Extracting Affordable Housing Data using GDELT

Igor Baranov

March 17, 2019

Introduction

This report is describing how to retrieve, process and present online news data to the issues related to Toronto affordable housing. The main source of infrmation is The GDELT Project which is a realtime network diagram and database of global human society for open research. GDELT Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day.

The news tone is an important information that in combination with actual pricing data could be used in analysis of the housing affordability and predictng of pricing trend.

Discussion

Even though GDELT contains lots of information usefull for our task, extracting actual historical economical information like housing prices in a particular area is a difficult and unreliable process. GDELT is designed to collect emotional and geolocational information related to the news. We will use it to access the news tonal information related specifically to Toronto's housing affordability crisis.

The following steps are suggested:

1. Retrieve historical news tone data.
2. Develop a method of scrited retrieval of detailed latest news tonal data.
3. Compare the news tone to major Canadian economical indices.

Data Retrieving

Retrieving historical tonal information using BigQuery API

The historical information has to be retrieved once, so we will do it using Google BigQuery API. The GDELT data structure is described in the DDELT GKG DATA FORMAT CODEBOOK. The following document provodes some examples of the queries to that API: Google BigQuery + GKG 2.0: Sample Queries. Access to the API provided browsing on over to the GKG table in BigQuery. Please note that this requeres a valid Google BigQuery account.

The SQL statement below will extract the news coverage average tone and number of articles related to Toronto's housing matters from the begining of 2015 to March 2019 aggregsted daily. The SQL was run and the resulting csv file saved locally. Note that this procedure could be done by running direct R or Python API accessing BigQuery from the script.

