GY7708_CW2

209036943

6/27/2022

```
#loading Libraries
library(tidyverse)
library(magrittr)
library(dplyr)
library(jsonlite)
library(httr)
library(sf)
library(tmap)
library(tidytext)
library(wordcloud)
library(stm)
library(tidytext)
library(reshape2)
library(quanteda)
library(tidylo)
library(stringr)
library(quanteda)
library(ggplot2)
library(gridExtra)
library(wordcloud2)
library(maptools)
library(spatstat)
library(ggraph)
library(igraph)
library(textdata)
```

Part 1

Scraping Data

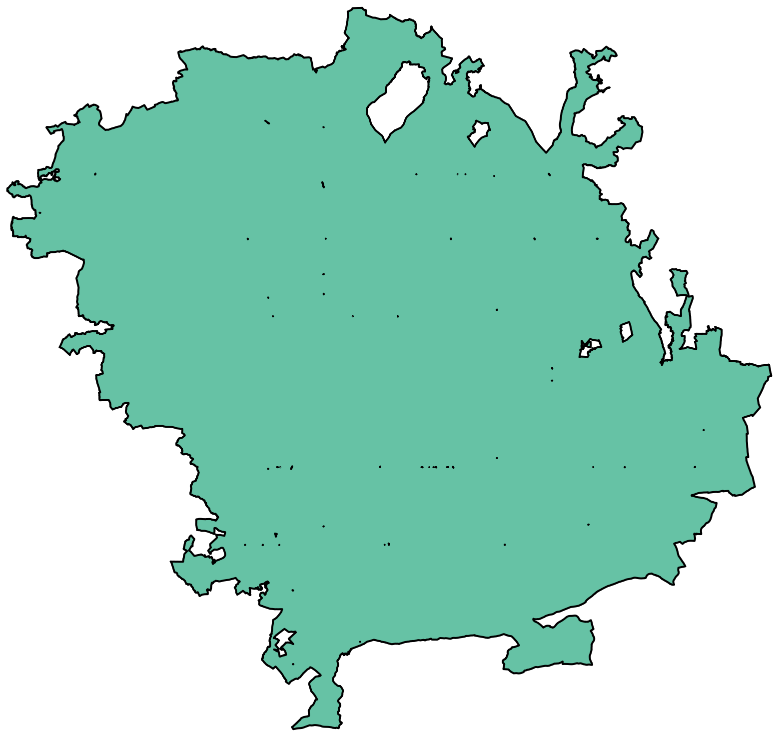
```
setwd('C:/Assignment_2-datapack')
# Part 1
## loading and filtering excel data
excel_data <- read.csv('wikipedia_geotags_in_UK.csv')

filter_excel <- excel_data %>% filter(LAD21NM == 'Ryedale') %>%
  filter(gt_primary == 1)

##Creating empty dataframe for the text
```

```
dfpage_n_W <- data_frame()

## Scraping Wikipedia Web data
for (page_name in filter_excel$page_title) {
  # Set a title
  # Retrieve the summary
  a_page_summary <-
    httr::GET(
      # Base API URL
      url = "https://en.wikipedia.org/w/api.php",
      # API query definition
      query = list(
        # Use JSON data format
        format = "json",
        action = "query",
        # Only retrieve the intro
        prop = "extracts",
        exintro = 1,
        explaintext = 1,
        redirects = 1,
        # Set the title
        titles = page_name
      )
    ) %>%
    # Get the content
    httr::content(
      as = "text",
      encoding = "UTF-8"
    ) %>%
    # Trasnform JSON content to R list
    jsonlite::fromJSON() %>%
    # Extract the summary from the list
    magrittr::extract2("query") %>%
    magrittr::extract2("pages") %>%
    magrittr::extract2(1) %>%
    magrittr::extract2("extract")

  summ <- data_frame(a_page_summary) %>%
    mutate(page_name = page_name)

  dfpage_n_W <- dfpage_n_W %>%
    bind_rows(summ)
}
```

Plotting Boundary Map

```
#loading in boundary shapefile of the study area
setwd('C:/Assignment_2-datapack')
boundary1 <- read_sf('Census 2022-06-17-03-56/census.shp')
boundary <- read_sf('Census 2022-06-17-03-56/ryedaledissolved.shp')
boundary <- boundary[1]
plot(boundary)
```

Column1



```
boundary_m <- tm_shape(boundary)+
  tm_borders()
```

Data Merging, Filtering and Subsetting

```
#removing '~' from the data
dfpage_n_W <- data.frame(lapply(dfpage_n_W, function(x){sub('~', '', x)}))

filter_excel2 <- filter_excel %>% select('gt_lat', 'gt_lon',
                                         'page_title')
#merging the data with the excel file to have coordinate
sentcs_w_pgs <- dfpage_n_W %>%
  left_join(filter_excel2, c('page_name' = 'page_title')) %>%
  st_as_sf(coords = c("gt_lon", "gt_lat"),
           crs = 4326, na.fail = TRUE) %>%
  st_transform(27700)
```

Text Tokenization

```r
#Tokenization of words and removing stopwords
T_W_n_Pg <- sentcs_w_pgs %>%
  unnest_tokens(word, a_page_summary) %>%
  anti_join(get_stopwords(),  c('word' = 'word'))

#Tokenizing Sentences and removing stopword
Snt_n_pg <- sentcs_w_pgs %>%
  unnest_tokens(sentence, a_page_summary,
               token = "sentences")
```

Part 2: Spatial Frequency Analysis

Spatial Frequency Analysis

Spatial frequency analysis is an offshoot of frequency analysis which measures how often an event occur. However, spatial frequency analysis combines the approach of frequency analysis with the coordinate or location of the event to understand the distribution or pattern of the event in the study area. It measures the frequency of an event per unit distance. In calculation, spatial frequency analysis is measured in cycle per distance. Within the Natural Language processing realm, some of the Spatial Frequency Analysis that can be used includes;

Ordinary Spatial Frequency (OSF): This includes measuring the oftenness of an occurrence with respect to its location on earth. This helps use understand the pattern. OSF was used evaluate the variation of words count in the Ryedale district in the following ways:

- **Total words count frequency:** Shows words with the highest frequency of occurrence.
- **Page word frequency:** examining the number of words on each page and their spatial pattern.
- **Per Page words count:** This shows the frequency of each word on each page, with their spatial pattern included.
- **Word Cloud:** This is also a frequency analysis. It presents a pictorial frequency of words, by making high frequency words bigger than lower one in the visualized picture. It can be used to show focus words.

Term Frequency Analysis: Unlike the ordinary frequency analysis, term frequency is used to measure the importance of a word. It measures the frequency of a word with respect to it subgroup, giving the event its relevance score.

TF-IDF: Another advanced level of frequency analysis which can implemented in spatial form. The full meaning of TF-IDF is term frequency-inverse document frequency. It combines the term frequency and inverse frequency approach to rate the relevance of a word to a document. It is used to measure phrase or word importance in the field of machine learning and information retrieval. Highly relevant words have high tfidf score and vice-a-vice.

Point Pattern Analysis (PPA): in order to understand better the distribution of words in the study area, PPA is performed. PPA helps understand the spatial arrangement of points.
The density, quadrant and K function analysis were performed to understand the pattern of words in the area.
- Density is used to measure the concentration of points across the area by evaluating the ratio of the global density to the local density. This helps shows the pattern: concentration and disperseness of points in the area.
- Quadrant is done by dividing the area into uniformly nshaped subregion quadrants and then examining

the number of points that falls in each quadrant.
- K function: this is used to examine the pattern of points, revealing if the observed pattern of the points are more or less clustered that the expected pattern.

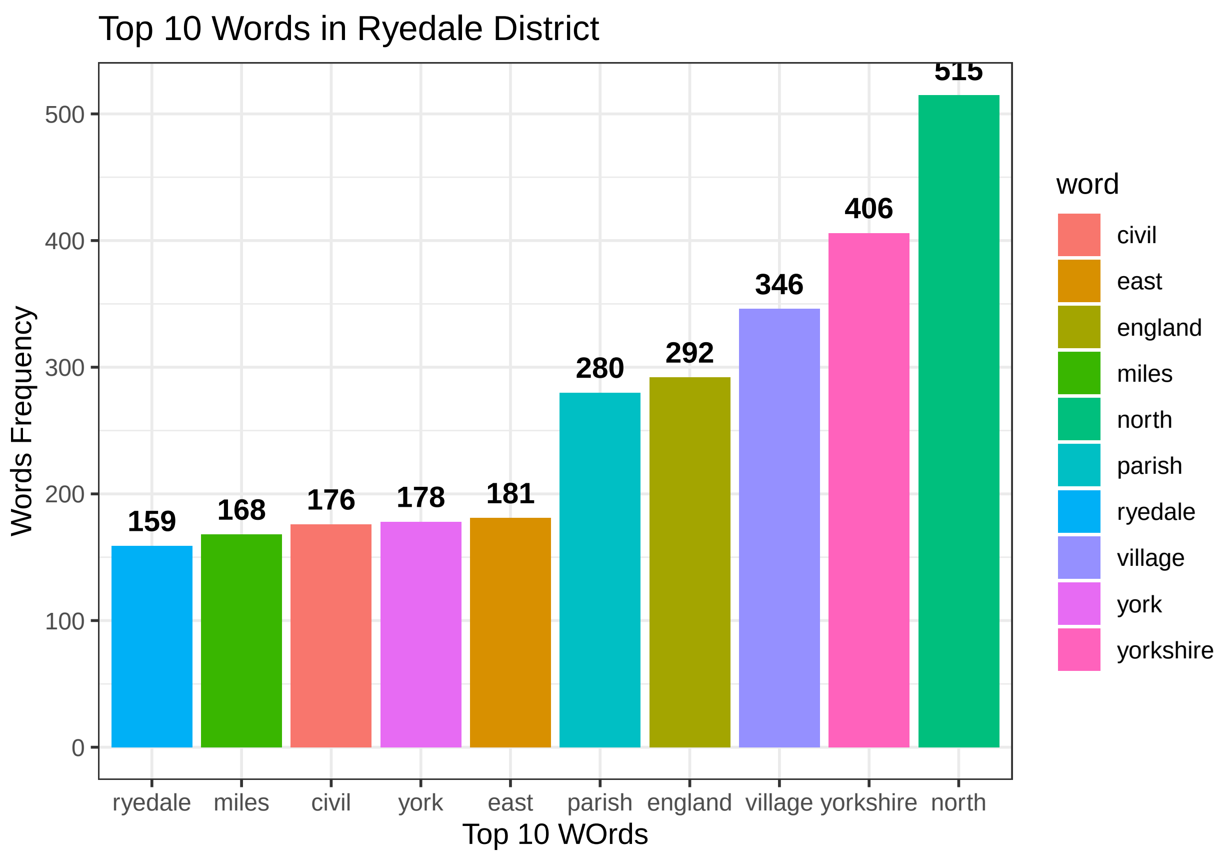
Word Frequency Analysis

```
word_freq <- T_W_n_Pg %>%
  count(word, sort = TRUE)
#There are 4172 unique word in Ryedale  Wikipedia page
```

```
top10_wds <- word_freq %>%
  slice_max(n, n = 10)

top10_wds %>%
  knitr::kable()
```

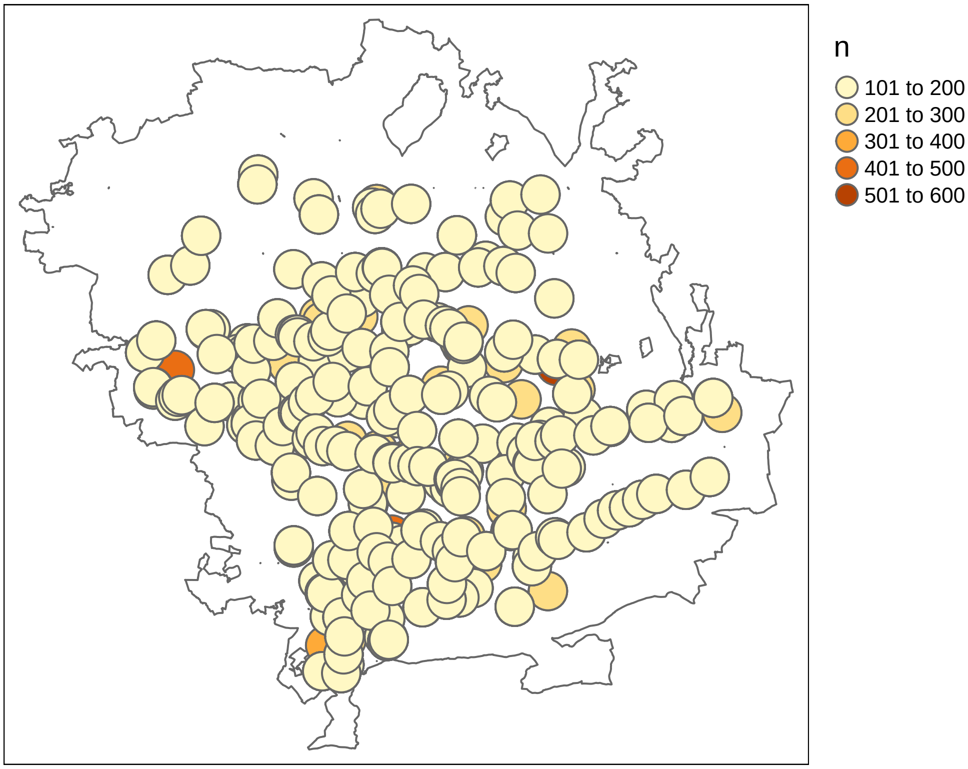
word	n	geometry
north	515	MULTIPOINT ((452348.1 48307...
yorkshire	406	MULTIPOINT ((452348.1 48307...
village	346	MULTIPOINT ((453045.3 48008...
england	292	MULTIPOINT ((452348.1 48307...
parish	280	MULTIPOINT ((452348.1 48307...
east	181	MULTIPOINT ((452348.1 48307...
york	178	MULTIPOINT ((452348.1 48307...
civil	176	MULTIPOINT ((452348.1 48307...
miles	168	MULTIPOINT ((453309.6 48412...
ryedale	159	MULTIPOINT ((452348.1 48307...

```
top10_wds %>%
  ggplot(aes(fct_reorder(word,n), n, fill = word)) +
  geom_col() +
  geom_text(aes(label = n), size = 4,
            fontface = "bold", vjust = -0.7) +
  labs(title = "Top 10 Words in Ryedale District",
       x = "Top 10 WOrds", y = "Words Frequency") +
  theme_bw()
```

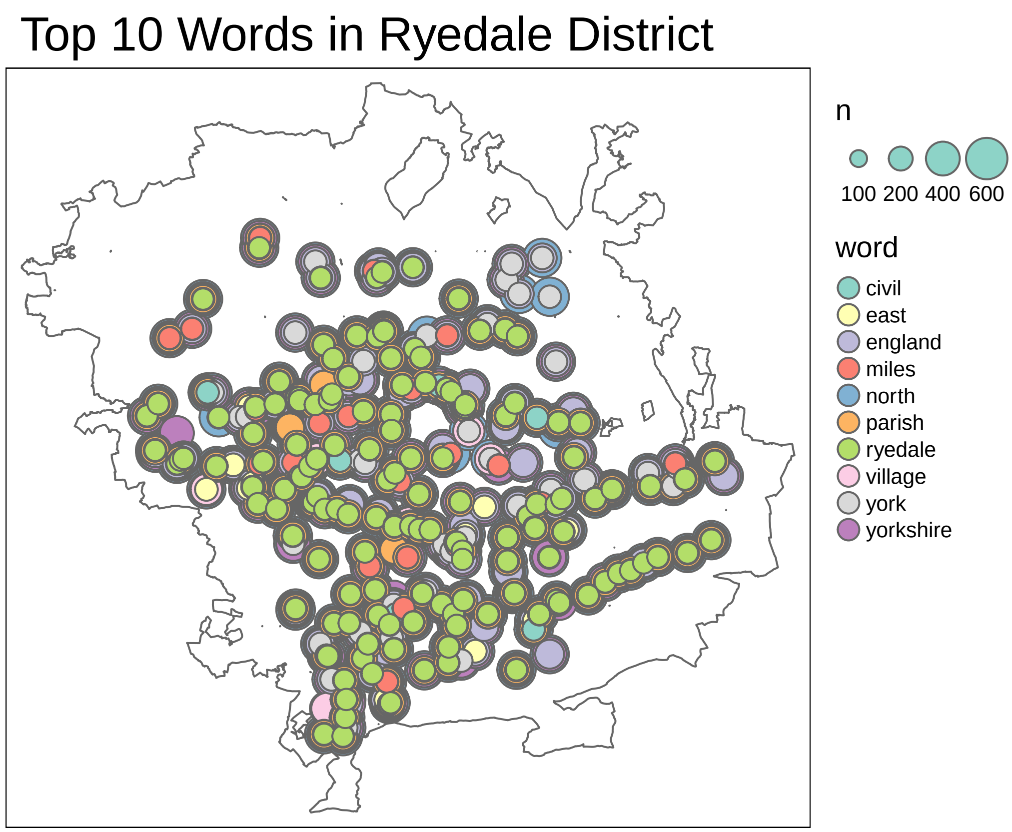



```
boundary_m +
  tm_shape(top10_wds)+
  tm_bubbles( col = 'n')+
  tm_layout(main.title = 'Top 10 Words in Ryedale District',
            title.size = 0.8,
            legend.outside=TRUE)
```

Top 10 Words in Ryedale District




```
boundary_m +
  tm_shape(top10_wds)+
  tm_bubbles( col = 'word', size = 'n')+
  tm_layout(main.title = 'Top 10 Words in Ryedale District',
            title.size = 0.8,
            legend.outside=TRUE)
```



“north” and “yorkshire” are the most common words in Ryedale district because Ryedale is spatially located in non-metropolitan district of north yorkshire.

```
#Word Cloud Map
word_freq %>%
  with(wordcloud(word, n, max.word = 100))
```

The word cloud also confirms that 'north' and 'yorkshire' are the most used in Ryedale district.

