



FAMINE & FOOD CRISIS FORECASTING CENTER

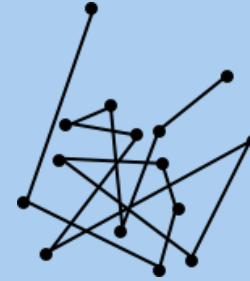
HACKATHON #1: YEMEN

Day 3: Identifying Secondary Data

Springboard Grant – Tier I
Sep 1, 2022 – August 31, 2023



**FAMINE & FOOD CRISIS
FORECASTING CENTER**



DISC

DATA REPLICABILITY AND TRANSPARENCY

ANNA RACHEL HAENCH

SENIOR DATA SCIENTIST, DATA INTENSIVE STUDIES CENTER

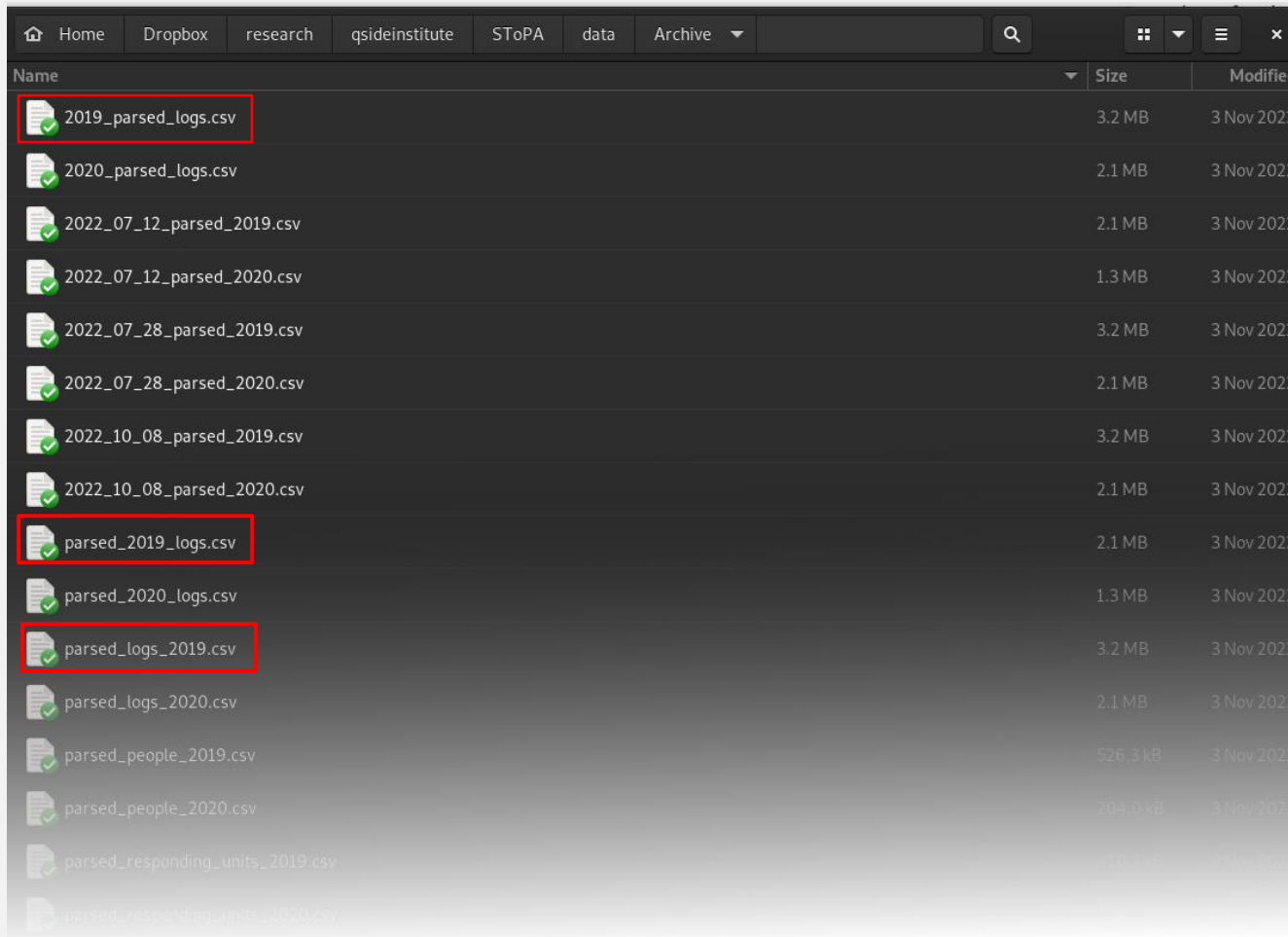
Experiences That Shaped Me

Overall “code hygiene”



- 1) Developer for open-source software
- 2) Industry software developer
- 3) Co-Director of Interdisciplinary/Community Lab
 - *Small Town Police Accountability Research Lab*
- 4) Academic researcher
 - Computational and Applied Mathematician
 - Data Scientist
 - Collaborator in Computational Social Sciences

A True Story



Name	Size	Modified
2019_parsed_logs.csv	3.2 MB	3 Nov 2022
2020_parsed_logs.csv	2.1 MB	3 Nov 2022
2022_07_12_parsed_2019.csv	2.1 MB	3 Nov 2022
2022_07_12_parsed_2020.csv	1.3 MB	3 Nov 2022
2022_07_28_parsed_2019.csv	3.2 MB	3 Nov 2022
2022_07_28_parsed_2020.csv	2.1 MB	3 Nov 2022
2022_10_08_parsed_2019.csv	3.2 MB	3 Nov 2022
2022_10_08_parsed_2020.csv	2.1 MB	3 Nov 2022
parsed_2019_logs.csv	2.1 MB	3 Nov 2022
parsed_2020_logs.csv	1.3 MB	3 Nov 2022
parsed_logs_2019.csv	3.2 MB	3 Nov 2022
parsed_logs_2020.csv	2.1 MB	3 Nov 2022
parsed_people_2019.csv	526.3 kB	3 Nov 2022
parsed_people_2020.csv	304.0 kB	3 Nov 2022
parsed_responding_units_2019.csv	100.3 kB	3 Nov 2022
parsed_responding_units_2020.csv	100.3 kB	3 Nov 2022

