Examining Illinois Population Decline Through Text Analysis

MScA 32018: Natural Language Processing and Cognitive Computing
Final Project
June 4, 2021

By: Rohit Satishchandra

Executive Summary

PROBLEM:

- According to latest 2020 Census estimates, **Illinois lost more people than any state** in the 2010s
  - Rural population loss in other states is typically offset by urban growth
  - However, **Chicago's growth rate itself is ranked 46th out of 50** in the nation, leading only former Rust Belt cities like Cleveland and Pittsburgh
- **Declining population is a major problem** for the state because it means a **reduced tax base** and **lower state revenues**, which in turn result in **fewer resources** to meet fixed (or rising) state costs

PROPOSED SOLUTION:

- Natural Language Processing (NLP) techniques were used on a large sample of recent news articles to identify root causes for population decline: high crime in Chicago, lack of economic opportunity outside of state, unfavorable taxes
- The following were identified as possible countermeasures:
  - Restore jobs downstate by investing in renewable energy and green technology in rural Illinois
  - Because it is frequently mentioned with positive sentiment, a graduated income tax is worth considering as a means to reduce economic inequality
  - Subsidize arts and culture scene in Chicago because these are major selling points for the city and a reason why people stay

Source Data Overview

400K

News Articles in Original Dataset

90K

News Articles After Filtering

Methodology

- Select relevant articles by applying keyword extraction to titles and text
- Tune LDA model to identify prevalent topics within filtered set of news articles
- For each topic, select top articles and perform sentiment analysis at both the individual sentence and whole article level
- Leverage text summarization before and after sentiment analysis to synthesize article content and identify reasons for population decline
- Apply named entity recognition (NER) to article titles and a sample of article texts and perform targeted sentiment analysis

Article Filtering

- Rapid Automatic Keyword Extraction (RAKE) algorithm was applied to both article titles and text in order to identify important words and phrases
    - Special characters (including new line ‘\n’) were removed from article titles and text

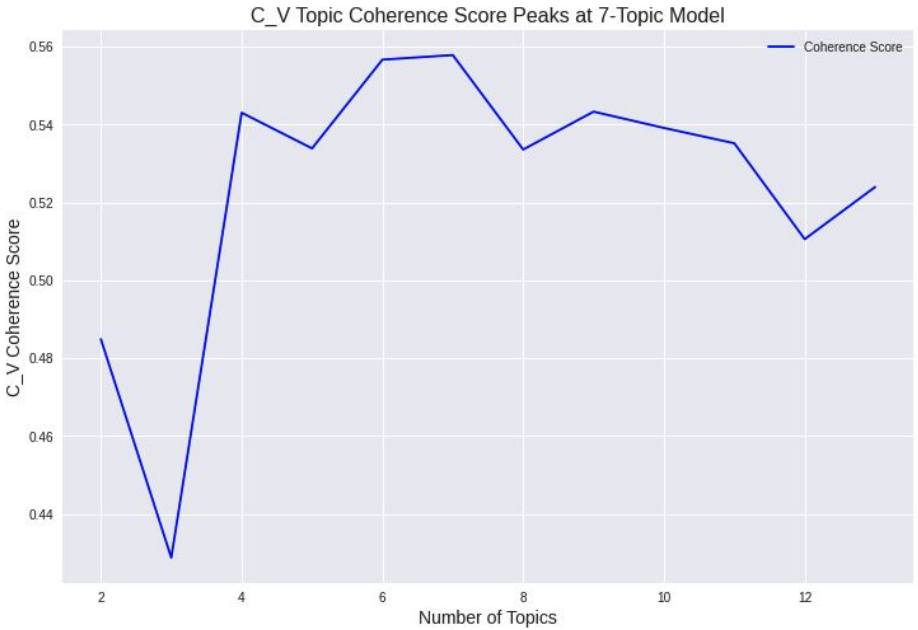
- Although the goal was to identify reasons for population decline, a broad set of keywords representing possible social, political, and economic factors were used as filters
    - Examples: *population*, *exodus*, *business*, *economy, economic growth, jobs, unemployment, tax, housing, education, crime, restaurants, culture*

