NLP PROJECT

Migration trends for Chicago / Illinois

Shikhar Madrecha

EXECUTIVE SUMMARY

The problem at hand was to gain relevant insights from a corpus of around 200k articles published in the news between 1st January and 6th May about why business and people are moving out of Illinois and Chicago and provide relevant recommendations to counter these trends.

In an attempt to come up with good recommendations, the approach followed was to clean the text of these articles as thoroughly as possible and pass the cleaned text through sentiment Analyzers, Named Entity recognisers and LDA to come up with the topics being discussed in these articles.

The main reason for leaving Illinois, as can be found from these articles, has been the increase in shooting violence in the state along with greater number of drug cases. Other reasons also included teachers protesting for hybrid learning methods, extreme and harsh winter climate and insider trading in companies.

The recommendations provided by the following report mainly focuses on stricter rules regarding gun ownership, drug consumption and distribution and more state involvement in public education and the financial markets. The analysis here can be improved, neverthless, it does some interesting observations to be considered.

TEXT CLEAN UP AND EDA

Text Clean Up

The article text and title were cleaned up using the texthero library. The following steps were followed to clean up the text:

1. remove whitespaces
2. remove digits
3. remove punctuations
4. remove stop words

```
#cleaning the data using text hero
%%time
def clean(df, column_name):
  custom_pipeline = [hero.preprocessing.fillna,
                     hero.preprocessing.remove_whitespace,
                     hero.preprocessing.remove_digits,
                     hero.preprocessing.remove_punctuation,
                     hero.preprocessing.remove_stopwords]
  for i in column_name:
    new_col = i+'_clean'
    df[new_col] = hero.clean(df[i], custom_pipeline)

clean(data, ['title', 'text'])
```

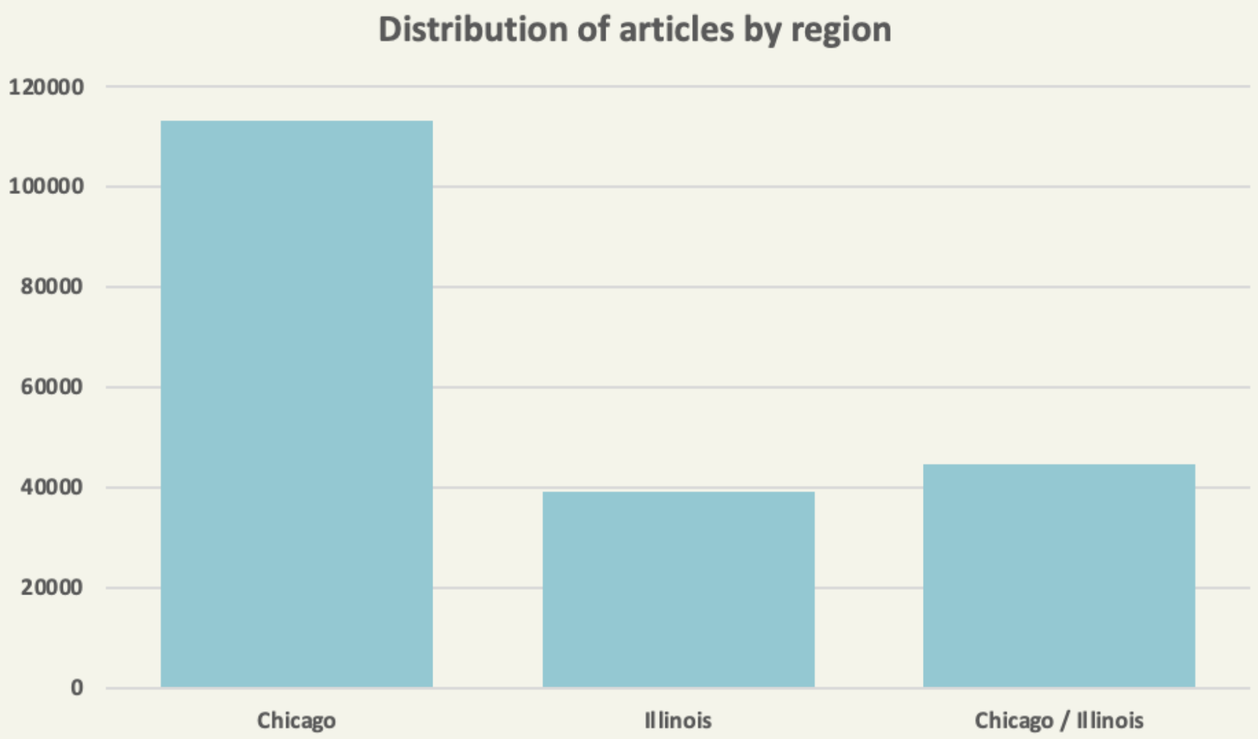
Sample text before clean up:

```
'Chicago PD 9×11 “Lies” Season 9 Episode 11 Promo — Voight employs his new informant to help solve a tricky drug trafficking case. At
water, who once again struggles to reconcile his personal and professional life, reaches a decision.\n@jesseleesoffer @marinasqu @trs
piridakos @NBCChicagoPD'
```

