PhilAudit.

Data Science tools for Policy Research

By Jack Vaughan | April, 2023

Introduction
Background

The data for this project comes from a data extraction/engineering application I developed here: PhilAuditSystem

Political and economic development researcher Mike Denly has been studying codified corruption in the governments of the developing world for years. His current work focuses on categorizing and quantifying corrupt practices found in government audit reports. This practice works exceptionally well in some cases.

Data Description

What's in these reports?

Column Name	Data Type	Description
`audit_observation`	`str`	The reported finding from the audit.
`recommendations`	`str`	The recommendation from the audit committee to the audited entity
`references`	`str`	The reference(s) to which the recommendation of the prior year pertains
`status_of_implementation`	`str`	Whether the recommendation was implemented or not. Can take on 3 values: not implemented, partially implemented, implemented.
`reasons_for_partial_or_non_implementation`	`str`	Reasons why the recommendation was not implemented
`management_action`	`str`	The action taken by the audited entity to address the recommendation, if any.

Example PDF:

What does this mean?

STATUS OF IMPLEMENTATION OF PRIOR YEARS' AUDIT RECOMMENDATIONS
As of December 31, 2013

Of the sixteen (16) audit recommendations embodied in the 2012 Annual Audit Report and other prior years, four (4) were fully implemented, six (6) were partially implemented while the remaining six (6) were not implemented at all, hence reiterated for implementation.

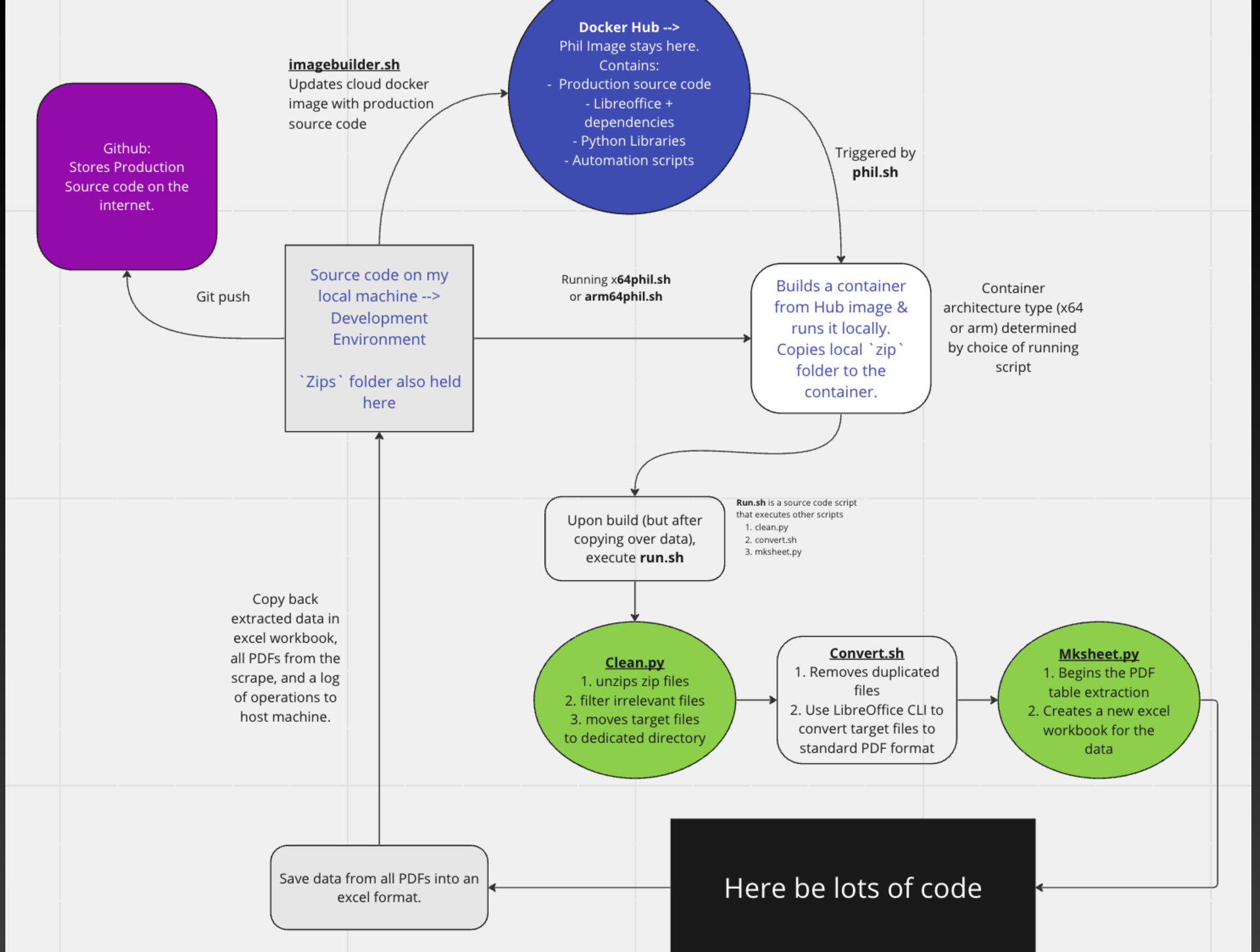
Observations and Comments	Recommendation	Ref	Specific Management Action/ Comment	Status of Implementation as of December 31, 2013	Reason for Partial or Non Implementation
1.The correctness, validity and condition of the Property, Plant and Equipment amounting to P59,580,576.29 and Inventories costing P2,289,331.68 were not established due to the failure of the Municipality to conduct a physical inventory and maintain stock cards, property/supplies ledger cards.	Initiate the conduct of a complete physical inventory of the Municipality's properties by creating an Inventory Committee composed of a representative each from the Mayor's Office, from the General Services Office (or the Municipal Treasurer's Office) and a representative from each department to prepare the necessary procedures in the proper and orderly conduct of the physical count of the all the property, plant and equipment and inventories owned by the Municipality.	AAR 2012	Some but not all offices have been issued AREs	Partially implemented	To be fully implemented in the ensuing year