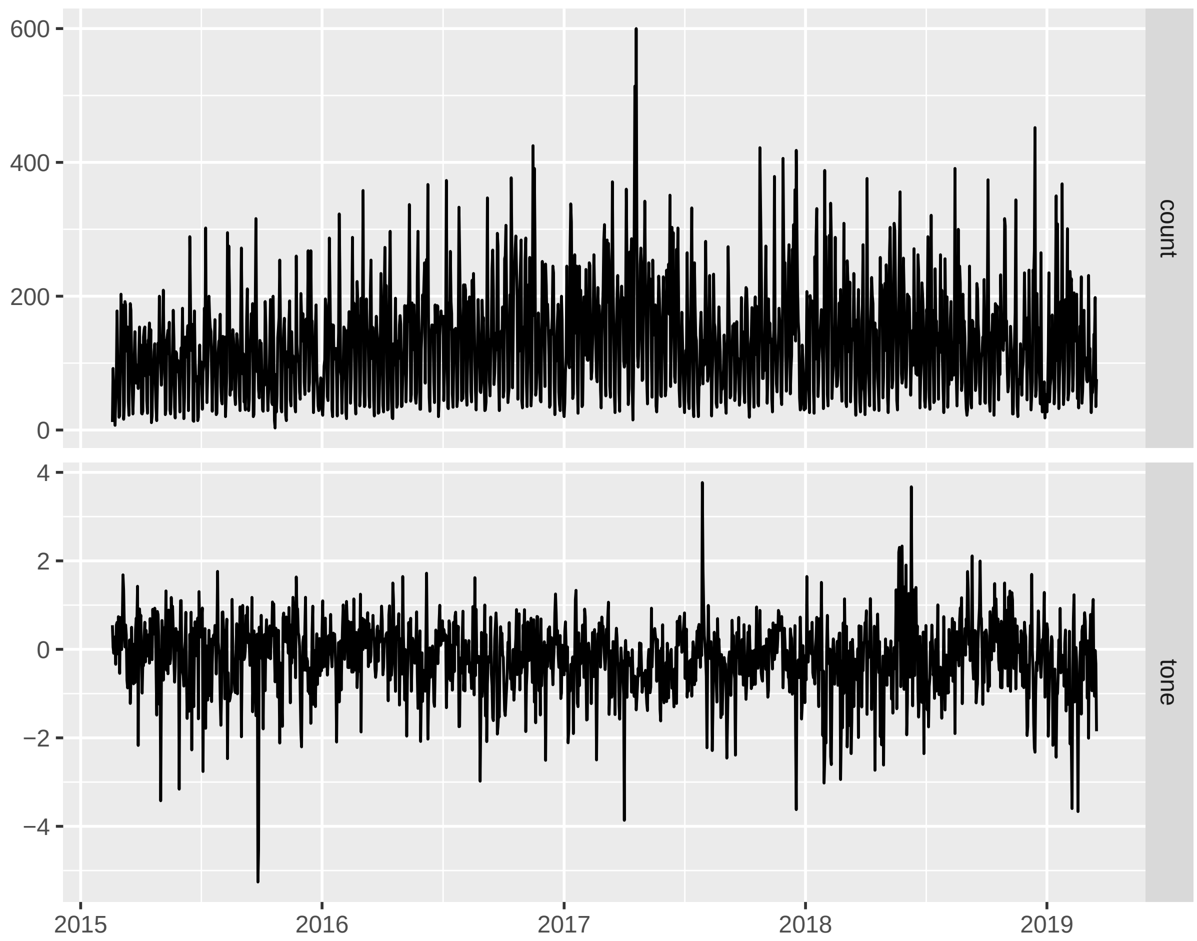

Figure 1: Toronto Housing Issue News Tone Time Series Aggregated Daily

```
SELECT
  substr(CAST(DATE AS STRING),0,8) as daydate, count(1) count,
  avg(CAST(REGEXP_REPLACE(V2Tone, r',.*', "") AS FLOAT64)) tone
FROM `gdelt-bq.gdeltv2.gkg_partitioned`
WHERE
    _PARTITIONTIME >= TIMESTAMP("2015-01-01")
  AND V2Locations like "4#Toronto%"
  AND V2Themes like "%_HOUSING%"
GROUP BY daydate
ORDER BY daydate
```

The following code loads the query results, converts it to time series and plots them (Fig. 1):

```
library(zoo)
tmp <- read.table("./toronto-housing-daily-tone-20190317-194930.csv",
                  sep = ",", header = T)
toneTS <- zoo(tmp[,2:3], as.Date(as.character(tmp[,1]), format="%Y%m%d"))
```

```
library(ggplot2)
autoplot(toneTS) + facet_free() + xlab(NULL)
```