Sample text after clean up:

```
'Chicago PD    Lies Season    Episode    Promo  Voight employs  new informant  help solve  tricky drug trafficking case  Atwater     st
ruggles  reconcile  personal  professional life  reaches  decision   jesseleesoffer  marinasqu  trspiridakos  NBCChicagoPD'
```

Exploratory Data Analysis



Many of the articles in the corpus also talk about an ongoing TV show about Chicago. As these are not relevant for our analysis, we exclude these from our corpus.

Additionally, the articles are split between 3 categories - articles talking only about "Chicago", articles talking about "Illinois" and articles talking about "Chicago and Illinois". We see that majority of the article are related to only Chicago.

ANALYSIS PIPELINE

Analysis Pipeline

The below steps are followed to get to the insights and recommendations

OVERALL NEGATIVE SENTIMENT ANALYSIS

EXTRACTING TOPICS FROM ENTIRE CORPUS

LDA is used to extract 11-12 topics from the entire corpus of text. The topics retrieved can be classified into the following buckets:

1. Sports
2. Employment
3. Education
4. Police / Crimes
5. Weather
6. Lawyers
7. Rehab
8. businesses

NEG SENTIMENT ANALYSIS ON SUB-TOPICS

The dataset is broken down into subsets based on the topics identified in the first step.
Using Sentiment Analysis, negative articles are filtered for each subtopic and LDA is run on these articles to extract the reasons for leaving Chicago / Illinois

TARGETED POSITIVE SENTIMENT ANALYSIS

NER ON ENTIRE CORPUS OF ARTICLES

For targeted sentiment Analysis, we identify the various entities in each body of article using Spacy's NER library.
This helps us identify different entities such as organisations, people, location, etc so as to filter the dataset and find articles only relating to business or people.

