Retention Analysis of Population in Illinois

NLP Project

Vritti Gandhi

June 4, 2022

Contents


Executive Summary


Data Cleaning


Detection of Topics


Reasons for Population Decline in Illinois – Sentiment Analysis


Corrective Actions and Sentiment Over Time


NER and Targeted Sentiment

Business Related Articles

Person Related Articles

Executive Summary

Illinois is one of the few states in the US with a declining population trend. Analyzing the news articles collected from different news sources, the below insights are generated:

- The main reasons for this declining trend seem to be crime, poverty, weather, increasing COVID cases, and commercial real estate not doing so well.

- On the other hand, the positives are good business, rehab center, dividends, education/school districts, sports teams, attorneys, work-life balance, and lots to explore in the city, especially restaurants.

Data Cleaning

We start with analyzing and removing duplicates in text. This leaves us with 187,910 articles.

We then look at articles with duplicate titles having same first 55 characters in text. These have some differences in the end of texts, but other than that are mostly similar. We drop these as well and get to 166,274 articles.

846 articles contain the word “ ads” and 55 contain “ css”. Inspecting those, we find them to be irrelevant and drop those.

Some articles also have a lot of Javascript content. We use regex to replace those starting with “var winW” and ending with “;” with space.

All this cleanup leaves us with a total of 165360 articles

We then use TextHero to create a custom pipeline for cleaning – filling missing values, removing whitespace, digits, HTML tags, URLs, punctuation, stopwords, and lowercasing the words.

Sample article containing the word “ ads”


```
'Thanks for visiting\nThe use of software that blocks ads hinders our ability to serve you the content you came here
to enjoy.\nWe ask that you consider turning off your ad blocker so we can deliver you the best experience possible wh
ile you are here.\nThank you for your support!'
```


Sample article containing the word “ css”


```
'/* custom css */ .tdi_54{ min-height: 0; } /* custom css */ .tdi_56, .tdi_56 .tdc-columns{ min-height: 0; }.tdi_56,
.tdi_56 .tdc-columns{ display: block; }.tdi_56 .tdc-columns{ width: 100%; }@media (min-width: 768px) { .tdi_56 { marg
in-left: -22px; margin-right: -22px; } .tdi_56 .tdc-row-video-background-error, .tdi_56 .vc_column { padding-left: 22
px; padding-right: 22px; } } /* landscape */ @media (min-width: 1019px) and (max-width: 1140px){ @media (min-width: 7
68px) { .tdi_56 { margin-left: -20px; margin-right: -20px; } .tdi_56 .tdc-row-video-background-error, .tdi_56 .vc_col
umn { padding-left: 20px; padding-right: 20px; } } } /* portrait */ @media (min-width: 768px) and (max-width: 1018px)
{ @media (min-width: 768px) { .tdi_56 { margin-left: -12px; margin-right: -12px; } .tdi_56 .tdc-row-video-background-
error, .tdi_56 .vc_column { padding-left: 12px; padding-right: 12px; } } } /* custom css */ .tdi_58{ vertical-align:
baseline; }.tdi_58 > .wpb_wrapper, .tdi_58 > .wpb_wrapper > .tdc-elements{ display: block; }.tdi_58 > .wpb_wrapper >
.tdc-elements{ width: 100%; }.tdi_58 > .wpb_wrapper > .vc_row_inner{ width: auto; }.tdi_58 > .wpb_wrapper{ width: aut
o; height: auto; } /* custom css */ .tdi_60{ position: relative !important; top: 0; transform: none; -webkit-transfor
m: none; }.tdi_60, .tdi_60 .tdc-inner-columns{ display: block; }.tdi_60 .tdc-inner-columns{ width: 100%; } /* inline
tdc_css att */ .tdi_60{ margin-right:0px !important; margin-bottom:30px !important; margin-left:0px !important; paddi
```

Topic Modeling – Using LDA and Zero Shot Learning (ZSL)

