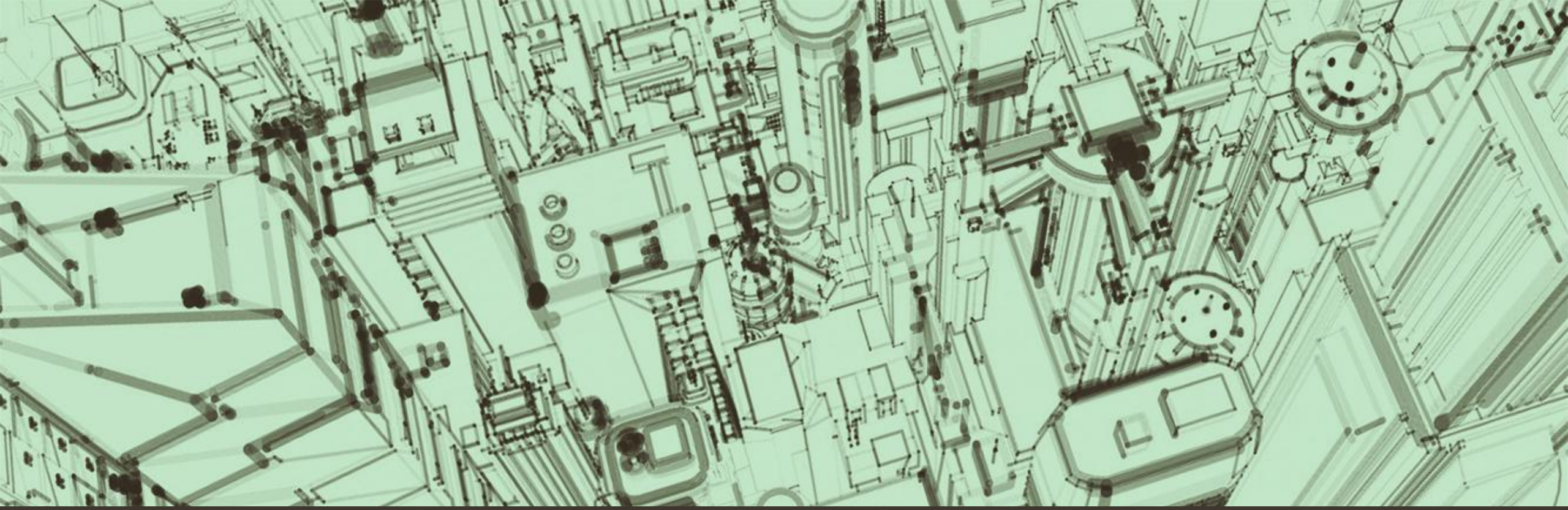


Chicago/Illinois News NLP Analysis

MSCA 32018 – NLP – Spring 2022

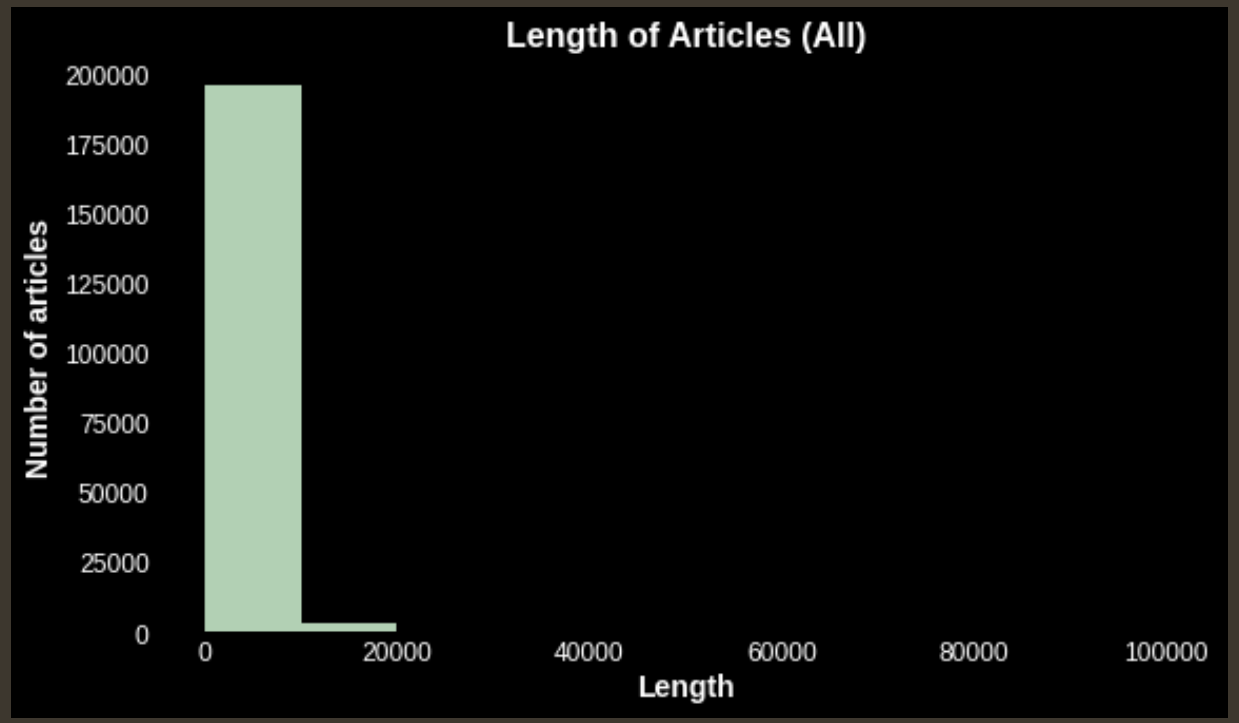
Bhadri Vaidhyanathan

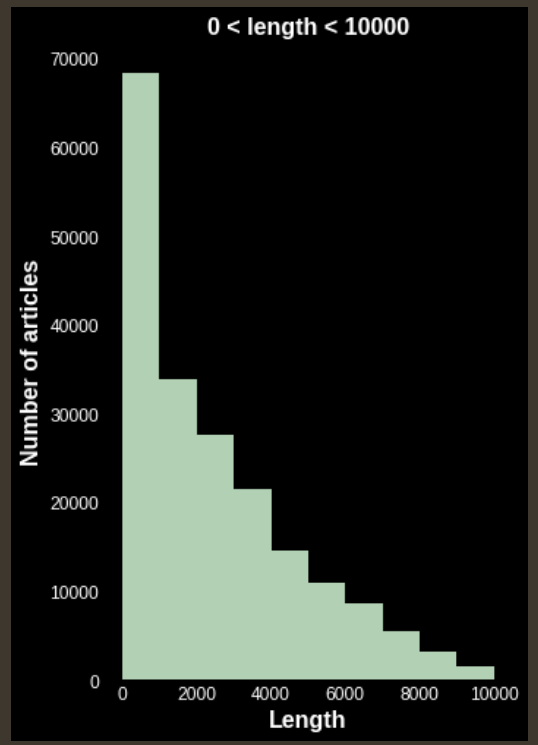
Contents

1. EDA
2. Filtering
3. Clustering and Topic Modeling