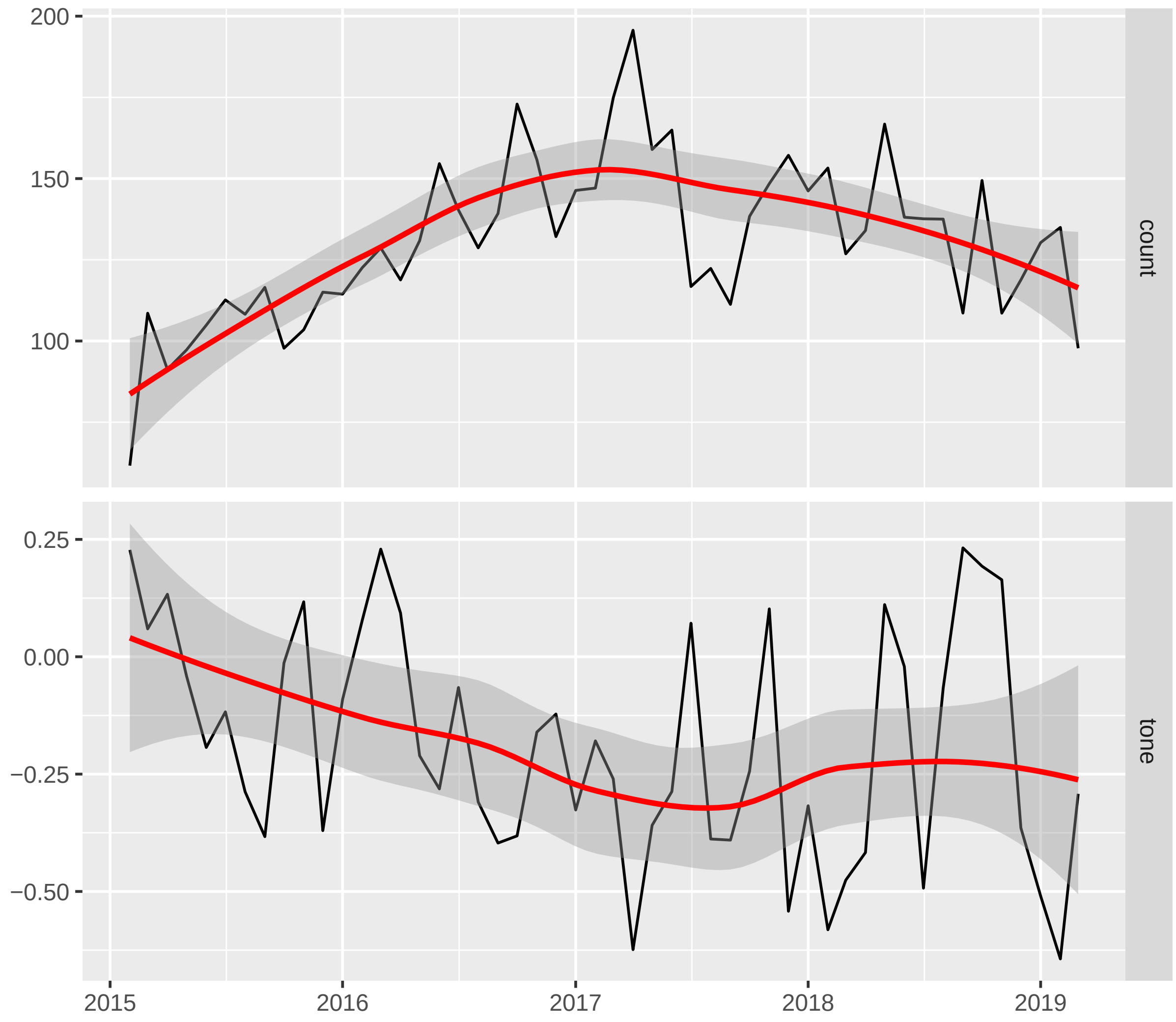

Figure 2: Toronto housing issue news tone time series, aggregated monthly

As we can see, the daily aggregation is not very representative. Let's aggregate the news tone and article count numbers monthly. Figure 2 displays the resulting data suplemented with trend lines.

```r
x2d <- aggregate(toneTS, as.Date(cut(time(toneTS), "month")), mean)
autoplot(x2d) + facet_free() + xlab(NULL) + geom_smooth(colour = "red")
```

