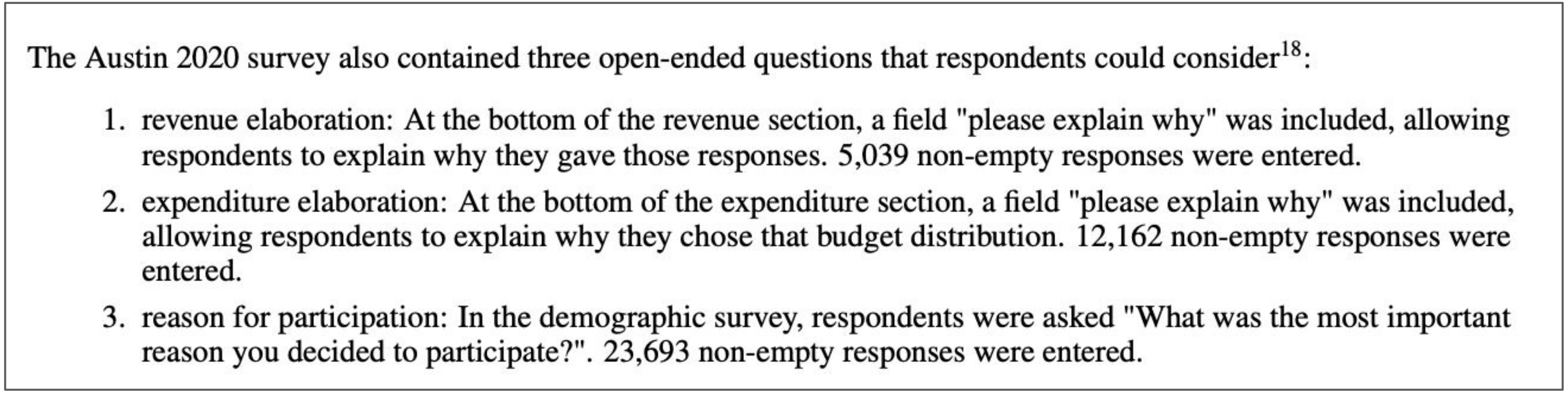
Project Title: Understanding and Correlating Quantitative and Qualitative Survey Clusters Using LLMs, Word2Vec, and Machine Learning

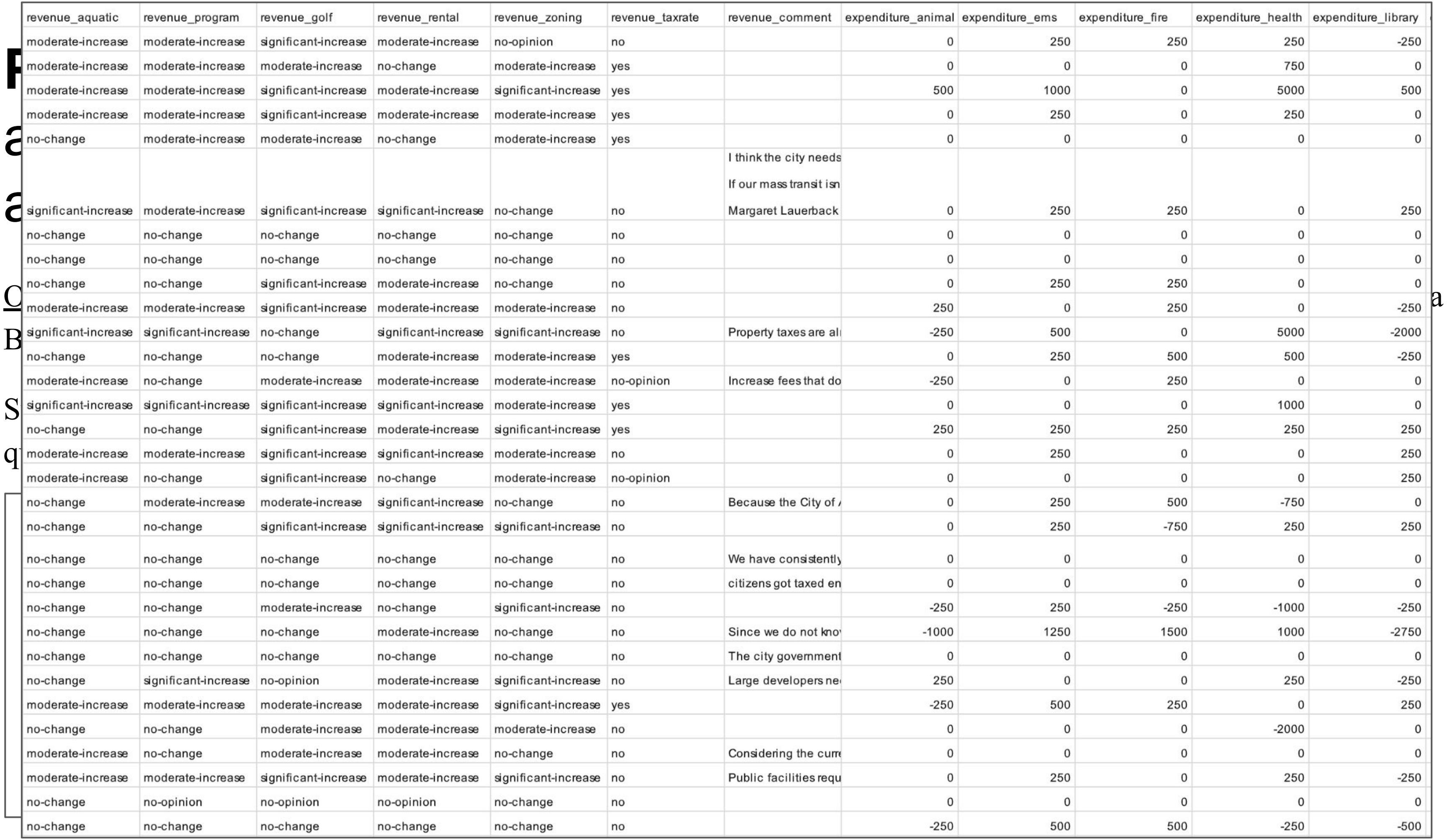
Original paper: “Opinion Change or Differential Turnout: Changing Opinions on the Austin Police Department in a Budget Feedback Process”