4. Sentiment Analysis
5. WordCloud
6. Entity Recognition
7. Targeted Entity Sentiment – Using Google Cloud NL API

EDA

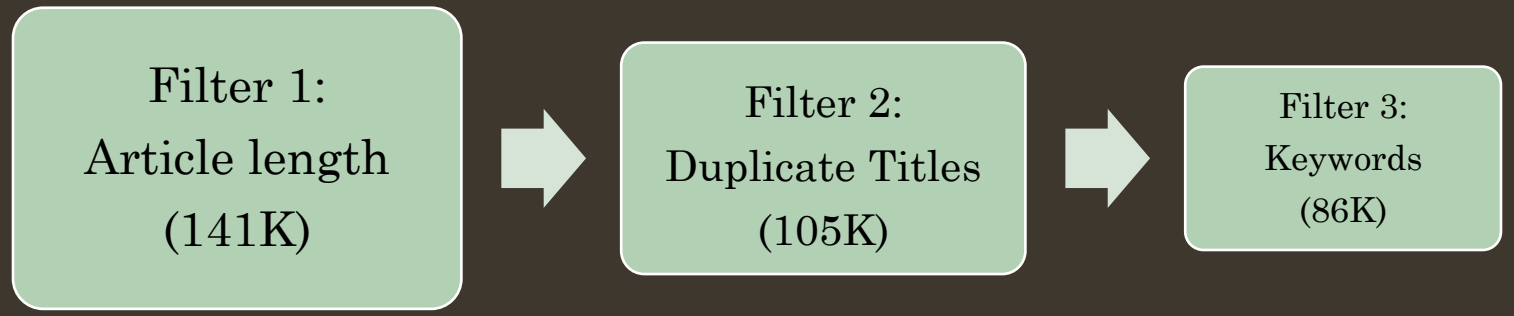
- Total number of documents = 200,119
- Average number of tokens/document = 537