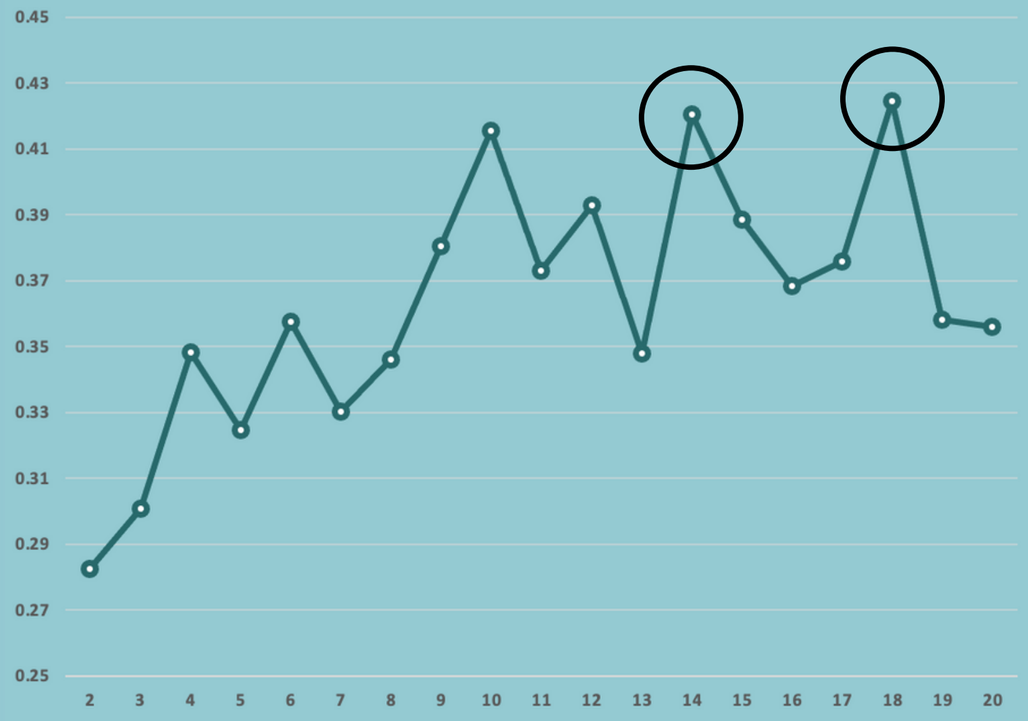
POSITIVE SENTIMENT ANALYSIS

Once articles are filtered for only businesses and people, we filter for articles with a positive sentiment and run topic modelling on these to identify the reasons wy businesses / people should move into Chicago / Illinois

TOPIC DETECTION

USING LDA

Coherence Score



TOPIC MODELLING

The LDA is run for multiple topic numbers to ascertain which provides the highest coherence score. **14 and 18** topics provide the highest coherence score of 0.42. Since, lesser number of topics provide the same coherence score, we have decided to use 14 topics to extract overall topics from all articles.

The topics derived from these are categorised in the following braod categories:

Sports Keyword: player, game, season	**Employement** Keyword: employ, poverty	**Education** Keyword: Student, teacher	**Police** Keyword: shoot, crime, violence
Weather Keyword: weather, snow	**Law** Keyword: Law, attorney, Injury	**Rehab** Keyword: drugs, detox, rehab	**Business** Keyword: company, work

POSSIBLE REASONS FOR DECLINE IN POPULATION - I

Using the topics that have been identified in the previous slide, I filter out the dataset for each of these topics and run sentiment analysis to identify the articles with negative sentiment in each of these articles. LDA is run again on these filtered articles to extract possible reasons for population decline

Police Factors

One of reasons which stands out when filtering for article with negative sentiment relating to Police and crime **seem to be about the shooting happening in Illinois.** The LDA Topic on the left highlights words such as "shoot", "man", "police" and "victim" which point to this reason

```
(1,
 '0.012*"say" + 0.009*"police" + 0.005*"shoot" + 0.005*"officer" + '
 '0.004*"man" + 0.003*"car" + 0.003*"report" + 0.003*"charge" + '
 '0.003*"victim" + 0.003*"vehicle"'),
```

Drug Factors

Another reason that comes up when we look at the sub-topic "rehab" is that there could be **high cases of drug overdoses or drug use / market in Chicago / Illinois** which is not appreciated by families and businesses and makes it unsafe for people.

```
[(0,
  '0.006*"say" + 0.003*"state" + 0.003*"year" + 0.002*"include" + 0.002*"use"
  '+ 0.002*"make" + 0.002*"also" + 0.002*"drug" + 0.002*"company" + '
  '0.002*"time"'),
```

School Factors

Some of the topics from education sub-topic were that **school teachers were forming unions and protesting to revert back to online medium of instructions.** If Chicago / Illinois state does not take actions then this could **discourage teachers from teaching in Chicago due to lower health standards** despite rising COVID cases.

```
(4,
 '0.019*"say" + 0.010*"school" + 0.007*"mask" + 0.006*"district" + '
 '0.006*"state" + 0.006*"student" + 0.004*"mandate" + 0.003*"case" + '
 '0.003*"include" + 0.003*"rule"'),
```

```
[(0,
  '0.021*"school" + 0.017*"say" + 0.013*"student" + 0.008*"teacher" + '
  '0.008*"district" + 0.005*"parent" + 0.005*"union" + 0.004*"state" + '
  '0.004*"city" + 0.003*"child"'),
```