Surveys taken from 2020, 2021, and 2022 (pre and post George Floyd shock). Original survey data contained both quantitative and qualitative responses



The Austin 2020 survey also contained three open-ended questions that respondents could consider[18]:

1. revenue elaboration: At the bottom of the revenue section, a field "please explain why" was included, allowing respondents to explain why they gave those responses. 5,039 non-empty responses were entered.
2. expenditure elaboration: At the bottom of the expenditure section, a field "please explain why" was included, allowing respondents to explain why they chose that budget distribution. 12,162 non-empty responses were entered.
3. reason for participation: In the demographic survey, respondents were asked "What was the most important reason you decided to participate?". 23,693 non-empty responses were entered.

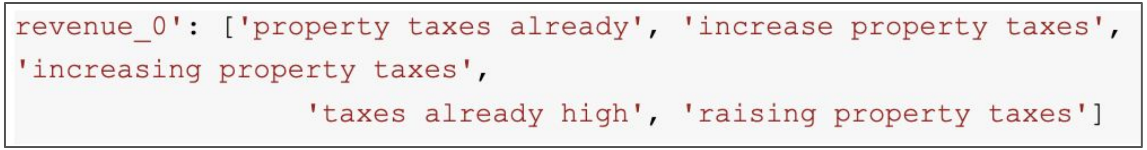


revenue_aquatic	revenue_program	revenue_golf	revenue_rental	revenue_zoning	revenue_taxrate	revenue_comment	expenditure_animal	expenditure_ems	expenditure_fire	expenditure_health	expenditure_library
moderate-increase	moderate-increase	significant-increase	moderate-increase	no-opinion	no		0	250	250	250	-250
moderate-increase	moderate-increase	moderate-increase	no-change	moderate-increase	yes		0	0	0	750	0
moderate-increase	moderate-increase	significant-increase	moderate-increase	significant-increase	yes		500	1000	0	5000	500
moderate-increase	moderate-increase	significant-increase	moderate-increase	moderate-increase	yes		0	250	0	250	0
no-change	moderate-increase	moderate-increase	no-change	moderate-increase	yes		0	0	0	0	0
significant-increase	moderate-increase	significant-increase	significant-increase	no-change	no	I think the city needs If our mass transit isn Margaret Lauerback	0	250	250	0	250
no-change	no-change	no-change	no-change	no-change	no		0	0	0	0	0
no-change	no-change	no-change	no-change	no-change	no		0	0	0	0	0
no-change	no-change	significant-increase	moderate-increase	no-change	no		0	250	250	0	0
moderate-increase	moderate-increase	significant-increase	moderate-increase	moderate-increase	no		250	0	250	0	-250
significant-increase	significant-increase	no-change	significant-increase	significant-increase	no	Property taxes are al	-250	500	0	5000	-2000
no-change	no-change	no-change	moderate-increase	moderate-increase	yes		0	250	500	500	-250
moderate-increase	no-change	moderate-increase	moderate-increase	moderate-increase	no-opinion	Increase fees that do	-250	0	250	0	0
significant-increase	significant-increase	significant-increase	significant-increase	moderate-increase	yes		0	0	0	1000	0
no-change	no-change	significant-increase	moderate-increase	significant-increase	yes		250	250	250	250	250
moderate-increase	moderate-increase	significant-increase	significant-increase	moderate-increase	no		0	250	0	0	250
moderate-increase	no-change	significant-increase	no-change	moderate-increase	no-opinion		0	0	0	0	250
no-change	moderate-increase	moderate-increase	significant-increase	no-change	no	Because the City of	0	250	500	-750	0
no-change	no-change	significant-increase	significant-increase	significant-increase	no		0	250	-750	250	250
no-change	no-change	no-change	no-change	no-change	no	We have consistently	0	0	0	0	0
no-change	no-change	no-change	no-change	no-change	no	citizens got taxed en	0	0	0	0	0