Article Filtering

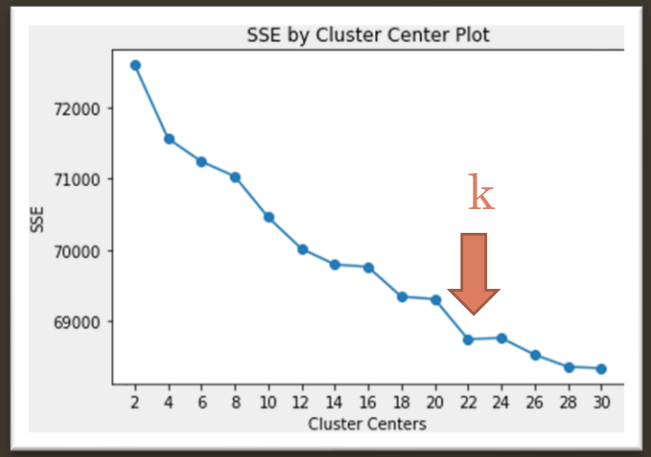
- Preprocessing:
  - Lowercase, punctuations, numbers, whitespace
  - Lemmatization not performed*
- Filtering:
  - Articles less than 750 characters dropped
  - Duplicate titles dropped
  - Keywords identified and articles filtered with keyword search
    - Keyword themes: sports terms, Chicago TV shows, popular spam terms



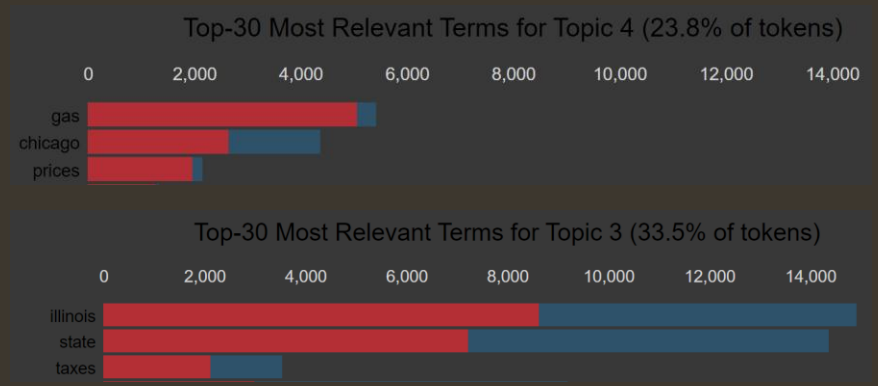
* Schofield, Alexandra et al. "Pre-Processing for Latent Dirichlet Allocation." (2017)