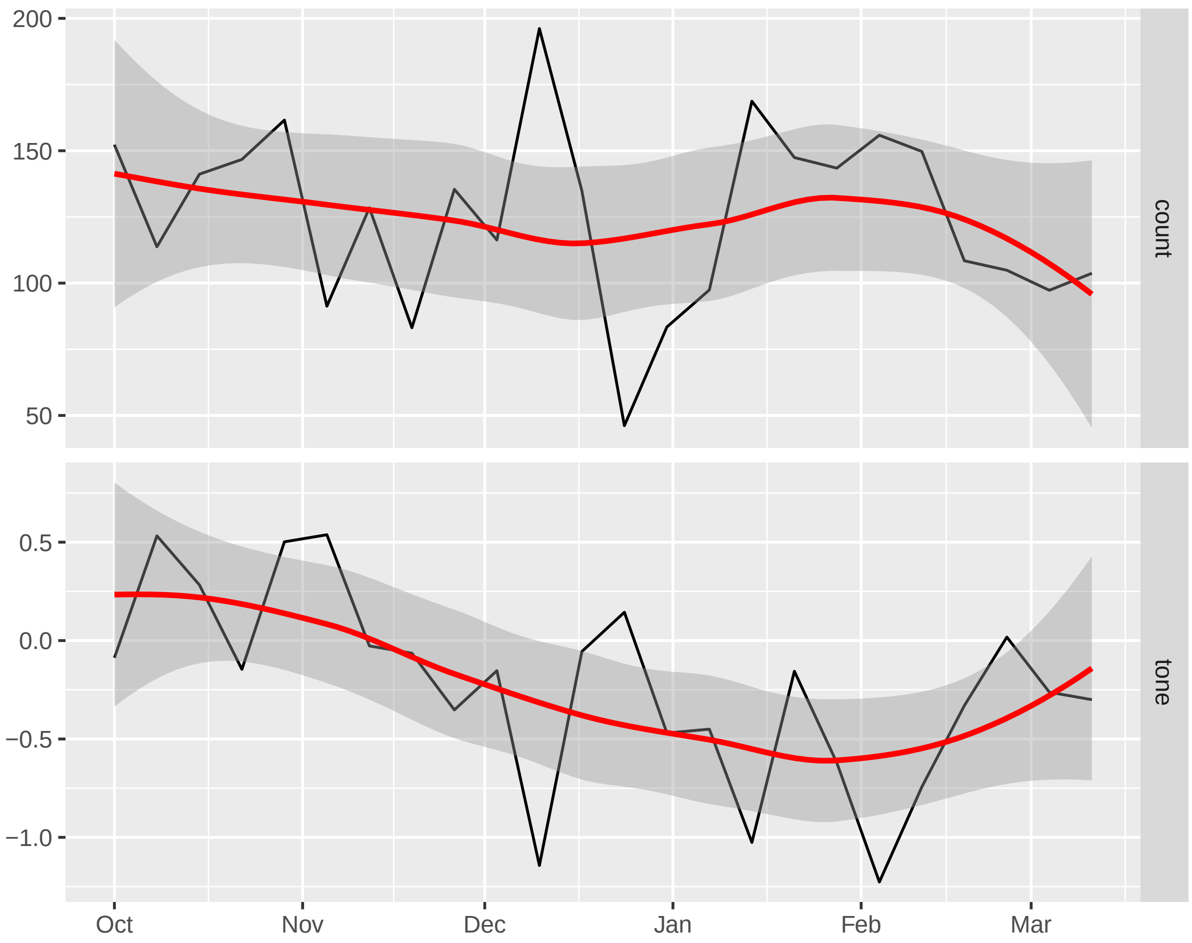

Figure 3: Toronto housing issue news tone time series, aggregated Weekly

For the purpose of analysis the tone time series can be zoomed to specific time window and the data can be aggregated differently. Figure 3 shows the last 6 month of data aggregated weekly.

```
w <- window(toneTS, start = as.Date("2018-10-01"))
w <- aggregate(w, as.Date(cut(time(w), "week")), mean)
autoplot(w) + facet_free() + xlab(NULL) + geom_smooth(colour = "red")
```

Compare News Tone to Economical Indices

Lets extract housing price index. This information was taken from Statistics Canada. Figure 4 shows Toronto housing index history from 2015.

```
tmp <- read.table("./toronto-houing-price-index.csv", sep = ",", header = T)
tmp <- tmp[2,2:50]
housing_index <- zoo(as.numeric(tmp), as.yearmon(names(tmp), format="%B.%Y"))
autoplot(housing_index) + scale_x_yearmon() +
  xlab(NULL) + ylab(NULL) + geom_smooth(colour = "green")
```

Another interesting index is Canada prime interest rate was taken from Bank of Canada site. Figure 5 shows the index history from 2015.

