Analysis of Population Decline in Illinois

Using unsupervised text processing

Master of Science in Analytics | University of Chicago

Executive Summary

This study applies various text processing techniques on a corpus of approximately 28,000 news and other media articles related to the State of Illinois or the City of Chicago.

I assume that the frequency of discussion of themes deemed supportive or deleterious to population growth relates to the extent to which these themes drive population growth or decline.

I make no assumption as to the particular themes of interest to the City or the State; rather, I employ an unsupervised approach to grouping the entire corpus of data into topics and subtopics.

Methodology

•After applying several text cleaning techniques, including lemmatization and the removal of stop words, I employ an unsupervised topic modeling technique called “Latent Dirichlet Allocation” (LDA), which assumes that each news article is comprised of a limited number of topics with various probabilities, which in turn, are correlated with a set number of terms.

•I determine an appropriate number of topics in which to split the corpus using ‘coherence scores’, which evaluate the quality of topics by measuring the extent to which pairs of words within each topic co-occur within the corpus.

•I ultimately decide to split the original corpus into 5 topics, assign a topic label to each document, and then split each of these 5 topics into 5 additional topics.

Methodology (Continued)

- During each application of LDA, I remove those documents that are assigned to any given topic with less than 30 percent probability in an attempt to remove outliers.

- For each of the 25 subtopics, I convert the resulting sub-corpus of labeled documents into ngrams of length 5, calculate the frequency of each ngram, and remove all but one of those ngrams that have the same frequency, given that said ngrams frequently (though not always) refer to the same theme.

- Finally, I visualize the frequency of these ngrams using word clouds and calculate the percentage of the corpus that each subtopic represents.

Results and Recommendations

1. One of the city's greatest assets is its cultural identity.

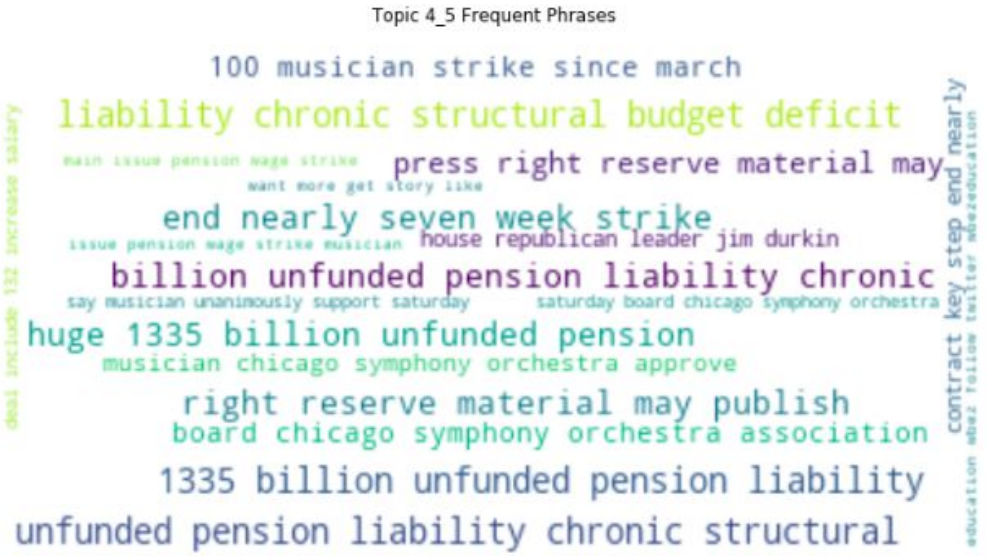
The City and the State should continue to market, support, and embrace its cultural identity, embodied in shows such as the Black Ink Crew as well as its art, music, festival, and hotel scenes.

That said, the City should be cognisant that its history of violence also shows up in this cultural identity, as discussed in the recently released, *American Summer: Love and Death in Chicago*.

Results and Recommendations

2. While violent crime is apparent in the news cycle, more attention is paid to the pension crisis and crimes of injustice (e.g. the Jussie Smollett case and the Wrigley Field racial gesture incident).

The City and the State should prioritize handling the pension crisis as a prerequisite to handling other issues and take the opportunity to tell success stories about when racial and other types of injustices are appropriately handled by local government and authorities.



Results and Recommendations

3. The legalization of marijuana is big news.

The City and the State should take advantage of the national attention to sell themselves as destinations for progressive, pro-business policies and for private investment in burgeoning industries.