Sentence Frequency Analysis

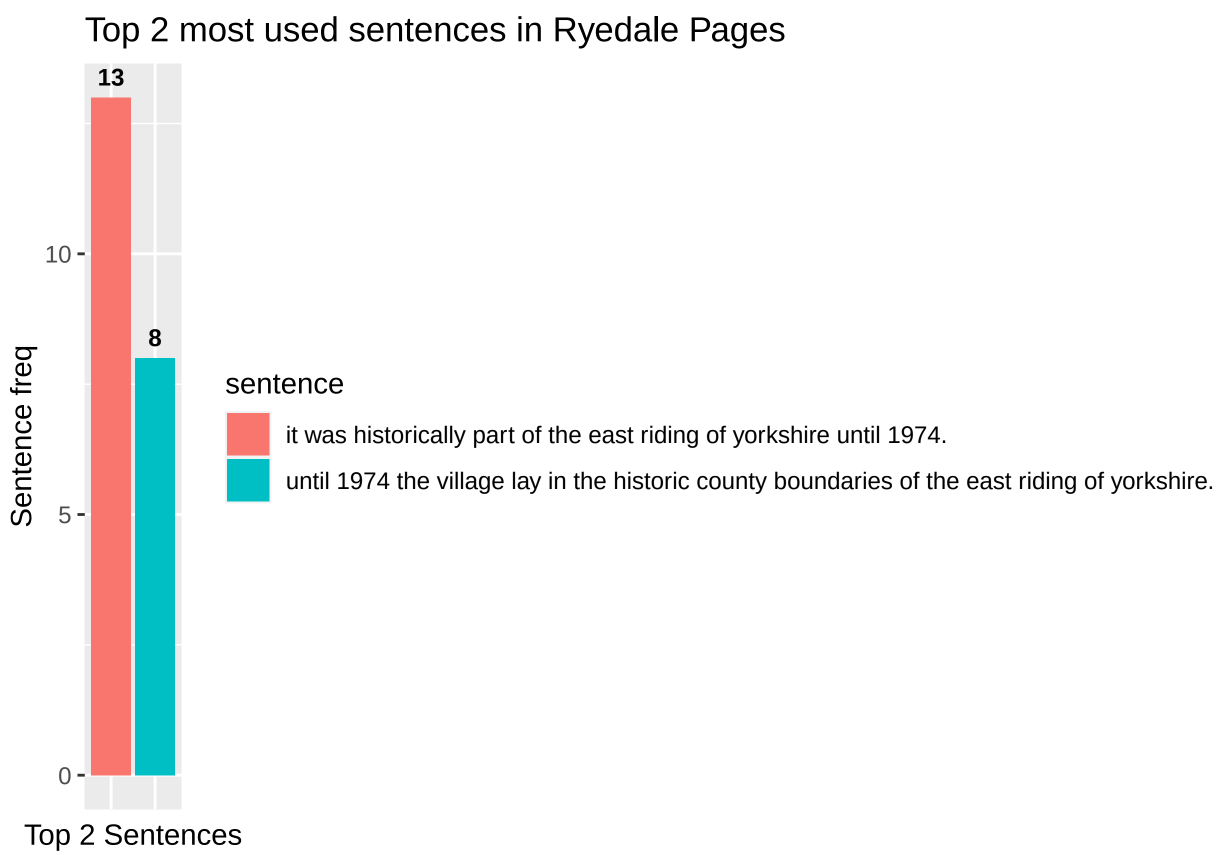
```
Setnc_freq <- Snt_n_pg %>%
  count(sentence, sort = TRUE)
```

There are 1504 unique sentences in Ryedale Wikipedia pages

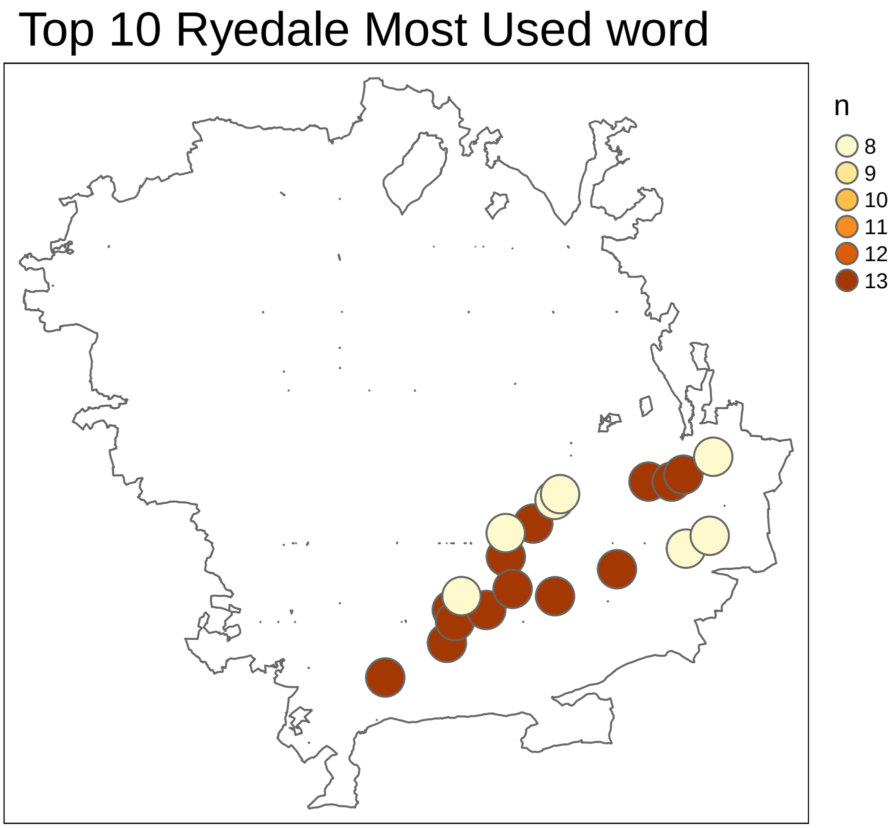
```
#Top 2 Sentences in Ryedale
tp_2_sent <- Setnc_freq %>%
  slice_max(n, n =2)

tp_2_sent %>%
  knitr::kable()
```

sentence	n	geometry
it was historically part of the east riding of yorkshire until 1974.	13	MULTIPOINT ((473116.6 46006...
until 1974 the village lay in the historic county boundaries of the east riding of yorkshire.	8	MULTIPOINT ((479709.8 46710...

```r
#Chart for top two Ryedale Sentences
tp_2_sent %>%
  ggplot(aes(sentence, n, fill = sentence)) +
  geom_col() +
  geom_text(aes(label = n), size = 3, fontface = "bold", vjust = -0.7) +
  labs(title = "Top 2 most used sentences in Ryedale Pages",
       x = "Top 2 Sentences", y = "Sentence freq") +
  theme(axis.ticks.x = element_blank(),
        axis.text.x = element_blank())
```




```r
#Plotting top 2 most frequent sentence in all Ryedale pages
boundary_m +
  tm_shape(tp_2_sent)+
  tm_bubbles( col = 'n', title.size="Word Freq")+
  tm_layout(main.title = 'Top 10 Ryedale Most Used word',
            title.size = 0.8,
            legend.outside=TRUE)
```



Places where these sentences are used are concentrated on the southern path of Ryedale.
The sentences are literately used to decribe Settrington which is located in the southern part if Ryedale.

Point Pattern Analysis of Top 10 word in Ryedale

```
RyedaleOwin <- as.owin(boundary)
class(RyedaleOwin)
```

```
## [1] "owin"
```

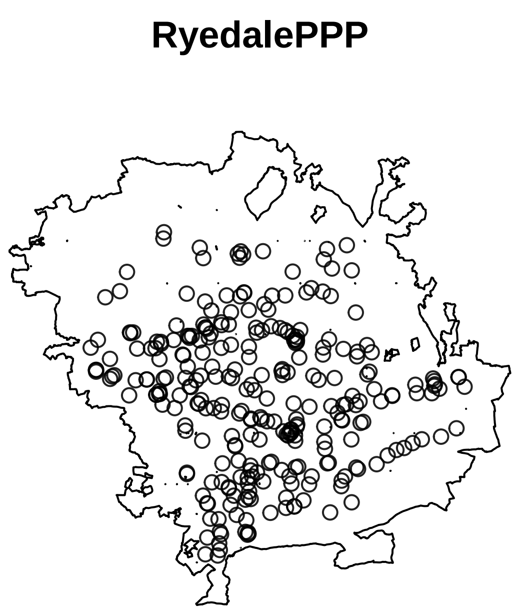
```
coord <- st_coordinates(top10_wds)[,c(1,2)]
RyedalePPP <- unique(ppp(coord[,1], coord[,2],
                         window = RyedaleOwin))
summary(RyedalePPP)
```

```
## Planar point pattern:  281 points
## Average intensity 1.185248e-07 points per square unit
##
## Coordinates are given to 2 decimal places
## i.e. rounded to the nearest multiple of 0.01 units
##
## Window: polygonal boundary
## 69 separate polygons (68 holes)
```

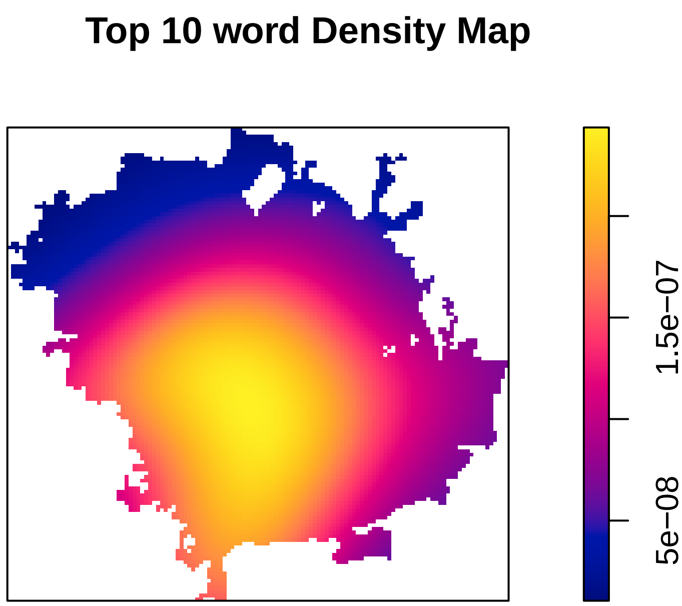
```
##                    vertices         area relative.area
## polygon 1             29896  2.39797e+09      1.01e+00
## polygon 2 (hole)       1473 -1.97676e+07     -8.34e-03
## polygon 3 (hole)        301 -2.14605e+06     -9.05e-04
## polygon 4 (hole)          3 -7.59565e-02     -3.20e-11
## polygon 5 (hole)          3 -5.47480e-01     -2.31e-10
## polygon 6 (hole)          3 -9.56565e-02     -4.03e-11
## polygon 7 (hole)        475 -7.28721e+05     -3.07e-04
## polygon 8 (hole)          3 -2.28694e-01     -9.65e-11
## polygon 9 (hole)          3 -9.14940e-02     -3.86e-11
## polygon 10 (hole)         3 -2.18432e-03     -9.21e-13
## polygon 11 (hole)         3 -9.89124e-03     -4.17e-12
## polygon 12 (hole)         3 -9.89123e-04     -4.17e-13
## polygon 13 (hole)         3 -5.08886e-09     -2.15e-18
## polygon 14 (hole)         3 -2.54411e-01     -1.07e-10
## polygon 15 (hole)         3 -6.12433e-02     -2.58e-11
## polygon 16 (hole)         3 -2.30177e-01     -9.71e-11
## polygon 17 (hole)         3 -1.97536e-01     -8.33e-11
## polygon 18 (hole)         3 -1.15810e-02     -4.88e-12
## polygon 19 (hole)         3 -1.71778e-01     -7.25e-11
## polygon 20 (hole)         3 -8.90212e-03     -3.75e-12
## polygon 21 (hole)         3 -9.47911e-03     -4.00e-12
## polygon 22 (hole)         3 -2.86846e-02     -1.21e-11
## polygon 23 (hole)         3 -3.10750e-02     -1.31e-11
## polygon 24 (hole)         3 -1.40456e-01     -5.92e-11
## polygon 25 (hole)         3 -2.61294e-02     -1.10e-11
## polygon 26 (hole)         3 -3.54436e-03     -1.49e-12
## polygon 27 (hole)         3 -9.34722e-02     -3.94e-11
## polygon 28 (hole)         3 -1.16346e-01     -4.91e-11
## polygon 29 (hole)         3 -6.61065e-02     -2.79e-11
## polygon 30 (hole)         3 -1.11276e-02     -4.69e-12
## polygon 31 (hole)         3 -3.24350e-02     -1.37e-11
## polygon 32 (hole)       165 -1.17695e+06     -4.96e-04
## polygon 33 (hole)       415 -1.37872e+06     -5.82e-04
## polygon 34 (hole)         3 -1.57765e-01     -6.65e-11
## polygon 35 (hole)         3 -8.98042e-02     -3.79e-11
## polygon 36 (hole)         3 -2.81900e-02     -1.19e-11
## polygon 37 (hole)         3 -8.18500e-02     -3.45e-11
## polygon 38 (hole)         3 -2.25026e-01     -9.49e-11
## polygon 39 (hole)         3 -2.46993e-01     -1.04e-10
## polygon 40 (hole)         3 -3.98947e-02     -1.68e-11
## polygon 41 (hole)         3 -4.15432e-02     -1.75e-11
## polygon 42 (hole)         3 -7.00630e-02     -2.96e-11
## polygon 43 (hole)         3 -2.15547e-02     -9.09e-12
## polygon 44 (hole)         3 -3.09101e-03     -1.30e-12
## polygon 45 (hole)         3 -2.32444e-02     -9.80e-12
## polygon 46 (hole)         3 -5.31654e-03     -2.24e-12
## polygon 47 (hole)         3 -1.37612e-01     -5.80e-11
## polygon 48 (hole)         3 -1.73097e-03     -7.30e-13
## polygon 49 (hole)         3 -5.30830e-02     -2.24e-11
## polygon 50 (hole)         3 -5.12284e-02     -2.16e-11
## polygon 51 (hole)         3 -1.48369e-03     -6.26e-13
## polygon 52 (hole)         3 -4.90441e-03     -2.07e-12
## polygon 53 (hole)         3 -1.11276e-02     -4.69e-12
```

```
## polygon 54 (hole)        3 -2.76130e-03     -1.16e-12
## polygon 55 (hole)        3 -4.87144e-02     -2.05e-11
## polygon 56 (hole)        3 -1.82988e-02     -7.72e-12
## polygon 57 (hole)        3 -5.07338e-02     -2.14e-11
## polygon 58 (hole)        5 -1.47646e+04     -6.23e-06
## polygon 59 (hole)        3 -3.63605e-10     -1.53e-19
## polygon 60 (hole)        3 -4.57470e-03     -1.93e-12
## polygon 61 (hole)        3 -1.05548e-01     -4.45e-11
## polygon 62 (hole)        3 -1.85461e-02     -7.82e-12
## polygon 63 (hole)        3 -1.40126e-02     -5.91e-12
## polygon 64 (hole)        3 -1.87934e-02     -7.93e-12
## polygon 65 (hole)        3 -1.13467e-02     -4.79e-12
## polygon 66 (hole)        3 -7.50498e-02     -3.17e-11
## polygon 67 (hole)      458 -1.94714e+06     -8.21e-04
## polygon 68 (hole)        3 -2.30795e-03     -9.73e-13
## polygon 69 (hole)        3 -6.40046e-02     -2.70e-11
## enclosing rectangle: [441503.8, 508368.7] x [448748.4, 511876.7] units
##                      (66860 x 63130 units)
## Window area = 2370810000 square units
## Fraction of frame area: 0.562
```

```r
#Top 10 Words Point Map
plot(RyedalePPP)
```

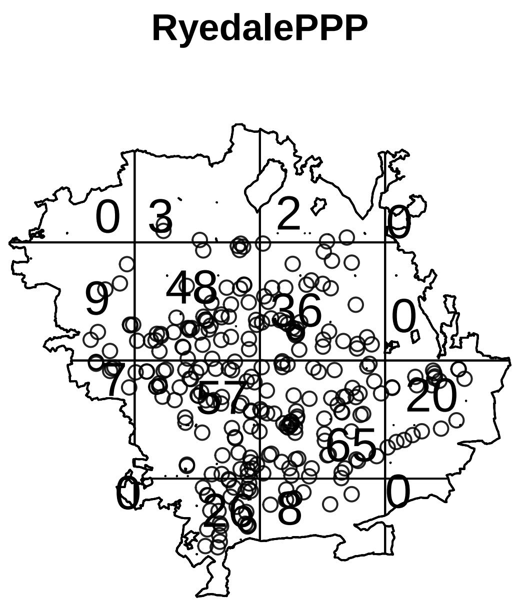


```r
# density plot
den <- density(RyedalePPP)
plot(den, main='Top 10 word Density Map')
```

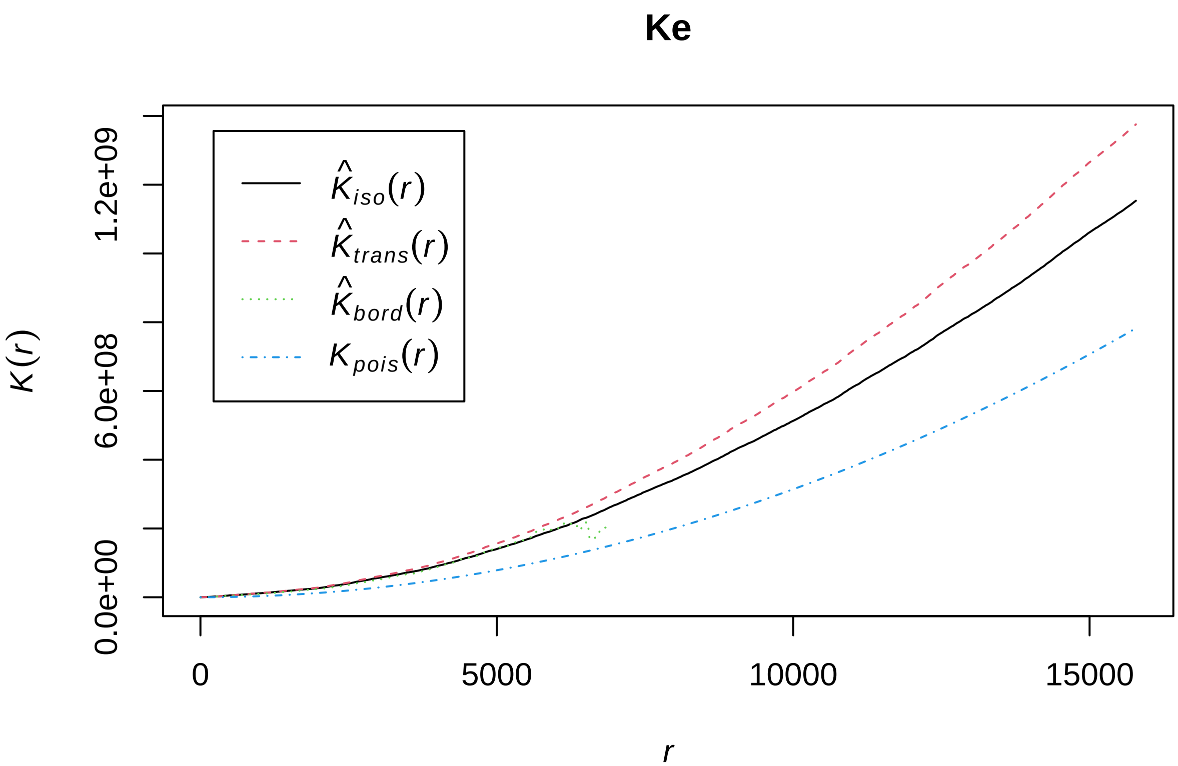


The above plot shows uneven distribution of the top 10 words with concentration at the particular point: from the central area to the south-western region of the area.

```r
#Quadrant point pattern analysis
quadrnt <-  quadratcount(RyedalePPP, nx = 4, ny = 4)
plot(RyedalePPP)
plot(quadrnt, add = TRUE, cex = 1.5)
```



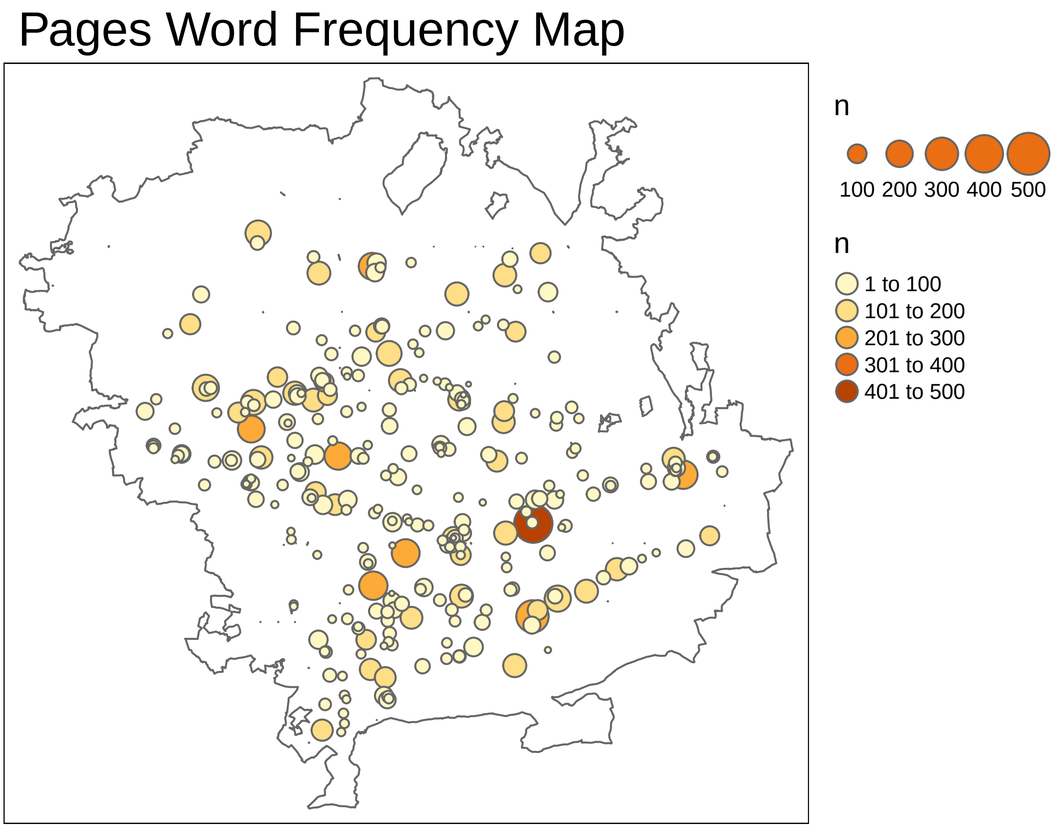

```
#The quadrant also confirms concentration of point at a particular area

Ke <- Kest(RyedalePPP)
plot(Ke)
```



While the K theoretical value for each radius under the assumption of complete randomness (Poisson) is represented by Kpois line and Kiso line represents the observed K values, we can say that the graph confirms clustering of points considering that the Kiso line is above the Kpois line.