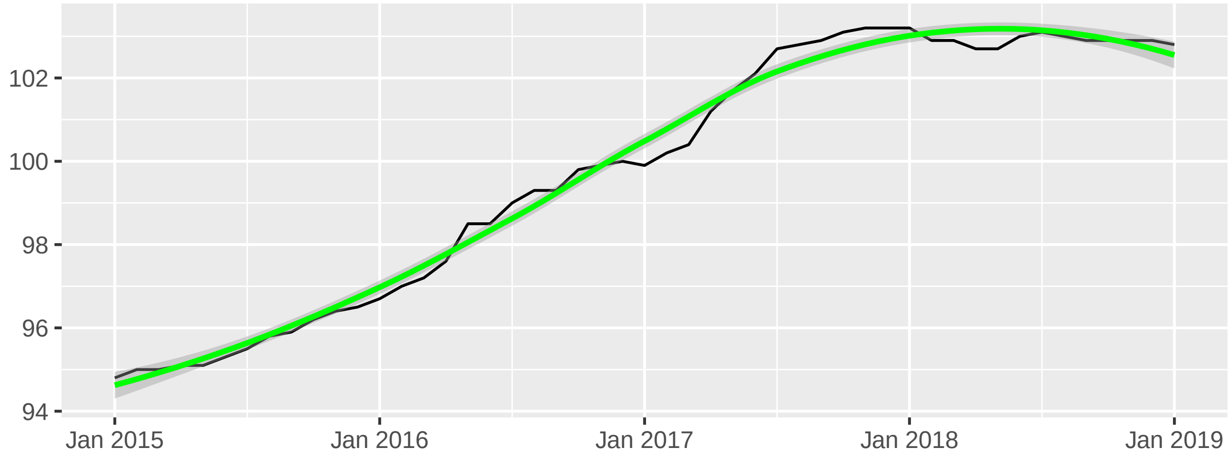

Figure 4: Toronto housing index history

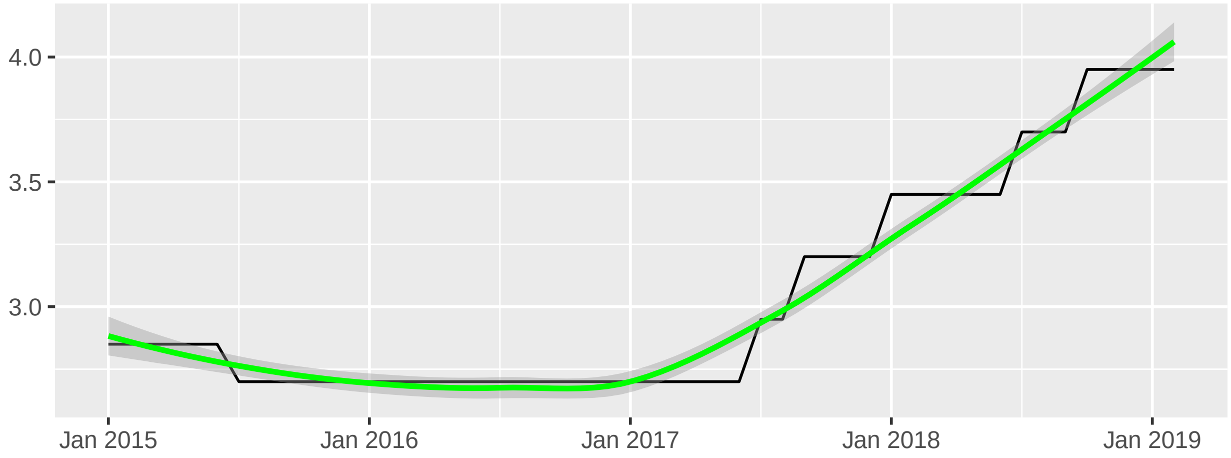

Figure 5: Bank of Canada prime rate history

```
tmp <- read.table("./prime-rate.csv", sep = ",", header = F)
prime_rates <- zoo(tmp[,2], as.yearmon(tmp[,1], format="%Y-%m"))
autoplot(prime_rates) + scale_x_yearmon() + geom_smooth(colour = "green") +
  xlab(NULL) + ylab(NULL)
```

Code below creates Figure 6 that comparing news tone to the economical indices since the begining of 2018.

```
x2d <- aggregate(toneTS, as.yearmon, mean)
a <- merge(x2d, housing_index, prime_rates)
wa <- window(a, start = as.yearmon(2018,1))
autoplot(wa) + scale_x_yearmon()+ facet_free()+ xlab(NULL) + geom_smooth(colour = "red")
```