no-change	no-change	moderate-increase	no-change	significant-increase	no		-250	250	-250	-1000	-250
no-change	no-change	no-change	moderate-increase	no-change	no	Since we do not kno	-1000	1250	1500	1000	-2750
no-change	no-change	no-change	no-change	no-change	no	The city government	0	0	0	0	0
no-change	significant-increase	no-opinion	moderate-increase	significant-increase	no	Large developers ne	250	0	0	250	-250
moderate-increase	moderate-increase	moderate-increase	moderate-increase	significant-increase	yes		-250	500	250	0	250
no-change	no-change	moderate-increase	moderate-increase	moderate-increase	no		0	0	0	-2000	0
moderate-increase	no-change	moderate-increase	moderate-increase	no-change	no	Considering the curr	0	0	0	0	0
moderate-increase	moderate-increase	significant-increase	moderate-increase	significant-increase	no	Public facilities requ	0	250	0	250	-250
no-change	no-opinion	no-opinion	no-opinion	no-change	no		0	0	0	0	0
no-change	no-change	no-change	no-change	no-change	no		-250	500	500	-250	-500

Summarizing qualitative clusters with trigrams and LLMs

Trigrams were a useful way to understand the intent behind each of the 3 qualitative clusters.

The trigram discrepancies to the right were remedied by removing custom stopwords via a my_custom_stopwords function.



```
revenue_0': ['property taxes already', 'increase property taxes',
'increasing property taxes',
                'taxes already high', 'raising property taxes']
```

Two types of LLM summaries: Abstractive and Extractive Text Summarization

Extractive summarizer with tokenized text:

```
Cluster 0 Summary: ems transportation fees increase
golf courses use gallons of water per day per course
golfers tend financially better able pay increased golf
fees animal parks rec except golf courses . use of
animal adoption fees increase significantly. eg parks
recreation aquatic fees seems reasonable increase fire
permit inspection fees less one time things people
stalls small businesses . emergency medical transport
fee healthcare costs already unmanageable.
```

Extractive summarizer with un-tokenized text:

```
Cluster 0 Summary: the city of houston has a budget of
400 MILLION DOLLARS a year . the city should not
increase life support fees . animal adoption fees
should be lower . a zoning increase would make it
harder for people to afford health care .
```

Abstractive (gpt-4o) summarizer:

```
1. Financial Accessibility: Many emphasize the importance of
keeping these fees affordable for low-income and working-class
families to ensure equitable access to essential services and
recreational activities.

2. Impact on Low-Income Families: There is a significant
concern that increased fees would disproportionately impact
lower-income and minority communities, potentially restricting
their access to important resources and activities.

3. Balancing Costs: While some activities like golf are seen as
luxury and could bear fee increases, essential services and
amenities such as pools, parks, and EMS should remain low-cost
or free to support community well-being and inclusion.

4. Social Equity: Many commenters advocate for fee structures
that do not exacerbate existing financial burdens on
disadvantaged populations and suggest alternative funding
through progressive taxation or reallocating existing
resources.
```