I'm really not sure...

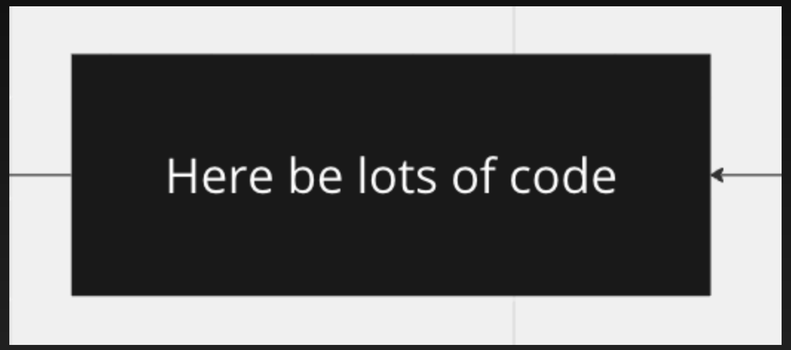
Introduction
Problem discovery

- In late 2022, we discovered a treasure trove of audit reports for the government of the Philippines.
- 3 Tasks:
  - ~~1) Automate scraping of files from the website~~
  - 2) Scrape the files for the data
  - 3a) Create NLP tools to analyze text data
  - 3b) Validate NLP-based modeling as an application for social science

Task 2


Docker Hub --> Phil Image stays here. Contains: - Production source code - Libreoffice + dependencies - Python Libraries - Automation scripts
imagebuilder.sh Updates cloud docker image with production source code
Triggered by phil.sh
Github: Stores Production Source code on the internet.
Git push
Source code on my local machine --> Development Environment
`Zips` folder also held here
Running x64phil.sh or arm64phil.sh
Builds a container from Hub image & runs it locally. Copies local `zip` folder to the container.
Container architecture type (x64 or arm) determined by choice of running script
Upon build (but after copying over data), execute run.sh
Run.sh is a source code script that executes other scripts
1. clean.py
2. convert.sh
3. mksheet.py
Clean.py
1. unzips zip files
2. filter irrelevant files
3. moves target files to dedicated directory
Convert.sh
1. Removes duplicated files
2. Use LibreOffice CLI to convert target files to standard PDF format
Mksheet.py
1. Begins the PDF table extraction
2. Creates a new excel workbook for the data
Copy back extracted data in excel workbook, all PDFs from the scrape, and a log of operations to host machine.
Save data from all PDFs into an excel format.
Here be lots of code

Data Engineering



Majority of project spent here

- What does the lots of code do?
  - 1) Reads PDF, locating relevant tables containing observations
  - 2) Establishes rules for dealing with tables and PDFs with variable headers and formats
  - 3) Coerces the final data to a canonical format for the output

Okay, so you’re scraping PDFs

That’s not so bad, right?

Task 2

Wrong.

Task 2

The Worst Tables You've Ever Seen

Audit Observation	Recommendation	Ref.	Management Action	Status of Implementio n	Reason for Partial/Non Implementation
settlement.					grant,utilization and liquidation of cash advances.
4.) The municipality failed to complete its physical count of the Property, Plant and Equipment (PPE) with a book value of P56,200,363.71 thus completed projects totaling P31,941,343.48 are still recorded in	It is recommended that a physical count of all properties be conducted by the inventory committee annually and a copy of the inventory report be submitted to the auditor within the time frame as required by the pertinent provision in the NGAS	AAR 2012 2011 2010 2009 2008 2007 and 2004	-Physical inventory was conducted on December, 2012. -Reconciliation of PPE account is still on-going. -Asset tagging is on-going.	Partially Implemented	Reconciliation is still on-going.

Task 2

Columns? Please?

system.	municipality's sea ports and waterworks system to generate income that could be used for projects that will benefit its constituents.				
16. Increase in RATA of municipal officials amounting to P156,580 was implemented even if the 55% personal services limitations was already exceeded contrary to the provisions in paragraph 3.0 of Local Budget Circular No. 84 dated April 13, 2007.	Mgt. should strictly observe the provisions cited in LBC No. 84 dated April 13, 2007 in granting the revised RATA rates and be more prudent in releasing additional personnel benefits until such time the PS limitation had been observed.	AAR 2007	The Mgnt granted RATA increase to boost the morale of the department heads and for humanitarian reasons. The LGU only exceeded the 55% PS limitation on the first year of implementation. On the succeeding years, the LGU is strictly observing and complying with the PS limitation and its rules	Not Implemented	It was given emphasis though that on the following year the PS limitation was not exceeded.

