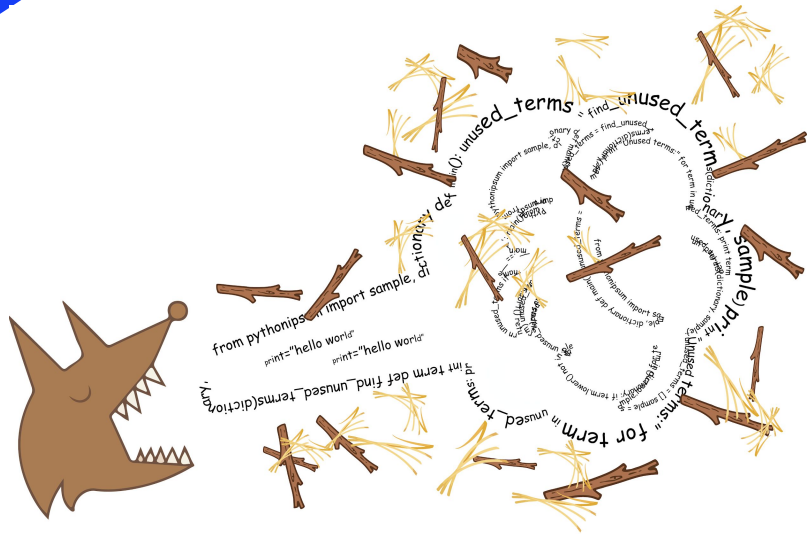


Huff n' Puff & Classify

Adrien Allorant, Ruofan Cai, Joseph Frostad, Kevin Hsu, Tianyi Li

Project Functional Specification

Background

Inadequate housing quality is a risk factor for a variety of negative health outcomes that are primarily experienced by people who live in developing countries. Properly identifying the geographic scope and magnitude of this problem is critical to designing effective interventions for those who experience the highest levels of risk.

Large-scale household survey datasets exist for many countries, but the variety of housing types used globally make it challenging to synthesize this data. Conversion of raw string descriptions of housing materials into ordinal values depicting the overall quality of a home will make quantitative analysis on a global scale feasible and allow for prediction modelling in order to identify high-risk areas.

User Profile

The intended users of this tool are researchers who study housing quality in developing countries.

They will have levels of Python experience that vary in between none and intermediate level. Those who cannot program in Python will be able to browse the web in a research setting and interpret data visualizations in order to draw conclusions within their topic area of expertise.

Given the variety of userbase experience, this toolkit will allow users to interact with both intermediate outputs and visualizations within Jupyter notebooks and also provide static outputs and visualizations that will increase user-friendliness - adding utility to novice users.

Data Sources

- Household survey datasets with housing quality classified solely as raw string descriptions
  - *Example from Reproductive Health Survey (Georgia 2010)*

13. **WHAT ARE THE MAIN MATERIALS USED IN THE ROOF? (RECORD OBSERVATION)**

1. **ROOF FROM NATURAL MATERIALS**
2. **RUDIMENTARY ROOF (PLASTIC, CARDBOARD, TAR-IMPREGNATED SHEETS)**
3. **CONCRETE ROOF**
4. **ASPHALT SHINGLES**
5. **CLAY TILES. CERAMIC TILES**
6. **CORRUGATED GALVANIZED IRON**
7. **SHEET METAL (COPPER, LEAD, ZINC, STEEL)**
8. **OTHER (SPECIFY)** _____

- Household survey datasets with housing quality classified as both raw string descriptions and as ordinal values which rank the relative quality
  - *Example from Demographic Health Survey v5*

NO.	QUESTIONS AND FILTERS	CODING CATEGORIES		SKIP
118	MAIN MATERIAL OF THE ROOF. (4) RECORD OBSERVATION.	NATURAL ROOFING		
		NO ROOF	11	
		THATCH/PALM LEAF	12	
		SOD	13	
		RUDIMENTARY ROOFING		
		RUSTIC MAT	21	
		PALM/BAMBOO	22	
		WOOD PLANKS	23	
		CARDBOARD	24	
		FINISHED ROOFING		
		METAL	31	
		WOOD	32	
		CALAMINE/CEMENT FIBER	33	
		CERAMIC TILES	34	
		CEMENT	35	
		ROOFING SHINGLES	36	
		OTHER _____ (SPECIFY)	96	