Page-Words Frequency Analysis

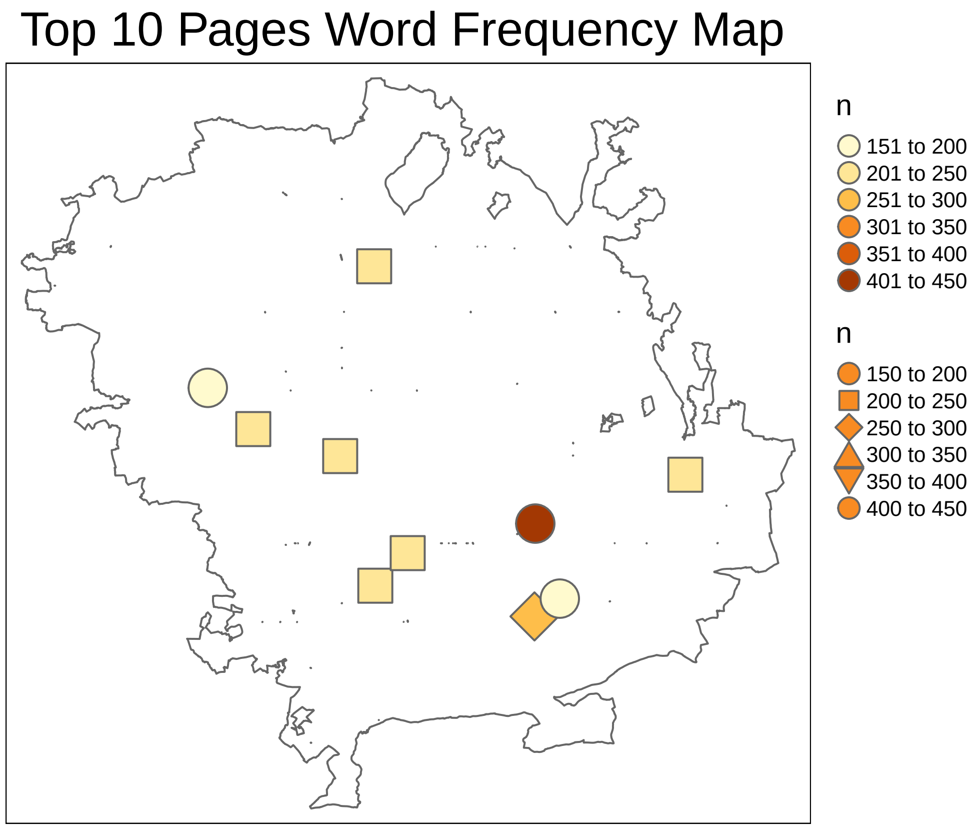
```
#Total word on each pages
wrds_on_pgs <- T_W_n_Pg %>%
  count(page_name, sort = TRUE)

Snt_on_pgs <- Snt_n_pg %>%
  count(page_name, sort = TRUE)

#Pages Word frequency maps
boundary_m +
  tm_shape(wrds_on_pgs) +
  tm_bubbles(col = "n", size = 'n')+
  tm_layout(main.title = 'Pages Word Frequency Map',
            title.size = 0.8,
            legend.outside=TRUE)
```



The map above indicates that the pattern of page words frequency is almost completely random. The map shows that there are more pages with lesser word frequency (between 0 and 200).

Top 10 Page-Words Frequency Analysis

```
boundary_m +
  tm_shape(wrds_on_pgs %>% slice_max(n, n=10)) +
  tm_bubbles(col = "n", shape = 'n')+
  tm_layout(main.title = 'Top 10 Pages Word Frequency Map',
            title.size = 0.8,
            legend.outside=TRUE)
```



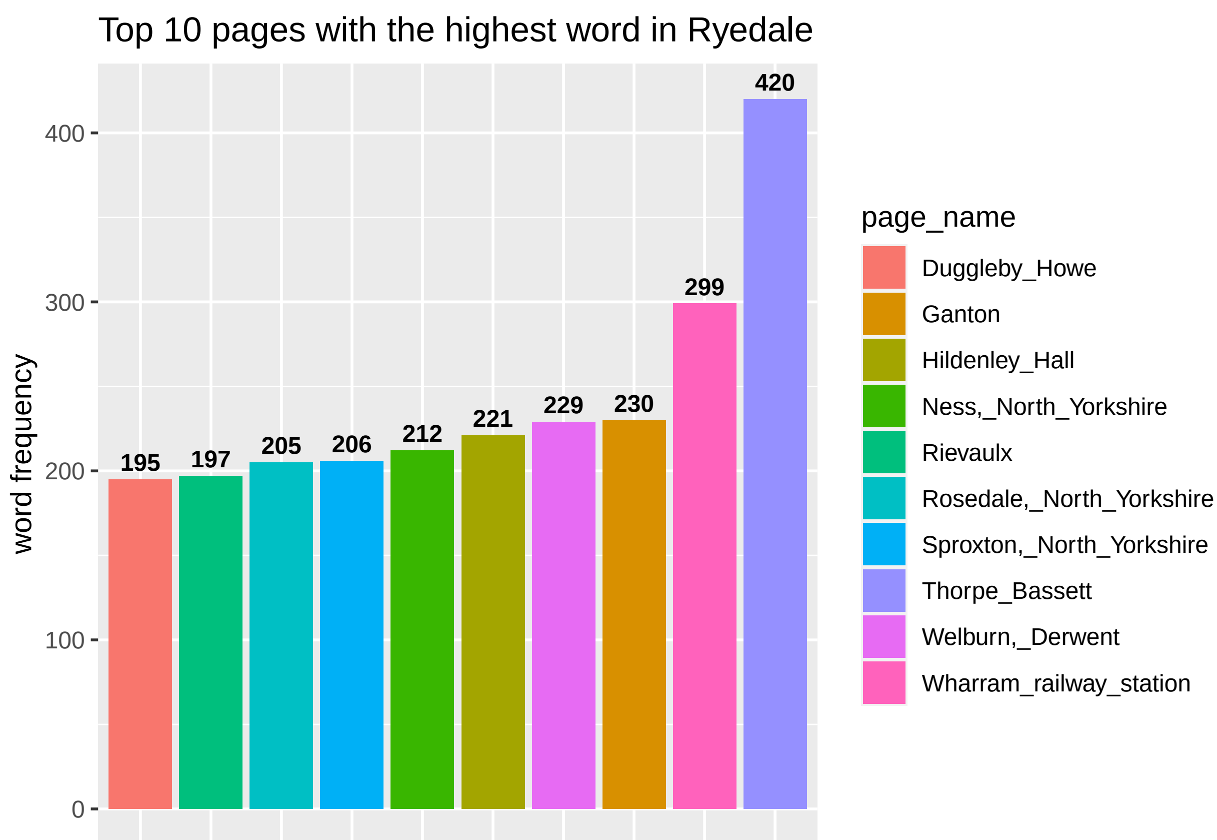
The top 10 pages with the highest words frequency are slightly clustered around a place (south-western region of the district).

```
wrds_on_pgs %>% slice_max(n, n=10) %>%
  knitr::kable()
```

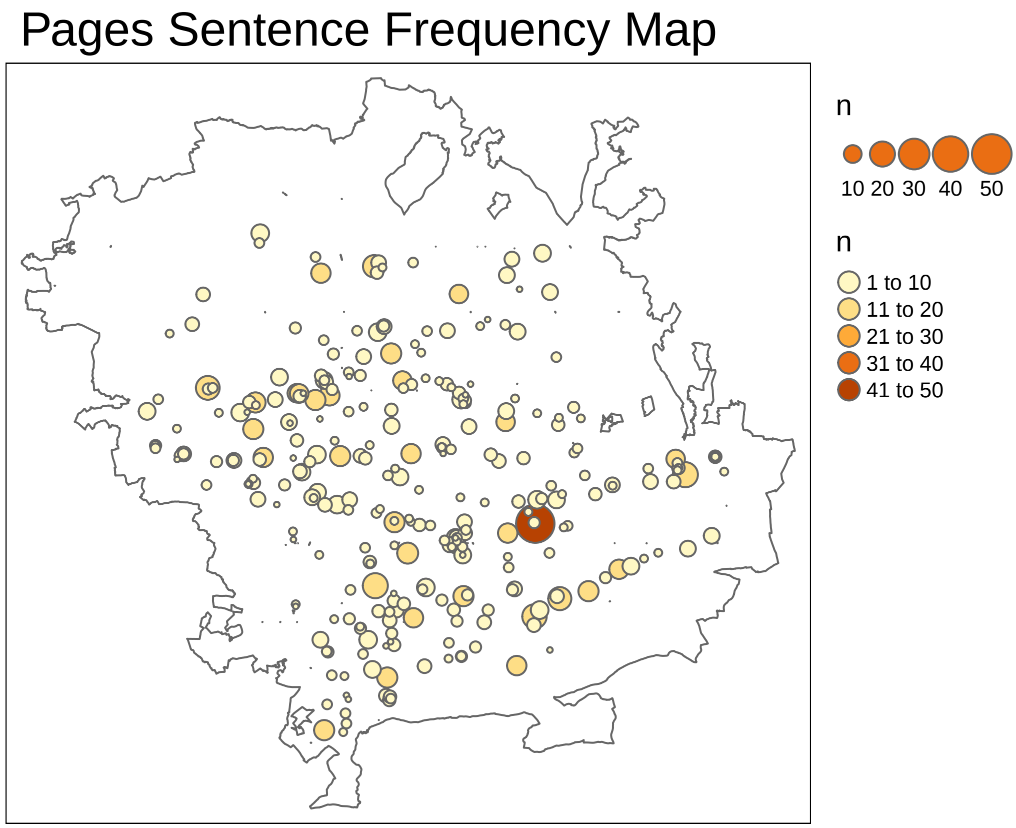
page_name	n	geometry
Thorpe_Bassett	420	POINT (485918.8 473374.6)
Wharram_railway_station	299	POINT (485850 465350.1)
Ganton	230	POINT (498899.3 477600)
Welburn,_Derwent	229	POINT (472083.9 468005.6)
Hildenley_Hall	221	POINT (474883.3 470821.1)
Ness,_North_Yorkshire	212	POINT (469045.2 479210.3)
Sproxton,_North_Yorkshire	206	POINT (461529.8 481545)
Rosedale,_North_Yorkshire	205	POINT (471985.1 495614.8)
Rievaulx	197	POINT (457593.9 485106.9)
Duggleby_Howe	195	POINT (488038.1 466890.3)

```
#Histogram of top 10 page-words frequency
wrds_on_pgs  %>%
  slice_max(n, n=10) %>%
  ggplot(aes(fct_reorder(page_name, n), n, fill = page_name)) +
  geom_col(show.legend = TRUE) +
  geom_text(aes(label = n), size = 3,
            fontface = "bold", vjust = -0.5) +
```

```r
  labs(title = "Top 10 pages with the highest word in Ryedale",
       y = "word frequency") +
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

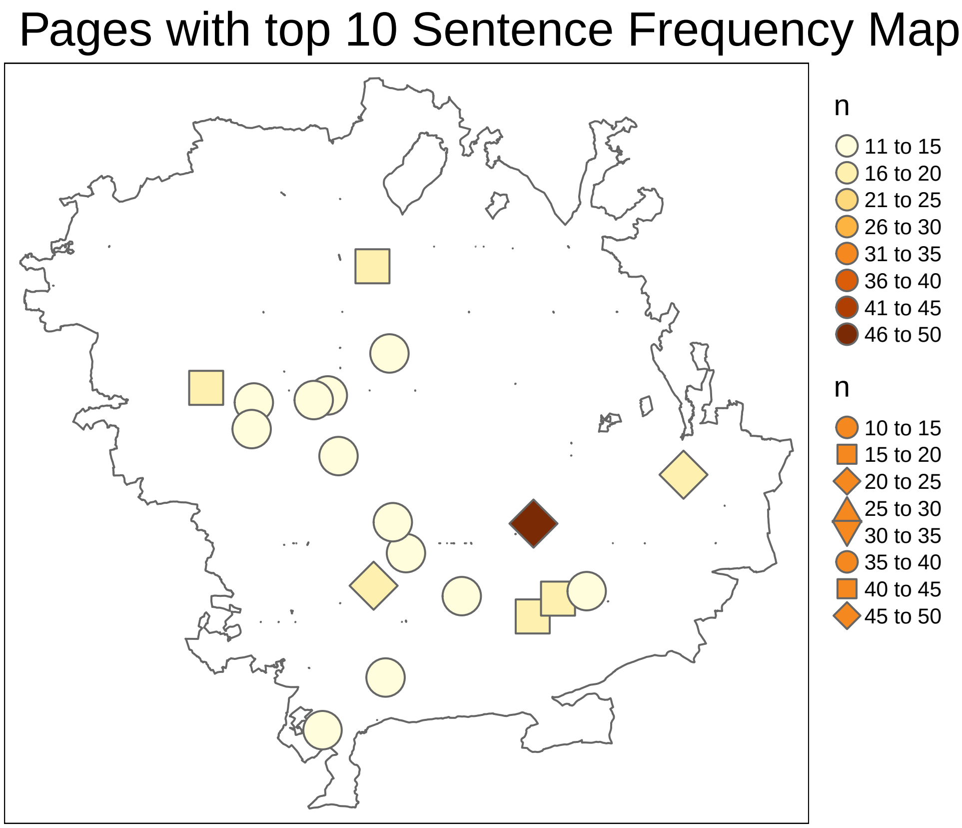



```r
#Pages sentence frequency map
boundary_m +
  tm_shape(Snt_on_pgs) +
  tm_bubbles(col = "n", size = 'n')+
  tm_layout(main.title = 'Pages Sentence Frequency Map',
            title.size = 0.7,
            legend.outside=TRUE)
```



The map above shows almost a random distribution of sentences frequency across the area.

```
boundary_m +
  tm_shape(Snt_on_pgs %>% slice_max(n, n=10)) +
  tm_bubbles(col = "n", shape = 'n')+
  tm_layout(main.title = 'Pages with top 10 Sentence Frequency Map', title.size = 0.7,
            legend.outside=TRUE)
```




```
#The top 10 map shows cluster of pages with 10 to 15 sentences.

Snt_on_pgs %>% slice_max(n, n=10) %>%
  knitr::kable()
```

page_name	n	geometry
Thorpe_Bassett	47	POINT (485918.8 473374.6)
Ganton	20	POINT (498899.3 477600)
Welburn,_Derwent	20	POINT (472083.9 468005.6)
Wharram_railway_station	19	POINT (485850 465350.1)
Rievaulx	18	POINT (457593.9 485106.9)
Duggleby_Howe	17	POINT (488038.1 466890.3)
Rosedale,_North_Yorkshire	16	POINT (471985.1 495614.8)
Hildenley_Hall	14	POINT (474883.3 470821.1)
Appleton-le-Moors	13	POINT (473453.1 488080.3)
Appleton-le-Street_with_Easthorpe	13	POINT (473743.7 473493.8)
Helmsley	13	POINT (461719.1 483844.1)
Kirby_Grindalythe	13	POINT (490520.9 467534.7)
Langton,_North_Yorkshire	13	POINT (479709.8 467104.5)
Ness,_North_Yorkshire	13	POINT (469045.2 479210.3)
Scrayingham	13	POINT (473116.6 460069.8)
Sproxton,_North_Yorkshire	13	POINT (461529.8 481545)
Warthill	13	POINT (467661.4 455519.8)
Welburn,_Kirkbymoorside	13	POINT (468120.3 484451.5)

page_name	n	geometry
Wombleton	13	POINT (466901.6 484051.9)

The pages with the highest sentence frequency are different from that of word frequency.

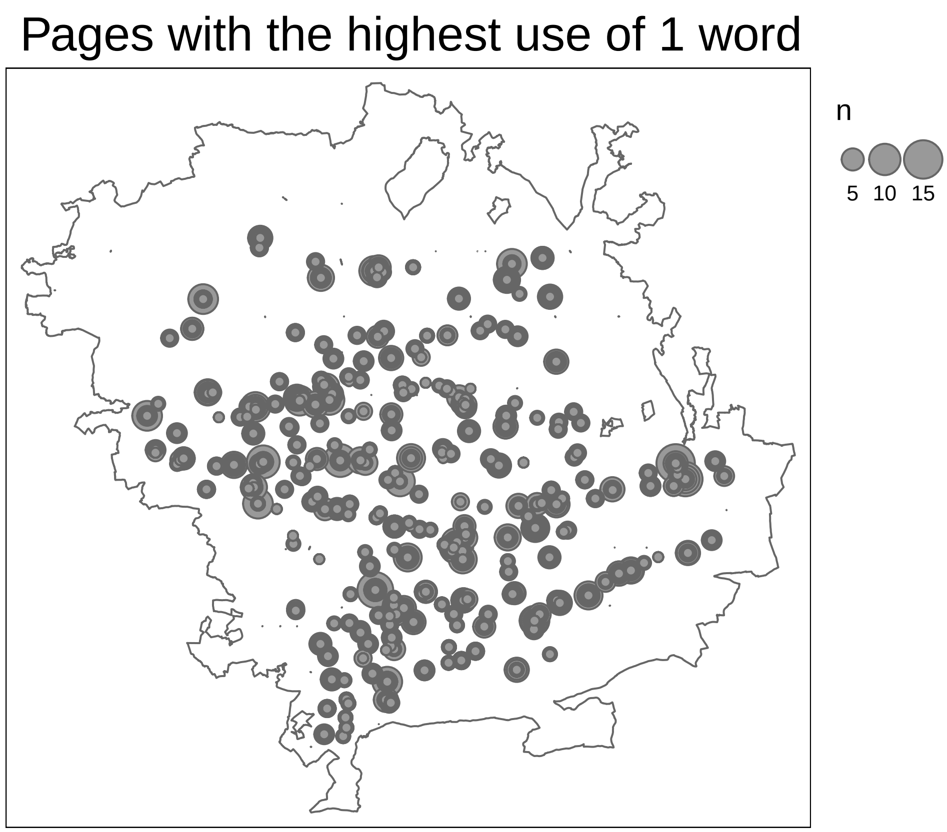
Per page word frequency

```
word_per_pg <- T_W_n_Pg %>%
  count(page_name, word,  sort = TRUE)

#Per page Sentence Frequency
Snt_per_pg <- Snt_n_pg %>%
  count(page_name, sentence, sort = TRUE)

print(Snt_per_pg %>% filter(n > 1))
```

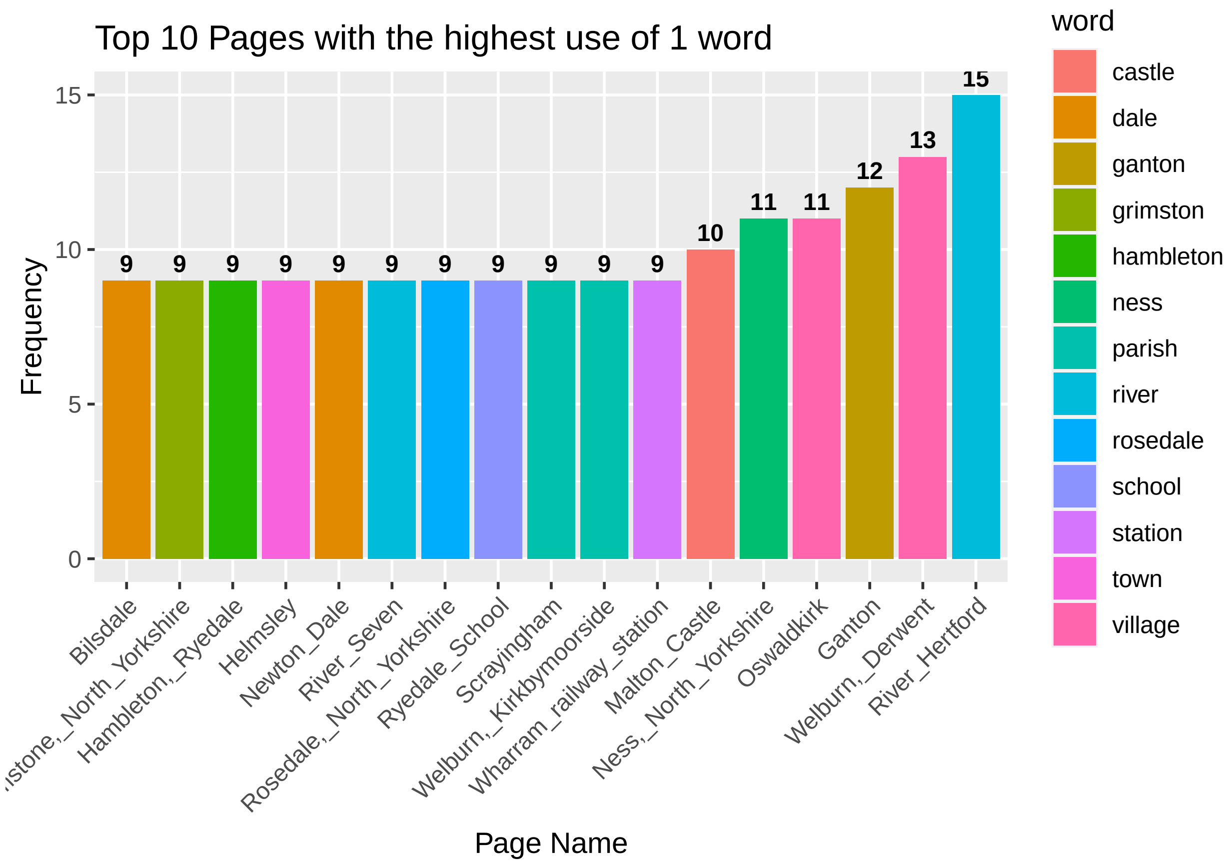
```
## Simple feature collection with 0 features and 3 fields
## Bounding box:  xmin: NA ymin: NA xmax: NA ymax: NA
## Projected CRS: OSGB 1936 / British National Grid
## [1] page_name sentence  n         geometry
## <0 rows> (or 0-length row.names)
```

The above result signifies that there is not two pages sharing the same sentence.