Clustering and Topic Modeling

Step 1: Deciding number of clusters



Step 2: Identify topics from TFIDF uni/bigrams and LDA / PyLDAvis Topic Modeling



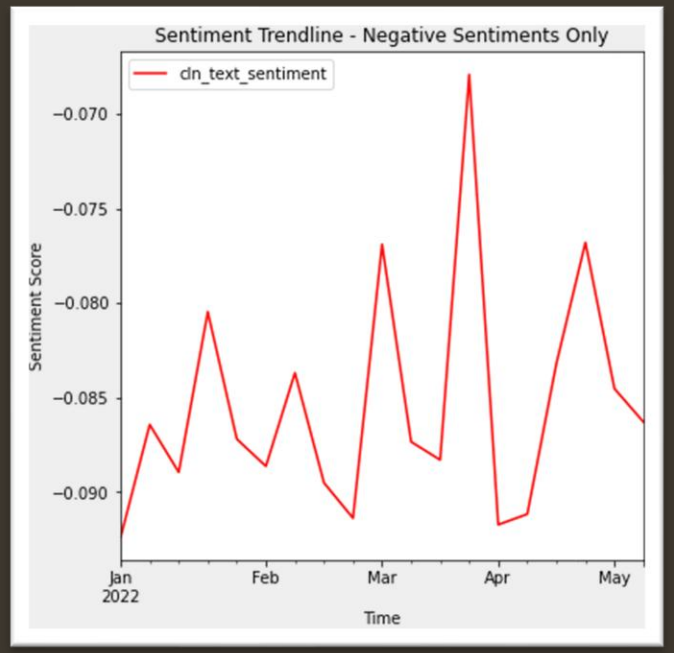
Topic Identified: gas price / state tax

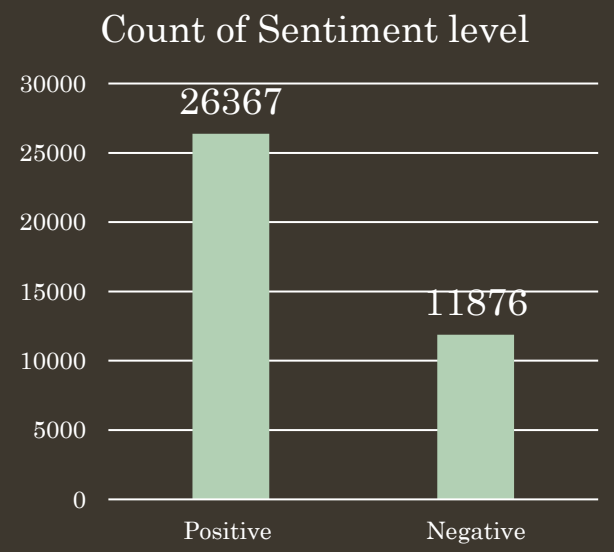
clusters	cluster_topic	count
18	**Crime / Policing**	7839
20	**Crime**	1706
22	**Crime / Violence**	2695
0	**About Chicago/IL**	5500
4	**Chicago/IL: Tax and Gas prices**	2056
12	**Corruption**	1135
5	Chicago/IL but less confidence	7048
6	Chicago/IL but less confidence	26617
19	Chicago/IL but less confidence	10984
26	Chicago/IL but less confidence	7684
1	Sports	512
2	headline and junk	765
3	Job posting	1686
7	Sports	4253
8	Covid	1466
9	Job posting - Medicine field	549
10	Ad block warning	401
11	Chicago weather	1840
13	junk	2470
14	Covid	2140
15	junk	1123
16	Investment firm	3880
17	Tv Show	475
21	junk	393
23	Ads	641
24	Schools & covid	2117
25	ukraine - russia war	2115

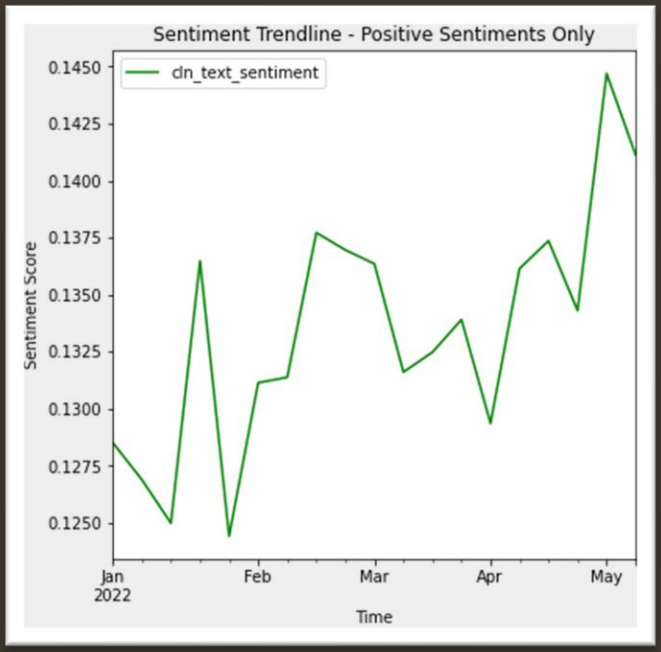
- Hybrid Model of using TfidfVectorizer clustering and LDA/PyLDAvis used to identify relevant and irrelevant topics
- The topic for some big clusters was not apparent and so above steps were repeated for that cluster alone to further identify relevant and irrelevant topics
- Using these methods, highly relevant articles were identified and are highlighted in green in the adjacent table

Sentiment Analysis

- Algorithm used:
  - TextBlob (PatternAnalyser)
- Negative sentiments have not changed much while Positive sentiments does have an upper slope
- The scores are regardless close to zero indicating neutrality as observed in other models as shown in later slides