	Cluster 0	Cluster 1	Cluster 2
Revenue	Bigrams: [“black lives”, “lives matter”, “recent events”, “george floyd”, “mike ramos”] Trigrams: [“black lives matter”, “lives matter movement”, “lives matter protests”]	Bigrams: [“would like”, “law enforcement”, “want see”, “like see”, “needs defunded”] Trigrams: [“would like see”, “improve quality life”, “would better spent”]	Bigrams: [“voice heard”, “make sure”, “tax dollars”, “voice opinion”, “property taxes”] Trigrams: [“make voice heard”, “tax dollars spent”, “want voice heard”]
Expenditure	Bigrams: [“low income”, “quality life”, “help people”, “parks rec”, “neighborhood development”] Trigrams: [“improve quality life”, “low income communities”, “actually help people”, “black brown communities”, “low income neighborhoods”]	Bigrams: [“tear gas”, “rubber bullets”, “military grade”, “de escalation”, “people color”] Trigrams: [“black lives matter”, “military grade equipment”, “tear gas rubber”, “gas rubber bullets”, “military grade weapons”]	Bigrams: [“would like”, “like see”, “neighborhood development”, “emergency medical”, “social workers”] Trigrams: [“would like see”, “cut cut cut”, “would better spent”, “improve quality life”, “reduce need policing”]
Survey	Bigrams: [“property taxes”, “low income”, “increase fees”, “lower income”, “emergency medical”] Trigrams: [“property taxes already”, “low income families”, “emergency medical transport”, “low income people”, “taxes already high”]	Bigrams: [“property taxes”, “increase fees”, “low income”] Trigrams: [“property taxes already”, “increase property taxes”, “low income families”, “raising property taxes”]	Bigrams: [“property taxes”, “cost living”, “increasing fees”, “cops cops”, “people color”] Trigrams: [“cops cops cops”, “increase property taxes”, “rubber bullets tear”, “bullets tear gas”, “serve black people”]

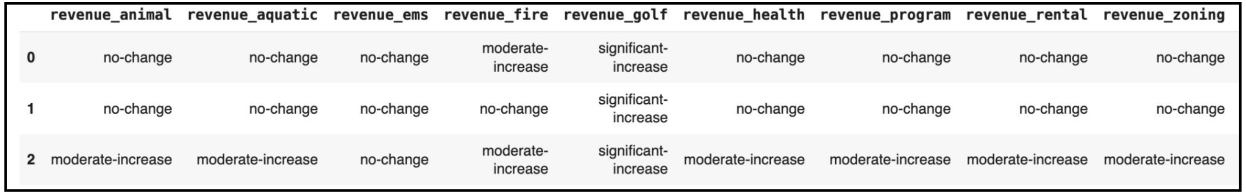
Correlating quantitative and qualitative clusters

Definition: 1:1 vs 3:3 clusters

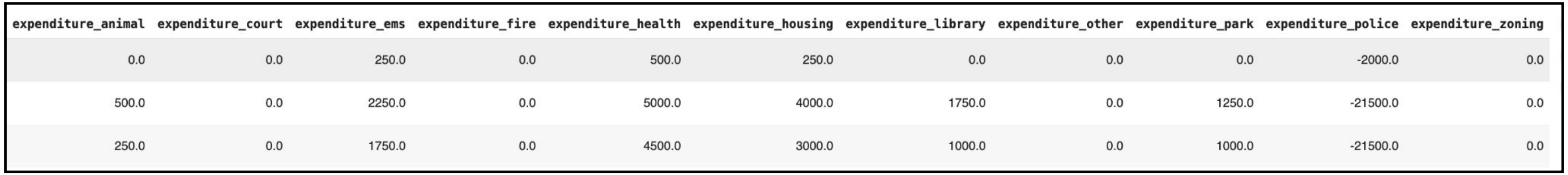
Steps:

1. Filtered the qualitative dataset to remove rows with more than one NaN prior to applying the cluster function. My cluster balancing results were: cluster 0: 4448, cluster 1: 2664, cluster 2: 5729. I performed sanity checks on this clustering to ensure that the documents in each cluster were valid.
2. The *clean_responses.csv* dataset contains all qualitative and quantitative responses. Quantitative clusters were re-generated after filtering the dataset the same way as Step 1.