- We use LDA and coherence score (keeping alpha=asymmetric and beta=auto) to find the optimal number of topics – these turn out to be 10.
- When looking at 10 topics, we see some overlap. We therefore go with 9 topics. Below are the broad topics we find:
  - Company shares
  - Food and restaurants
  - Games
  - Weather and lawyers
  - Schools
  - Poverty
  - Rehab centers
  - Divorces
  - Business
- We also use Zero Shot Learning and choose 8 broad topics to classify each article into – business, crime, education, sports, food, divorce, drugs, attorney. We will use these later in our analysis.

```
[(0,
  '0.004*"product" + 0.004*"additional_shares_last_quarter" + '
  '0.003*"owns_shares_industrial_product" + '
  '0.003*"industrial_products_company_stock" + 0.003*"total_transaction" + '
  '0.003*"shares_illinois_tool_work" + 0.003*"share" + '
  '0.002*"ratio_debt_equity_ratio" + 0.002*"ex_dividend_date_dividend" + '
  '0.002*"represents_dividend_annualized_basis"'),
 (1,
  '0.008*"also" + 0.004*"restaurant" + 0.004*"roof_covere" + 0.004*"roof" + '
  '0.004*"well" + 0.004*"area" + 0.003*"home" + 0.003*"food" + 0.003*"city" + '
  '0.003*"park"'),
 (2,
  '0.010*"say" + 0.007*"go" + 0.006*"team" + 0.005*"time" + 0.005*"also" + '
  '0.005*"make" + 0.005*"year" + 0.005*"get" + 0.005*"game" + 0.005*"come"'),
 (3,
  '0.002*"time" + 0.002*"order" + 0.001*"family_lawyers_near" + 0.001*"and" + '
  '0.001*"section" + 0.001*"pay" + 0.001*"child" + '
  '0.001*"report_created_automatically_weather" + 0.001*"court" + '
  '0.001*"points_rebounds_assists_steal"'),
 (4,
  '0.029*"say" + 0.008*"state" + 0.006*"school" + 0.005*"city" + '
  '0.005*"people" + 0.004*"student" + 0.004*"report" + 0.004*"also" + '
  '0.004*"time" + 0.004*"case"'),
 (5,
  '0.005*"also" + 0.002*"percent_individuals_living_poverty" + 0.002*"well" + '
  '0.002*"propose" + 0.002*"poverty_line_total_individual" + 0.002*"suitable" '
  '+ 0.002*"advised_check_bookmaker_depende" + '
  '0.002*"funding_account_order_view" + 0.002*"bookmaker_live_stream_strongly" '
  '+ 0.001*"live"'),
 (6,
  '0.007*"fact" + 0.006*"alcohol_drug_rehab_center" + 0.005*"also" + '
  '0.005*"new_beginnings_alcohol_drug" + 0.004*"get" + '
  '0.004*"best_drug_rehab_center" + 0.004*"chicago_content_section" + '
  '0.003*"center" + 0.003*"work" + 0.003*"well"'),
 (7,
  '0.003*"news_search_news_emovie" + 0.003*"state" + '
  '0.002*"family_lawyers_near" + 0.002*"divorce" + '
  '0.002*"videos_videostagged_cbs_chicago" + 0.002*"chicago_videos_news_new" + '
  '0.002*"court" + 0.002*"lawyer" + 0.002*"emusic_ebooks_search_search" + '
  '0.002*"chicago_news_videos"'),
 (8,
  '0.008*"company" + 0.005*"use" + 0.005*"also" + 0.005*"work" + 0.005*"may" + '
  '0.004*"business" + 0.004*"include" + 0.004*"need" + 0.004*"service" + '
  '0.004*"get"')]
```

Sentiment Analysis – Using ZSL and BERT

- We tried out sentiment analysis trained on Yelp data, using Logistic Regression and Naïve Bayes as two classifiers. However, when looking at topics within them for positive and negative sentiments, our results were not very interpretable.
- We next tried using Zero Shot Learning to classify articles as positive or negative, using nli_template = 'the sentiment of this news article is {}', normalizing the relevance scores such that they summed up to 1.
- We then use the Hugging Face transformer – BERT for Sequence Classification with distilbert-base-uncased – to fine tune our model. The tokenizer we use is the AutoTokenizer from pretrained bert-base-uncased.
- Training the model for 5 epochs, we get accuracies of 61%, 60%, and 60% for train, validation, and test data respectively.
- Using LDA to look at top 7 topics within the predicted positive and negative sentiments, we see that they seem better.