POSSIBLE REASONS FOR DECLINE IN POPULATION - II

Weather Factors

Weather is quite extreme and this also seems to be a factor for negative sentiment for Chicago / Illinois. **Heavy snowfall and extreme weather** is another reason which could potentially lead to decline in population

```
(2,
 '0.008*"say" + 0.003*"state" + 0.003*"time" + 0.003*"include" + '
 '0.003*"people" + 0.003*"day" + 0.002*"snow" + 0.002*"year" + 0.002*"also" +
 '0.002*"report"'),
```

Law Factors

Law related factors for negative sentiment include **high attorney fees in Chicago / Illinois**

```
(1,
 '0.003*"the_sponsor_reserves_right" + '
 '0.003*"certain_terms_conditions_specifie" + 0.002*"also" + '
 '0.001*"attorneys_fees_court_cost" + '
 '0.001*"construction_validity_interpretation_enforceability" + '
 '0.001*"fluctuations_any_difference_state" + '
 '0.001*"the_sponsor_responsible_late" + 0.001*"lost_damaged_stolen_luggage" '
 '+ 0.001*"sponsor_well_released_partie" + 0.001*"new_york_ny_this"'),
```

Business Factors

Relating to the business sub-topic, one of the things that was highlighted by the topic was **insider trading**. This could be one of the reasons why business would not want to set up operation in Chicago

```
(0,
 '0.009*"shares_industrial_products_company" + '
 '0.007*"industrial_products_company_stock" + '
 '0.006*"shares_illinois_tool_work" + 0.006*"illinois_tool_work" + '
 '0.004*"works_inc_nyse_itw" + 0.003*"owns_shares_industrial_product" + '
 '0.003*"products_company_stock_worth" + 0.003*"insider_trades_illinois_tool" '
 '+ 0.003*"tool_works_third_quarter" + 0.003*"products_company_stock_value"'),
```

Actionable Recommendations

Below are some of the recommendations based on topics identified earlier

- **Stricter gun regulation** in the state can reduce the misguided purchase of guns by youth and reduce violence.

- **Harsher rules around drug distribution** based on age of the consumer and stricter rules regarding non-medical drug consumption in the state.

- **Establishing more well maintained rehabilitation centre** to deal with drugs and other substance abuse.

- An increase focus on stopping the spread of virus by making **masking compulsory throughout the city**. Implementing new **hybrid teaching medium** for public schools to accommodate teacher's interests.