Task 2

How does this happen?

This belongs to the observation on the previous table

Audit Observation	Recommendation	Ref.	Management Action	Status of Implementation	Reason for Partial/ Non Implementation
overstating the Assets and understating the Expenses accounts.	a) Henceforth, direct the Municipal Accountant to monitor closely the liquidation of all cash advances and see to it that travel advances are liquidated within thirty (30) days after return to permanent official station in accordance with the guidelines issued under COA Circular No. 96-004, implementing the aforecited Executive Order No. 248, as amended.		No additional cash advance granted without liquidation of the previous cash advance.	Implemented	No employee was assigned to prepare and submit the report to COA.
28. Expenses incurred for catering services provided to officials and employees, guests and visitors during staff meetings, conferences and special occasions in the total amount of P1,309,074.25 were not related to training expenses and considered personal in nature and unnecessary, contrary to Section 343 of Republic Act No. 7160 and COA Circular No. 85-55A dated September 8, 1985, thereby resulting to irregular disbursements of government funds and exposing public funds to wastage.	We recommend that Management: a) Ensure that meals and snacks served during staff meetings or monthly conferences and for the entertainment of guests or visitors shall be properly charged against the representation allowance/s of the concerned LGU officials granted the same. b) Require the Municipal Accountant to check the	AAR CY 2010	Expenses for entertaining guests/visitors were charged to the appropriation for representation expenses. Appropriations and charges were thoroughly verified.	Partially Implemented Implemented	Served only on occasional basis.

And we're supposed to believe this is a single row.

Solution?

Deal with it.

- Enter the data cleaning olympics:

Fuzzy Logic
A cool tangent

- Fuzzywuzzy: library for implementing fuzzy string logic.
- Used to clean up strings in the target!

Searches the status of implementation column for all unique instances of a string like “implemented”

(Removes those like ‘not implemented’)

```
# Demonstration:
matches = find_fuzzy_matches(
    df,
    'status_of_implementation',
    'implemented',
    92)
matches = list(set(
    match.replace('not ', '')
    for match in matches
    ))
matches
```

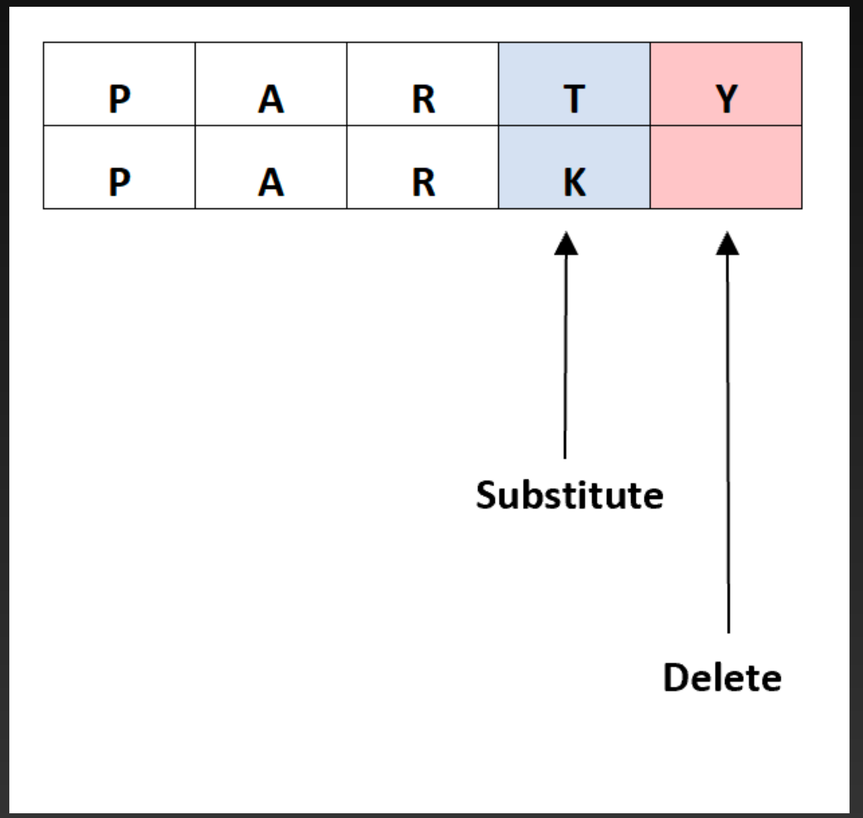
✓ 0.1s

```
['implemente d',
 'unimplemented.',
 'implemented .',
 'unimplemented',
 'implemen ted',
 'impleme nted',
 'implemented',
 'implement ed',
 'implemen- ted',
 'implemented.',
 'imple mented']
```

Fuzzy Logic
Levenshtein Distance

- String metric for measuring the difference between two sequences.
- “Informally, the Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other.”