```
boundary_m +
  tm_shape(word_per_pg) +
  tm_bubbles(size = 'n')+
  tm_layout(main.title = 'Pages with the highest use of 1 word',
            title.size = 0.7,
            legend.outside=TRUE)
```




```
#The map displays random pattern.


word_per_pg %>%
  slice_max(n, n=10) %>%
  ggplot(aes(reorder(page_name, n), n, fill = word)) +
  geom_col(show.legend = TRUE) +
  geom_text(aes(label = n), size = 3, fontface = "bold",
            vjust = -0.5) +
  labs(title = "Top 10 Pages with the highest use of 1 word",
       x = 'Page Name',
       y = "Frequency") +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```




```
word_per_pg %>%
  slice_max(n, n=10) %>%
  knitr::kable()
```

page_name	word	n	geometry
River_Hertford	river	15	POINT (498060.2 478963.4)
Welburn,_Derwent	village	13	POINT (472083.9 468005.6)
Ganton	ganton	12	POINT (498899.3 477600)
Ness,_North_Yorkshire	ness	11	POINT (469045.2 479210.3)
Oswaldkirk	village	11	POINT (462392.3 479093.3)
Malton_Castle	castle	10	POINT (479031.5 471656.9)
Bilsdale	dale	9	POINT (457193.7 493181)
Grimstone,_North_Yorkshire	grimston	9	POINT (461935.2 475477.2)
Hambleton,_Ryedale	hambleton	9	POINT (452348.1 483078)
Helmsley	town	9	POINT (461719.1 483844.1)
Newton_Dale	dale	9	POINT (483900.3 496231.5)
River_Seven	river	9	POINT (474215.5 477398.7)
Rosedale,_North_Yorkshire	rosedale	9	POINT (471985.1 495614.8)
Ryedale_School	school	9	POINT (465543.2 484389.6)
Scrayingham	parish	9	POINT (473116.6 460069.8)
Welburn,_Kirkbymoorside	parish	9	POINT (468120.3 484451.5)
Wharram_railway_station	station	9	POINT (485850 465350.1)

Most of the pages which one word was repeated used teh words nine times in the above analysis.

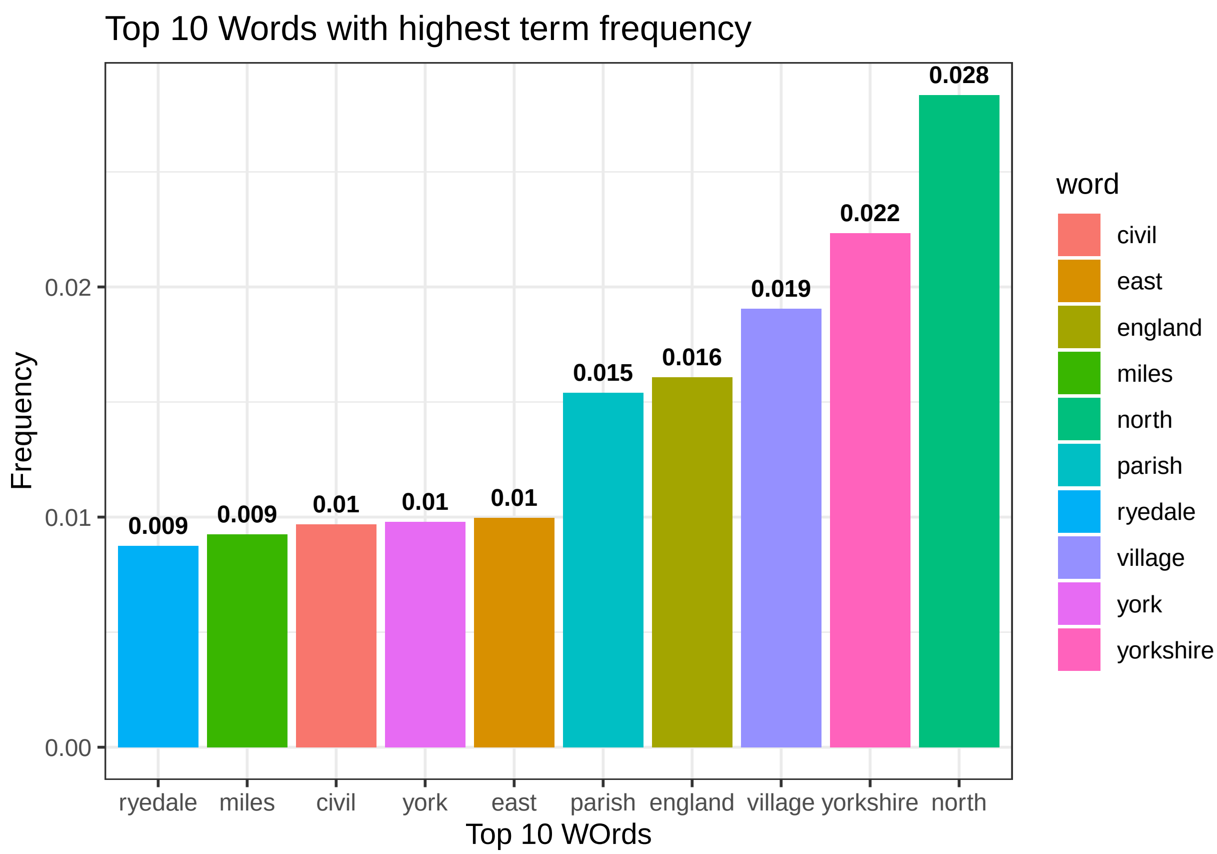
Term Freqency Analysis

```
#Per word term frequency analysis
term_words <- word_freq %>%
  mutate(term = n/sum(n))

top10_termwords <- term_words %>%
  slice_max(n, n = 10)

top10_termwords %>%
  knitr::kable()
```

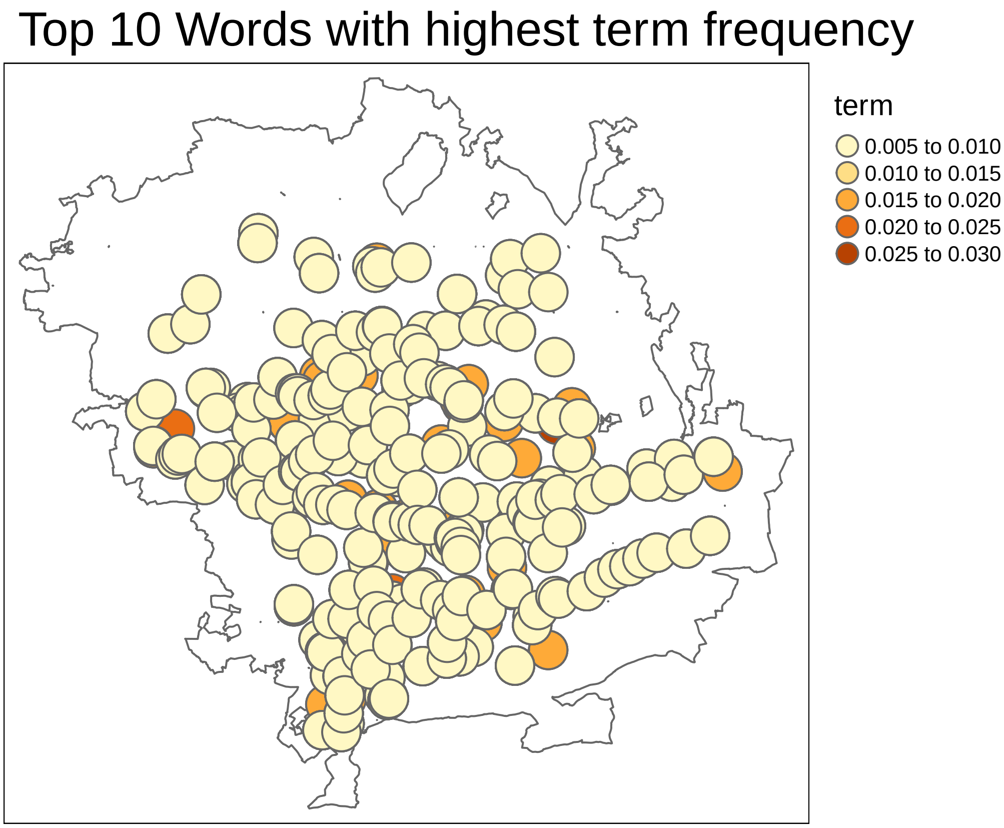
word	n	term	geometry
north	515	0.0283450	MULTIPOINT ((452348.1 48307...
yorkshire	406	0.0223458	MULTIPOINT ((452348.1 48307...
village	346	0.0190434	MULTIPOINT ((453045.3 48008...
england	292	0.0160713	MULTIPOINT ((452348.1 48307...
parish	280	0.0154109	MULTIPOINT ((452348.1 48307...
east	181	0.0099620	MULTIPOINT ((452348.1 48307...
york	178	0.0097969	MULTIPOINT ((452348.1 48307...
civil	176	0.0096868	MULTIPOINT ((452348.1 48307...
miles	168	0.0092465	MULTIPOINT ((453309.6 48412...
ryedale	159	0.0087512	MULTIPOINT ((452348.1 48307...

Term frequency analysis also reveales that North and Yorkshire remain the most important words.

```
top10_termwords %>%
  ggplot(aes(fct_reorder(word,term), term, fill = word)) +
  geom_col() +
  geom_text(aes(label = round(term, 3)), size = 3,
            fontface = "bold", vjust = -0.7) +
  labs(title = "Top 10 Words with highest term frequency",
       x = "Top 10 WOrds", y = "Frequency") +
  theme_bw()
```




```
boundary_m +
  tm_shape(top10_termwords)+
  tm_bubbles( col = 'term')+
  tm_layout(main.title = 'Top 10 Words with highest term frequency',
            title.size = 0.7,
            legend.outside=TRUE)
```



The pattern in the map reveals cluster of lesser term frequency words

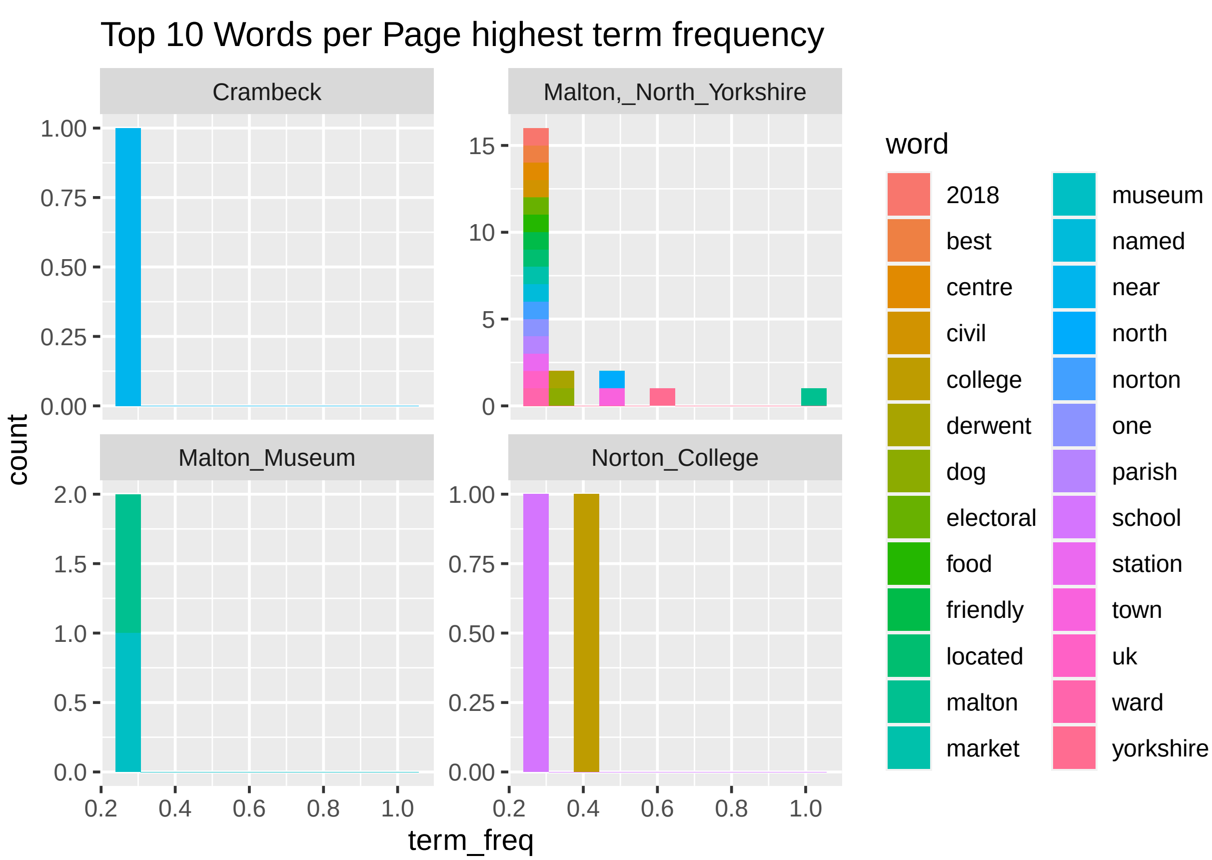
#Words per page Term frequency

```
sum_total_word <- word_per_pg %>%
  group_by(page_name) %>%
  summarize(total = sum(n)) %>%
  arrange(desc(total))

#adding sum column
word_per_pg1 <- st_join(word_per_pg, sum_total_word, left = TRUE)

#Creating term frequency
word_per_pg1 %>%
  mutate(term_freq = n/total) %>%
  slice_max(term_freq, n= 10)
```

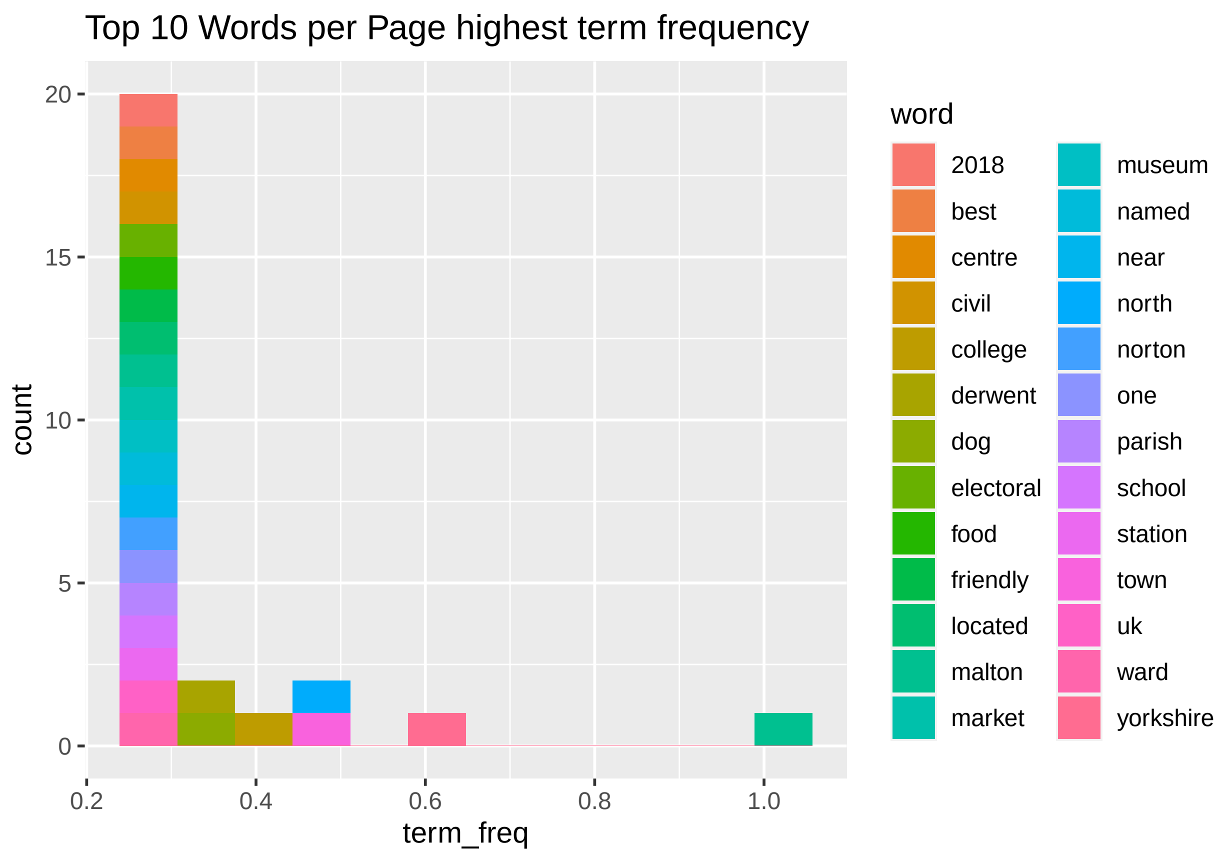
```
## Simple feature collection with 27 features and 6 fields
## Geometry type: POINT
## Dimension:     XY
## Bounding box:  xmin: 473682.8 ymin: 467340.8 xmax: 479644.7 ymax: 472136.4
## Projected CRS: OSGB 1936 / British National Grid
## First 10 features:
##                      page_name.x      word n   page_name.y total term_freq
## 22.1    Malton,_North_Yorkshire    malton 8 Malton_Museum     8     1.000
## 106.1   Malton,_North_Yorkshire yorkshire 5 Malton_Museum     8     0.625
## 215.1   Malton,_North_Yorkshire     north 4 Malton_Museum     8     0.500
```

```
## 216.1  Malton,_North_Yorkshire      town 4 Malton_Museum     8      0.500
## 23.1             Norton_College   college 8 Malton_School    20      0.400
## 514.1  Malton,_North_Yorkshire   derwent 3 Malton_Museum     8      0.375
## 515.1  Malton,_North_Yorkshire       dog 3 Malton_Museum     8      0.375
## 111.1            Norton_College    school 5 Malton_School    20      0.250
## 1002                   Crambeck      near 2      Crambeck     8      0.250
## 1462.1 Malton,_North_Yorkshire      2018 2 Malton_Museum     8      0.250
##                          geometry
## 22.1     POINT (479018 472136.4)
## 106.1    POINT (479018 472136.4)
## 215.1    POINT (479018 472136.4)
## 216.1    POINT (479018 472136.4)
## 23.1   POINT (479644.7 470655.8)
## 514.1    POINT (479018 472136.4)
## 515.1    POINT (479018 472136.4)
## 111.1  POINT (479644.7 470655.8)
## 1002   POINT (473682.8 467340.8)
## 1462.1   POINT (479018 472136.4)
```