Quantitative Cluster Summaries



	revenue_animal	revenue_aquatic	revenue_ems	revenue_fire	revenue_golf	revenue_health	revenue_program	revenue_rental	revenue_zoning
0	no-change	no-change	no-change	moderate-increase	significant-increase	no-change	no-change	no-change	no-change
1	no-change	no-change	no-change	no-change	significant-increase	no-change	no-change	no-change	no-change
2	moderate-increase	moderate-increase	no-change	moderate-increase	significant-increase	moderate-increase	moderate-increase	moderate-increase	moderate-increase



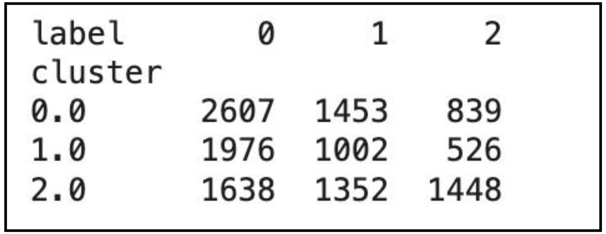
expenditure_animal	expenditure_court	expenditure_ems	expenditure_fire	expenditure_health	expenditure_housing	expenditure_library	expenditure_other	expenditure_park	expenditure_police	expenditure_zoning
0.0	0.0	250.0	0.0	500.0	250.0	0.0	0.0	0.0	-2000.0	0.0
500.0	0.0	2250.0	0.0	5000.0	4000.0	1750.0	0.0	1250.0	-21500.0	0.0
250.0	0.0	1750.0	0.0	4500.0	3000.0	1000.0	0.0	1000.0	-21500.0	0.0

Correlating quantitative and qualitative clusters

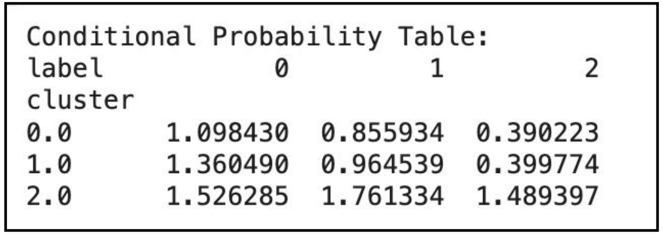
Definitions: 1:1 vs 3:3 clusters, contingency tables, and conditional probability tables

$$\text{conditional}[c][l] = \frac{\text{contingency\_table}[c][l] \text{ x } \text{total\_sum}}{\text{row\_sum}(l) \text{ x } \text{col\_sum}(c)}$$

1:1 Clusters

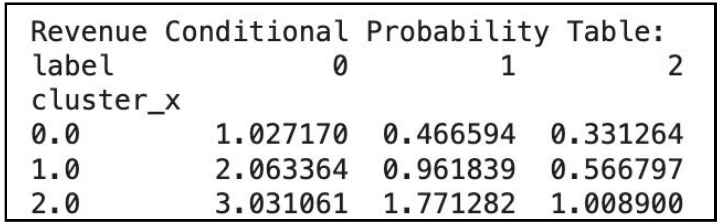


```
label      0     1     2
cluster
0.0     2607  1453   839
1.0     1976  1002   526
2.0     1638  1352  1448
```

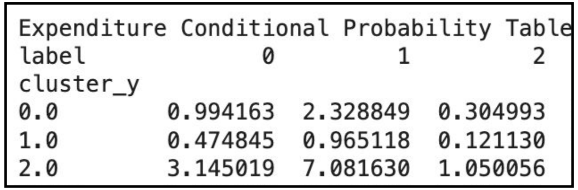


```
Conditional Probability Table:
label           0          1          2
cluster
0.0      1.098430   0.855934   0.390223
1.0      1.360490   0.964539   0.399774
2.0      1.526285   1.761334   1.489397
```