https://en.wikipedia.org/wiki/Levenshtein_distance



https://www.statology.org/levenshtein-distance-in-python/

Final Dataset

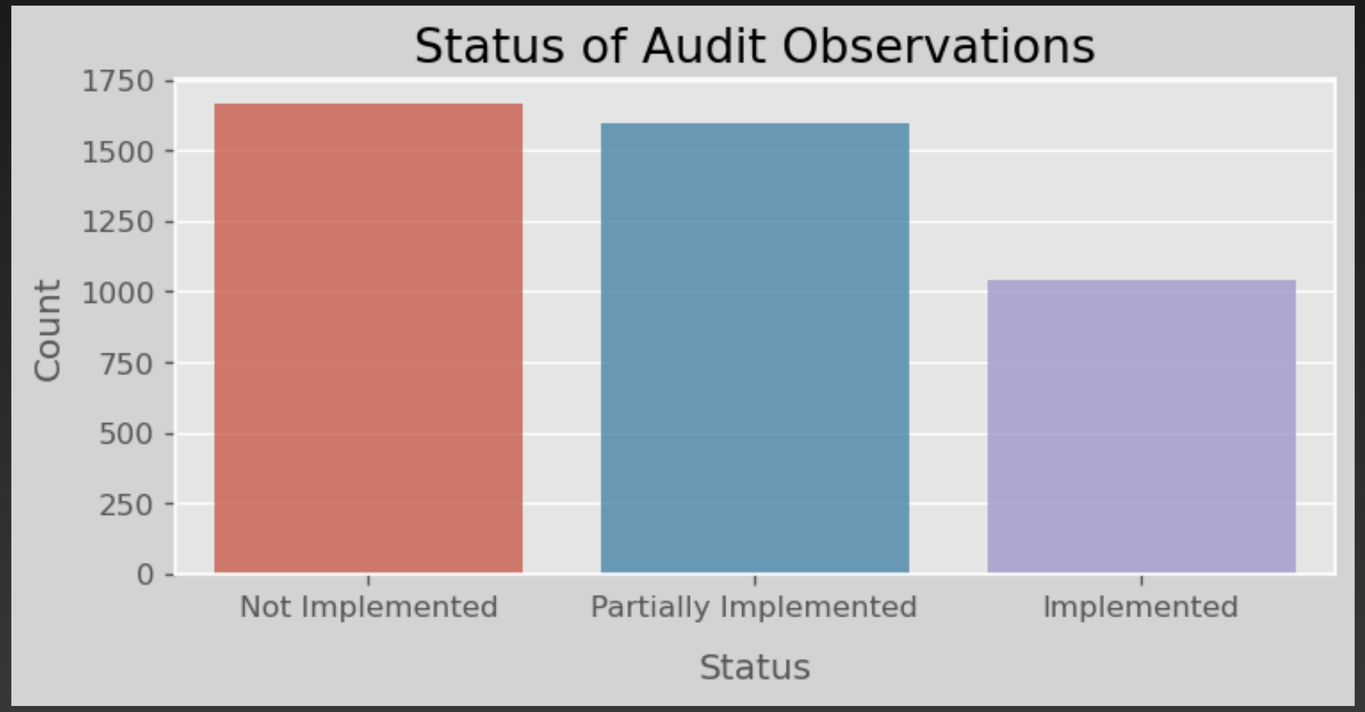
It’s not final ‘till it’s final.

- Conglomerate `audit_observation`, `recommendations`, and `management_action` text
- Coalesce `status_of_implementation` with `reasons_for_partial_or_non_implementation` where appropriate
- Calculate % implementation for each observation
- Map the completion value to a category:

Completion Percentage	Categorical Value
0	0
0 < x < 1	1
1	2

EDA
Sleeper-fest

- EDA for this project was not that interesting, but here's a graph of the class balance!