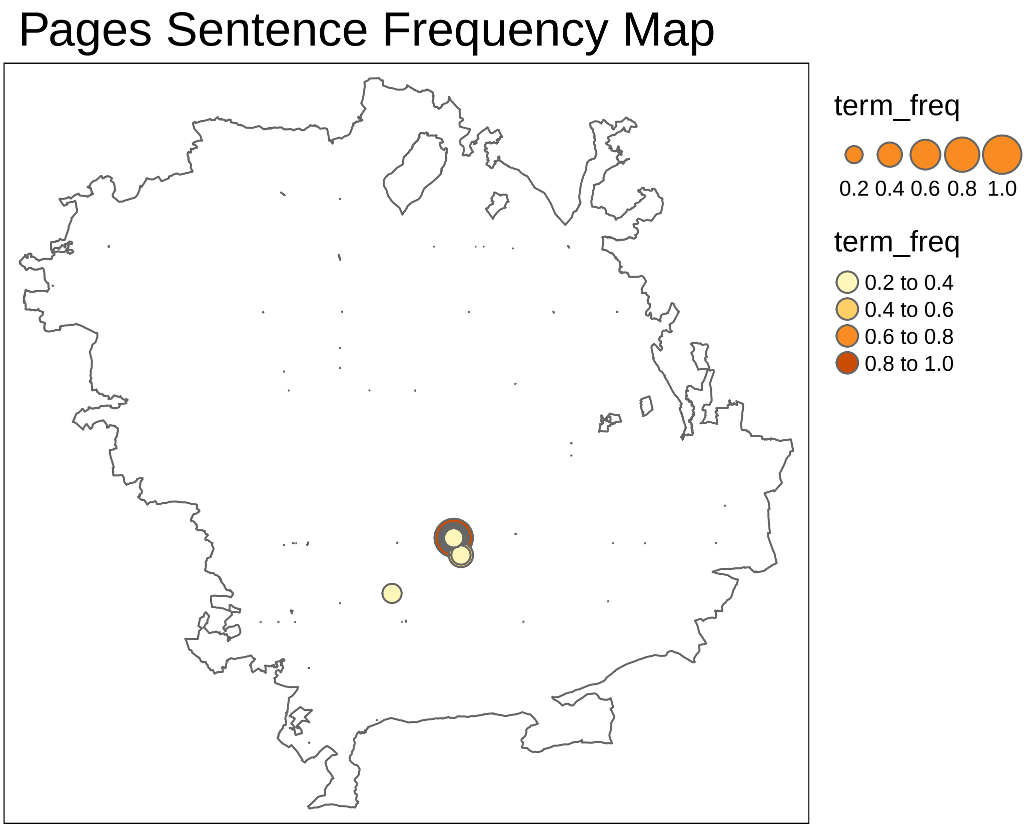
```
word_per_pg1 %>%
  mutate(term_freq = n/total) %>%
  slice_max(term_freq, n= 10) %>%
  ggplot( aes(term_freq,  fill = word)) +
  geom_histogram(show.legend = TRUE, bins = 12) +
  facet_wrap(~page_name.x, ncol = 2, scales = "free_y")+
  labs(title = "Top 10 Words per Page highest term frequency")
```



```
word_per_pg1 %>%
  mutate(term_freq = n/total) %>%
  slice_max(term_freq, n= 10) %>%
  ggplot( aes(term_freq,  fill = word)) +
  geom_histogram(show.legend = TRUE, bins = 12) +
  labs(title = "Top 10 Words per Page highest term frequency")
```




```
#Term Frequency: Malton is the most important word per page term frequency

term_tb <- word_per_pg1 %>%
  mutate(term_freq = n/total) %>%
  slice_max(term_freq, n= 10)

boundary_m +
  tm_shape(term_tb) +
  tm_bubbles(size = 'term_freq', col = 'term_freq')+
  tm_layout(main.title = 'Pages Sentence Frequency Map', title.size = 0.7,
            legend.outside=TRUE)
```



This is an indication that most of the high ranked word per term frequency ranking in are located in almost the same area. The word are located in the Malton_Museum pages, Malton School and Crambeck. No doubt, the word are related to the pages.

Tfidf analysis

```
tfidf <- word_per_pg %>%
  bind_tf_idf(word, page_name, n) %>%
  arrange(desc(tf_idf))

top10_tfidf <- tfidf %>%
  slice_max(tf_idf, n=10)

top10_tfidf %>% knitr::kable()
```

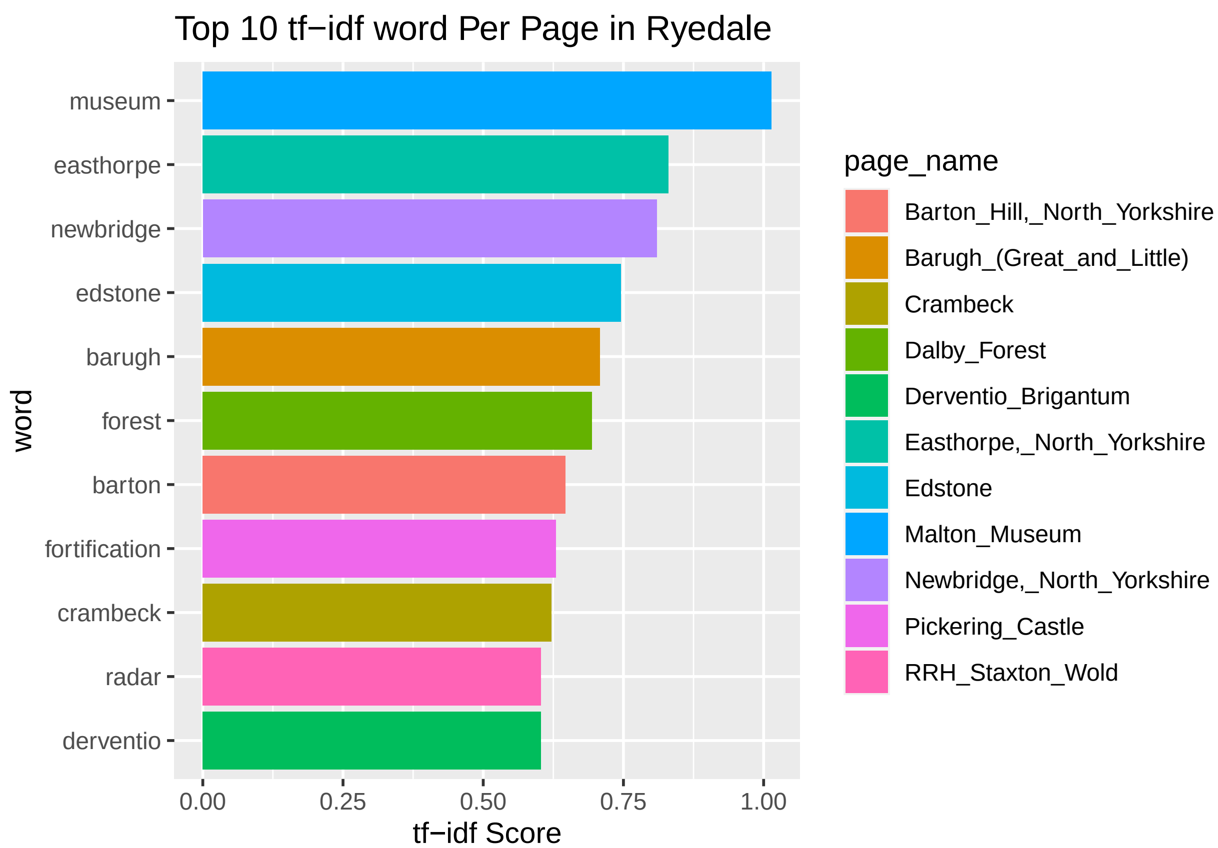
page_name	word	n	tf	idf	tf_idf	geometry
Malton_Museum	museum	2	0.2500000	4.056989	1.0142472	POINT (479018 472136.4)
Easthorpe,_North_Yorkshire	easthorpe	2	0.1818182	4.567814	0.8305117	POINT (473734 471490.4)
Newbridge,_North_Yorkshire	newbridge	1	0.1428571	5.666427	0.8094895	POINT (480320 485427)

page_name	word	n	tf	idf	tf_idf	geometry
Edstone	edstone	3	0.1500000	4.973280	0.7459919	POINT (471067.3 483470)
Barugh_(Great_and_Little)	barugh	8	0.1250000	5.666427	0.7083033	POINT (475174.7 479417.4)
Dalby_Forest	forest	6	0.1621622	4.280132	0.6940755	POINT (487738.6 487764)
Barton_Hill,_North_Yorkshire	barton	4	0.1739130	3.720517	0.6470464	POINT (470774.9 464465.5)
Pickering_Castle	fortification	1	0.1111111	5.666427	0.6296030	POINT (479878.3 484504.6)
Crambeck	crambeck	1	0.1250000	4.973280	0.6216599	POINT (473682.8 467340.8)
Derventio_Brigantum	derventio	4	0.1212121	4.973280	0.6028218	POINT (479133.8 471825.6)
RRH_Staxton_Wold	radar	4	0.1212121	4.973280	0.6028218	POINT (502259.2 477866.7)

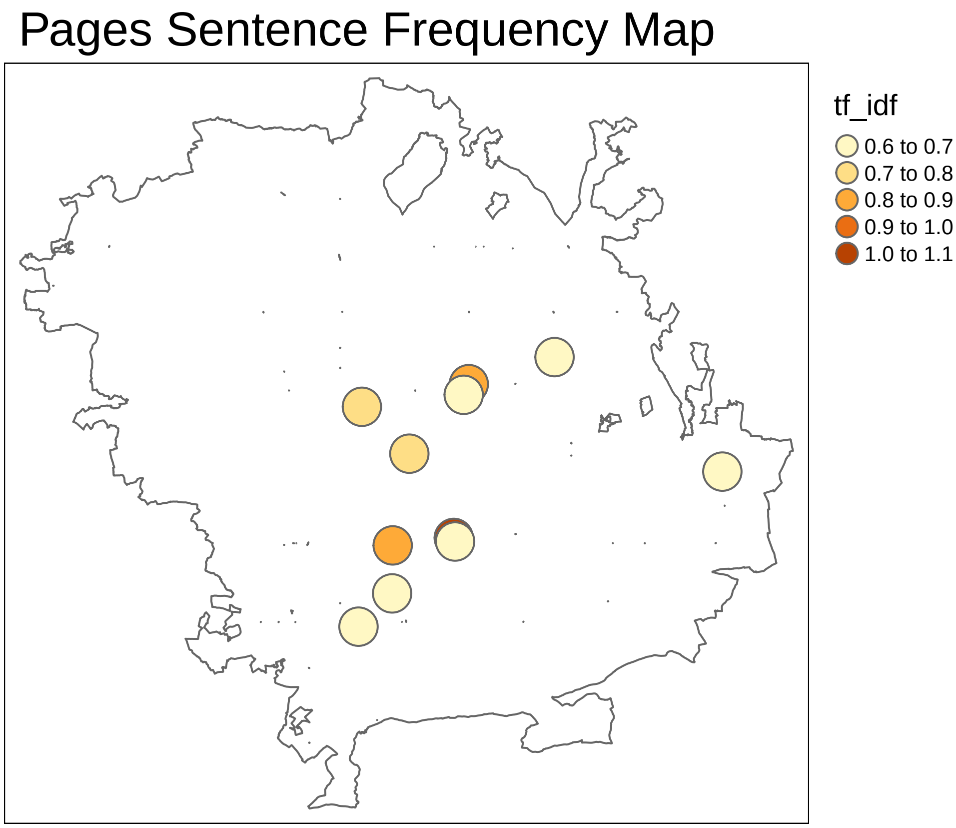
Museum is also among the high rank word in the tf_idf analysis

Top 10 Tfidf analysis

```
top10_tfidf %>%
  ggplot(aes(tf_idf, fct_reorder(word, tf_idf), fill = page_name)) +
  geom_col() +
  labs(title = 'Top 10 tf-idf word Per Page in Ryedale',
       x = 'tf-idf Score', y = 'word')
```




```
boundary_m +
  tm_shape(top10_tfidf) +
  tm_bubbles( col = 'tf_idf')+
  tm_layout(main.title = 'Pages Sentence Frequency Map',
            title.size = 0.7,
            legend.outside=TRUE)
```




```
tfidf %>%
  group_by(page_name) %>%
  slice_max(tf_idf, n=1) %>%
  ungroup()
```

```
## Simple feature collection with 362 features and 6 fields
## Geometry type: POINT
## Dimension:     XY
## Bounding box:  xmin: 452348.1 ymin: 455354.8 xmax: 502259.2 ymax: 498475.6
## Projected CRS: OSGB 1936 / British National Grid
## # A tibble: 362 x 7
##    page_name       word                n     tf   idf tf_idf             geometry
##    <chr>           <chr>           <int>  <dbl> <dbl>  <dbl>          <POINT [m]>
##  1 1949_Ryder_Cup  u.s                 3 0.0353  5.67  0.200 (498299.4 478182.6)
##  2 2000_Curtis_Cup ganton              2 0.0952  3.47  0.330 (498299.4 478182.6)
##  3 2003_Walker_Cup ireland             2 0.0588  5.67  0.333 (498225.3 477980.7)
##  4 A170_road       avoided             1 0.0213  5.67  0.121 (475220.7 485394.4)
##  5 A170_road       drovers             1 0.0213  5.67  0.121 (475220.7 485394.4)
##  6 A170_road       enough              1 0.0213  5.67  0.121 (475220.7 485394.4)
##  7 A170_road       folk                1 0.0213  5.67  0.121 (475220.7 485394.4)
##  8 A170_road       kirkbysmoorside     1 0.0213  5.67  0.121 (475220.7 485394.4)
##  9 A170_road       paying              1 0.0213  5.67  0.121 (475220.7 485394.4)
## 10 A170_road       prehistoric         1 0.0213  5.67  0.121 (475220.7 485394.4)
## # ... with 352 more rows
```

```
dim(sentcs_w_pgs %>%
  filter(str_detect(a_page_summary, 'u.s')) %>%
  select(a_page_summary) )
```

```
## [1] 61  2
```

The word 'u.s' appears in 61 pages. We can see 'u.s' in the 1949_Ryder_Cup page. Let's investigate.

Bigrams Analysis

```
Ryedale_bigrams <- sentcs_w_pgs %>%
  unnest_tokens(bigrams, a_page_summary, token = 'ngrams', n = 2)


Ryedale_bigrams[c('word1', 'word2')] <- str_split_fixed(
  string = Ryedale_bigrams$bigrams, pattern = " ", n=2
)

anti_Ryedale_bigrams <- Ryedale_bigrams %>%
  anti_join(get_stopwords(), c('word1' = 'word')) %>%
  anti_join(get_stopwords(), c('word2' = 'word'))

count_bigram <- anti_Ryedale_bigrams %>%
  count(word1, word2, sort = TRUE)

count_bigram
```

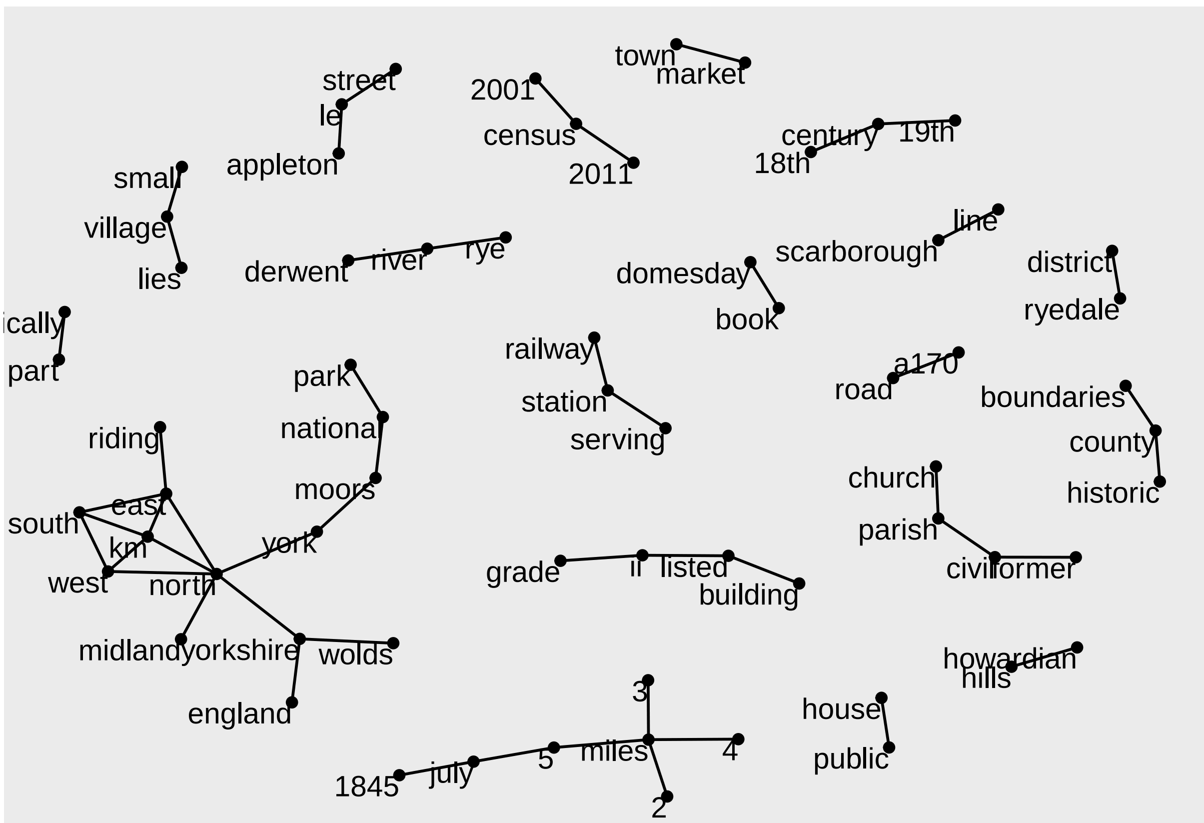
```
## Simple feature collection with 5812 features and 3 fields
## Geometry type: GEOMETRY
## Dimension:     XY
## Bounding box:  xmin: 452348.1 ymin: 455354.8 xmax: 502259.2 ymax: 498475.6
## Projected CRS: OSGB 1936 / British National Grid
## First 10 features:
##        word1     word2   n                       geometry
## 1      north yorkshire 295 MULTIPOINT ((452348.1 48307...
## 2  yorkshire   england 244 MULTIPOINT ((452348.1 48307...
## 3      civil    parish 172 MULTIPOINT ((452348.1 48307...
## 4    ryedale  district 136 MULTIPOINT ((452348.1 48307...
## 5      north      york  78 MULTIPOINT ((453045.3 48008...
## 6       york     moors  78 MULTIPOINT ((453045.3 48008...
## 7    railway   station  61 MULTIPOINT ((457483.3 47671...
## 8   national      park  58 MULTIPOINT ((453045.3 48008...
## 9       east    riding  57 MULTIPOINT ((469306.6 45535...
## 10     moors  national  55 MULTIPOINT ((453045.3 48008...
```

```
top10_count_bigram <- count_bigram %>%
  filter(n > 10)

top10_count_bigram %>%
  knitr::kable()
```

word1	word2	n	geometry
north	yorkshire	295	MULTIPOINT ((452348.1 48307...
yorkshire	england	244	MULTIPOINT ((452348.1 48307...
civil	parish	172	MULTIPOINT ((452348.1 48307...
ryedale	district	136	MULTIPOINT ((452348.1 48307...
north	york	78	MULTIPOINT ((453045.3 48008...
york	moors	78	MULTIPOINT ((453045.3 48008...
railway	station	61	MULTIPOINT ((457483.3 47671...
national	park	58	MULTIPOINT ((453045.3 48008...
east	riding	57	MULTIPOINT ((469306.6 45535...
moors	national	55	MULTIPOINT ((453045.3 48008...
2011	census	46	MULTIPOINT ((453309.6 48412...
km	north	42	MULTIPOINT ((454299.8 48979...
grade	ii	37	MULTIPOINT ((461187.2 48387...
ii	listed	34	MULTIPOINT ((461187.2 48387...
north	east	29	MULTIPOINT ((463796.7 48603...
km	south	26	MULTIPOINT ((457193.7 49318...
small	village	26	MULTIPOINT ((454944.2 47892...
2001	census	25	MULTIPOINT ((455214.3 47924...
listed	building	22	MULTIPOINT ((461187.2 48387...
historically	part	21	MULTIPOINT ((461719.1 48384...
le	street	21	MULTIPOINT ((472185.1 47429...
river	derwent	21	MULTIPOINT ((470799.3 47922...
km	west	19	MULTIPOINT ((453309.6 48412...
south	west	18	MULTIPOINT ((460402.5 48297...
4	miles	17	MULTIPOINT ((462392.3 47909...
parish	church	17	MULTIPOINT ((461187.2 48387...
river	rye	17	MULTIPOINT ((457193.7 49318...
scarborough	line	17	MULTIPOINT ((467337.5 46333...
yorkshire	wolds	17	MULTIPOINT ((464999.9 47200...
km	east	16	MULTIPOINT ((461719.1 48384...
howardian	hills	15	MULTIPOINT ((461479.8 47692...
village	lies	15	MULTIPOINT ((455505.4 47939...
2	miles	14	MULTIPOINT ((461479.8 47692...
3	miles	14	MULTIPOINT ((468457.4 48803...
domesday	book	13	MULTIPOINT ((461935.2 47547...
market	town	13	MULTIPOINT ((461003.2 48300...
18th	century	12	MULTIPOINT ((458022 485103....
5	miles	12	MULTIPOINT ((453309.6 48412...
a170	road	12	MULTIPOINT ((452348.1 48307...
july	1845	12	MULTIPOINT ((467337.5 46333...
north	midland	12	MULTIPOINT ((470800.1 46432...
north	west	12	MULTIPOINT ((454299.8 48979...
station	serving	12	MULTIPOINT ((465860.1 48464...
19th	century	11	MULTIPOINT ((462130 498475....
5	july	11	MULTIPOINT ((470800.1 46432...
appleton	le	11	MULTIPOINT ((472185.1 47429...
county	boundaries	11	MULTIPOINT ((479709.8 46710...
former	civil	11	MULTIPOINT ((468568.8 48053...
historic	county	11	MULTIPOINT ((479709.8 46710...
public	house	11	MULTIPOINT ((466901.6 48405...
south	east	11	MULTIPOINT ((462130 498475....