- This was done to account for possible *latent* reasons for leaving or staying in Illinois and/or Chicago

Topic Modeling

Latent Dirichlet Allocation (LDA) Was Implemented ; 7-Topic Model was selected based on coherence score

- Article text was prepared via removal of stopwords, punctuation, and special characters
    - Remaining text was **lemmatized** using NLTK Lemmatizer
- N-topic **LDA topic models**, ranging from N=2 to N= 13 were fit to cleaned article text using gensim
- **CV Topic Coherence** peaked for 6 and 7 Topic Model
- The **most relevant topic for each article** was identified with confidence score
- 6-topic model was also trained using **KTrain** module for comparison
    - Identified topics also related to crime, covid pandemic, and business, but also leisure and 2020 election



Topic Modeling (cont.)

Topics were inferred from collection of important terms

Top Words From Each LDA Topic

“share”, “company”, “stock”, “illinois”, “quarter”, “inc” → **Illinois Company Financials**

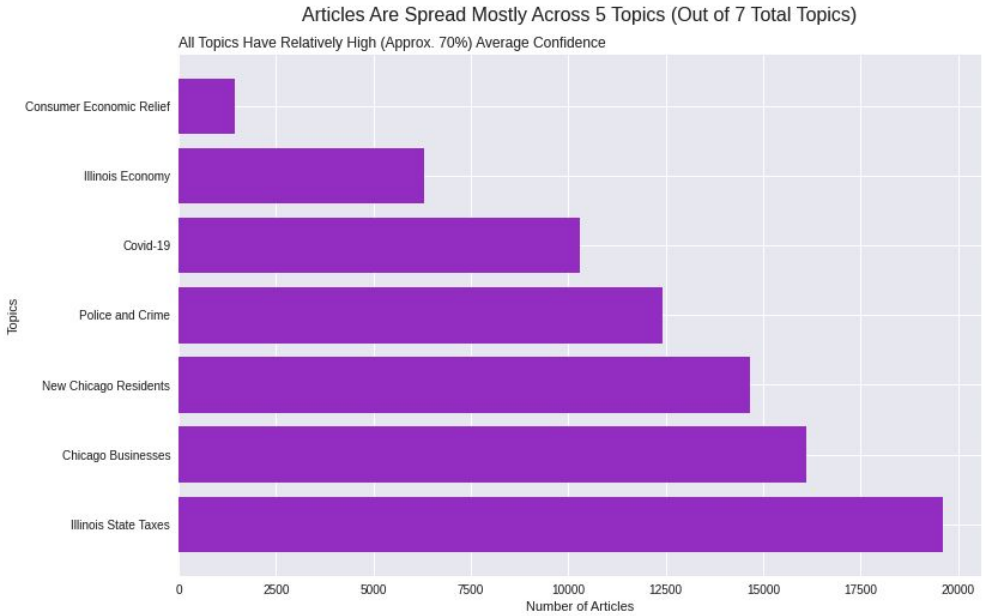
“illinois” “covid19” “state” “health” “case” “coronavirus” “chicago” “people” → **Covid-19**

“police”, “chicago”, “officer”, “city”, “shot” “crime” “shooting” → **Police & Crime**

“chicago”, “year”, “people”, “first”, “new”, “city” → **New Chicago Residents**

“chicago”, “business”, “company”, “service” → **Chicago Businesses**

“illinois”, “state”, “tax”, “year”, “school”, “million” → **Illinois Taxes/Government**



Topic Modeling (cont.)