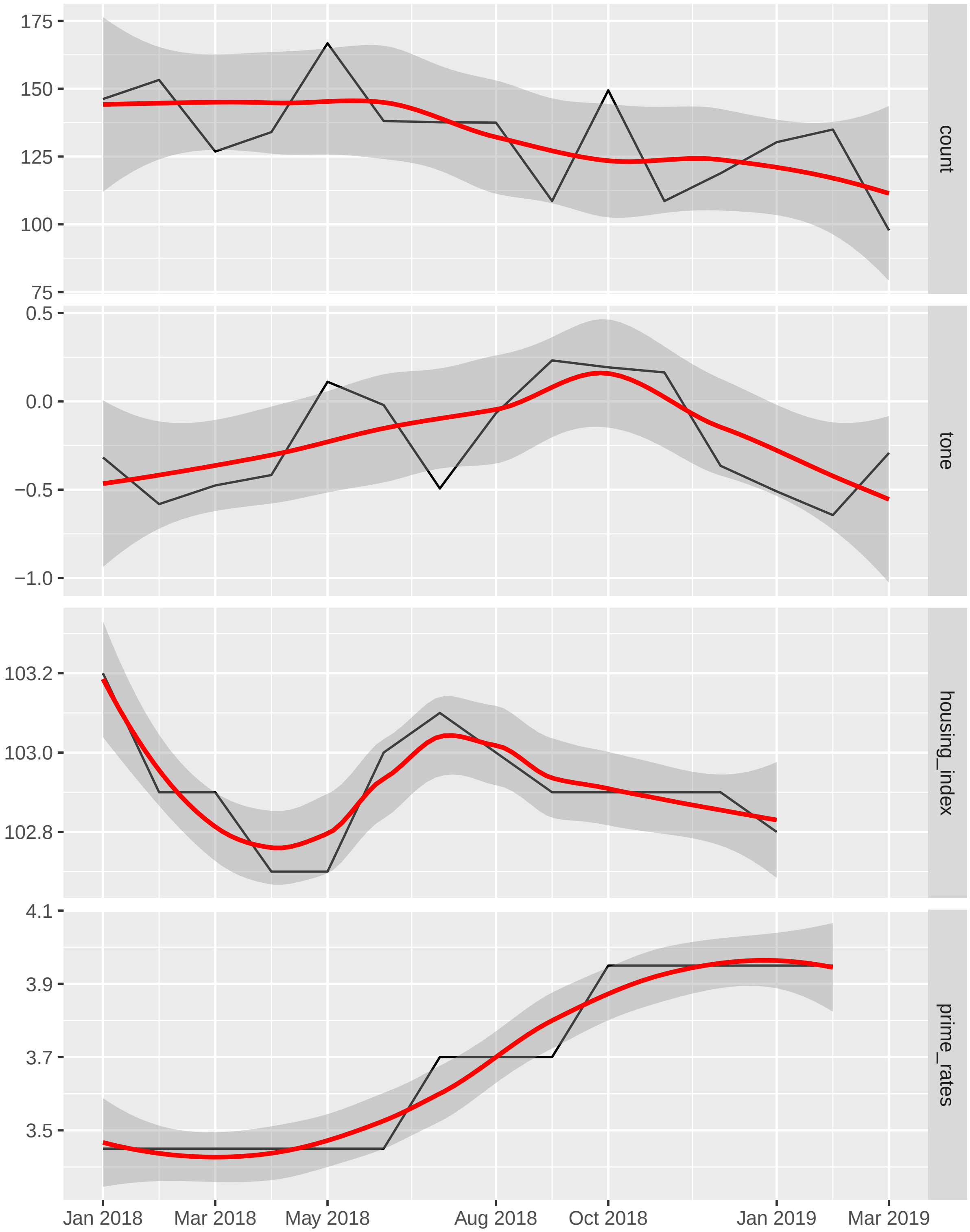

Figure 6: Comparing news tone to major economical Indices

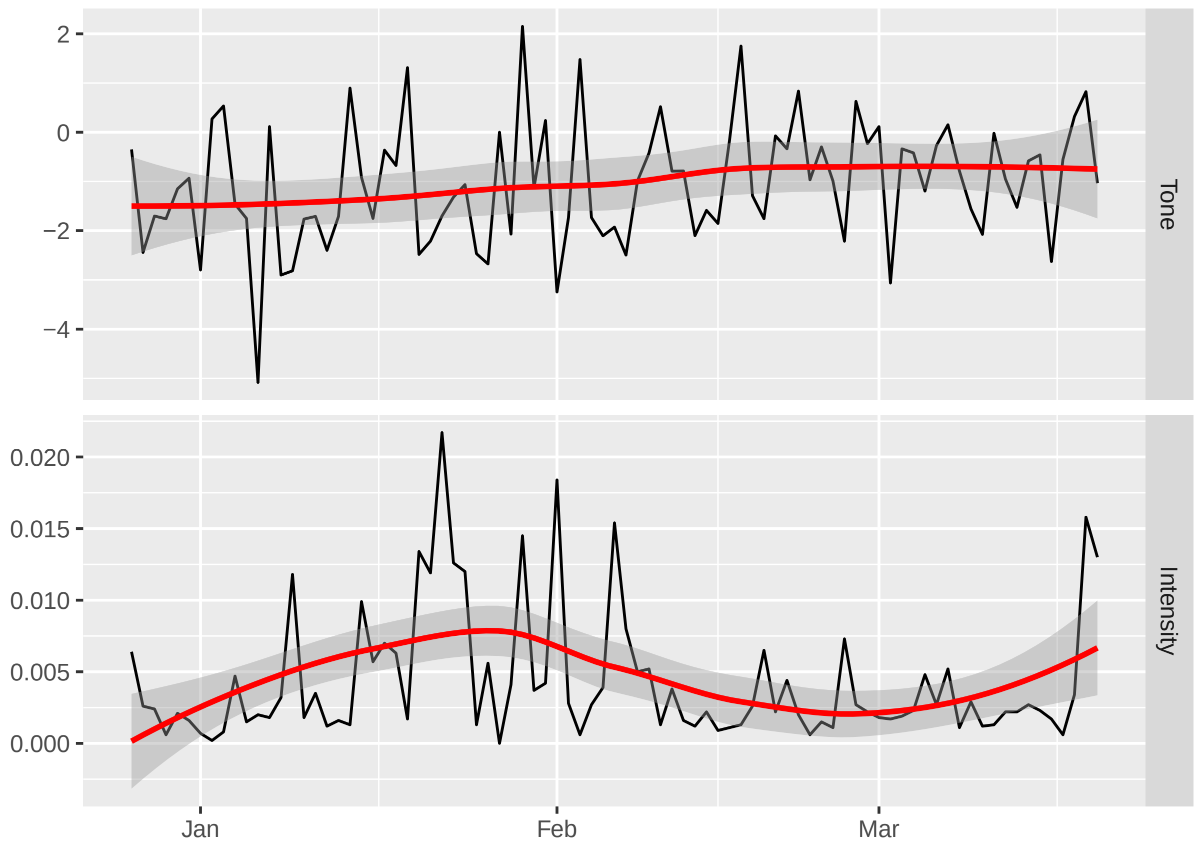

Figure 7: Toronto Affordable Housing - last 3 month of news coverage

Using GDELT full-text search API

Another possibility to access GDELT tone information is to use GDELT full-text search API. It allows to access a 3-month window of latest information updated every 15 minutes. The advantage of this API over the one described before is ability to do a full-text search instead of quering tags and by that not relying on questionable accuracy of GDELT algorithms to recognise those tags in the news articles. Let's try to extract the news tone information via full-text search. The code below gets the average "tone" of all matching coverage, from extremely negative to extremely positive. Figure 7 shows the 30 days history of the news tome and intensity.

```
api   <- "https://api.gdeltproject.org/api/v2/doc/doc?format=csv&query="

terms <- "near10:%22toronto%20housing%22"

tmp <- read.table(paste0(api,terms,"&mode=TimelineTone"), sep = ",", header = T)
Tone <- zoo(tmp[,3], as.Date(as.character(tmp[,1]), format="%Y-%m-%d"))

tmp <- read.table(paste0(api,terms,"&mode=TimelineVol"), sep = ",", header = T)
Intensity <- zoo(tmp[,3], as.Date(as.character(tmp[,1]), format="%Y-%m-%d"))

tone30 <- merge(Tone, Intensity)

autoplot(tone30) + facet_free()+ xlab(NULL) + geom_smooth(colour = "red")
```