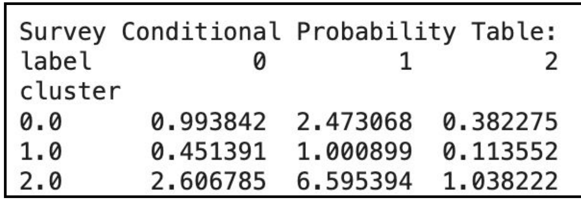
3:3 Clusters



```
Revenue Conditional Probability Table:
label              0          1          2
cluster_x
0.0         1.027170   0.466594   0.331264
1.0         2.063364   0.961839   0.566797
2.0         3.031061   1.771282   1.008900
```



```
Expenditure Conditional Probability Table
label              0          1          2
cluster_y
0.0         0.994163   2.328849   0.304993
1.0         0.474845   0.965118   0.121130
2.0         3.145019   7.081630   1.050056
```

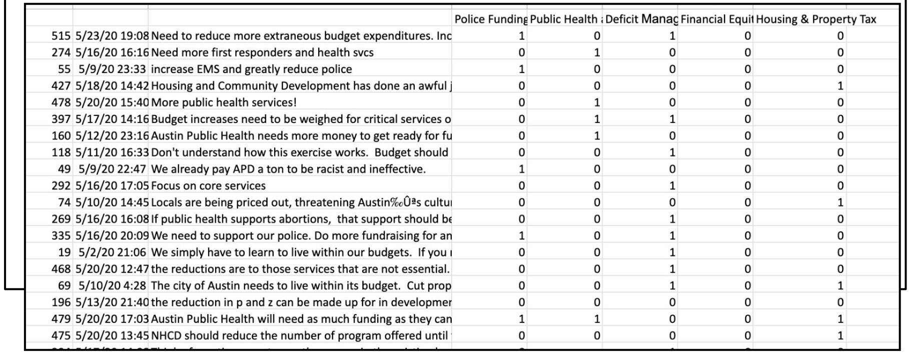


```
Survey Conditional Probability Table:
label              0          1          2
cluster
0.0         0.993842   2.473068   0.382275
1.0         0.451391   1.000899   0.113552
2.0         2.606785   6.595394   1.038222
```

Using ML models on hand-labeled datasets and to evaluate correlation

Hand-labeled dataset:

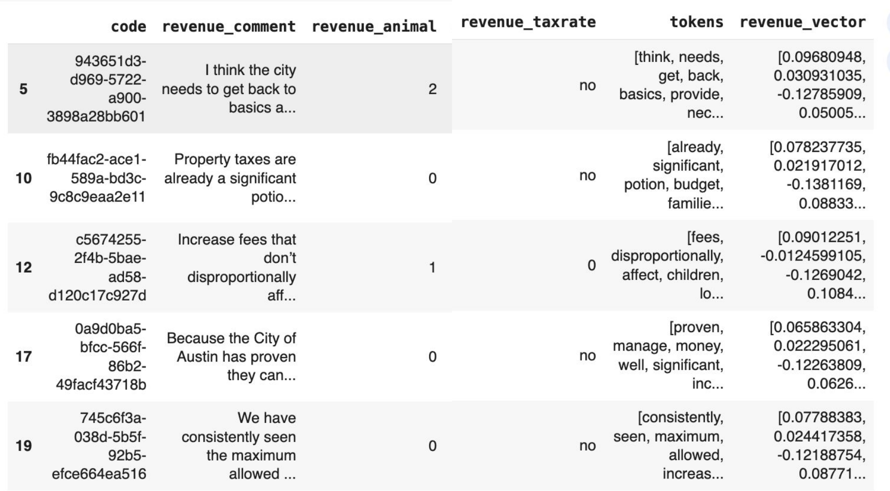
- Trained a MultiLabel Classifier to predict labels for each response
- Initial model: Hamming loss of 1.38 per row
    - Random would be 1.4167
- Added cosine-similarity features, saw false negative rates of 1 for labels 3 and 4 (false positive = 0)
- Added co-occurrence features, Hamming loss of 0.21 x 5 = 1.05