Sentiment Analysis – Using ZSL and BERT

Positive Sentiment Topics

```
[(0,
  '0.006*"company" + 0.004*"may" + 0.004*"use" + 0.004*"also" + 0.004*"work" + '
  '0.004*"include" + 0.004*"need" + 0.004*"business" + 0.003*"get" + '
  '0.003*"time"'),
 (1,
  '0.008*"also" + 0.005*"fact" + 0.004*"well" + 0.003*"get" + '
  '0.003*"alcohol_drug_rehab_center" + 0.003*"work" + 0.003*"section" + '
  '0.003*"roof_covere" + 0.003*"new_beginnings_alcohol_drug" + '
  '0.002*"best_drug_rehab_center"'),
 (2,
  '0.004*"product" + 0.004*"additional_shares_last_quarter" + '
  '0.003*"owns_shares_industrial_product" + 0.003*"total_transaction" + '
  '0.003*"industrial_products_company_stock" + '
  '0.003*"shares_illinois_tool_work" + 0.002*"ratio_debt_equity_ratio" + '
  '0.002*"ex_dividend_date_dividend" + '
  '0.002*"represents_dividend_annualized_basis" + '
  '0.002*"stock_sold_average_price"'),
 (3,
  '0.006*"say" + 0.003*"state" + 0.002*"school" + 0.002*"work" + 0.001*"child" '
  '+ 0.001*"get" + 0.001*"day" + 0.001*"year" + 0.001*"time" + '
  '0.001*"student"'),
 (4,
  '0.011*"team" + 0.010*"game" + 0.008*"season" + 0.006*"play" + 0.006*"go" + '
  '0.005*"bear" + 0.005*"get" + 0.005*"player" + 0.005*"say" + 0.005*"bull"'),
 (5,
  '0.003*"also" + 0.003*"chicago_atlantic_real_estate" + 0.002*"say" + '
  '0.002*"well" + 0.001*"chicago_bankruptcy_law_firm" + 0.001*"attorney" + '
  '0.001*"bankruptcy" + 0.001*"time" + 0.001*"com" + '
  '0.001*"personal_bankruptcy"'),
 (6,
  '0.013*"say" + 0.005*"also" + 0.004*"make" + 0.004*"time" + 0.004*"go" + '
  '0.004*"work" + 0.004*"year" + 0.004*"see" + 0.004*"show" + 0.003*"get"')]
```

Good business, best rehab center, good dividends, good education, good sports teams, good attorneys, time to explore

Negative Sentiment Topics

```
[(0,
  '0.008*"say" + 0.006*"go" + 0.006*"team" + 0.005*"year" + 0.005*"game" + '
  '0.005*"time" + 0.004*"get" + 0.004*"season" + 0.004*"see" + 0.004*"show"'),
 (1,
  '0.030*"say" + 0.007*"school" + 0.005*"police" + 0.005*"case" + '
  '0.005*"report" + 0.005*"read" + 0.005*"city" + 0.004*"time" + 0.004*"state" '
  '+ 0.004*"student"'),
 (2,
  '0.012*"say" + 0.008*"state" + 0.004*"company" + 0.004*"include" + '
  '0.004*"use" + 0.003*"year" + 0.003*"people" + 0.003*"work" + 0.003*"new" + '
  '0.003*"service"'),
 (3,
  '0.003*"close_modal_suggest_correction" + 0.003*"news_post_world_new" + '
  '0.003*"usa_news_washington_celebrity" + '
  '0.002*"suggest_correction_file_source" + '
  '0.002*"points_rebounds_assists_steal" + 0.002*"say" + '
  '0.002*"field_opponent" + 0.002*"fully_vaccinated_fully_vaccinate" + '
  '0.002*"population_fully_vaccinated_fully" + 0.002*"game_shoote"'),
 (4,
  '0.006*"say" + 0.002*"time" + 0.002*"rights_reserve" + 0.002*"police" + '
  '0.002*"shoot" + 0.001*"officer" + 0.001*"windy_city_time" + '
  '0.001*"news_windy_city_times" + 0.001*"nightspots_chicago_glbt_nightlife" + '
  '0.001*"department"'),
 (5,
  '0.005*"say" + 0.003*"ap_college_basketball_coverage" + '
  '0.003*"rights_reserve" + 0.002*"report_created_automatically_weather" + '
  '0.002*"state" + 0.002*"data_stats_llc" + '
  '0.002*"generated_automated_insights_use" + 0.001*"data_provide" + '
  '0.001*"press" + 0.001*"rewritten_redistribute"'),
 (6,
  '0.003*"covid" + 0.003*"say" + 0.003*"test" + '
  '0.002*"real_estate_real_estate" + 0.002*"buyer_address_sale_price" + '
  '0.002*"school" + 0.002*"betterads_el_width_betterad" + '
  '0.002*"betterads_el_width_else" + 0.002*"show" + 0.001*"report"')]
```