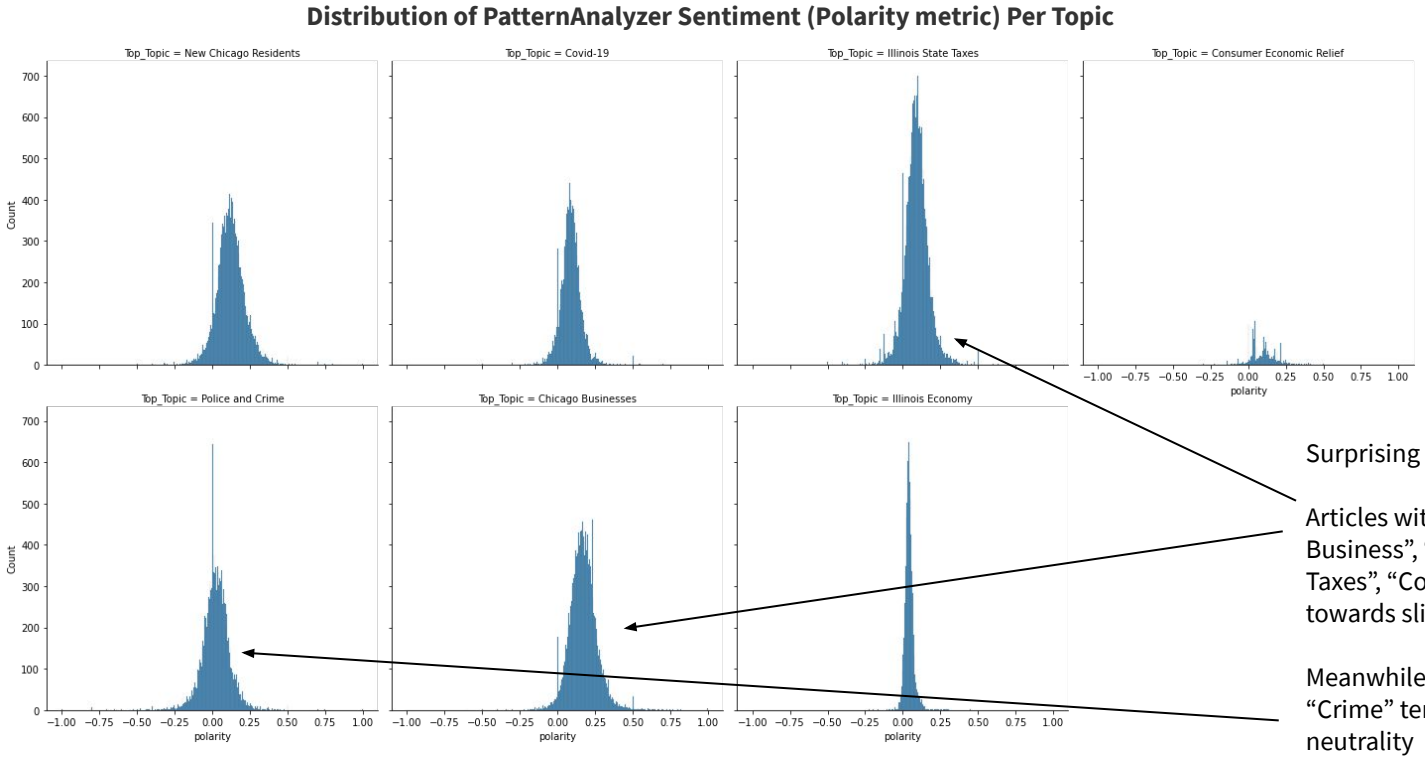
Topic confidence scores (from 0 to 1) were fairly high for most topics, with the exception of “consumer relief” and “new chicago residents”

Articles with “Consumer Economic Relief” as top topic had *low confidence*

Meanwhile, articles with “Illinois Economy” as top topic were predominantly *high confidence*

Sentiment Analysis

TextBlob PatternAnalyzer was applied to to both whole articles and sentences from a sample of high confidence articles



Surprising results:

Articles with “Chicago Business”, “Illinois State Taxes”, “Covid-19” skew towards slightly positive

