Classy Data Analysis

Hansen | Nagda | Pimentel | Reyes | Tai

Challenge

We have a 1-D understanding of NPOs...

- More often than not, the services of an NPO span multiple sectors, e.g Health, Education, etc.
- Financially speaking, small local charities operate very differently from multi-million dollar organizations.

We need a more complex solution to find like-minded organizations.

IRS Form 990 + 

Solution

... but NPOs are multi-faceted.

- Lay out a common “social space”, where the organizations that drive social change and potential donors can connect, find organizations and be offered recommendations, and where discovery of new causes - and events within causes (i.e. fundraisers) - could be facilitated.

- Use combination of government IRS form 990 (returns for nonprofits) data along with external textual information (i.e. social media) to create a robust semantic space.

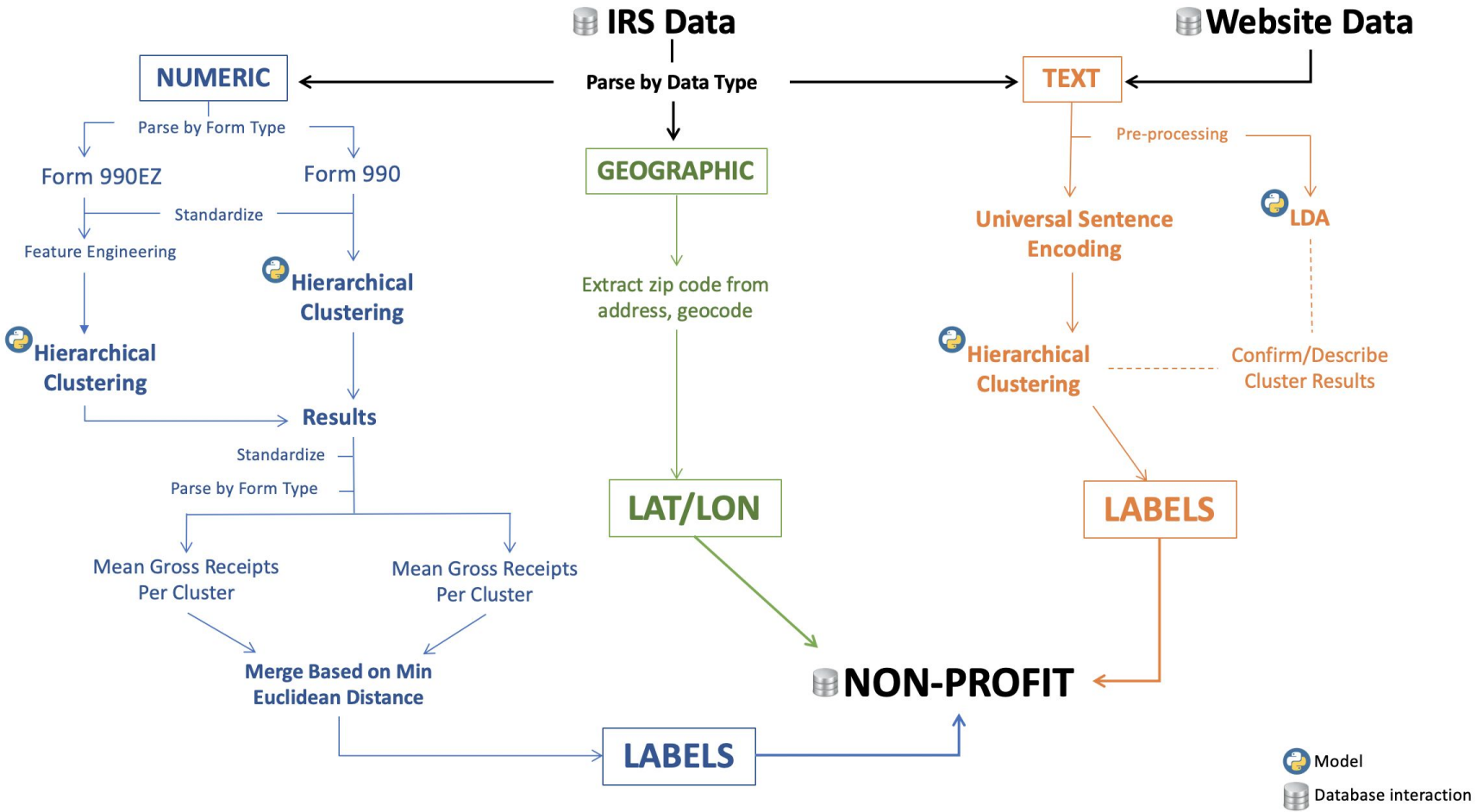
IRS

Form 990

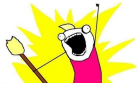
Agenda

1. Introduction
2. Solution
3. Pipeline
4. Communication Plan
5. Visualization
6. Next Steps

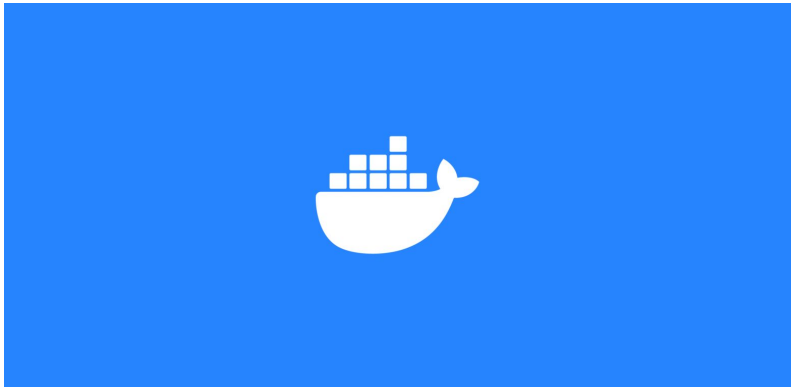
Pipeline


IRS Data
Website Data
NUMERIC
Parse by Data Type
TEXT
Parse by Form Type
Pre-processing
Form 990EZ
Form 990
GEOGRAPHIC
Standardize
Universal Sentence Encoding
LDA
Feature Engineering
Hierarchical Clustering
Extract zip code from address, geocode
Hierarchical Clustering
Hierarchical Clustering
Confirm/Describe Cluster Results
Results
Standardize
Parse by Form Type
LAT/LON
LABELS
Mean Gross Receipts Per Cluster
Mean Gross Receipts Per Cluster
Merge Based on Min Euclidean Distance
NON-PROFIT
LABELS
Model
Database interaction