Negative sentiment towards sports teams, crime and shooting, weather, increase in covid cases

Sentiment Analysis – Using ZSL and BERT

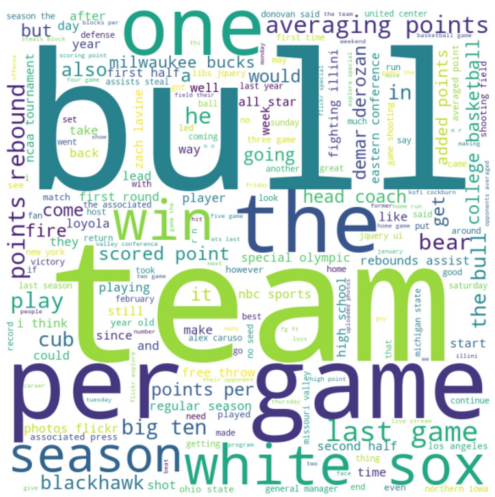
- We also use the topics that we generated using ZSL to look at the top reasons for negative sentiment in Illinois. Below is the distribution of articles across topics within negative sentiment articles. Crime, as expected, tops the list. We generate some wordclouds to get the most significant words within crime, business, and sports.

crime	33663
business	26064
sports	15812
attorney	12737
education	8342
food	4130
drugs	1573
divorce	564

Crime

Business

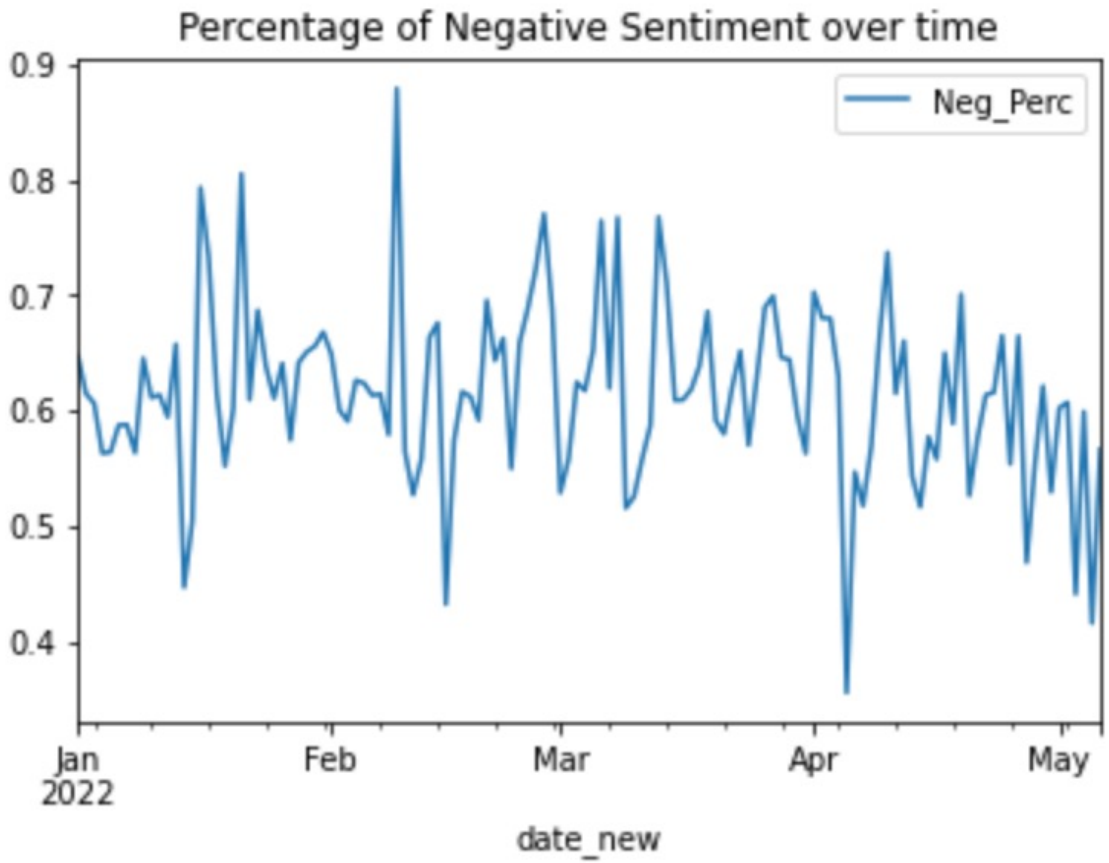
Sports



While there are some irrelevant words, we see that for crime, some words are south block, police, which indicates presence of high crime in south side Chicago. For Business, we see words like real estate, bedroom, bathroom, land size - which could mean that the expensive commercial real estate or its short supply could be another reason for the negative sentiment. Also, for games, apparently bulls did not have a good playoff season this year, which could be a reason for negative sentiment towards them.