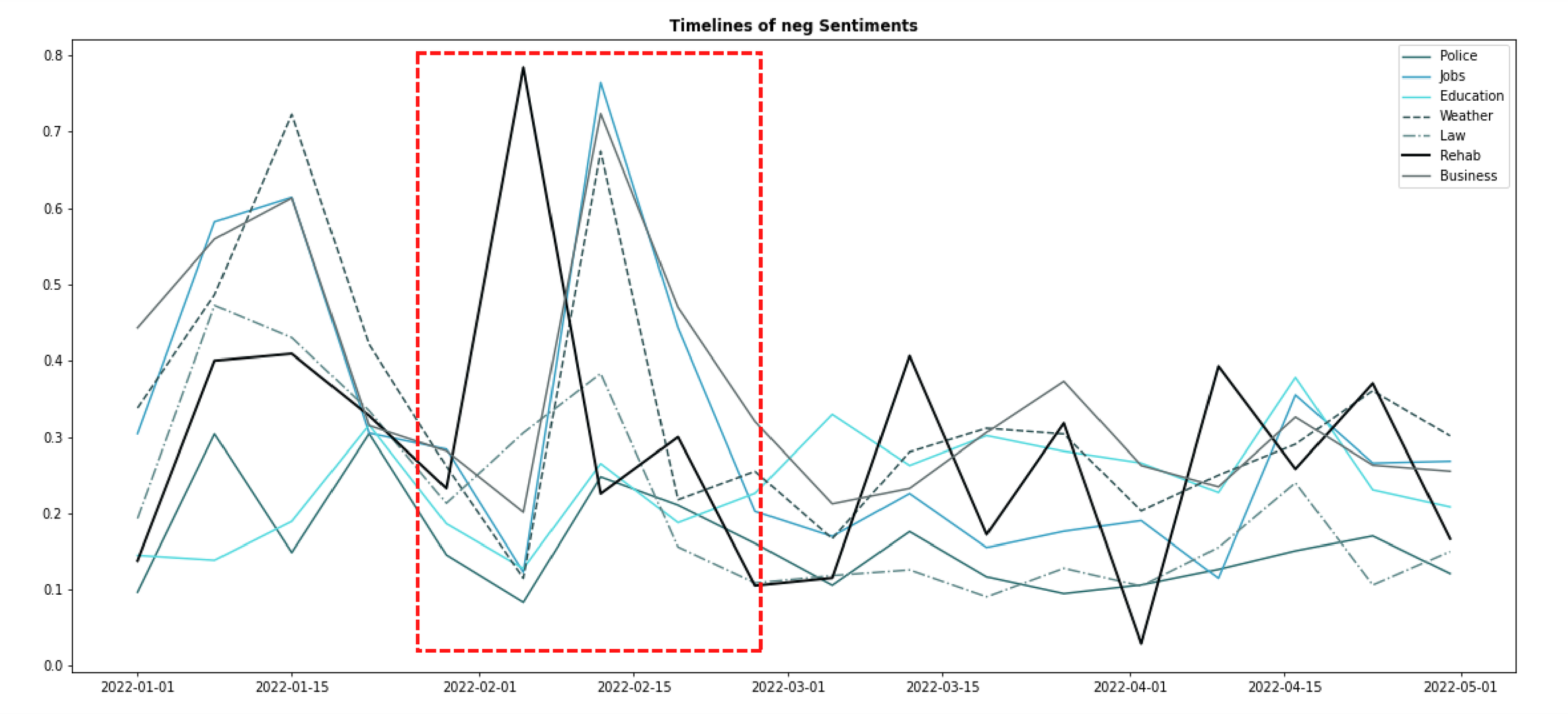
- **Harsher and more regulated securities market** to ensure insider trading is kept under control.

SENTIMENT ANALYSIS - TREND OVER TIME

USING YELP REVIEW PRE-TRAINED LOGISTIC REG MODEL FOR SENTIMENT DETECTION

Timeline of Negative Sentiment for each topic

We see a sudden surge of negative sentiments In February which again dies down in March. The proportion of articles which are classified as negative are in the range of 70-85% in February and this falls to around 40-10% for all sub-topics