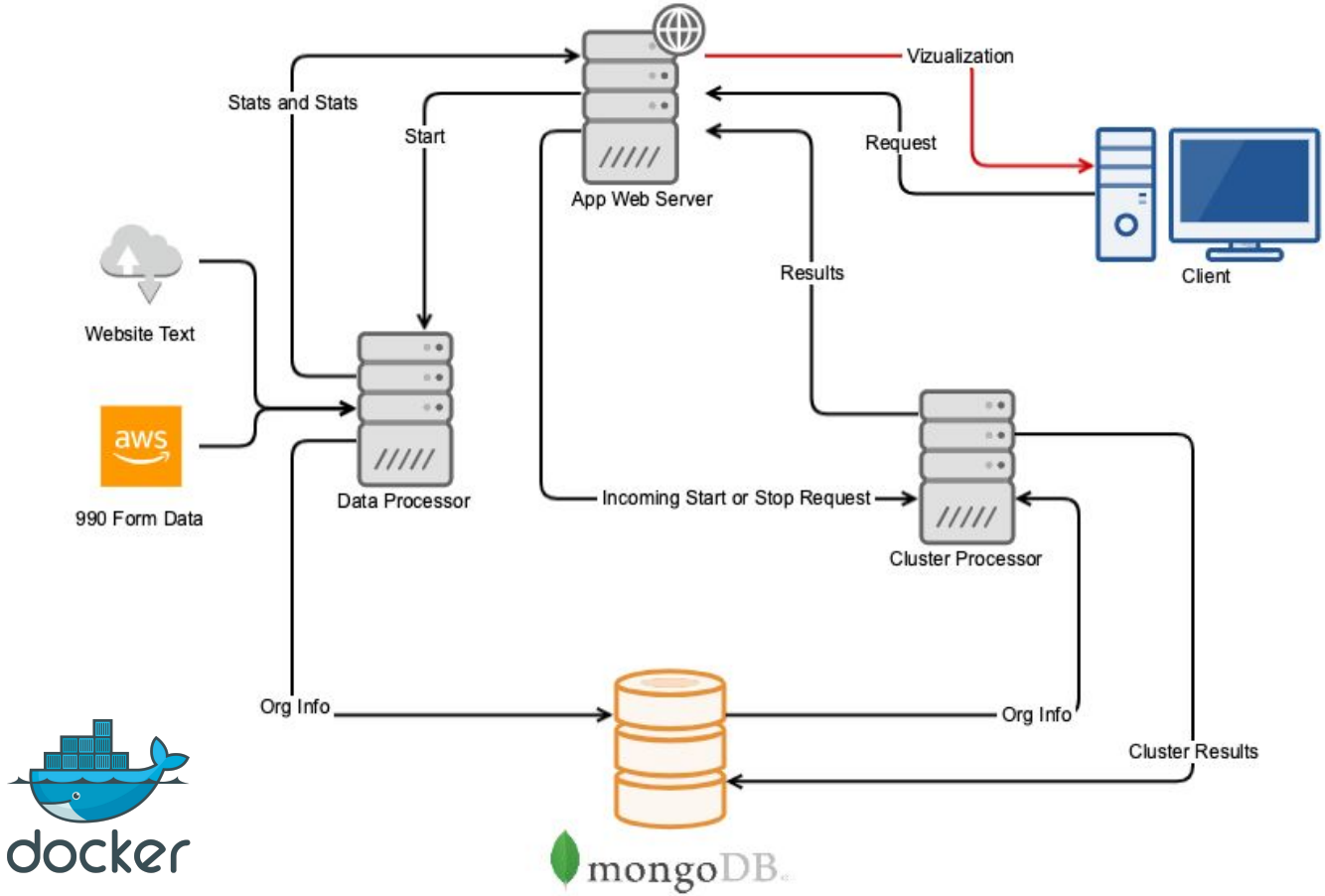
Docker Everything



- Created a complete docker environment for each module
- Allowed our code to be platform independent and easily scalable.



Communication Plan



Visualization

MOTIVATION:

A **single**, easy-to-understand and **interactive** visual that answers the question:

“Which organizations are most similar to an organization that I care about?”

While tying together the three main descriptive categories of nonprofit organizations (text, finance, geography)

Visualization

A Story of Usage

A very passionate supporter of the World Wide Fund for Nature (WWF) wants to reach out to other like-minded nonprofits to open discussions about creating a powerful “alliance” for nature conservancy. She desires to reach out to similarly well-funded organizations like the WWF and plans on personally meeting with the heads of such organizations.

Visualization

- Text Data - 200 features after embedding + PCA
- Financial Data - 34 features
- Geographical Data - 2 features (latitude / longitude)

To represent these 3 categorical dimensions in a human-interpretable space, we collapse these 3 data sources into a single feature, each:

Distance to centroid!