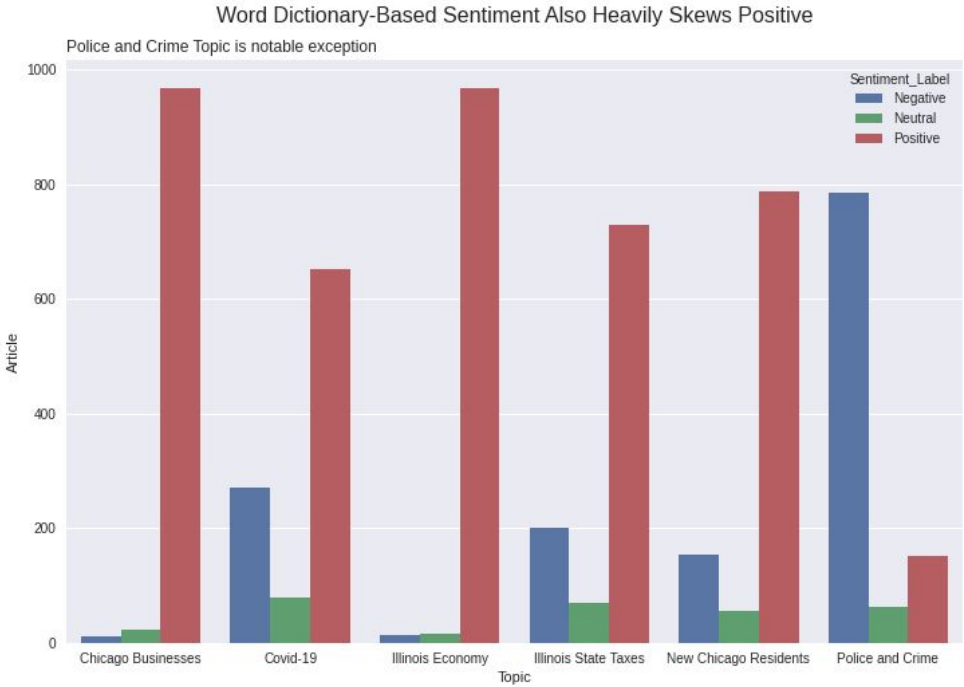
Meanwhile, articles with “Crime” tend towards neutrality

Sentiment Analysis (Tuned)

Positive/Negative Word Dictionary-Based Sentiment Analysis

- This approach computed **sentiment based on the presence of positive and negative words** (from a lexicon developed by researchers at University of Pittsburgh)

- The **algorithm was tuned** by adding a small sample of words specific to our context:
  - Positive: opportunity, restoration, restore
  - Negative: deficit, joblessness, homeless, declining, violent



Sentiment Analysis

Selected statements from positive and negative articles in relevant topics:

Crime

“Chicago has rampant awful gun crime”

“Violence is unfortunately ingrained in the culture out here”

“Residents, businesses now looking to leave Magnificent Mile due to looting”

“Neighbors Worry About Rise In Violent Crime In Ravenswood”

“I’m not a fan of big government and tend to want states and counties to handle their own business but the violence Chicago has gone on for too long and swept under the rug”

“Community members speak out after hate crime”

Government/Economy

“Open doors for more minority suppliers to do business with us, and build the clean energy future in our service territory”

“Facing steep budget deficits and escalating income inequality, Mayor Lori Lightfoot should consider a graduated payroll tax and other ways to make corporations pay their fair share.”

“Expand consumer affordability protections”

“Graduated income tax is good for Illinois”

“Thousands temporarily laid off at Chicago Assembly Plant”

“Exelon Generation announced last year that the two-unit Byron and Dresden nuclear power plants will both be retired in 2021, citing unfavourable market rules in the PJM capacity auction.

“Cook, Madison and St. Clair counties just put Illinois on the map for being one of the worst Judicial Hellholes in the nation again”

Culture

“When the sun is shining, the lakefront is thriving and patios are still open for all of us to take advantage of, Chicago is moving and grooving”

“Street food remains a vital part of the culture”