ENTITY IDENTIFICATION AND TARGETED SENTIMENT ANALYSIS

USING SPACY, VADER SENTIMENT AND LDA

Identifying Entities in Text

Named Entity Recognition

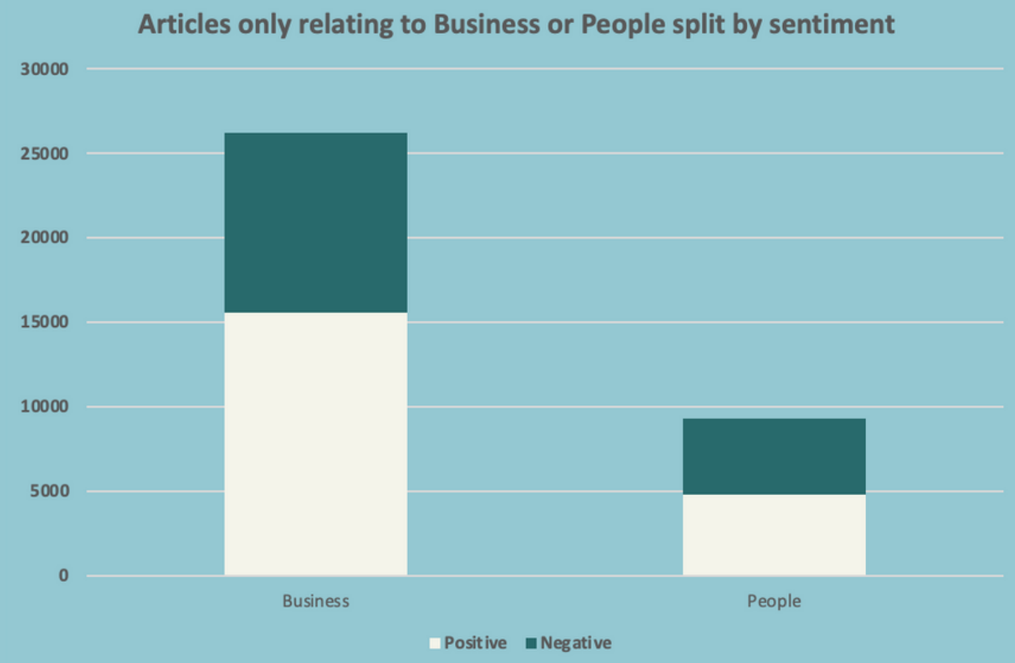
I use Spacy's Named Entity Recognition Package to detect different entities in the text corpus. I have used the Spacy package without sentence segmentation for this exercise.

4 entities are identified in for each article - Organisations, People, Location and National or Religious or Political Groups.

After this, I ran a sentiment analysis on each of these articles using the vader sentiment package. The overall sentiment for each article is calculated as follows:

- The article is first broken down into sentences.
- A sentiment score ranging from -1 to 1 is calculated for each sentence. Score above 1 indicate positive sentiment and vice versa.
- A score is calculated for the entire article by averaging out the compound score for each sentence.

Based on the NER Analysis, article are filtered on the condition that they only contain references to organisation or people. We get around 26k articles referring to only organisations and 9k articles referring only to people. Out of which 15k and 4.8k have a positive sentiment.



REASONS FOR PEOPLE / BUSINESS TO MOVE INTO IL

BUSINESS

Some of the topics that stood out for "business" which could be the reasons why businesses move to Illinois are:

- State is supportive of business and companies which set up operations in Illinois.
- Opening up a lawyer office / law company is encouraged in the state, especially family law attorney. As this topic is extracted from positive sentiments, it is likely that people in IL generally need family law attorneys.
- Businesses and companies here also support team work, experience and service.

```
(2,
 '0.004*"get" + 0.004*"use" + 0.004*"may" + 0.004*"company" + 0.004*"also" + '
 '0.004*"time" + 0.004*"make" + 0.003*"business" + 0.003*"state" + '
 '0.003*"go"'),
```

```
(3,
 '0.004*"also" + 0.003*"business" + 0.002*"time" + '
 '0.002*"family_law_attorneys_near" + 0.002*"company" + 0.002*"lawyer" + '
 '0.002*"well" + 0.002*"com" + 0.001*"state" + 0.001*"work"'),
```

```
(6,
 '0.004*"company" + 0.004*"additional_shares_last_quarter" + '
 '0.003*"shares_industrial_products_company" + 0.003*"quarter" + '
 '0.003*"owns_shares_company_stock" + 0.003*"also" + '
 '0.003*"industrial_products_company_stock" + 0.003*"share" + '
 '0.003*"stock_valued_purchasing_additional" + '
 '0.002*"owns_shares_industrial_product"'),
```

```
(7,
 '0.007*"program" + 0.006*"work" + 0.005*"also" + 0.005*"support" + '
 '0.004*"experience" + 0.004*"service" + 0.004*"well" + 0.004*"team" +
```

PEOPLE

Some of the topics that stood out for "people" which could be the reasons why people move to Illinois are:

- There are good work opportunities in Chicago
- People are very friendly and hence, it is easy to find company in the city
- There are a lot of games and shows which keep happening in Chicago which can be good for entertainment and attract more tourists.

