Combating Negative Population Growth in Illinois -
Text Analysis on News Articles

Natural Language Processing & Cognitive Computing
Targoon Siripanichpong

Executive Summary & Methodology

Executive Summary

- **Objectives** - Identify reasons for IL population decline & suggest potential actions to the state of Illinois to reduce decline
- **Dataset** - 30 days of news articles related to Illinois were collected and filtered using multiple methods to eliminate irrelevant articles
- **Analysis & Techniques**
  - **Topic Modeling** - Provides a high-level overview of different topics of the news articles
  - **Sentiment Analysis** - Implemented with a custom lexicon. Analyzes sentiments for the relevant news articles to provide insights
  - **Named Entity Recognition -** Targets people and organizations for more specific sentiment analysis
- **Observations**
  - **Negative Sentiments** - taxation, law enforcement, crime/violence
  - **Positive Sentiments -** clean energy, healthcare, recreational policies
- **Recommendations**
  - Consider population growth when making decisions on law & regulation changes, especially around taxation, recreational policies, and healthcare.
  - Promote clean energy consumption and production (e.g., tax credit/deduction for business/residents, subsidies for businesses in the clean energy sector).

Methodology Overview

01	Text Preprocessing	- Text Tokenization - Stopword & Punctuation Removal (with custom dictionary) - Lemmatization & Singularization
02	Article Cleanup/Filtering & Topic Modeling	- Removing duplicate articles with SimHash - Filtering out articles that are irrelevant based on topics from Topic Modeling & human-defined topics - Re-performing Topic Modeling to obtain post-filtered topics
03	Sentiment Analysis	- Leveraging Vader Sentiment Analysis library to initiate original sentiment score articles related to Illinois residents & businesses - Customizing the Vader lexicon based on manual label inspections
04	Targeted Sentiment Analysis	- Identifying entities (people & businesses) using Named Entity Recognition - Predicting targeted sentiments on articles specific to people or businesses

Text Preprocessing - News Article Cleanup/Filtering

1 Text Preprocessing

Perform text tokenization, extended stopword removal, lemmatization, and singularization

2 Near-Duplicate Articles Removal

Use SimHash algorithm to filter out near-duplicate articles

3 Filtering via Topic Modeling

Perform LDA to cluster news into different topics. Identify topics that are irrelevant to analysis (e.g. Job Postings, Home Sales & Rentals). Filter out these topics.

4 Additional Article Filtering

Remove articles with very high/low word counts in the titles

5 Keyword-based Article Filtering

Remove articles related to sports, weather, TV shows & movies, job postings, sales & promotions, and one-time occurrence news. Keywords are being used in this instance, as topic modeling was not able to extract these topics explicitly.
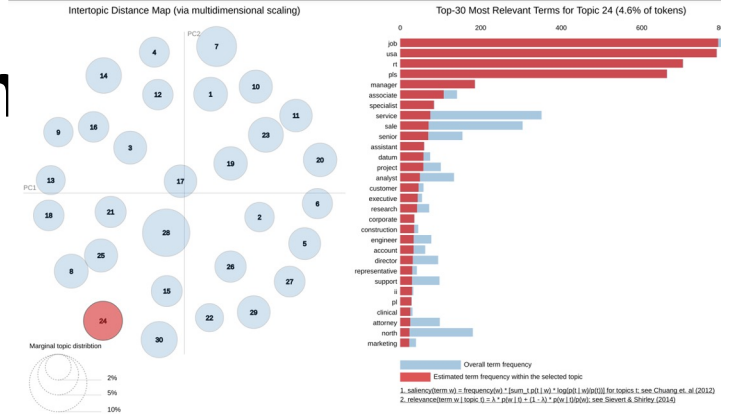
News Article Topic Detection

After multiple combinations and iterations of **Topic Modeling with LDA** and **manual keyword inspections** were performed, a few topics were identified, notably:

- Topic related to **sports**
- Topic related to **weather**
- Topic related to **entertainment** (Movie, TV shows, festivals, actor/actress, etc.)
- Topic related to **home sales/rentals, flight promotions, ads**
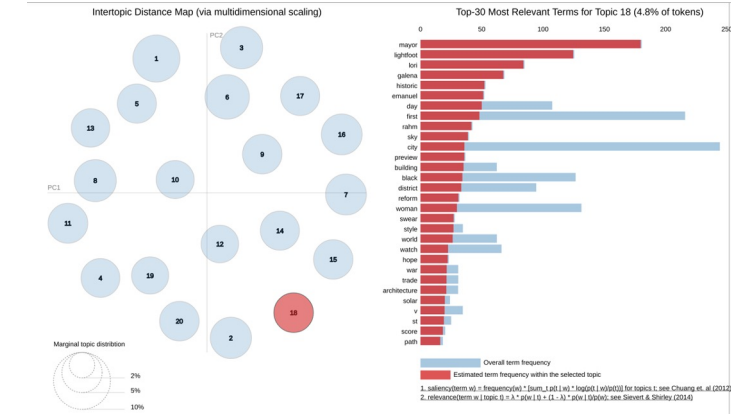- Topic related to **job postings & news related to specific companies**

The majority of these topics are not quite relevant to this analysis i.e. articles about TV shows, Chicago Bulls performance, or NYC-ORD flight promotion should not impact IL population growth. These topics are filtered out.