Data

Merge Datasets (?)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1		ihme_loc_id	nid	survey_serie	hhweight	urban	hh_size	year_start	int_year	housing_roo	housing_wal	housing_floo	housing_roo	housing_wal	housing_floo	iso3	cluster_id	index	floor_rank	region_name	super_region_name	
2	1	AFG	56830	UNICEF_MIC	0.719489	1	8	2010	2011	Rustic mat	Dirt	Vinyl or asph	21	13	32	AFG	2033	2973531	3	North Africa	North Africa and Middle East	
3	2	AFG	56830	UNICEF_MIC	0.719489	1	6	2010	2011	Wood planks	Bricks	Carpet	23	33	35	AFG	2033	2973532	3	North Africa	North Africa and Middle East	
4	3	AFG	56830	UNICEF_MIC	0.719489	1	7	2010	2011	Cement	Cement	Carpet	35	31	35	AFG	2033	2973533	3	North Africa	North Africa and Middle East	
5	4	AFG	56830	UNICEF_MIC	0.719489	1	6	2010	2011	Roofing shin	Uncovered a	Carpet	36	23	35	AFG	2033	2973534	3	North Africa	North Africa and Middle East	
6	5	AFG	56830	UNICEF_MIC	0.719489	1	12	2010	2011	Rustic mat	Dirt	Carpet	21	13	35	AFG	2033	2973535	3	North Africa	North Africa and Middle East	
7	6	AFG	56830	UNICEF_MIC	0.719489	1	6	2010	2011	Cement	Cement	Carpet	35	31	35	AFG	2033	2973536	3	North Africa	North Africa and Middle East	
8	7	AFG	56830	UNICEF_MIC	0.719489	1	4	2010	2011	Ceramic tiles	Stone with li	Carpet	34	32	35	AFG	2033	2973537	3	North Africa	North Africa and Middle East	
9	8	AFG	56830	UNICEF_MIC	0.719489	1	9	2010	2011	Cement	Cement	Carpet	35	31	35	AFG	2033	2973538	3	North Africa	North Africa and Middle East	
10	9	AFG	56830	UNICEF_MIC	0.719489	1	3	2010	2011	Cement	Bricks	Cement	35	33	34	AFG	2033	2973539	3	North Africa	North Africa and Middle East	
11	10	AFG	56830	UNICEF_MIC	0.719489	1	5	2010	2011	Wood	Dirt	Earth / sand	32	13	11	AFG	2033	2973540	1	North Africa	North Africa and Middle East	
12	11	AFG	56830	UNICEF_MIC	0.719489	1	8	2010	2011	Wood	Mud wall/Ba	Carpet	32	21	35	AFG	2033	2973541	3	North Africa	North Africa and Middle East	
13	12	AFG	56830	UNICEF_MIC	0.719489	1	8	2010	2011	Wood	Bricks	Vinyl or asph	32	33	32	AFG	2033	2973542	3	North Africa	North Africa and Middle East	
14	13	AFG	56830	UNICEF_MIC	0.719489	1	7	2010	2011	Rustic mat	Covered ado	Cement	21	35	34	AFG	2033	2973543	3	North Africa	North Africa and Middle East	
15	14	AFG	56830	UNICEF_MIC	0	1	NA	2010	2011				NA	NA	NA	AFG	2033	2973544	NA	North Africa	North Africa and Middle East	
16	15	AFG	56830	UNICEF_MIC	0.719489	1	4	2010	2011	Metal	Dirt	Carpet	31	13	35	AFG	2033	2973545	3	North Africa	North Africa and Middle East	
17	16	AFG	56830	UNICEF_MIC	0.719489	1	5	2010	2011	Wood	Dirt	Vinyl or asph	32	13	32	AFG	2033	2973546	3	North Africa	North Africa and Middle East	
18	17	AFG	56830	UNICEF_MIC	0.719489	1	3	2010	2011	Rustic mat	Dirt	Carpet	21	13	35	AFG	2033	2973547	3	North Africa	North Africa and Middle East	
19	18	AFG	56830	UNICEF_MIC	0.719489	1	4	2010	2011	Wood	Dirt	Vinyl or asph	32	13	32	AFG	2033	2973548	3	North Africa	North Africa and Middle East	
20	19	AFG	56830	UNICEF_MIC	0.719489	1	5	2010	2011	Rustic mat	Uncovered a	Vinyl or asph	21	23	32	AFG	2033	2973549	3	North Africa	North Africa and Middle East	
21	20	AFG	56830	UNICEF_MIC	0.719489	1	8	2010	2011	Metal	Bricks	Carpet	31	33	35	AFG	2033	2973550	3	North Africa	North Africa and Middle East	