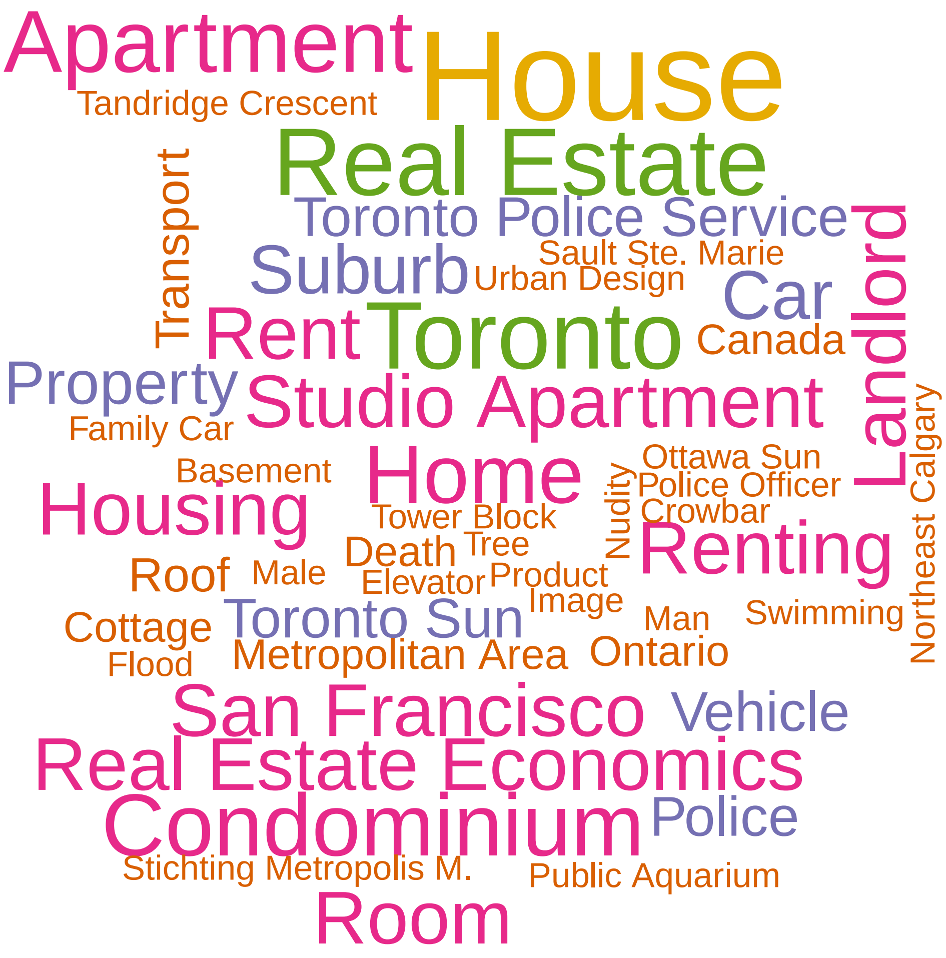

Figure 8: Toronto Affordable Housing Word cloud

Use GDELT to create word clouds

GDELT API allows creating word clouds. Please note that by applying time and date parameters those clouds could be tied to the timeline. Code below creates a word cloud (Fig. 8) of the Toronto housing related news for the last 3 month.

```
library(RColorBrewer)
library(wordcloud)

set.seed(11)
Words <- read.table(paste0(api,terms,"&mode=WordCloudImageWebTags"),
                    sep = ",", header = T)
wordcloud(Words[,1], Words[,2], colors = brewer.pal(6,"Dark2"), max.words = 50)
```

Conclusion

This report covered several possibilities of Toronto housing problem news coverage analysis. This coverage is presented in a quantified form of time series that allows to use this information along with other housing related data like rent prices for different forms of analysis. In this report the following was done:

- Proposed method of one time extracting and presenting of historical news tone and intencity data.
- Proposed method of regular scripted extraction of up to 3 months of real time news tone and coverage intencity based of full text search of the all news data.
- Demonstrated a method of comparing news tone and intencity time series to major economical indices like interest rate and housing index.
- Demonstrated a method of creating date-specific word clouds based on full text search of related news articles.

The next steps in the research could be:

- Creating an interavtive online application allowing to present and download the information presented in the report.
- Developing a prediction model based on data described in the report and other related Toronto housing data that should be scraped from the web.