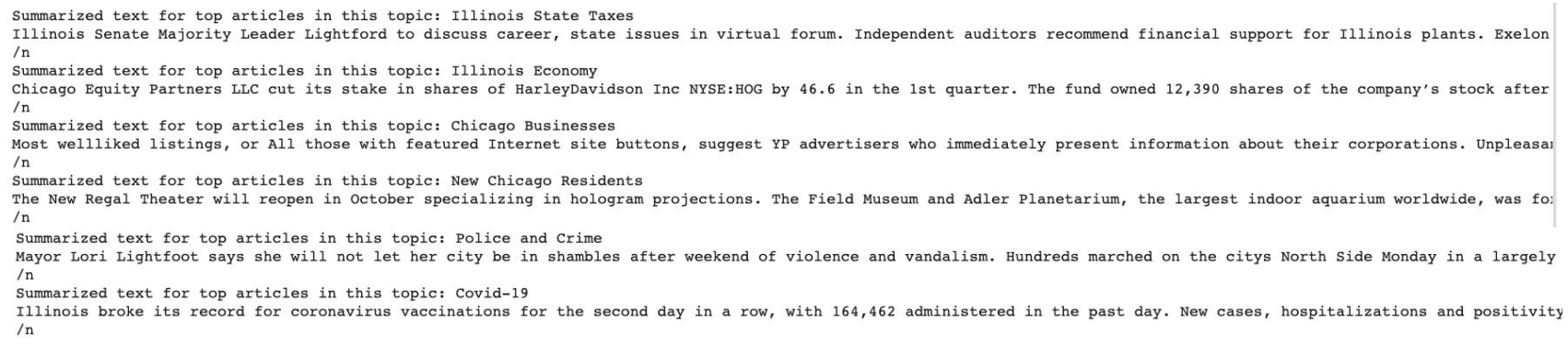
“Enjoy beautiful riverwalk views”

“Chicago may be the city that is beautiful the shores of Lake Michigan. It really is famous as a us social hub”

Text Summarization

Text summaries from ktrain hint at possible revision of topics, but high variation makes interpretation difficult (repeated runs result in very different summaries)

- For example, based on summarization, the following topics could be more broadly interpreted since results cover a wide range:
  - Illinois State Taxes → Illinois Government, Policy, and Elections
  - New Chicago Residents → Chicago Culture (sports, food, music)
- Chicago Business topics contains a lot of advertisements
- Illinois Economy features many articles reporting financial performance of public companies



Example Text Summarization For Articles Using KTrain

NER and Targeted Sentiment

Notable Entities (Person or Organization)	Number of Mentions
Lori Lightfoot	1180
J.B. Pritzker	692
Michael Madigan (IL State Representative)	110
Kim Foxx (State's Attorney for Cook County)	57
Illinois Tool Works Inc.	4058
Illinois Municipal Retirement Fund	1007

0.08

Mean polarity for articles referencing Governor Pritzker (more or less neutral)

0.04

Mean polarity for articles referencing Mayor Lori Lightfoot (more or less neutral)

-0.06

Mean polarity for "Crime" articles referencing Mayor Lightfoot

Possible Future Work

Keyword Extraction

Expand the working corpus of text by including more keywords in the filter list. This will account for other latent factors explaining population decline.

Possible words to add are “infrastructure”, “corruption”, “weather”, “property cost”

Topic Modeling

LDA was implemented using gensim and ktrain modules.

Other topic coherence scores could be explored, such as ‘umass’.

Another approach altogether would be to try “Zero Shot” in NLI module.

Sentiment Analysis

TextBlob PatternAnalyzer (polarity and sentiment) and word dictionary-based approach were used.

Zero-shot custom classifier could be used.

Alternatively, label articles with snorkel.

Text Summary

Experiment with the use of embeddings like Word2Vec and GloVE

Named Entity Recognition

Out-of-the-box spaCy models perform quite well at recognizing individuals and organizations.

Not a high area of prioritization for future work.

Targeted sentiment can be done with IBM Watson API

Conclusion

- Illinois and Chicago have a lot of challenges to think about as they begin to bounce back from the Covid-19 pandemic
- High crime in Chicago, budget deficit, lack of employment opportunities downstate, income inequality, and corruption are major drivers of negative sentiment
- Our recommendation is to reform tax policy, invest in talent development, take advantage of remote work culture to and transition from traditional industry (coal) to renewable and nuclear energy
    - The priority should be to keep residents in rural Illinois rather than continue to consolidate Chicago's economic dominance of the state
- On the other hand, Chicago has a vibrant culture with popular food, arts, and other cultural pursuits that should continue to receive support from city government