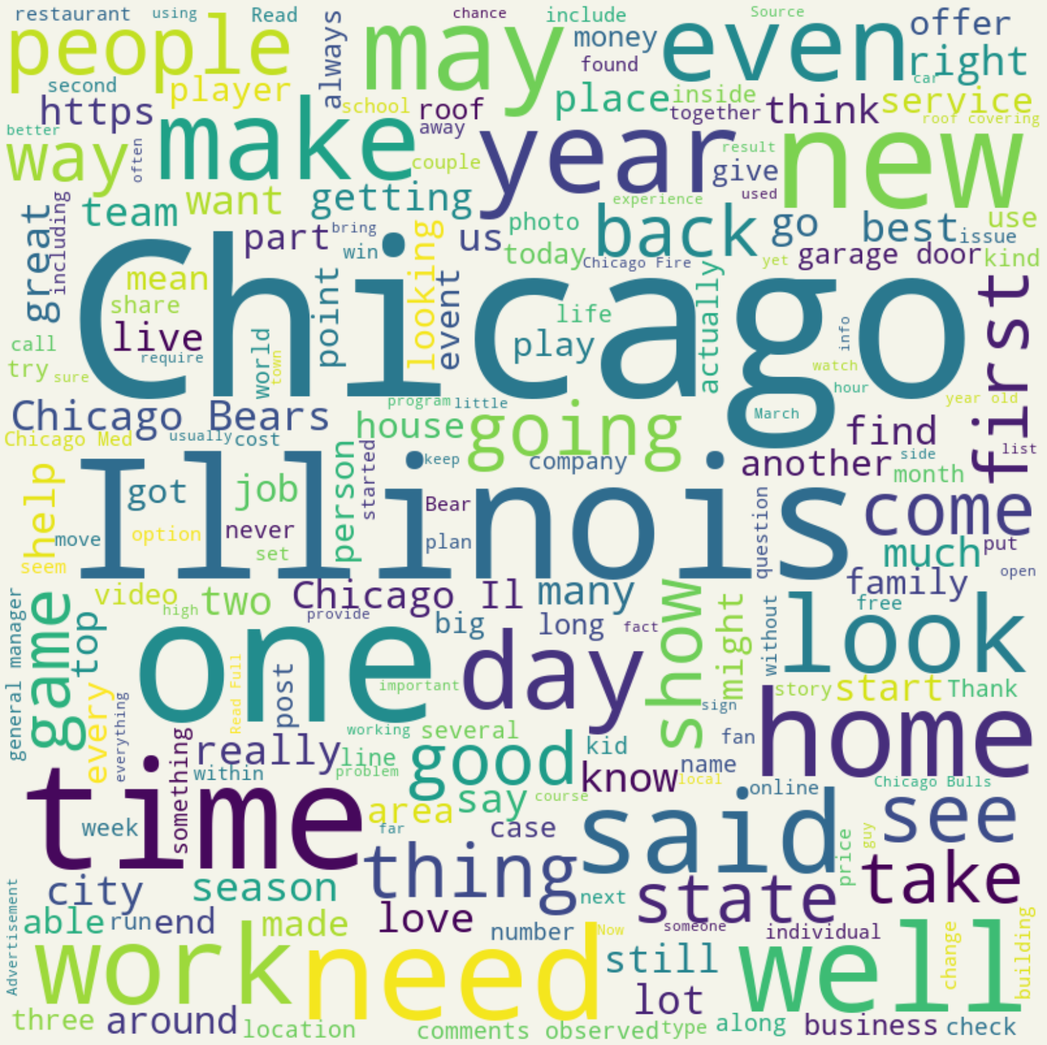
```
[(0,
  '0.004*"go" + 0.004*"get" + 0.003*"make" + 0.003*"also" + 0.003*"work" + '
  '0.003*"see" + 0.003*"good" + 0.003*"help" + 0.003*"may" + 0.003*"come"'),
```

```
(5,
 '0.004*"may" + 0.004*"go" + 0.004*"get" + 0.003*"company" + 0.003*"make" + '
 '0.003*"need" + 0.003*"good" + 0.003*"find" + 0.003*"use" + 0.002*"live"'),
```