2021 Datathon For Justice

Hey, do we have a copy of the parsed 2019 data?

Sure! It's in the team dropbox.

...

Be Kind To Your Future Selves

Some Things That Helped Me

1. Talk with your team about file naming conventions
 - Protip: use dates and initials like *parsed_2019_logs_AH_20230212.csv*
2. Have a conversation with your team about project folder structure
 - Protip: document the outcome of this conversation and make it your README.
3. Give people the tools to run the data processing pipeline.
 - Protip: if you're using Github, provide a Jupyter notebook that walks through the workflow.
 - Protip: if you're not using Github, write it all down and include lots of good pictures.
4. If you're writing code, be excessively rigid about style guidelines
 - Protip: if you're using Python, the PEP8 is a really helpful resource.



**FAMINE & FOOD CRISIS
FORECASTING CENTER**

DATALAB

INNOVATE , ANALYZE, VISUALIZE

DATA BEST PRACTICES

UKU-KASPAR UUSTALU DATA SCIENCE SPECIALIST, TUFTS DATA LAB

Three simple rules...

1. Each column is a variable.
2. Each row is an observation.
3. Each cell is a single value.

Two more things to remember...

documentation = important

(excessive) redundancy = evil



**FAMINE & FOOD CRISIS
FORECASTING CENTER**

DATALAB

INNOVATE , ANALYZE, VISUALIZE

MERGING PLACES AND NAMES

PETER NADEL

DIGITAL HUMANITIES NLP SPECIALIST

AISHWARYA VENKAT

DOCTORAL STUDENT, FRIEDMAN SCHOOL (AFE)

Are we talking about the same place?