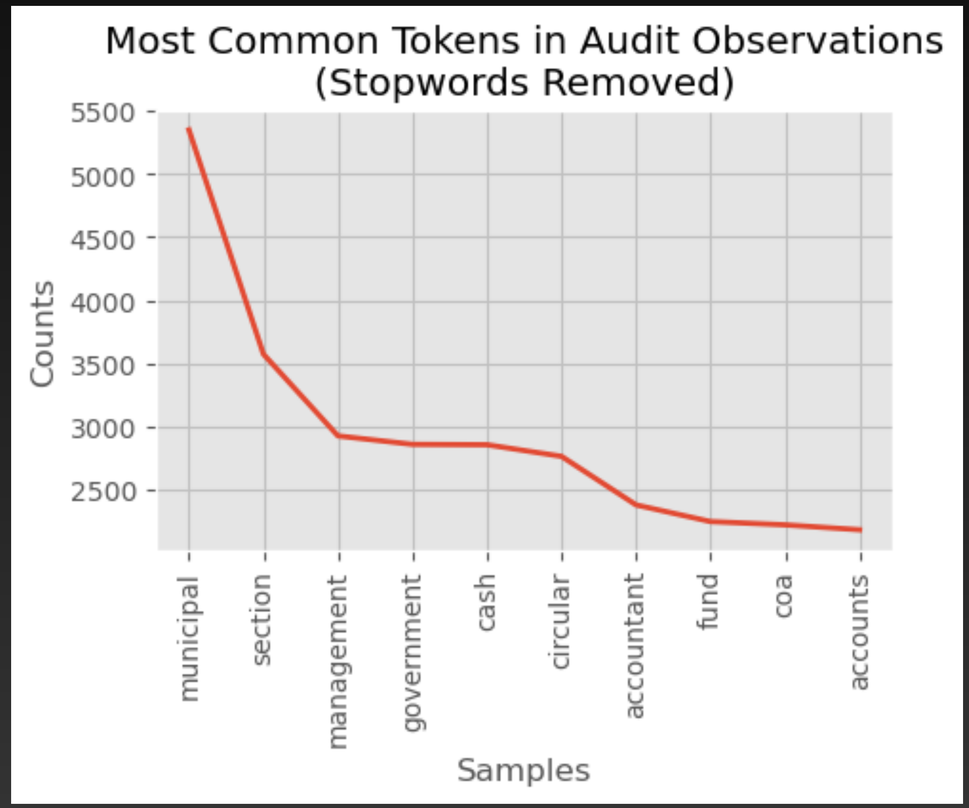


Task 3

EDA/Processing

What was done:

- Audit observation word and character length distributions
- Tokenization & Token frequency analysis
- Phrase (2&3-gram) distributions with and without stop words
- Lemmatization
- Stemming



Task 2 Finished!
Problems

- 3 Tasks:
  - ~~1) Automate scraping of files from the website~~
  - ~~2) Scrape the files for the data~~
  - 3a) Create NLP tools to analyze text data
  - 3b) Validate NLP-based modeling as an application for social science

Task 3.

Defining the Scope

Why do NLP?

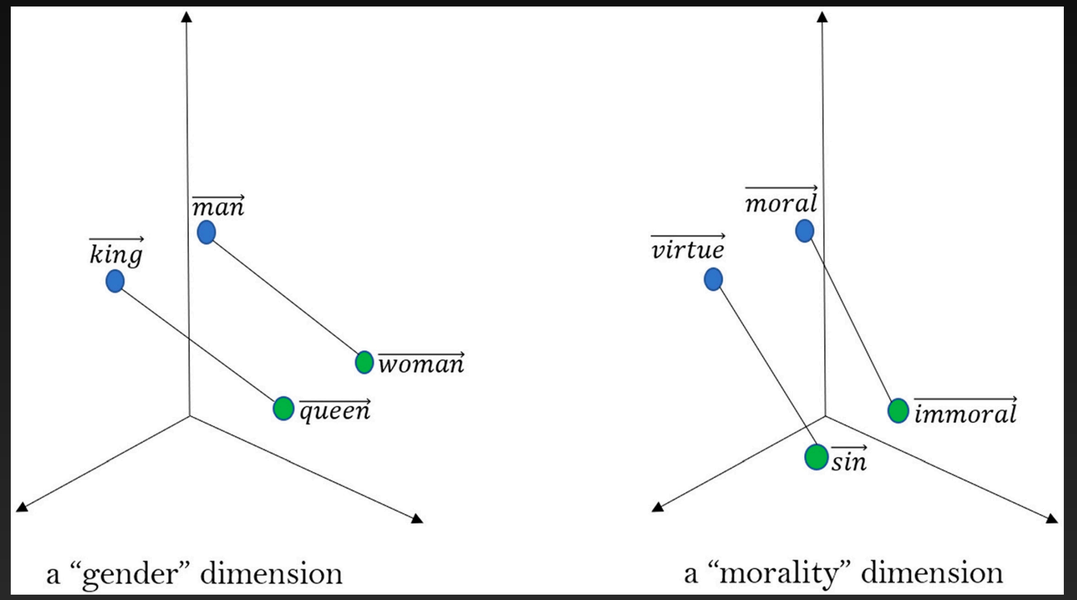
- Applying computational techniques like word vectorization, dimensionality reduction, and clustering helps to uncover hidden patterns.
- Can we go further?
- Word Embeddings!
- Can we use these word embeddings of the audit observation data to predict the status of implementation?

Task 3

Unsupervised Learning

Word Embeddings

A technique that converts words into numerical representations, making it easier for computers to process and analyze text.



These representations of words as vectors in high dimensional space capture syntactical relationships and similarities between words, allowing computers to understand context and meaning.