After filtering out topics that are considered irrelevant to the analysis, LDA was performed again. A clearer picture has been established with 20 topics post-filtering. These topics include:

- Chicago mayoral election
- Legislation on electric vehicle registration fee
- State taxes initiatives
- Marijuana legalization plan
- Police-related affairs

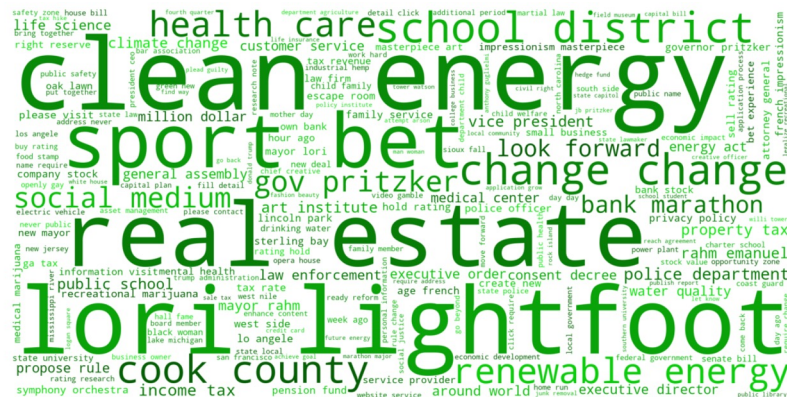Topic related to job postings extracted from LDA prior to filtering

Topic related to a new Chicago mayor election extracted from LDA after filtering
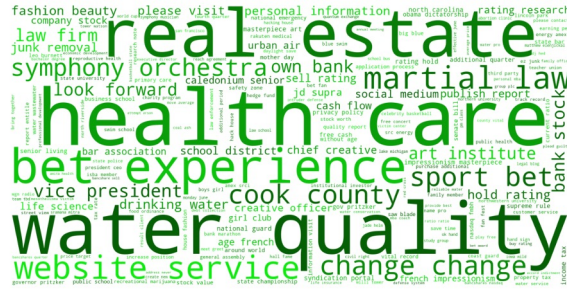
Sentiment Analysis

- **Custom sentiment lexicon** (based on **Vader**) is leveraged to produce sentiments associated with the filtered news articles
- Based on the sentiment analysis, the following inferences can be made regarding the status of IL population and businesses, notably:
  - **Potential reasons for population decline**
    - **Taxation** - Negative sentiment is observed for news articles related to **property taxes**, **income taxes**, **taxes on electric vehicles**, and **general tax increases**
    - **Law Enforcement** - Police & law enforcement are heavily associated with negative sentiment. While this could be biased due to the nature of news articles, it can be an indicator of the amount of crimes and violence in IL.
  - **Potential reasons for people & businesses to be attracted to IL**
    - **Clean Energy** - Positive sentiment is heavily associated with clean & renewable energy.
    - **Recreational Policies on sports betting** & **rec. marijuana** legalization make the list as well. Liberal recreational policies may potentially retain & attract more residents to IL.

Wordcloud for Overall Negative Sentiment

Wordcloud for Overall Positive Sentiment

Targeted Sentiment

Wordcloud for Business-focused Negative Sentiment

Wordcloud for Business-focused Positive Sentiment

Wordcloud for People-focused Negative Sentiment

Wordcloud for People-focused Positive Sentiment

- **Named Entity Recognition (NER) extraction** was performed using **Stanford NLP NER Tagger** to target articles specifically related to **persons & organizations.**
- Business/Organization
  - **Water Quality, Health Care,** and **Real Estate** have been heavily associated with positive sentiment.
  - Upon further inspection, it is found that businesses & orgs in IL are taking multiple initiatives toward water quality improvement, health care promotion, and public housing expansion.
  - Topics of **tax** and **law enforcement** are still receiving negative sentiments
- People
  - Unfortunately, not a lot of information can be derived from the sentiment analysis that would provide additional insight to IL population growth/decline.
  - It does however, recap the sentiment toward the Chicago mayoral election, in which Lori Lightfoot won the race.

Insights & Recommendations

Wordcloud for Overall Negative Sentiment

Wordcloud for Overall Positive Sentiment

Why should businesses stay in IL or move to IL?
- Take advantage of upcoming laws & regulations
  - With the focus on clean energy production & recreational policies, there are multiple opportunities to expand business in the clean energy sector and the sports betting industry

Why should residents stay in IL or move to IL?
- Focus on quality of life improvement
  - Businesses & orgs in IL are taking multiple initiatives toward water quality improvement, health care promotion, and real estate/public housing expansion
- Liberal recreational policies

Recommendations to the State of Illinois
- Take **population growth & public sentiment** into account when introducing changes in laws & regulations, especially on **taxation, recreational policies, and healthcare**
  - Sentiments on these topics are very strong, which could lead to population growth/decline. **Tax hikes** are heavily associated with negative sentiment, while **recreational policies and healthcare** are associated with positive sentiment.
- Focus on Law Enforcement
  - Continue to focus on reducing crime rates and negative perceptions of law enforcement
- Continue to promote clean energy consumption and production
  - e.g. tax credit/deduction for business/residents, subsidies for business in the clean energy sector.