Creating igraph object

```
#visualizing relationship between word
bigram_graph <- top10_count_bigram%>%
  graph_from_data_frame()


ggraph(bigram_graph, layout = "fr") +
  geom_edge_link() +
  geom_node_point() +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1)
```




```
#Source: https://www.tidytextmining.com/ngrams.html
```

The word 'north' had most connection with other word and forms most centre nodes, which are connected bt the name of places. 'km' and miles also have high connections. However, most word that connect with 'km' are also interconnected, hence, would likely form a cycle as part of phrases.

3. 0 Sentiment Analysis

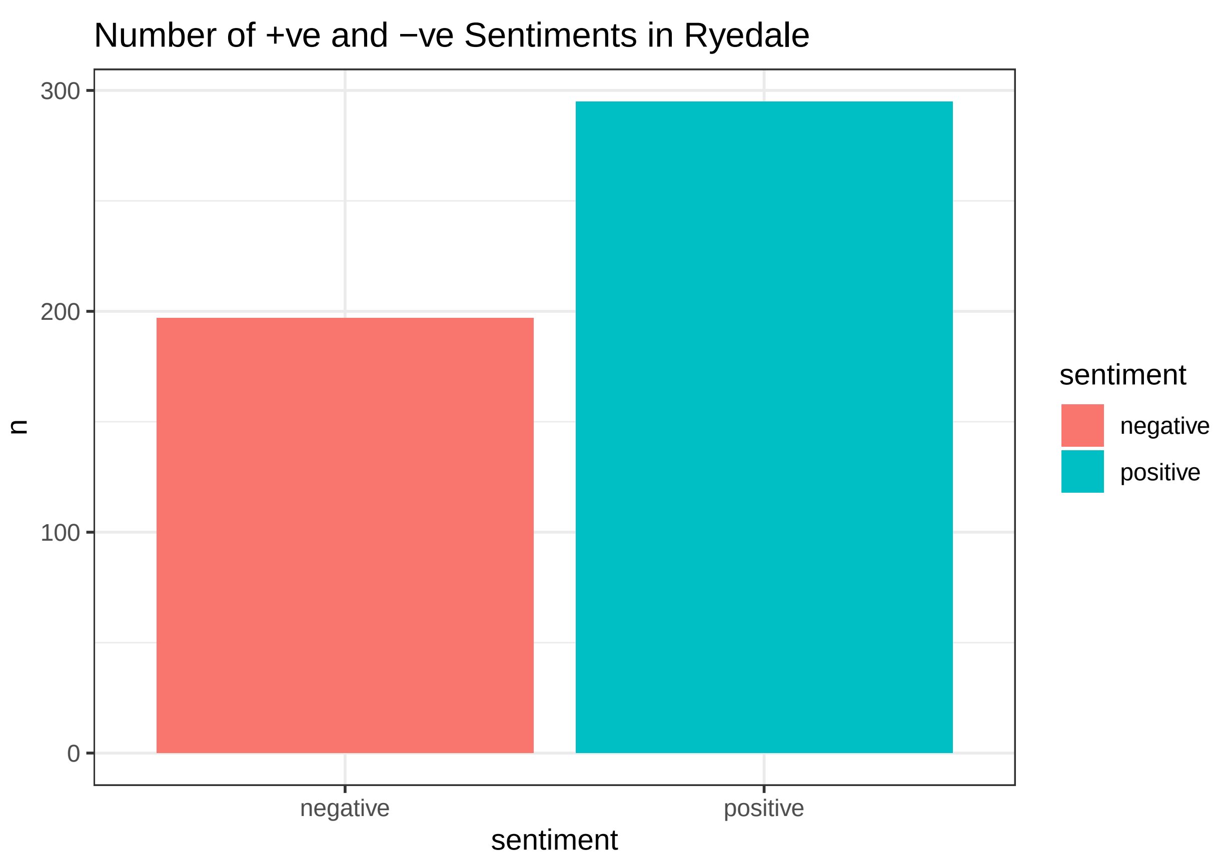
Sentiment Analysis: Sentiment analysis is used to extract subjective information about text in form of opinion, mood, feelings and perception. It is used by companies to quickly understand and analyze feedbacks from customers based on information posted online. Sentiment analysis has different lexicons: Bing, NRC, AFINN. Bing offers negative and positive sentiment analysis format, NRC offers emotional sentiment analysis, while AFINN helps rank the sentiment based on score between -5 and 5.

- Bing Sentiment Analysis

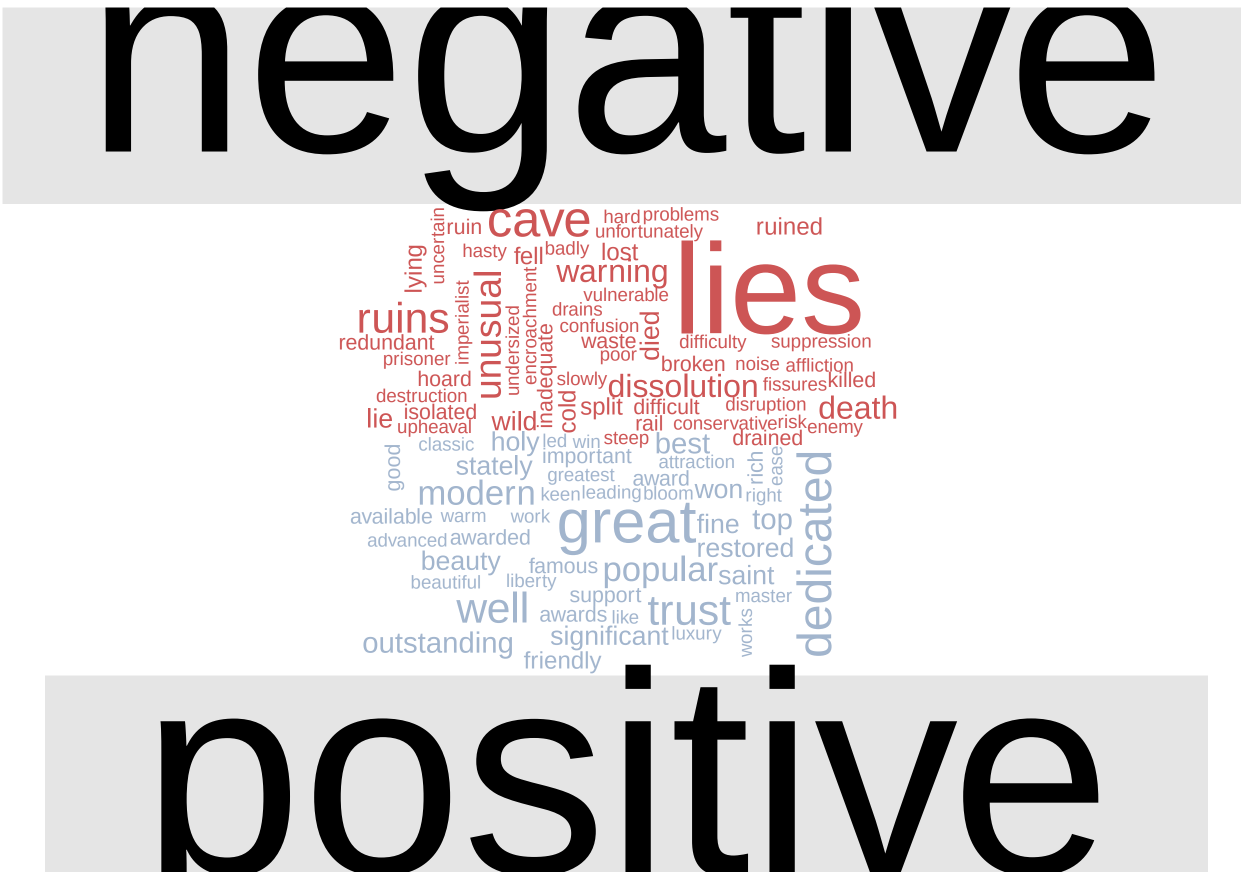
```
bing_sent <- T_W_n_Pg %>%
  inner_join(get_sentiments('bing'))
```

```
## Joining, by = "word"
```

```
bing_sent_count <- bing_sent %>%
  count(sentiment)

#Bing Sentiment Frequency Chart
bing_sent_count %>%
  ggplot(aes(sentiment, n, fill = sentiment))+
  geom_col()+
  ggtitle(label = 'Number of +ve and -ve Sentiments in Ryedale')+
  theme_bw()
```



```r
#Top 100 Bing sentimental word cloud
bing_sent %>%
  count(word, sentiment, sort = TRUE) %>%
  acast(word ~ sentiment, value.var = "n", fill = 0) %>%
  comparison.cloud(random.order=TRUE, title.size=9, fixed.asp=TRUE,
                   colors = c("indianred3","lightsteelblue3"),
                   max.word = 100)
```



There are more postive words than negative words in Ryedale

- Afinn Sentinment Analysis

```r
afinn_sent <- T_W_n_Pg %>%
  inner_join(get_sentiments('afinn'))
```

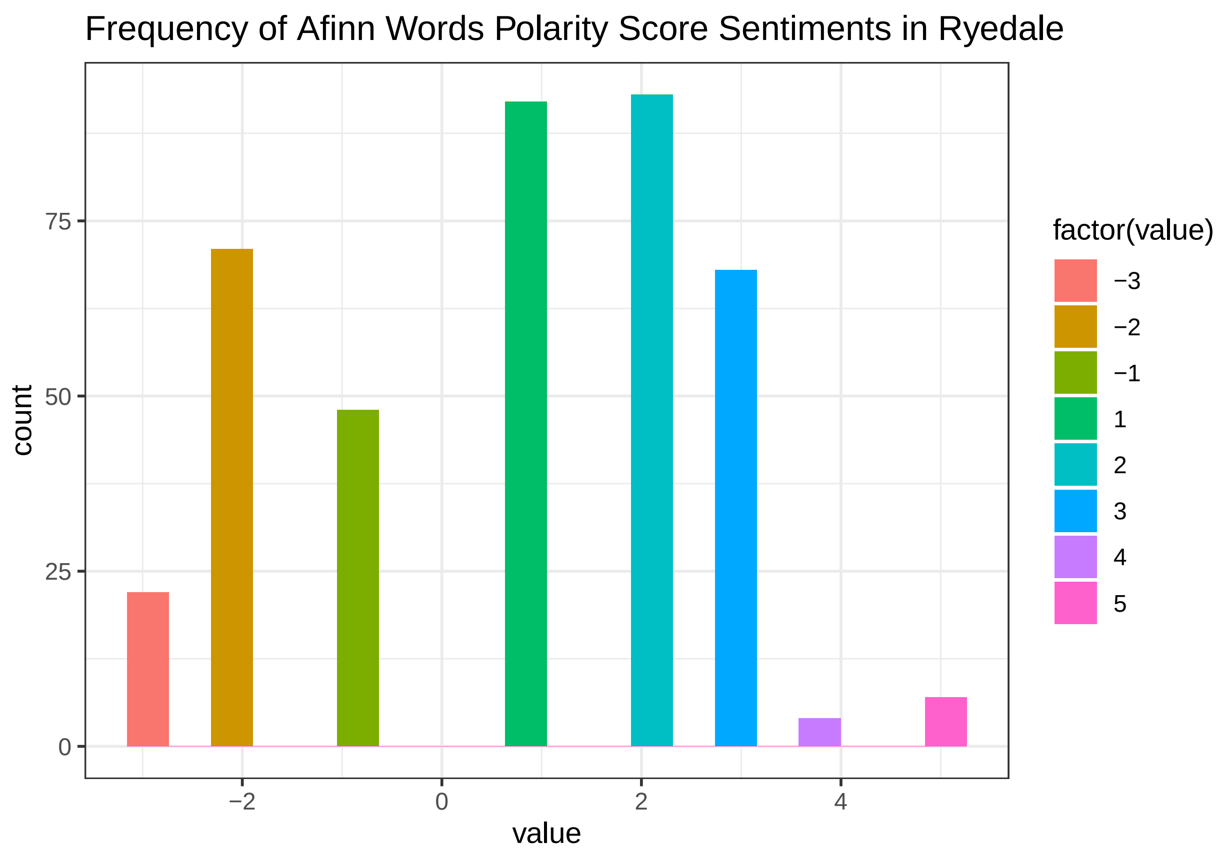
```
## Joining, by = "word"
```

```r
afinn_sent %>%
  count(value) %>%
  slice_max(n, n=10) %>%
  knitr::kable()
```

value	n	geometry
2	93	MULTIPOINT ((455505.4 47939...
1	92	MULTIPOINT ((453045.3 48008...
-2	71	MULTIPOINT ((454944.2 47892...
3	68	MULTIPOINT ((453068.7 47988...
-1	48	MULTIPOINT ((454925.6 48158...
-3	22	MULTIPOINT ((456249.6 49060...
5	7	MULTIPOINT ((462392.3 47909...
4	4	MULTIPOINT ((466901.6 48405...

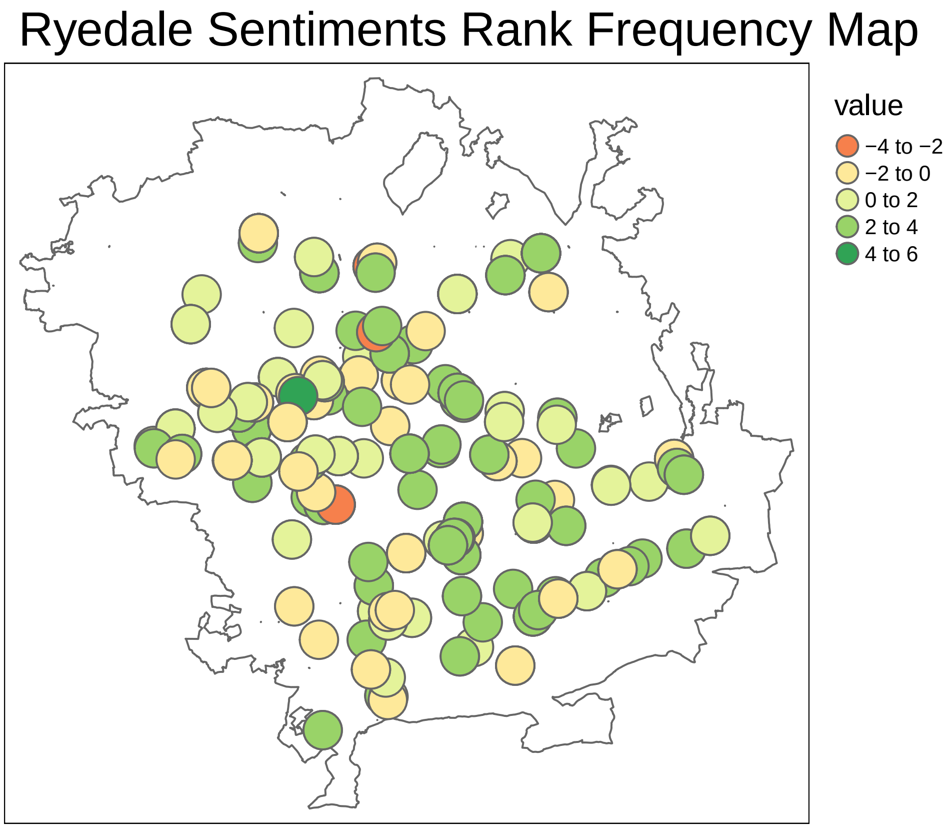
```
#Top 100 Afinn ranking sentiment word cloud
afinn_sent %>%
  count(word, value, sort = TRUE) %>%
  acast(word ~ value, value.var = "n", fill = 0) %>%
  comparison.cloud(random.order=TRUE, title.size=3, fixed.asp=TRUE,
                   max.word = 100)
```


```
#Afinn Sentiment Ranking Frequency Chart
afinn_sent %>%
  ggplot(aes(value, fill= factor(value)))+
  geom_histogram(bins = 20)+
  ggtitle(
    label = 'Frequency of Afinn Words Polarity Score Sentiments in Ryedale')+
  theme_bw()
```




```
#Ryedale Sentiment Rank Map
boundary_m +
  tm_shape(afinn_sent) +
  tm_bubbles(col = 'value')+
  tm_layout(main.title = 'Ryedale Sentiments Rank Frequency Map',
            title.size = 0.7,
            legend.outside=TRUE)
```

```
## Variable(s) "value" contains positive and negative values, so midpoint is set to 0. Set midpoint = N
```



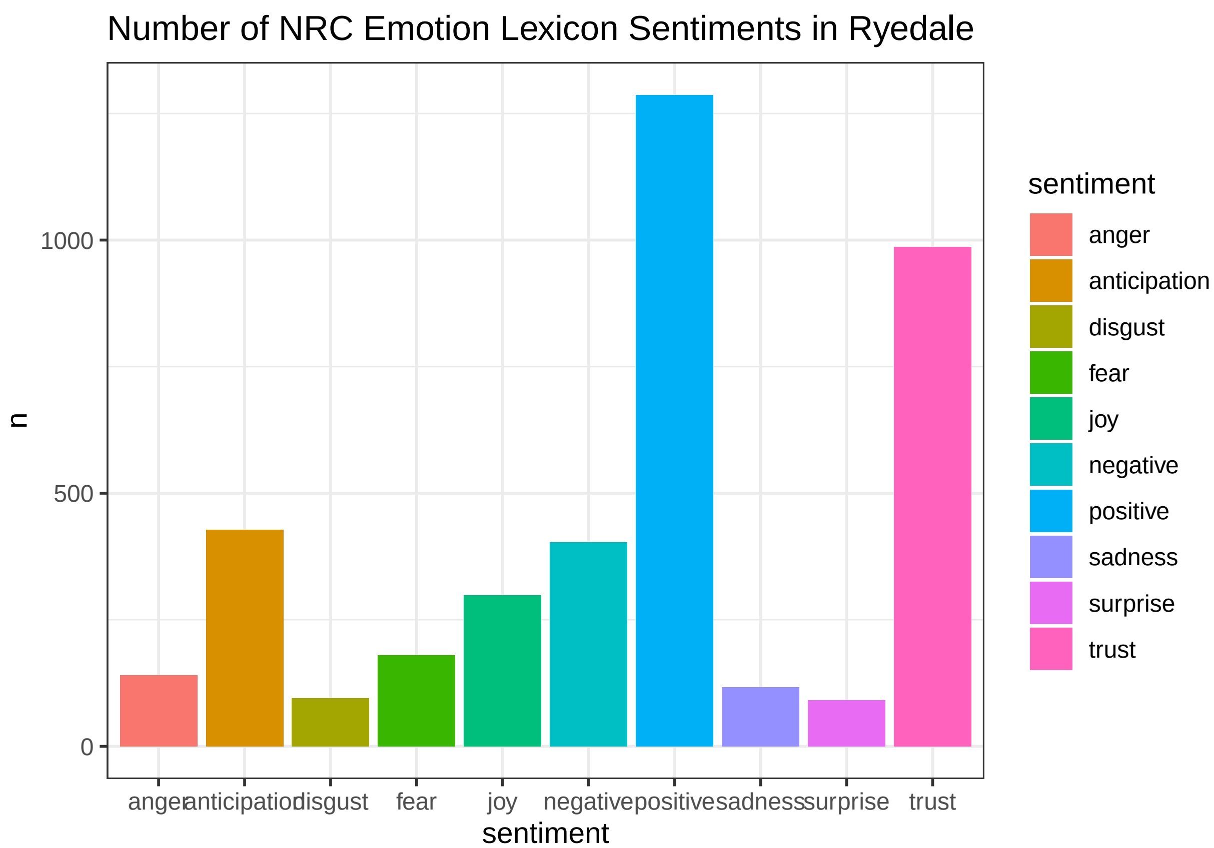
Rank 2 is the most frequenct sentiment rank. it is positive; hence, somewhat tallies with Bing sentiment analysis. The map shows that there is somewhat no cluster pattern in the distribution of sentiment rank

