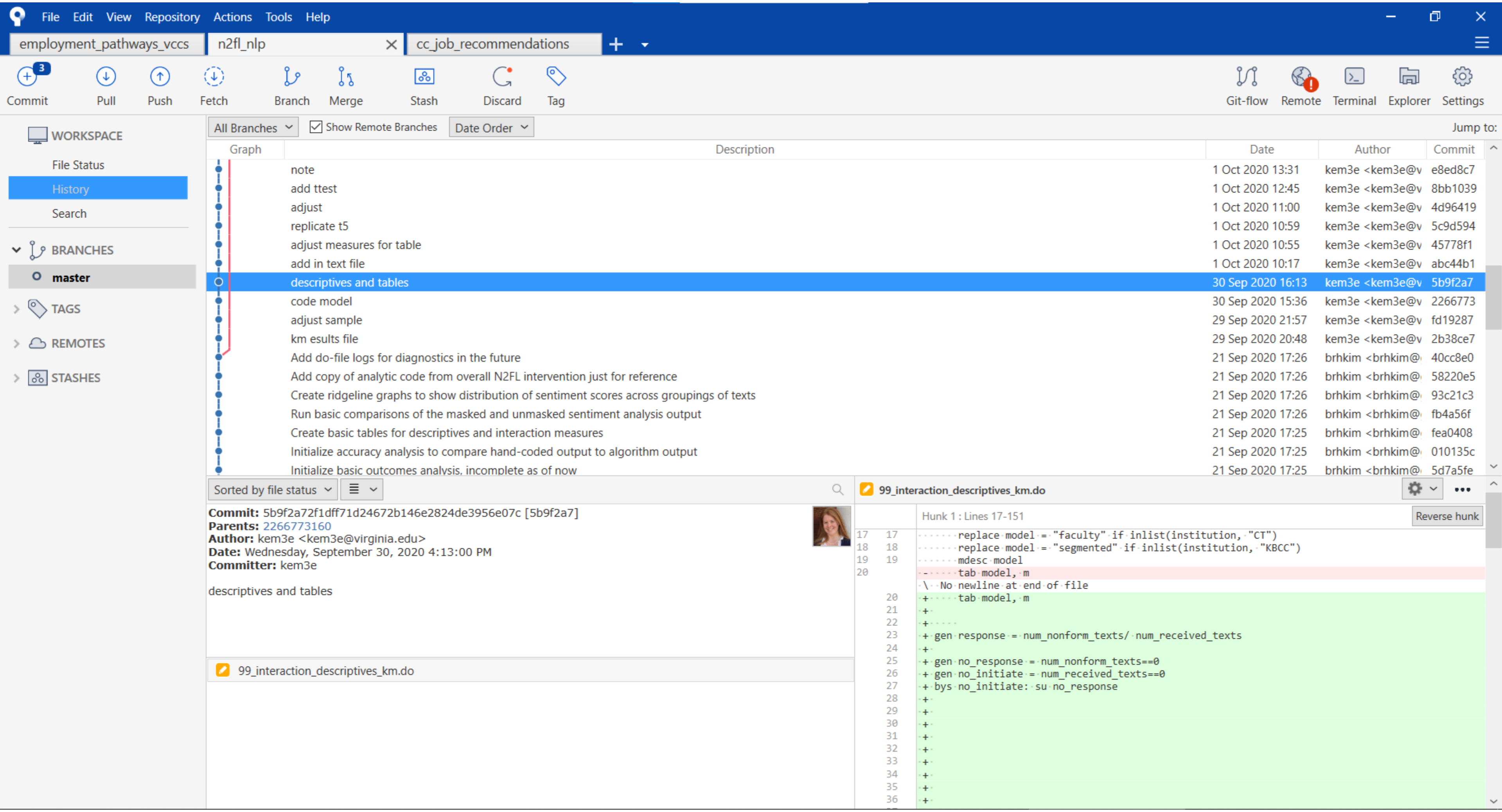
04_process_analytic_data.do

Hunk 1 : Lines 96-103

brhkim <brhkim@gmail.com>

Create automated process to create individual qual-coding validation datasets

☐ Push changes immediately to origin/master



☑ Show Remote Branches

Graph	Description	Date	Author	Commit
	note	1 Oct 2020 13:31	kem3e <kem3e@v	e8ed8c7
	add ttest	1 Oct 2020 12:45	kem3e <kem3e@v	8bb1039
	adjust	1 Oct 2020 11:00	kem3e <kem3e@v	4d96419
	replicate t5	1 Oct 2020 10:59	kem3e <kem3e@v	5c9d594
	adjust measures for table	1 Oct 2020 10:55	kem3e <kem3e@v	45778f1
	add in text file	1 Oct 2020 10:17	kem3e <kem3e@v	abc44b1
	descriptives and tables	30 Sep 2020 16:13	kem3e <kem3e@v	5b9f2a7
	code model	30 Sep 2020 15:36	kem3e <kem3e@v	2266773
	adjust sample	29 Sep 2020 21:57	kem3e <kem3e@v	fd19287
	km esults file	29 Sep 2020 20:48	kem3e <kem3e@v	2b38ce7
	Add do-file logs for diagnostics in the future	21 Sep 2020 17:26	brhkim <brhkim@	40cc8e0
	Add copy of analytic code from overall N2FL intervention just for reference	21 Sep 2020 17:26	brhkim <brhkim@	58220e5
	Create ridgeline graphs to show distribution of sentiment scores across groupings of texts	21 Sep 2020 17:26	brhkim <brhkim@	93c21c3
	Run basic comparisons of the masked and unmasked sentiment analysis output	21 Sep 2020 17:26	brhkim <brhkim@	fb4a56f
	Create basic tables for descriptives and interaction measures	21 Sep 2020 17:25	brhkim <brhkim@	fea0408
	Initialize accuracy analysis to compare hand-coded output to algorithm output	21 Sep 2020 17:25	brhkim <brhkim@	010135c
	Initialize basic outcomes analysis, incomplete as of now	21 Sep 2020 17:25	brhkim <brhkim@	5d7a5fe

Commit: 5b9f2a72f1dff71d24672b146e2824de3956e07c [5b9f2a7]
Parents: 2266773160
Author: kem3e <kem3e@virginia.edu>
Date: Wednesday, September 30, 2020 4:13:00 PM
Committer: kem3e

descriptives and tables

99_interaction_descriptives_km.do

99_interaction_descriptives_km.do

Hunk 1 : Lines 17-151 Reverse hunk

```
      replace model = "faculty" if inlist(institution, "CT")
      replace model = "segmented" if inlist(institution, "KBCC")
      mdesc model
-     tab model, m
\  No newline at end of file
+     tab model, m
+ 
+     
+ gen response = num_nonform_texts/ num_received_texts
+ 
+ gen no_response = num_nonform_texts==0
+ gen no_initiate = num_received_texts==0
+ bys no_initiate: su no_response
+ 
+ 
+ 
+ 
+ 
+ 
+ 
+ 
+ 
```