			Police Funding	Public Health :	Deficit Manag	Financial Equit	Housing & Property Tax
515	5/23/20 19:08	Need to reduce more extraneous budget expenditures. Inc	1	0	1	0	0
274	5/16/20 16:16	Need more first responders and health svcs	0	1	0	0	0
55	5/9/20 23:33	increase EMS and greatly reduce police	1	0	0	0	0
427	5/18/20 14:42	Housing and Community Development has done an awful j	0	0	0	0	1
478	5/20/20 15:40	More public health services!	0	1	0	0	0
397	5/17/20 14:16	Budget increases need to be weighed for critical services o	0	1	1	0	0
160	5/12/20 23:16	Austin Public Health needs more money to get ready for fu	0	1	0	0	0
118	5/11/20 16:33	Don't understand how this exercise works. Budget should	0	0	1	0	0
49	5/9/20 22:47	We already pay APD a ton to be racist and ineffective.	1	0	0	0	0
292	5/16/20 17:05	Focus on core services	0	0	1	0	0
74	5/10/20 14:45	Locals are being priced out, threatening Austin‰Ûªs cultu	0	0	0	0	1
269	5/16/20 16:08	If public health supports abortions, that support should be	0	0	1	0	0
335	5/16/20 20:09	We need to support our police. Do more fundraising for an	1	0	1	0	0
19	5/2/20 21:06	We simply have to learn to live within our budgets. If you	0	0	1	0	0
468	5/20/20 12:47	the reductions are to those services that are not essential.	0	0	1	0	0
69	5/10/20 4:28	The city of Austin needs to live within its budget. Cut prop	0	0	1	0	1
196	5/13/20 21:40	the reduction in p and z can be made up for in developmer	0	0	0	0	0
479	5/20/20 17:03	Austin Public Health will need as much funding as they can	1	1	0	0	1
475	5/20/20 13:45	NHCD should reduce the number of program offered until	0	0	0	0	1

ML-model on revenue:

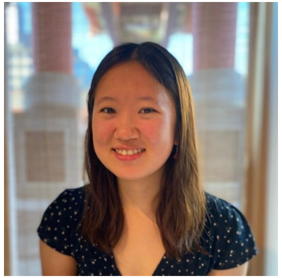
- Predicting *revenue_fire* and *revenue_zoning* numerical categories from revenue_vector
- Used np.vstack to convert revenue_vector into features for RandomForestClassifier.
- Achieved accuracy of 0.48 and 0.46 respectively



	code	revenue_comment	revenue_animal	revenue_taxrate	tokens	revenue_vector
5	943651d3-d969-5722-a900-3898a28bb601	I think the city needs to get back to basics a...	2	no	[think, needs, get, back, basics, provide, nec...	[0.09680948, 0.030931035, -0.12785909, 0.05005...
10	fb44fac2-ace1-589a-bd3c-9c8c9eaa2e11	Property taxes are already a significant potio...	0	no	[already, significant, potion, budget, familie...	[0.078237735, 0.021917012, -0.1381169, 0.08833...
12	c5674255-2f4b-5bae-ad58-d120c17c927d	Increase fees that don't disproportionally aff...	1	0	[fees, disproportionally, affect, children, lo...	[0.09012251, -0.0124599105, -0.1269042, 0.1084...
17	0a9d0ba5-bfcc-566f-86b2-49facf43718b	Because the City of Austin has proven they can...	0	no	[proven, manage, money, well, significant, inc...	[0.065863304, 0.022295061, -0.12263809, 0.0626...
19	745c6f3a-038d-5b5f-92b5-efce664ea516	We have consistently seen the maximum allowed ...	0	no	[consistently, seen, maximum, allowed, increas...	[0.07788383, 0.024417358, -0.12188754, 0.08771...

Cynthia Chen

B.S. Computer Science, minor in Management Science & Engineering, Stanford University '25



Related papers:

1. Text Summarization using LLMs: https://arxiv.org/pdf/2310.10449
2. Human-Interpretable Clustering: https://arxiv.org/html/2405.07278v2

Results/Limitations:

- Trigrams (and bigrams) and removing custom stopwords create meaningful cluster summaries, along with abstractive LLM analysis. However, the qualitative data may be intrinsically un-clusterable
- Quantitative responses are meaningful
- Austin 2020 survey qual and quant clusters are inherently uncorrelated, and verified via ML models. Certain features like co-occurrence and cosine similarity improved loss.
- Various preprocessing methods are needed to ensure validity of clustering
- Size of hand-labeled dataset was only 300

Future Work:

- Creating **expenditure** and **survey** category ML models and predictors
- Understanding 2021 and 2022 data
- Explore alternative clustering methods (GMMs), fine-tuning LLMs using hand-labeled dataset
- Applications to other types of short-form surveys like Long Beach, identifying traits of data that provide information about clusterability prior to performing any analyses