Corrective Actions & Sentiment Over Time

- With this analysis, some corrective actions to reduce the negative sentiment could be:
  - Increasing police patrolling and working hard towards curbing crime
  - Reinforce COVID safety protocols as soon as cases start increasing
  - Work towards improvements in commercial real estate
- Looking at sentiment over time, while the trend is very volatile, overall it looks like as we edge towards May, the negative sentiment % is declining (even though it is still close to 50%). This could have something to do with good weather as we approach summer.
- The spike in negative sentiment in early Feb could have something to do with winter storm around Feb 3, or even with Bulls losing against Philadelphia and Phoenix.



NER and Targeted Sentiment – Business Related Articles

- We use spaCy's 'en_core_web_md' pipeline to extract the named entities from the text, and count the number of ORGs and PERSONs. Articles with higher number of ORGs are classified as business related articles, and those with higher number of PERSONs are classifier as person related articles. This brings the total number of business and person related articles to 54,102 and 76,503 respectively.
- We then apply LDA to positive business articles. The sentiments, since they are at an article level, remain the same.
- On the right are the top 5 topics for positive business articles. Broadly, they revolve around shares, teamwork.

```
[(0,
  '0.007*"also" + 0.004*"well" + 0.004*"say" + 0.003*"year" + '
  '0.002*"roof_covere" + 0.002*"time" + 0.002*"game" + 0.002*"team" + '
  '0.002*"use" + 0.002*"get"'),
 (1,
  '0.003*"product" + 0.002*"owns_shares_industrial_product" + 0.002*"go" + '
  '0.002*"year" + 0.002*"industrial_products_company_stock" + '
  '0.002*"shares_illinois_tool_work" + 0.002*"total_transaction" + '
  '0.002*"bull" + 0.002*"say" + 0.002*"additional_shares_last_quarter"'),
 (2,
  '0.005*"say" + 0.004*"get" + 0.004*"also" + 0.004*"team" + 0.004*"time" + '
  '0.003*"make" + 0.003*"go" + 0.003*"work" + 0.003*"year" + 0.003*"company"'),
 (3,
  '0.004*"company" + 0.004*"use" + 0.004*"say" + 0.003*"work" + 0.003*"also" + '
  '0.003*"time" + 0.003*"include" + 0.003*"home" + 0.003*"service" + '
  '0.003*"may"'),
 (4,
  '0.009*"say" + 0.004*"also" + 0.004*"show" + 0.003*"go" + 0.003*"work" + '
  '0.003*"time" + 0.003*"make" + 0.003*"see" + 0.003*"get" + 0.003*"year"')]
```

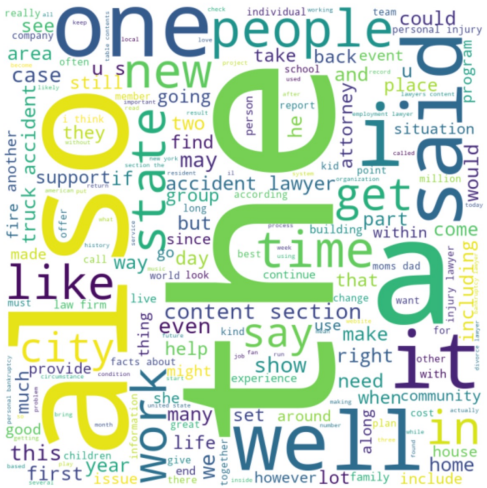
NER and Targeted Sentiment – Business Related Articles

- For a more comprehensive understanding, we look at the Zero Shot topics. Below is the distribution of articles across topics within positive sentiment business related articles. We generate some wordclouds to get the most significant words within business, attorney, and education topics.