- NRC Sentinment Analysis

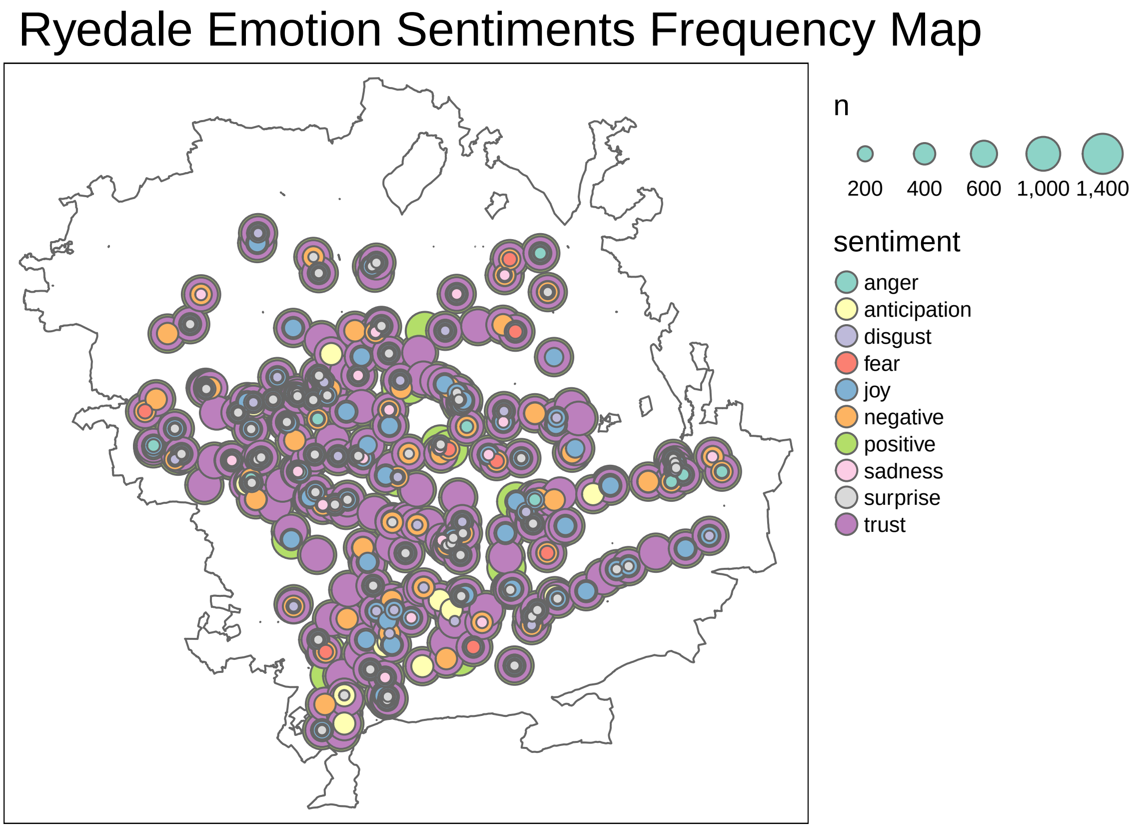
```
nrc_sent <- T_W_n_Pg %>%
  inner_join(get_sentiments('nrc'))
```

```
## Joining, by = "word"
```

```
nrc_sent_count <- nrc_sent %>%
  count(sentiment)

#NRC Emotion Sentiment Frequency Chart
nrc_sent_count %>%
  ggplot(aes(sentiment, n, fill = sentiment))+
  geom_col()+
  ggtitle(label = 'Number of NRC Emotion Lexicon Sentiments in Ryedale')+
  theme_bw()
```




```r
#Ryedale Emotion Sentiment Map
boundary_m +
  tm_shape(nrc_sent_count) +
  tm_bubbles(size = 'n', col = 'sentiment')+
  tm_layout(main.title = 'Ryedale Emotion Sentiments Frequency Map',
            title.size = 0.7,
            legend.outside=TRUE)
```



Positive words have the higest freqiency according to NRC sentiment analysis. The map shows that there is somewhat cluster of 'Trust' Sentiment with an average count of around 600.

Per page Sentiment Analysis

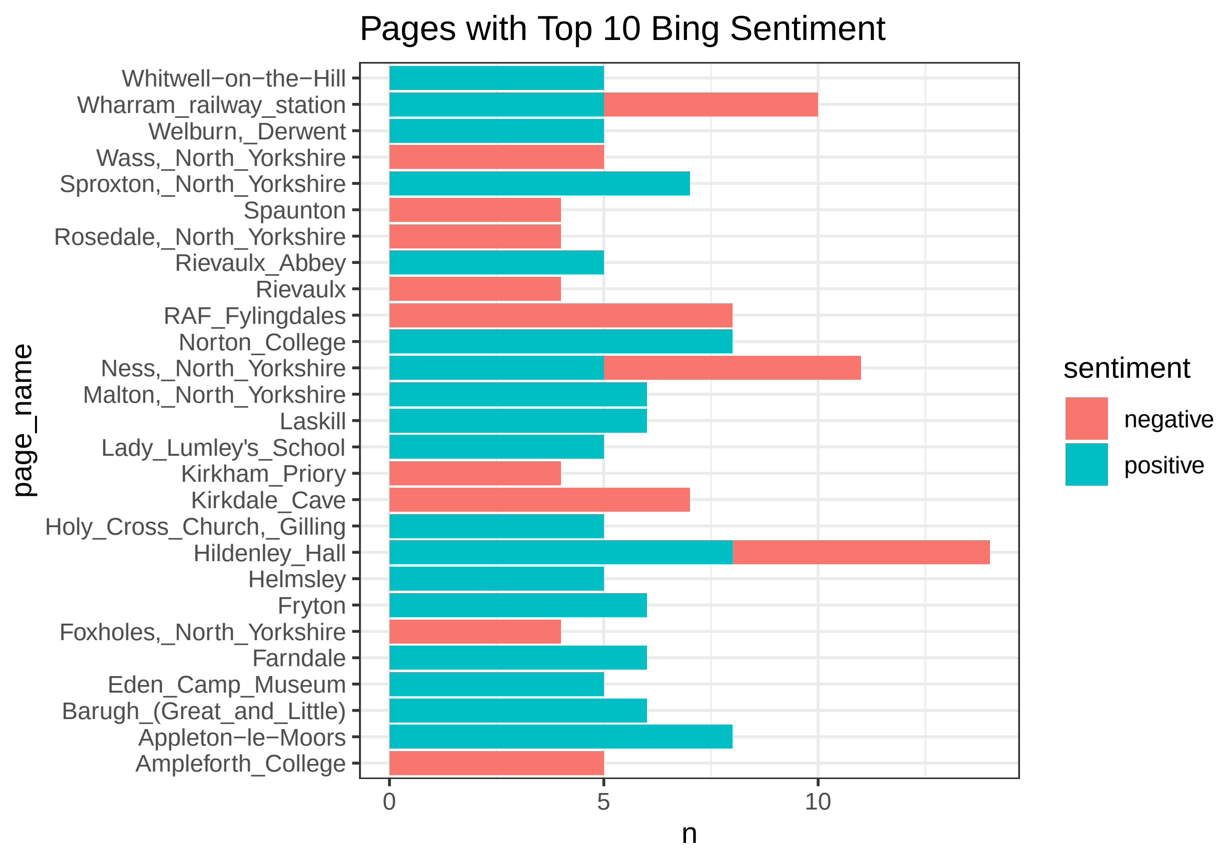
- Bing

```
bing_per_page_count <- bing_sent %>%
  count(page_name, sentiment)

top10_bing_per_pg <- bing_per_page_count %>%
  group_by(sentiment) %>%
  slice_max(n, n= 10)

top10_bing_per_pg %>%
  knitr::kable()
```

page_name	sentiment	n	geometry
RAF_Fylingdales	negative	8	POINT (486545.2 496744.3)
Kirkdale_Cave	negative	7	POINT (467836.3 485601.1)
Hildenley_Hall	negative	6	POINT (474883.3 470821.1)
Ness,_North_Yorkshire	negative	6	POINT (469045.2 479210.3)
Ampleforth_College	negative	5	POINT (459857.2 478833.9)

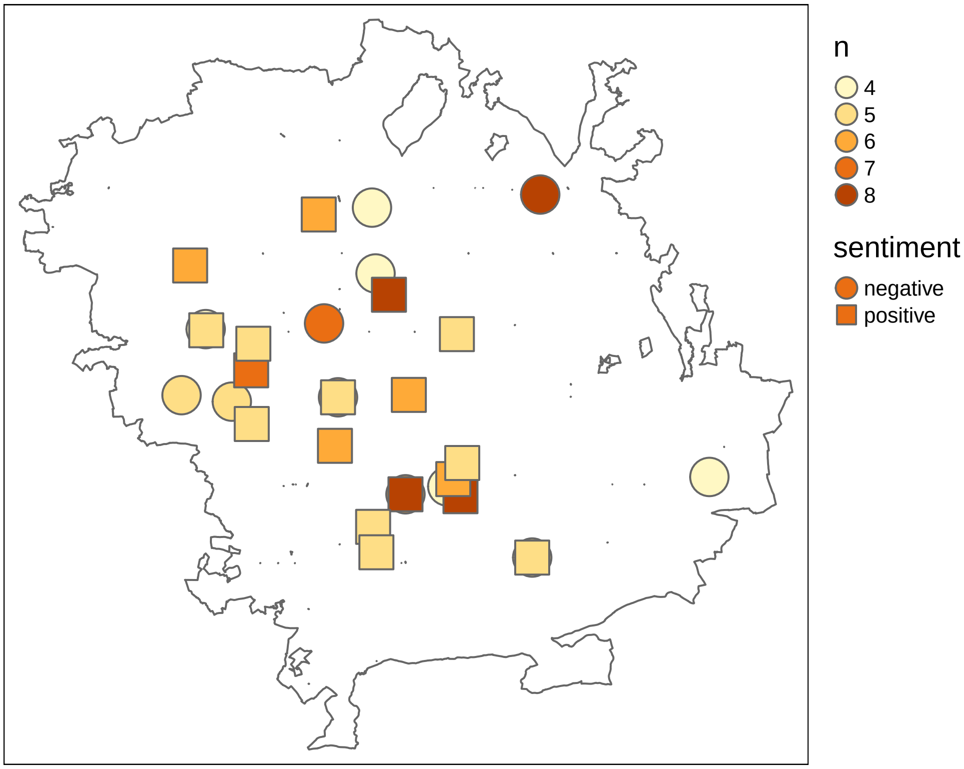
page_name	sentiment	n	geometry
Wass,_North_Yorkshire	negative	5	POINT (455505.4 479394.9)
Wharram_railway_station	negative	5	POINT (485850 465350.1)
Foxholes,_North_Yorkshire	negative	4	POINT (501179.8 472320)
Kirkham_Priory	negative	4	POINT (478500 471500)
Rievaulx	negative	4	POINT (457593.9 485106.9)
Rosedale,_North_Yorkshire	negative	4	POINT (471985.1 495614.8)
Spaunton	negative	4	POINT (472293 489921)
Appleton-le-Moors	positive	8	POINT (473453.1 488080.3)
Hildenley_Hall	positive	8	POINT (474883.3 470821.1)
Norton_College	positive	8	POINT (479644.7 470655.8)
Sproxton,_North_Yorkshire	positive	7	POINT (461529.8 481545)
Barugh_(Great_and_Little)	positive	6	POINT (475174.7 479417.4)
Farndale	positive	6	POINT (467378.1 495021.8)
Fryton	positive	6	POINT (468780.2 475015.3)
Laskill	positive	6	POINT (456249.6 490609.9)
Malton,_North_Yorkshire	positive	6	POINT (479018 472136.4)
Eden_Camp_Museum	positive	5	POINT (479796.3 473528.5)
Helmsley	positive	5	POINT (461719.1 483844.1)
Holy_Cross_Church,_Gilling	positive	5	POINT (461579.9 476901.5)
Lady_Lumley's_School	positive	5	POINT (479331.6 484673.2)
Ness,_North_Yorkshire	positive	5	POINT (469045.2 479210.3)
Rievaulx_Abbey	positive	5	POINT (457642.9 485007.4)
Welburn,_Derwent	positive	5	POINT (472083.9 468005.6)
Wharram_railway_station	positive	5	POINT (485850 465350.1)
Whitwell-on-the-Hill	positive	5	POINT (472366.7 465806.6)

```
top10_bing_per_pg %>%
  ggplot(aes(n, page_name, fill = sentiment))+
  geom_col()+
  ggtitle(label = 'Pages with Top 10 Bing Sentiment')+
  theme_bw()
```




```
boundary_m +
  tm_shape(top10_bing_per_pg) +
  tm_bubbles(shape = 'sentiment', col = 'n')+
  tm_layout(main.title = 'Ryedale Top 10 Bing Sentiments Map',
            title.size = 0.7,
            legend.outside=TRUE)
```

Ryedale Top 10 Bing Sentiments Map



The Map shows that there is cluster of positive words in Ryedale among top 10 Bing Sentiment.

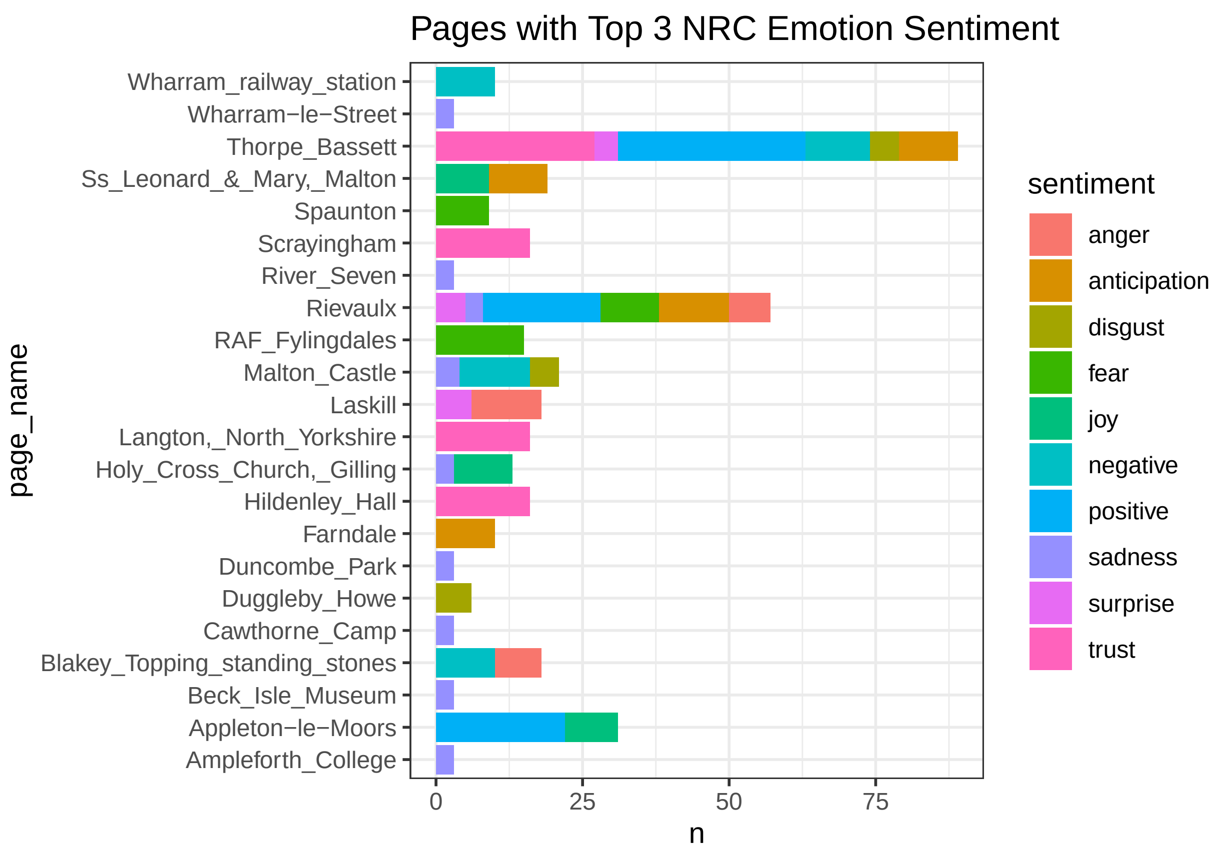
- NRC

```
nrc_per_page_count <- nrc_sent %>%
  count(page_name, sentiment)

top3_nrc_per_pg <- nrc_per_page_count %>%
  group_by(sentiment) %>%
  slice_max(n, n= 3)

top3_nrc_per_pg %>%
  knitr::kable()
```

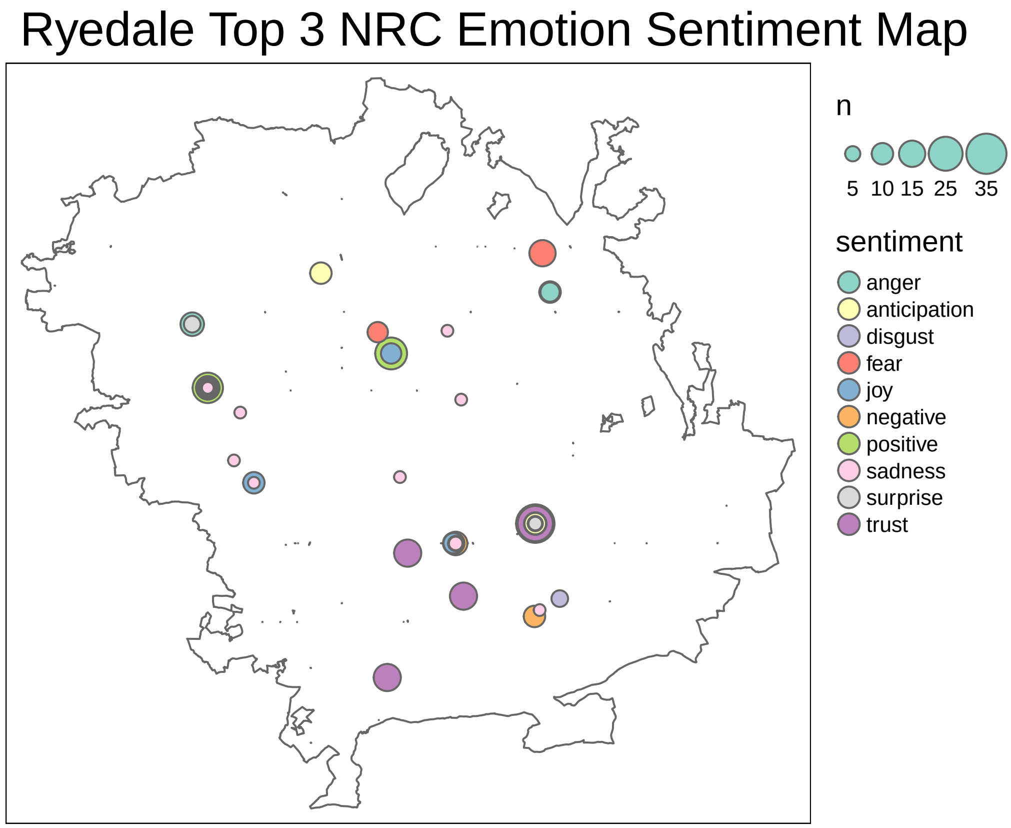
page_name	sentiment	n	geometry
Laskill	anger	12	POINT (456249.6 490609.9)
Blakey_Topping_standing_stones	anger	8	POINT (487195.5 493385.5)
Rievaulx	anger	7	POINT (457593.9 485106.9)
Rievaulx	anticipation	12	POINT (457593.9 485106.9)
Farndale	anticipation	10	POINT (467378.1 495021.8)
Ss_Leonard_&_Mary,_Malton	anticipation	10	POINT (478861.9 471676.3)
Thorpe_Bassett	anticipation	10	POINT (485918.8 473374.6)
Duggleby_Howe	disgust	6	POINT (488038.1 466890.3)

page_name	sentiment	n	geometry
Malton_Castle	disgust	5	POINT (479031.5 471656.9)
Thorpe_Bassett	disgust	5	POINT (485918.8 473374.6)
RAF_Fylingdales	fear	15	POINT (486545.2 496744.3)
Rievaulx	fear	10	POINT (457593.9 485106.9)
Spaunton	fear	9	POINT (472293 489921)
Holy_Cross_Church,_Gilling	joy	10	POINT (461579.9 476901.5)
Appleton-le-Moors	joy	9	POINT (473453.1 488080.3)
Ss_Leonard_&_Mary,_Malton	joy	9	POINT (478861.9 471676.3)
Malton_Castle	negative	12	POINT (479031.5 471656.9)
Thorpe_Bassett	negative	11	POINT (485918.8 473374.6)
Blakey_Topping_standing_stones	negative	10	POINT (487195.5 493385.5)
Wharram_railway_station	negative	10	POINT (485850 465350.1)
Thorpe_Bassett	positive	32	POINT (485918.8 473374.6)
Appleton-le-Moors	positive	22	POINT (473453.1 488080.3)
Rievaulx	positive	20	POINT (457593.9 485106.9)
Malton_Castle	sadness	4	POINT (479031.5 471656.9)
Ampleforth_College	sadness	3	POINT (459857.2 478833.9)
Beck_Isle_Museum	sadness	3	POINT (479510 484090)
Cawthorne_Camp	sadness	3	POINT (478321.2 490042.6)
Duncombe_Park	sadness	3	POINT (460402.5 482971.6)
Holy_Cross_Church,_Gilling	sadness	3	POINT (461579.9 476901.5)
Rievaulx	sadness	3	POINT (457593.9 485106.9)
River_Seven	sadness	3	POINT (474215.5 477398.7)
Wharram-le-Street	sadness	3	POINT (486296.5 465900.4)
Laskill	surprise	6	POINT (456249.6 490609.9)
Rievaulx	surprise	5	POINT (457593.9 485106.9)
Thorpe_Bassett	surprise	4	POINT (485918.8 473374.6)
Thorpe_Bassett	trust	27	POINT (485918.8 473374.6)
Hildenley_Hall	trust	16	POINT (474883.3 470821.1)
Langton,_North_Yorkshire	trust	16	POINT (479709.8 467104.5)
Scrayingham	trust	16	POINT (473116.6 460069.8)

```
top3_nrc_per_pg %>%
  ggplot(aes(n, page_name, fill = sentiment))+
  geom_col()+
  ggtitle(label = 'Pages with Top 3 NRC Emotion Sentiment')+
  theme_bw()
```




```
boundary_m +
  tm_shape(top3_nrc_per_pg) +
  tm_bubbles(col = 'sentiment', size = 'n')+
  tm_layout(main.title = 'Ryedale Top 3 NRC Emotion Sentiment Map',
            title.size = 0.7,
            legend.outside=TRUE)
```



No form of cluster was notice. However, sentiment 'Trust' seems to show slight cluster.