Use Cases

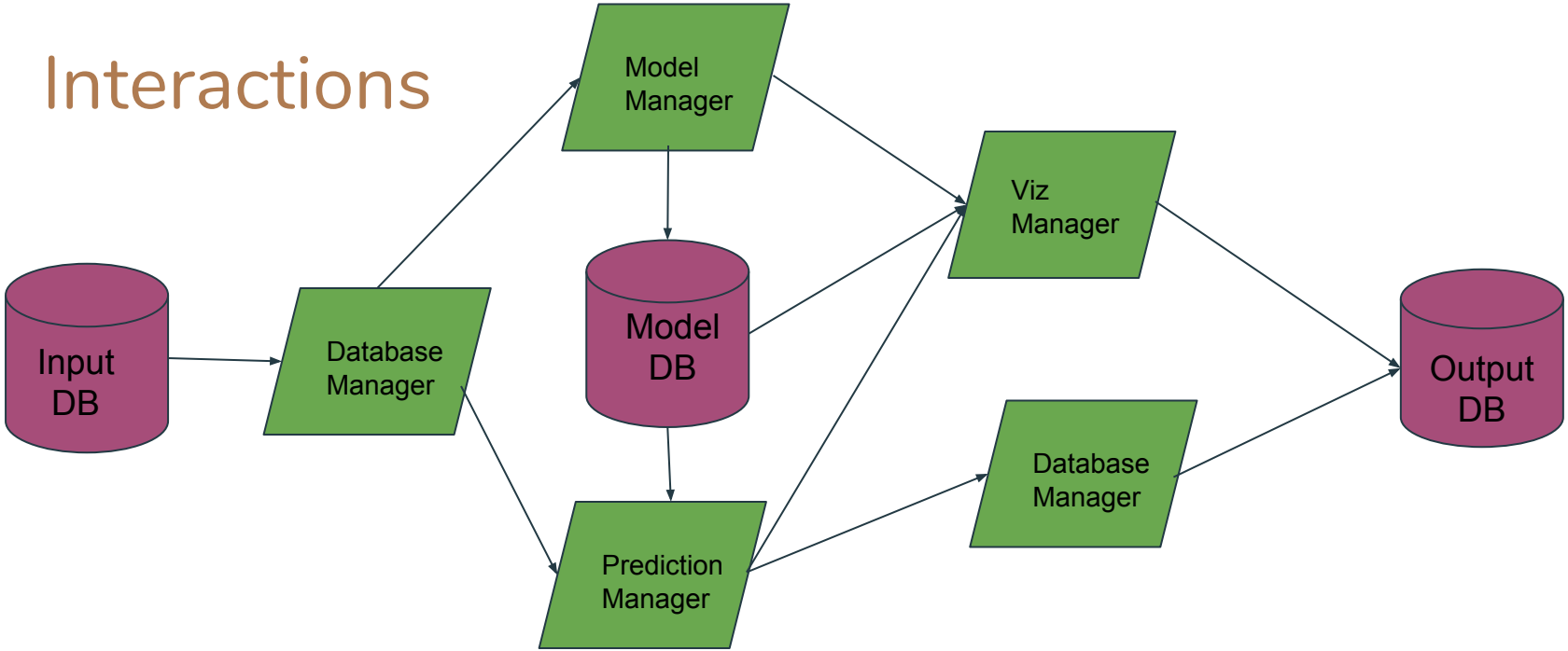
- Quantitative researchers trying to model the ordinal scores for housing quality across time and space
  - They may be asking questions such as:
    - In which areas of Nigeria are people most likely to be living in rudimentary homes?
    - What is the average housing quality score for people in Peru? Has this value changed from 2000-present?
  - They will be most interested in using the predicted values from this software (final output)
- Qualitative researchers who want to better understand the types of housing that are prevalent in certain places
  - They may be asking questions such as:
    - In Southeast Asia, what are the most common categories of housing materials used when constructing a modern home?
    - What are the different ways in which stone is used to construct walls and how do they vary regionally?
  - They will be most interested in using the data visualization piece of this software in order to explore the relationships between strings inputs (intermediate output/visualization)

Project Component Specification

Software Components

- Database manager:
  - Simplified interface to access the databases containing the raw input data and the output predictions of ordinal score for housing quality
  - This component will be responsible for data pre-processing and for building representative test and training datasets in order to inform a cross-validation modelling approach
- Visualization manager:
  - Accesses intermediate and final outputs in order to create maps of housing quality and visualize the networks resulting from interdependencies across the universe of potential string and ordinal values
- Model manager
  - Accesses pre-processed input data as training/test subdata and uses ML toolkits in order to analyze relationships between string keywords and ordinal values relating to quality
- Prediction manager:
  - Input the characteristics of housing (material of the roof/walls/floor), and results of model built/optimized using training dataset, then produces as output an ordinal score for housing quality

Interactions



Preliminary Plan

1. Design logo (prototype made)
2. Clean data
3. Technology Review
4. Decide on preferred NLP toolkit
5. Divide dataset into training/test data for cross-validation
6. Use NLP to model relationship between short text values and ordinal scores using training data
7. Predict ordinal scores in test data
8. Refine model, repeat 4-6
9. Visualize results