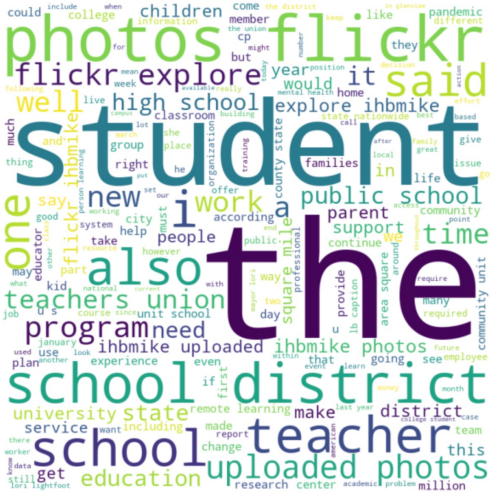
business	6799
sports	4234
attorney	3686
education	2011
crime	1734
food	959
drugs	689
divorce	447

Business

Attorney



Education



While there are some irrelevant words, we can infer that good financial performance of companies, good attorneys, and good school districts (that might help in recruiting well performing students) are some reasons for businesses to stay/move into Illinois.

NER and Targeted Sentiment – Person Related Articles

- We now apply LDA to positive person related articles.
- On the right are the top 5 topics for positive person related articles. Broadly, they revolve around good companies to work for, time/good work-life balance, good for lawyers, good sports events.

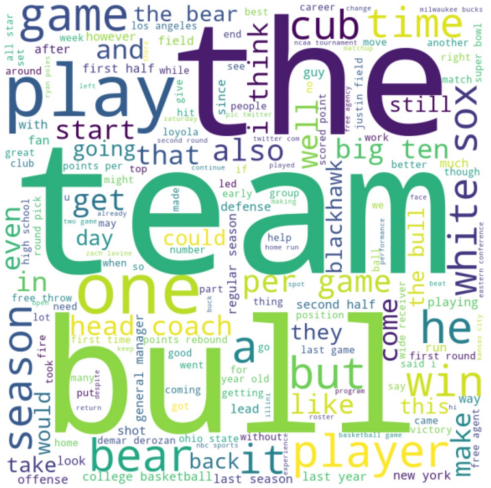
```
[(0,
  '0.006*"also" + 0.003*"well" + 0.003*"fact" + 0.002*"use" + 0.002*"get" + '
  '0.002*"product" + 0.002*"work" + 0.002*"company" + 0.002*"section" + '
  '0.002*"additional_shares_last_quarter"'),
 (1,
  '0.012*"say" + 0.004*"state" + 0.003*"year" + 0.003*"work" + 0.003*"also" + '
  '0.003*"time" + 0.003*"make" + 0.003*"include" + 0.003*"company" + '
  '0.003*"report"'),
 (2,
  '0.004*"also" + 0.004*"lawyer" + 0.003*"may" + 0.003*"use" + 0.003*"company" '
  '+ 0.003*"get" + 0.003*"work" + 0.003*"need" + 0.003*"include" + '
  '0.003*"well"'),
 (3,
  '0.005*"go" + 0.005*"team" + 0.005*"say" + 0.005*"get" + 0.004*"also" + '
  '0.004*"time" + 0.004*"make" + 0.004*"game" + 0.003*"year" + 0.003*"season"'),
 (4,
  '0.004*"also" + 0.003*"work" + 0.002*"team" + 0.002*"include" + 0.002*"well" '
  '+ 0.002*"time" + 0.002*"use" + 0.002*"say" + 0.002*"program" + '
  '0.002*"year"')]
```

NER and Targeted Sentiment – Person Related Articles

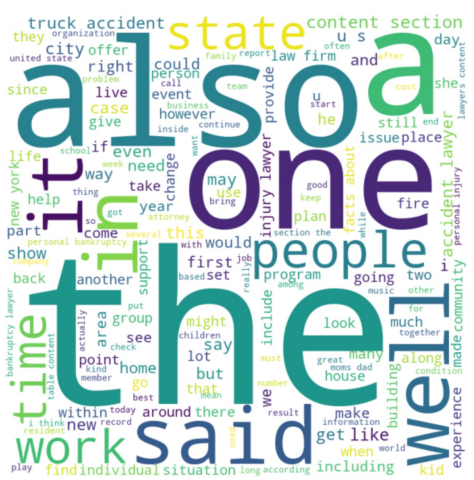
- For a more comprehensive understanding, we look at the Zero Shot topics. Below is the distribution of articles across topics within positive sentiment person related articles. We generate some.
- wordclouds to get the most significant words within sports, attorney, education, food, drugs, and divorce topics

business	9559
sports	6023
attorney	5449
education	2593
crime	2404
food	1324
drugs	946
divorce	574