HuggingFace

Doing it the easy (best) way

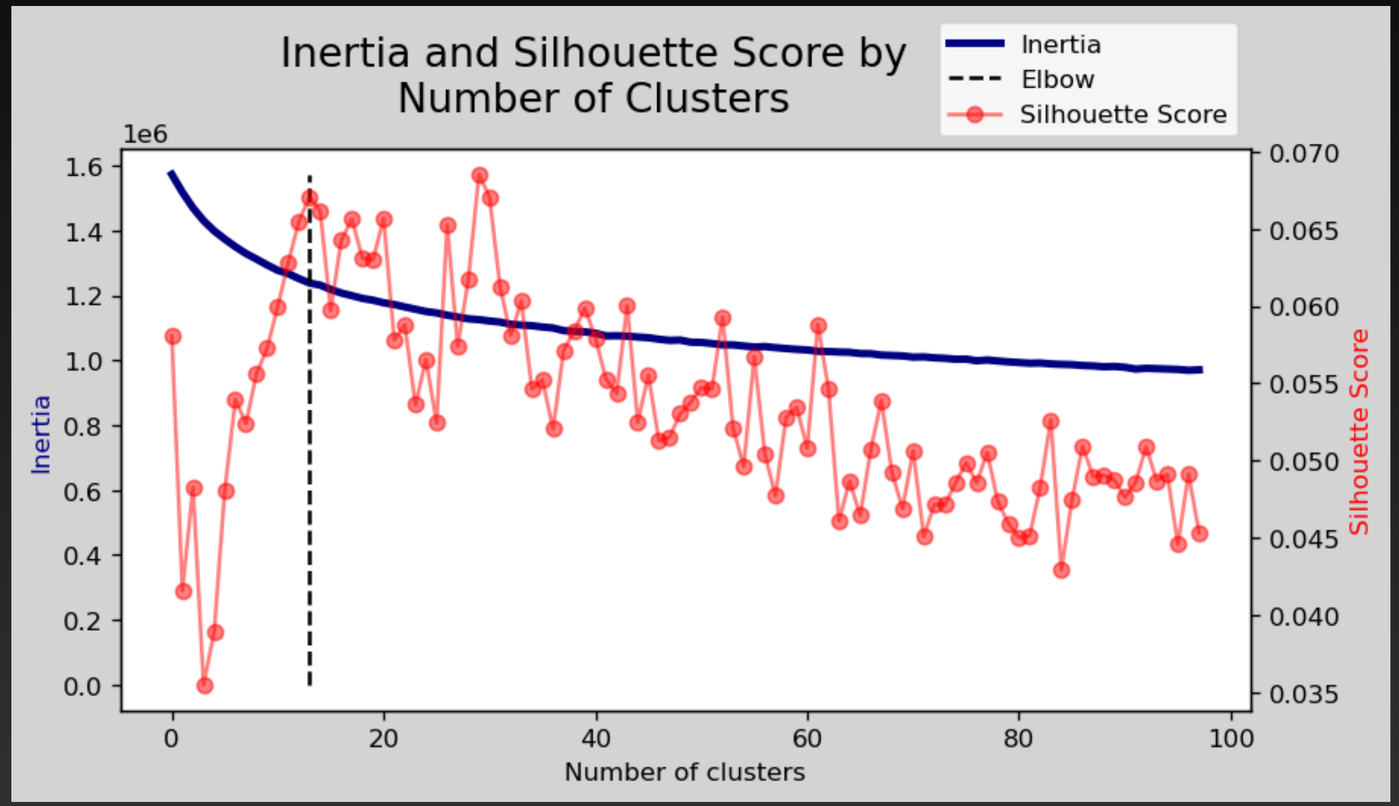
- Just load up a `SentenceTransformer` from the `sentence_transformer` library and encode the text!
- HuggingFace `all-MiniLM-L6-v2` sentence transformer used here
- Smaller model that is optimized for speed and memory usage,
- Chosen because of the small corpus and limited compute resources

Hugging Face

Task 3

Clustering

K-Means Failure :(



Even after conducting principal component analysis, attempts at clustering the data using only the word embeddings were unsuccessful.

Task 3a Finished!
Problems

- 3 Tasks:
  - ~~1) Automate scraping of files from the website~~
  - ~~2) Scrape the files for the data~~
  - ~~3a) Create NLP tools to analyze text data~~
  - 3b) Validate NLP-based modeling as an application for social science

Predictive Models

Philosophy

- Focusing on building a model with the best predictive ability.
- No care for interpretability, as our feature space is already only barely human interpretable anyways.
- Evaluation metrics:
  - **Accuracy**: Classic, straightforward, reliable.
  - **F1-Score**: Balances load between precision and recall. Helps judge model's decision-making
  - **ROC AUC**: Demonstrates model's ability to distinguish between classes