- ▶ Missing letters, varied spellings, abbreviations
 - ▶ E.g. Guatemala: Totonicapan vs. Totonicapn
 - ▶ E.g. Pakistan: Tando Mohammad Khan vs. Tando M. Khan
- ▶ Administrative differences
 - ▶ E.g. Phillippines: Zamboanga del Sur vs. Region IX (Zamboanga Peninsula)
 - ▶ E.g. Afghanistan: Shahrak Tula vs Sharak
- ▶ Linguistic differences, encodings
 - ▶ E.g. Benin: Segbana vs. Ségbana
 - ▶ E.g. Somalia: Dhusa Mareb vs. Dhuusamareeb
 - ▶ E.g. Ethiopia: West Hararghe vs. Mirab Hararghe vs. Ouest Hararghe

Can algorithms help resolve place-names?

Classification with HuggingFace transformers

In this notebook, we are tasked to find and replace misspelled or misidentified placenames in geographic data. In addition to the placenames, we also have scores that come from a fuzzymatching process, which will give us a starting place for building our model.

This notebook should be a place of experimentation and exploration. I have tried to mark places where you could try other approaches and implementations.

A few notes on running this yourself:

1. It requires a GPU. You can access a GPU on Google Cloud. Then find GPU in the Hardware area.
2. I tried to make it so that there were no version or installation issue.

```
# pred will be 1 if they are refering to the same place
# pred will be 0 if they are refering to a different place
p = preds_df.sample(5).apply(lambda x: print(x['input'], x['preds'], '\n'),axis=1) # sampling 5 at a time
```

```
TEXT1: kenya - mandera - el wak or central; TEXT2: kenya - meru - igembe central 0.0
```

```
TEXT1: uganda - northern - amudat; TEXT2: sudan - northern - addabah 0.0
```

```
TEXT1: uganda - northern - moroto; TEXT2: uganda - moroto - moroto 0.0
```

```
TEXT1: cote divoire - woroba - bafing; TEXT2: côte d'ivoire - woroba - bafing 1.0
```

```
TEXT1: myanmar - west - rakhine; TEXT2: myanmar - rakhine - sittwe 0.0
```

Can we automate place-name matching?

- Limit range of possible matches by geography or regions
- Validation across multiple NLP models and spatial datasets

Country	Name of Search Region	Name of Matched Region	Decision
South Sudan	Greater Upper Nile	Upper Nile	CORRECT
Philippines	Northern Mindanao	Lanao del Norte	INCORRECT
India	Odisha	Orissa	CORRECT; State renamed
Somalia	Erigavo	Ceerigaabo	CORRECT; c sound no English eq
Ethiopia	Dollo Odo	Doolo	UNCLEAR. Could be Doollo (zone in Somali region) or Doolow (woreda in the Liben zone). Needs spatial verification.

Geographer's questions

- Are we mapping the right boundaries?
 - New states, redistricting
 - Contested territories
 - Independence movements
 - Conflict
- Is this place where it is supposed to be?
 - Data collection/entry error
 - Coordinate system problem



Source: [The Economist](#), April 2022

Detailed map available on [PolGeoNow](#)

HACKING GROUPS!

- ▶ Work on your slides
- ▶ Identify relevant secondary datasets
 - ▶ Submit details about secondary data on portal
 - ▶ Create data wish list
- ▶ If time, merge nutrition indicators with secondary data

- ▶ Create 3 slides with charts, visualizations, and any preliminary findings





FAMINE & FOOD CRISIS FORECASTING CENTER

HACKATHON #1: YEMEN

Team Presentations, Day 3



**FAMINE & FOOD CRISIS
FORECASTING CENTER**



Gerald J. and Dorothy R.
Friedman School of
Nutrition Science and Policy

CLOSING

ELENA N. NAUMOVA

PROFESSOR, NUTRITION EPIDEMIOLOGY & DATA SCIENCE