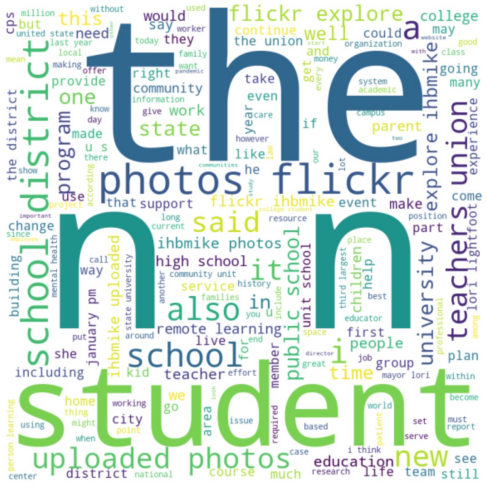
Sports



Attorney



Education

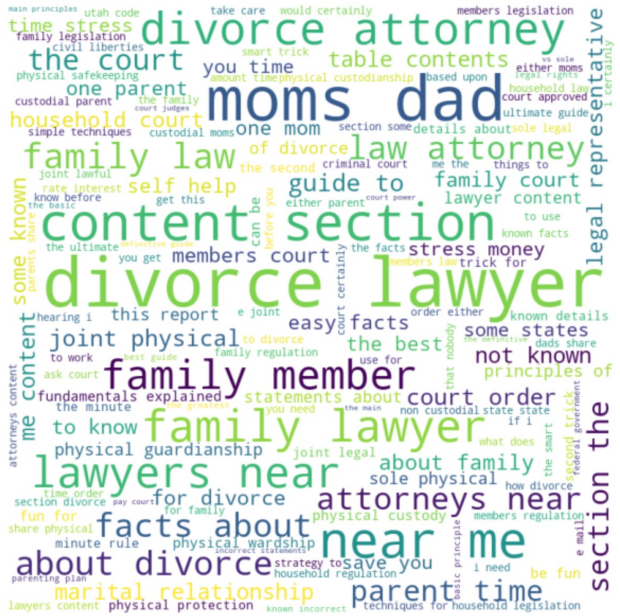


Food

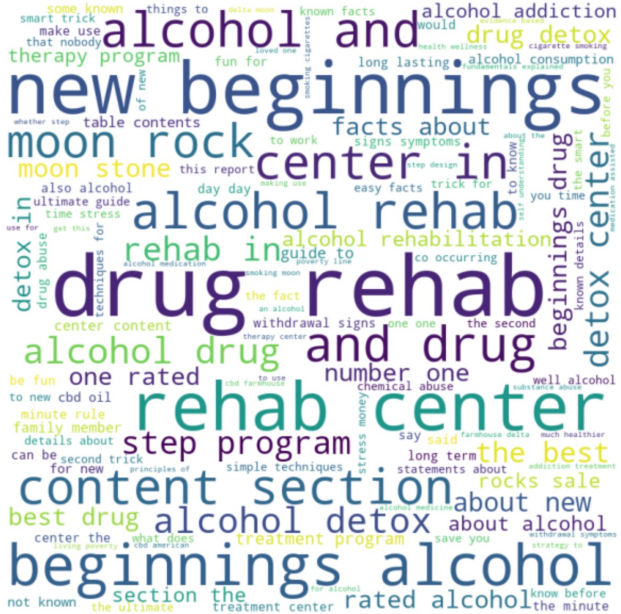
NER and Targeted Sentiment – Person Related Articles

- While there are some irrelevant words, we can infer that good sports teams, good attorneys, good school districts, Chicago food and restaurants scene, and good rehab centers are some reasons for residents to stay/move into Illinois.

Divorce



Drugs



Actionable Recommendations

- Some recommendations for corrective actions are:
  - Increasing police patrolling and working hard towards curbing crime
  - Reinforcing COVID safety protocols as soon as cases start increasing
  - Working towards improvements in commercial real estate

- Additionally, way to attract businesses would be highlighting good financial performance of companies, good attorneys, and good school districts.

- Way to attract residents would be good sports teams, good attorneys, good school districts, Chicago food and restaurants scene, and good rehab centers.

Thank You