Baseline Model

Dummy Classifier

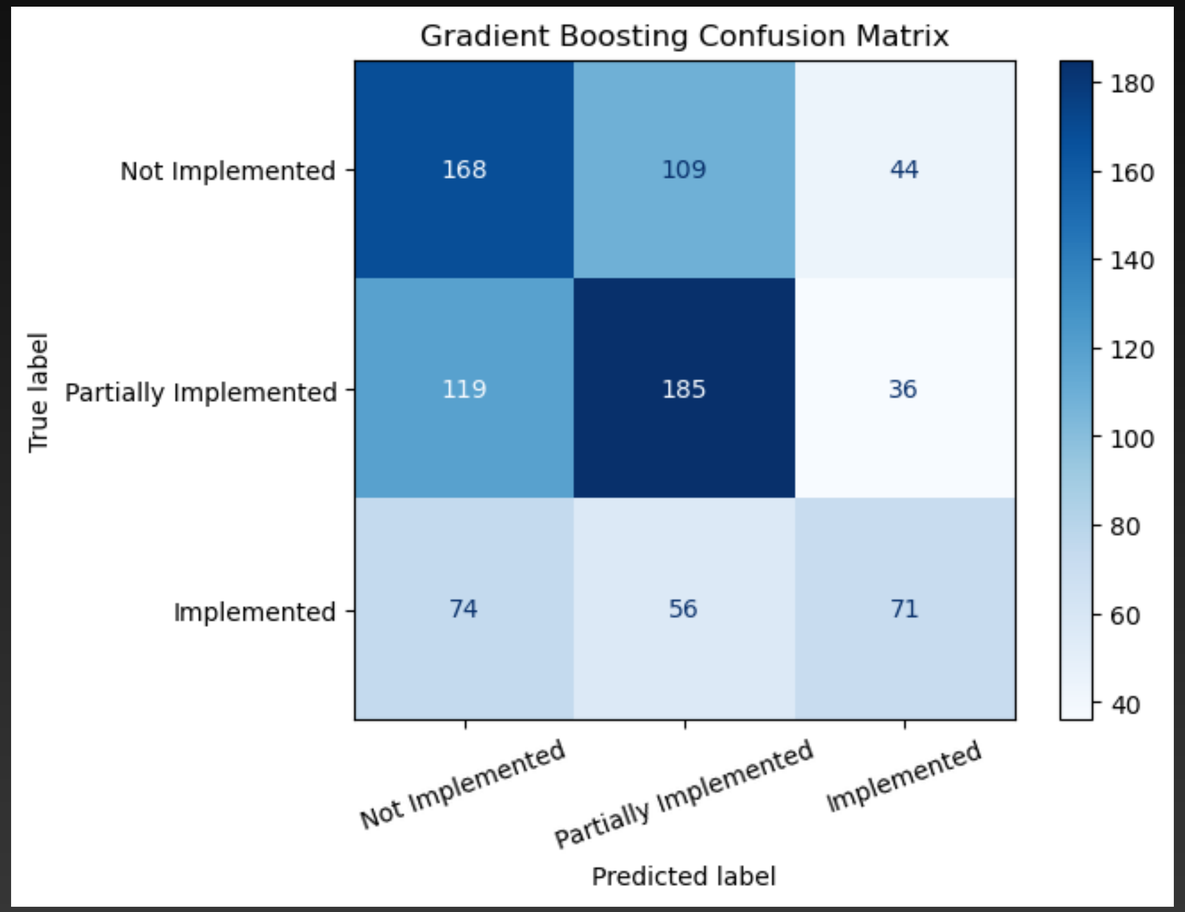
	Precision	Recall	F1-Score
Not implemented	0.37	1.0	0.54
Partially Implemented	0	0	0
Implemented	0	0	0
Weighted Avg.	0.14	0.37	0.20

```
dummy = DummyClassifier(
    strategy='most_frequent',
    random_state=42)
```

Not too difficult to beat, right?