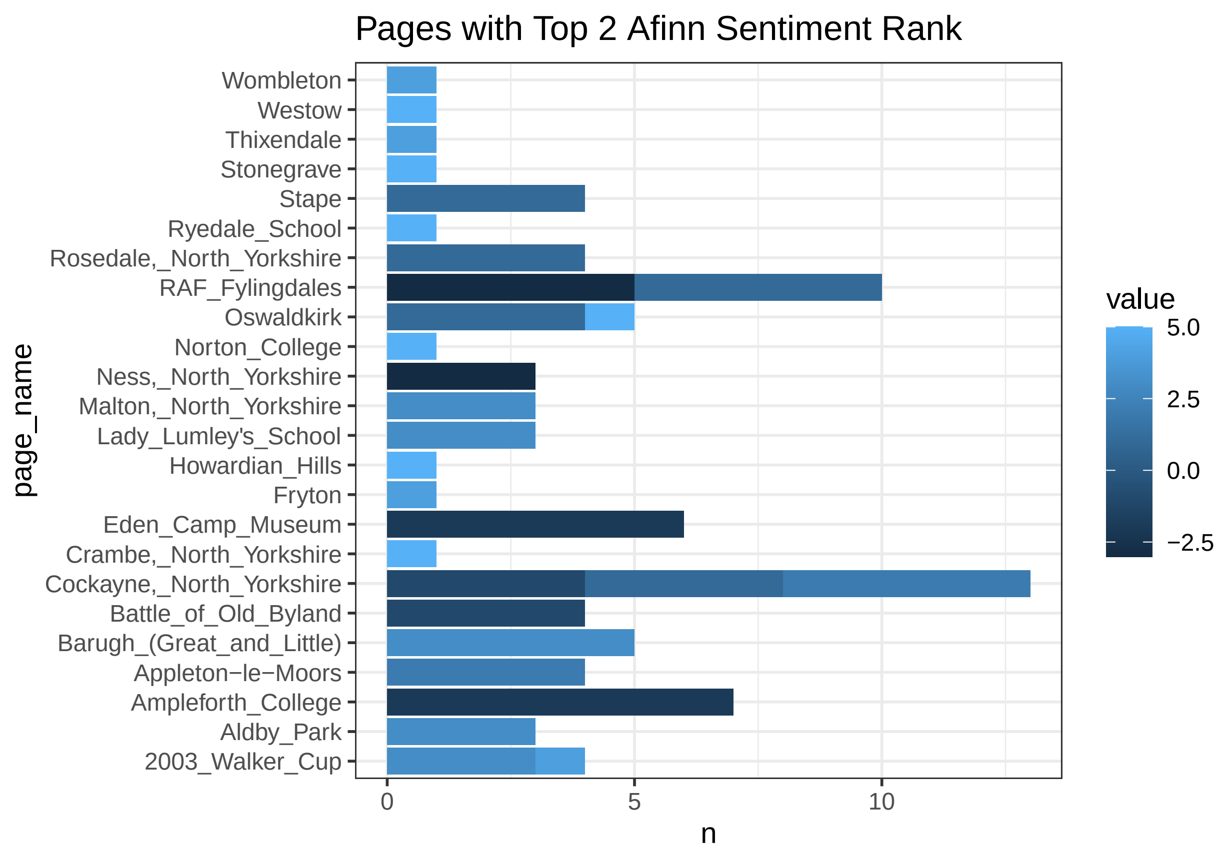
- AFINN

```
afinn_per_page_count <- afinn_sent %>%
  count(page_name, value)


top2_afinn_per_pg <- afinn_per_page_count %>%
  group_by(value) %>%
  slice_max(n, n= 2)

top2_afinn_per_pg %>%
  knitr::kable()
```

page_name	value	n	geometry
RAF_Fylingdales	-3	5	POINT (486545.2 496744.3)
Ness,_North_Yorkshire	-3	3	POINT (469045.2 479210.3)
Ampleforth_College	-2	7	POINT (459857.2 478833.9)
Eden_Camp_Museum	-2	6	POINT (479796.3 473528.5)
Battle_of_Old_Byland	-1	4	POINT (454925.6 481580.1)
Cockayne,_North_Yorkshire	-1	4	POINT (462130 498475.6)
RAF_Fylingdales	1	5	POINT (486545.2 496744.3)

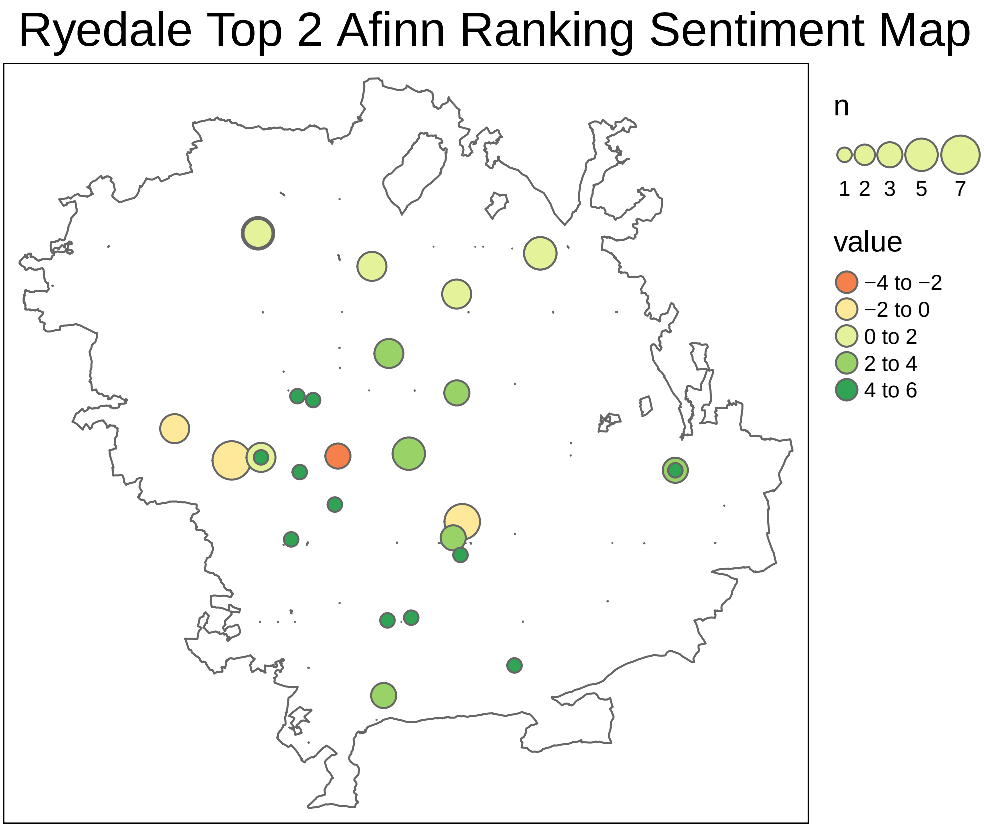
page_name	value	n	geometry
Cockayne,_North_Yorkshire	1	4	POINT (462130 498475.6)
Oswaldkirk	1	4	POINT (462392.3 479093.3)
Rosedale,_North_Yorkshire	1	4	POINT (471985.1 495614.8)
Stape	1	4	POINT (479314.4 493220.4)
Cockayne,_North_Yorkshire	2	5	POINT (462130 498475.6)
Appleton-le-Moors	2	4	POINT (473453.1 488080.3)
Barugh_(Great_and_Little)	3	5	POINT (475174.7 479417.4)
2003_Walker_Cup	3	3	POINT (498225.3 477980.7)
Aldby_Park	3	3	POINT (472999.8 458499.5)
Lady_Lumley's_School	3	3	POINT (479331.6 484673.2)
Malton,_North_Yorkshire	3	3	POINT (479018 472136.4)
2003_Walker_Cup	4	1	POINT (498225.3 477980.7)
Fryton	4	1	POINT (468780.2 475015.3)
Thixendale	4	1	POINT (484315.9 461100.6)
Wombleton	4	1	POINT (466901.6 484051.9)
Crambe,_North_Yorkshire	5	1	POINT (473328.3 464998.2)
Howardian_Hills	5	1	POINT (464999.9 472000.5)
Norton_College	5	1	POINT (479644.7 470655.8)
Oswaldkirk	5	1	POINT (462392.3 479093.3)
Ryedale_School	5	1	POINT (465543.2 484389.6)
Stonegrave	5	1	POINT (465737.6 477826.6)
Westow	5	1	POINT (475384.1 465231.5)

```
top2_afinn_per_pg %>%
  ggplot(aes(n, page_name, fill = value))+
  geom_col()+
  ggtitle(label = 'Pages with Top 2 Afinn Sentiment Rank')+
  theme_bw()
```




```
boundary_m +
  tm_shape(top2_afinn_per_pg) +
  tm_bubbles(col = 'value', size = 'n')+
  tm_layout(main.title = 'Ryedale Top 2 Afinn Ranking Sentiment Map',
            title.size = 0.7,
            legend.outside=TRUE)
```

```
## Variable(s) "value" contains positive and negative values, so midpoint is set to 0. Set midpoint = N
```



Only rank value '4 to 6' shows slight cluster in the map.

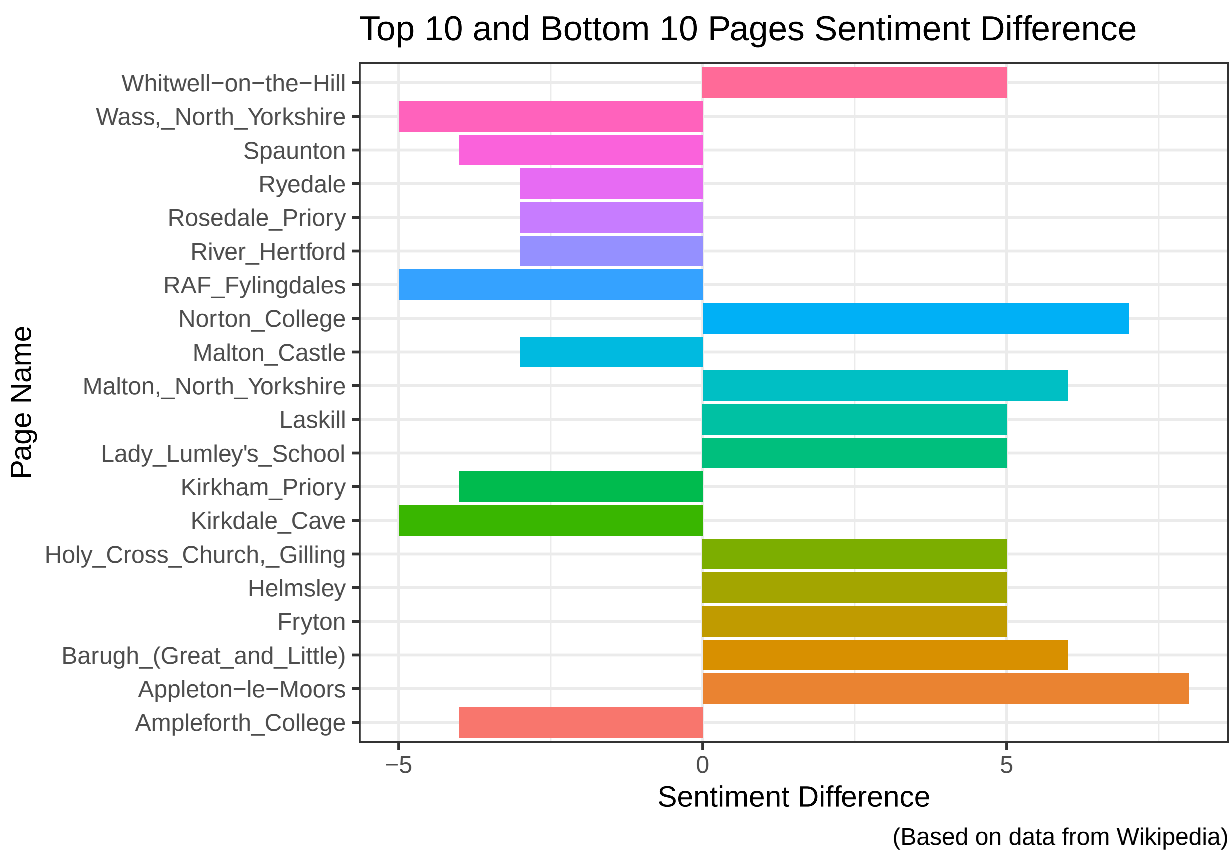
Sentiment Difference Analysis: (Positive - Negative)

```r
bing_sent_diff <- bing_per_page_count %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sent_diff = positive - negative)

bing_sent_diff %>%
  slice_max(sent_diff, n=20)
```

```
## Simple feature collection with 28 features and 4 fields
## Geometry type: POINT
## Dimension:     XY
## Bounding box:  xmin: 454925.6 ymin: 458499.5 xmax: 498225.3 ymax: 495021.8
## Projected CRS: OSGB 1936 / British National Grid
## First 10 features:
##                       page_name negative positive sent_diff
## 1           Appleton-le-Moors        0        8         8
## 2              Norton_College        1        8         7
## 3   Barugh_(Great_and_Little)        0        6         6
## 4      Malton,_North_Yorkshire        0        6         6
## 5                      Fryton        1        6         5
## 6                    Helmsley        0        5         5
```

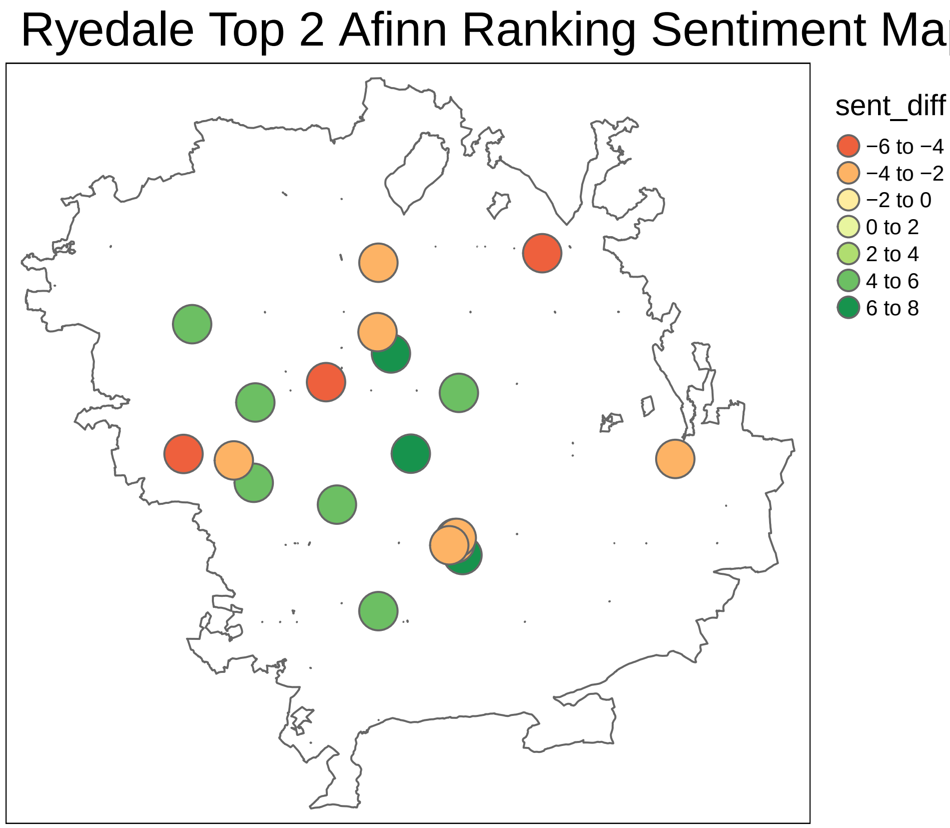
```
## 7  Holy_Cross_Church,_Gilling        0        5          5
## 8        Lady_Lumley's_School        0        5          5
## 9                     Laskill        1        6          5
## 10       Whitwell-on-the-Hill        0        5          5
##                        geometry
## 1  POINT (473453.1 488080.3)
## 2  POINT (479644.7 470655.8)
## 3  POINT (475174.7 479417.4)
## 4    POINT (479018 472136.4)
## 5  POINT (468780.2 475015.3)
## 6  POINT (461719.1 483844.1)
## 7  POINT (461579.9 476901.5)
## 8  POINT (479331.6 484673.2)
## 9  POINT (456249.6 490609.9)
## 10 POINT (472366.7 465806.6)
```

```
min_max_diff <- rbind(
bing_sent_diff %>%
  arrange(desc(sent_diff)) %>%
  head(10),
bing_sent_diff %>%
  arrange(desc(sent_diff)) %>%
  tail(10))

min_max_diff %>%
  ggplot(aes(sent_diff, page_name, fill = page_name)) +
  geom_col(show.legend = FALSE) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        plot.title = element_text(size=9),
        axis.text.y = element_text(size=7, angle = 20)) +
  ylab("Page Name")+
  xlab('Sentiment Difference')+
  labs(title = 'Top 10 and Bottom 10 Pages Sentiment Difference',
       caption = '(Based on data from Wikipedia)')+
  theme_bw()
```




```
boundary_m +
  tm_shape(min_max_diff) +
  tm_bubbles(col = 'sent_diff')+
  tm_layout(main.title = 'Ryedale Top 2 Afinn Ranking Sentiment Map',
            title.size = 0.7,
            legend.outside=TRUE)
```

```
## Variable(s) "sent_diff" contains positive and negative values, so midpoint is set to 0. Set midpoint
```



The positive sentiment difference points are slightly clustered while the negative are dispersed. Places with more positive sentiment words tends to be located around the hearth of the study area. While places with more negative values are distributed in a dispersed way.

Refrence

Atenstaedt, R., 2012. Word cloud analysis of theBJGP. British Journal of General Practice, 62(596), pp.148-148.

Letico, M., 2022. RPubs - Ngrams analysis and NPL modelling. [online] Rpubs.com. Available at: https://rpubs.com/mletico/361214 [Accessed 28 June 2022].

Robinson, J., 2022. 4 Relationships between words: n-grams and correlations | Text Mining with R. [online] Tidytextmining.com. Available at: https://www.tidytextmining.com/ngrams.html [Accessed 27 June 2022].