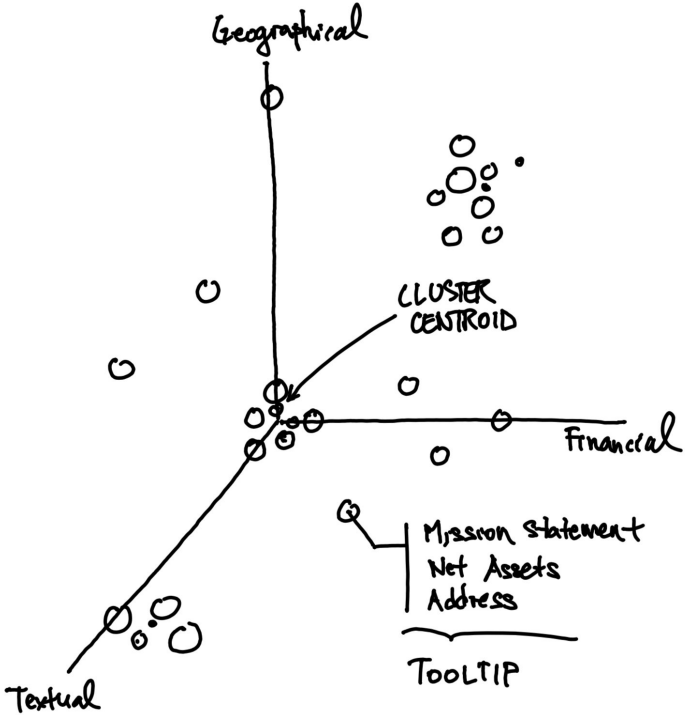
Visualization

How it Works:

1. Select an organization (Org. A)
2. Identify Org. A's cluster in "text space" and find all samples belonging to that cluster (Set A)
3. Identify Org. A's cluster in "financial space", find all samples belonging to that cluster (Set B), and create the union of Set A and Set B (Set C)
4. For all samples in Set C:
    a. Calculate their distance in "text space" to the cluster centroid of Org. A
    b. Calculate their distance in "financial space" to the cluster centroid of Org. A
    c. Calculate their geographical distance from the location of Org. A
5. Plot distances on a 3-axis visualization (each axis for the 3 metrics in Step 4)

Visualization

- Organizations close to the origin are similar to Org. A
- Organizations close to each other that are far away from the origin are close to each other and represent pockets of similar organizations that are loosely related to Org. A



Visualization Demo

Next Steps

- Finish visualization dashboard web application
  - Integrate tooltip interactivity
  - Build feature for selective filtering along descriptive dimensions
- Build out docker container that will include the entire data pipeline and code base

Questions?