Model Results

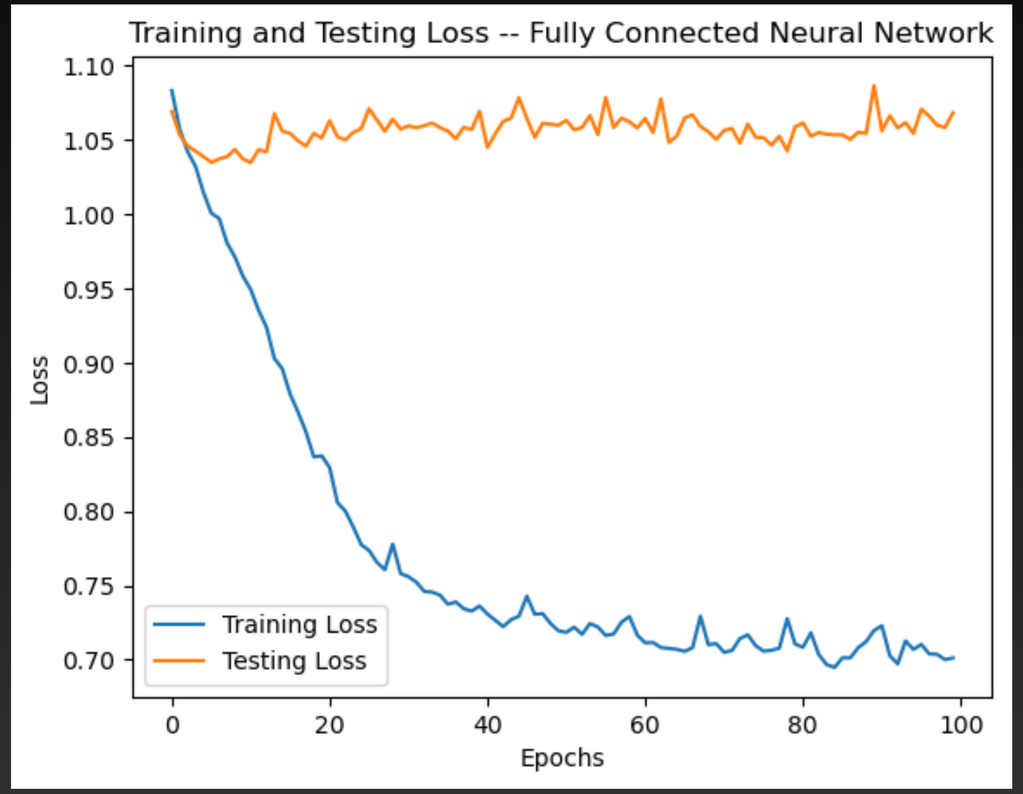
Gradient Boosted Tree



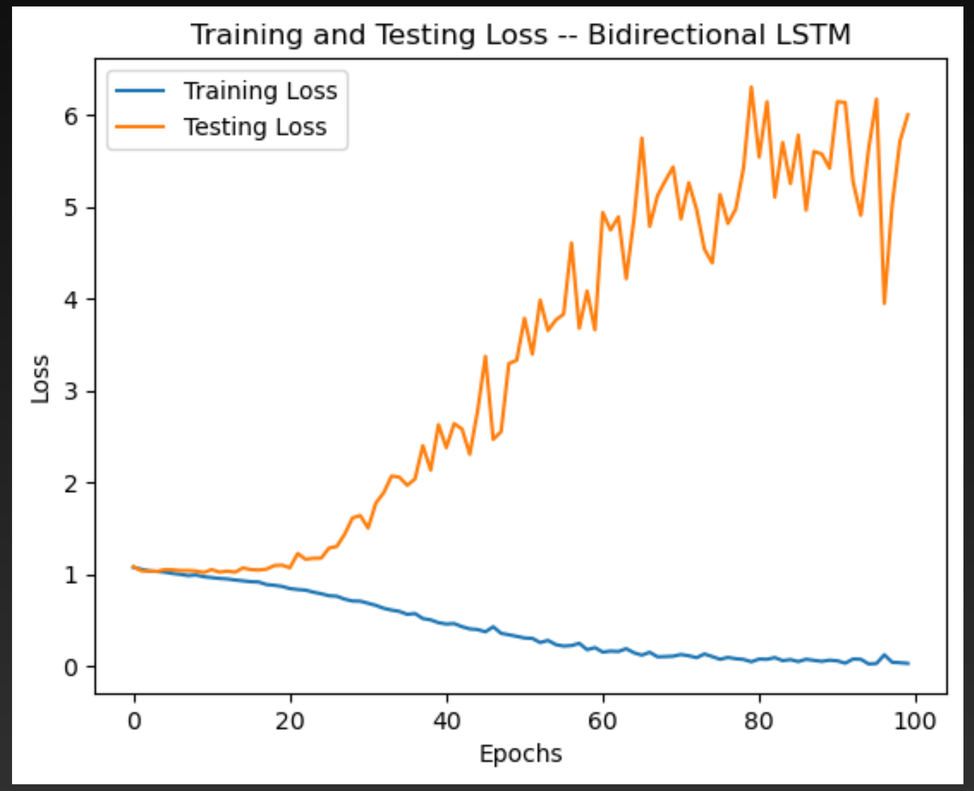
- Train Acc: 1.0
- Test Acc: 0.49
- **Best model.**

	precision	recall	f1-score	support
0	0.47	0.52	0.49	321
1	0.53	0.54	0.54	340
2	0.47	0.35	0.40	201
accuracy			0.49	862
macro avg	0.49	0.47	0.48	862
weighted avg	0.49	0.49	0.49	862

Model Results:
Neural Networks



Architecture: 3 hidden layers of sizes (256, 128, 64)



Architecture: BLSTM with 2 hidden layers size (256, 128)

Task 3

Model Results

Queue Radiohead

Model	Train Acc	Test Acc	F1-Score	ROC AUC
Baseline	-	0.37	0.20	0.50
Logistic Regression	0.46	0.44	0.46	0.58
KNN Classifier	0.44	0.43	0.45	0.58
Gradient Boosted Tree	1	0.49	0.49	0.58
Fully Connected Neural Network	0.48	0.48	0.48	0.50
Bidirectional LSTM	0.48	0.48	0.48	0.50

Model Results

Summary

- Gradient Boosting Classifier has the best Test Accuracy, F1 Score, and ROC AUC among all models.
- However, its overfitting issue should be addressed to improve generalization.
- The Logistic Regression, KNN Classifier, and Bidirectional LSTM models show similar performance levels but are better than the baseline model.
- The Fully Connected Neural Network has a slightly higher Test Accuracy than the aforementioned models but a worse ROC AUC.

Interpreting the Results

Pity party, or a new hope?

- Reservations about inference:
  - Models better than baseline... why?
- The nature of choice.
- If the models are learning, it mostly has to do with the doubling in F1-Score
  - I suspect that transforming the problem into binary classification would yield significant improvements in model performance!

Task 3b Started..?
Problems

- 3 Tasks:
  - ~~1) Automate scraping of files from the website~~
  - ~~2) Scrape the files for the data~~
  - ~~3a) Create NLP tools to analyze text data~~
  - 3b) Validate NLP-based modeling as an application for social science

Goals for the Future

Silicon Valley here I come.

- Final dataset for this analysis consisted of only ~4000 observations. Full potential dataset contains ~50,000, so there is plenty of room to grow here.
- Pivot application to using word embeddings to extend corruption typology scheme.
- Web interface for researchers to interact with directly
- Auditing more countries (I want to be on all the lists)