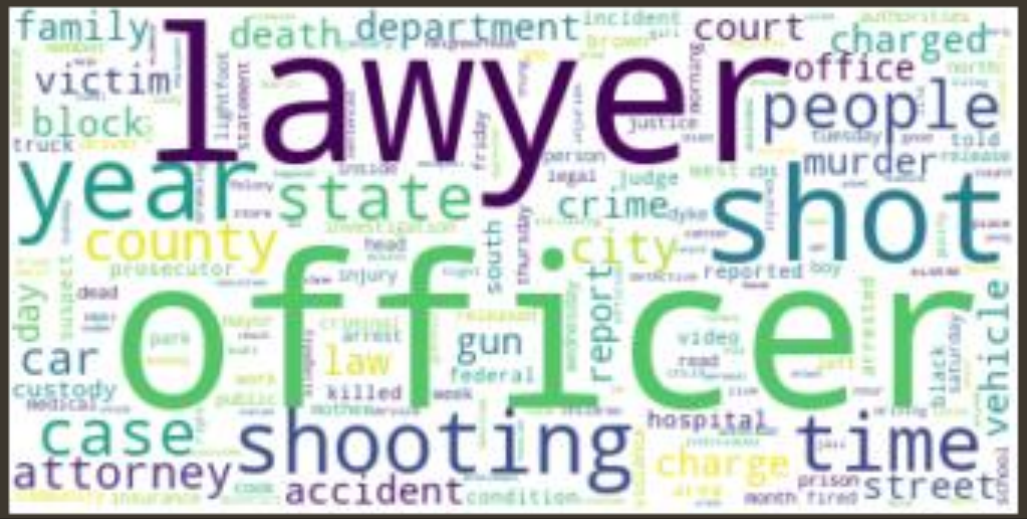


Wordclouds

Key articles on Tax and Gas price cluster

Key articles on population decline

Key articles on crime



Entity Recognition

- Spacy NER
  - Medium pipeline was used
- Top NERs were related ‘Person’ for the topics
- Very few organizations showed up in the list although TFIDF Trigrams should many organizations

cluster_topic	NER (People & ORG)	Identification
Crime / Policing	('mata', 112), ('draheim', 33), ('gloria', 28), ('hobley', 27), ('karen', 22), ('mendoza', 20),	criminals, victims, common names
Crime	('russell knight', 16), ('larry rogers', 14), ('rogers', 9), ('joseph power', 6), ('rahul iyer', 6)	Popular Lawyers/firms in Chicago
Crime / Violence	('evans', 11), ('balogh', 6), ('cunningham', 6), ('beltran', 6), ('gyovanny arzuaga', 6)	criminals, victims, common names
About Chicago/IL	('madeleine albright', 43), ('chris oberheim', 34), ('kunstler', 32), ('abbie', 30), ('thompsons', 29), ('chris oberheim maranatha', 29), ('american legislative exchange council', 29), ('clay jackson herald', 28),	Politicians, Police Officer, Activist
Chicago/IL: Tax and Gas prices	('scorp', 20), ('epa', 18), ('intuit', 12), ('irs', 9), ('garnishment', 8), ('avg gas price', 8),	Tax Org
corruption	('madigan', 20), ('michael', 15), ('thompson', 15), ('mike frerichs', 14) ('irs', 9), ('ameren', 9)	Politicians

Targeted Entity Sentiment – Google API

- Google Cloud NL API provides “Entity sentiment analysis” and can also be used to target few entities
- Entities, “Chicago” and “Illinois” were selected for analysis
- Substantial proportion are ‘Neutral”
- Further evaluation needed for Targeted Entity Sentiment Analysis with specific articles

Interpreting Entity Sentiment Results

Sentiment	Sample Values
Clearly Positive*	"score": 0.8, "magnitude": 3.0
Clearly Negative*	"score": -0.6, "magnitude": 4.0
Neutral	"score": 0.1, "magnitude": 0.0
Mixed	"score": 0.0, "magnitude": 4.0

Sentiment Interpretation	Illinois	Chicago
Mixed	4.89%	6.50%
Neutral	95.11%	93.50%

https://cloud.google.com/natural-language/docs/sentiment-tutorial
https://codelabs.developers.google.com/codelabs/cloud-natural-language-python3
https://cloud.google.com/natural-language/docs/basics#interpreting_sentiment_analysis_values

Recommendations

General Goals:

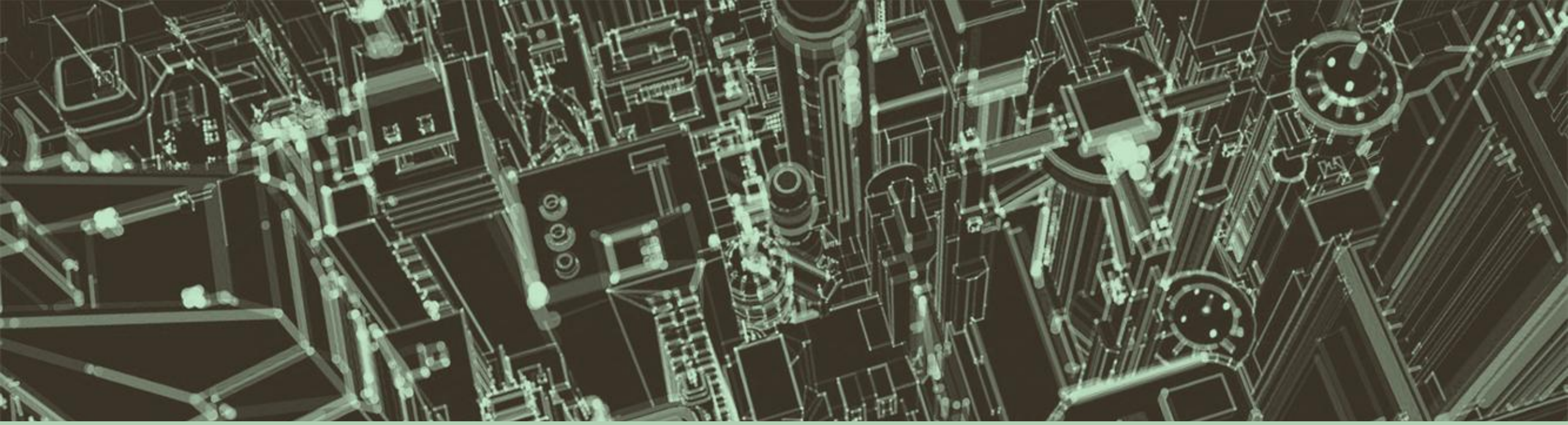
- Reduce corruption

For General Population

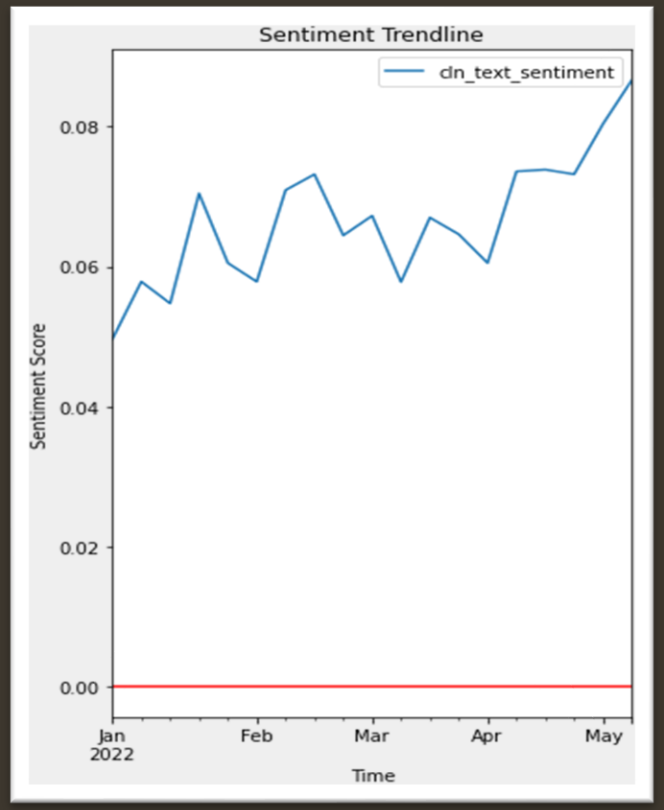
- Reduce crime
- Police should work to regain the trust of the people
- Reduce cost of living (lower tax burden, lower gas price, property tax)

For Business

- Reduce property tax to encourage business owners



Thank you!


Sentiment Trendline
dn_text_sentiment
0.08
0.06
0.04
0.02
0.00
Sentiment Score
Jan
2022
Feb
Mar
Apr
May
Time

Appendix