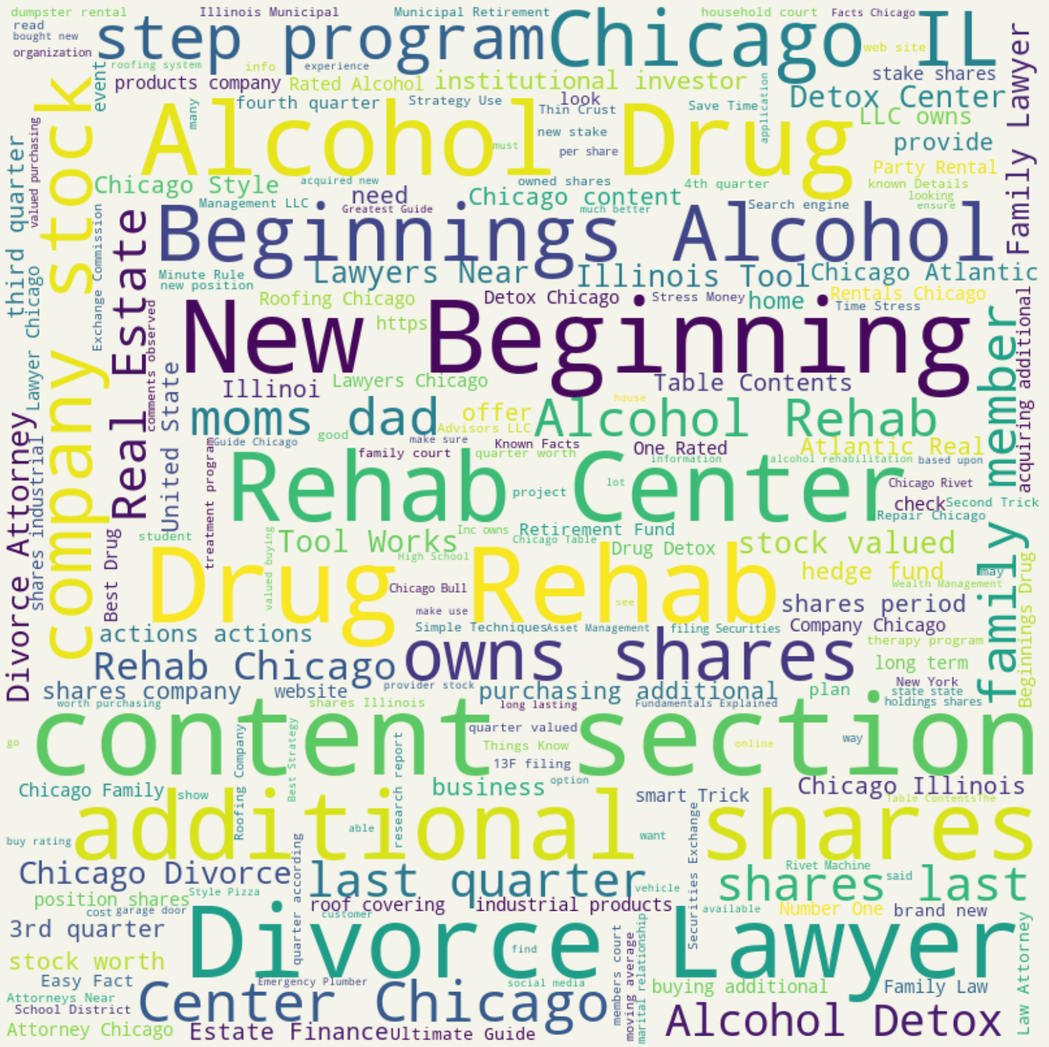
```
(4,
 '0.005*"come" + 0.004*"make" + 0.004*"also" + 0.004*"time" + 0.003*"go" + '
 '0.003*"see" + 0.003*"show" + 0.003*"get" + 0.003*"game" + 0.003*"start"'),
```

WORD CLOUDS FOR BUSINESS AND PEOPLE

PEOPLE



BUSINESS



THANK YOU