

Leak Detection N-Grams

Kat Bardash

Purpose

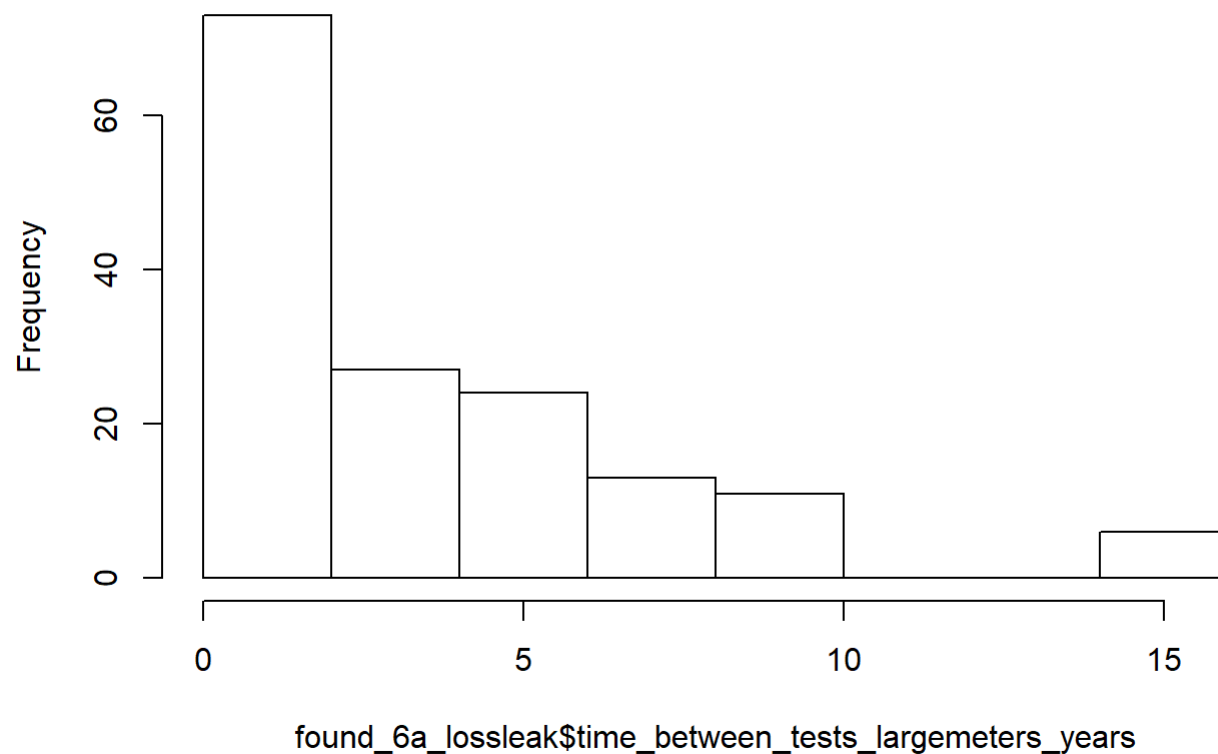
- The purpose of this file is conduct text analysis on the leak_detection_field_technology column in the foundational_06a_loss_and_leak_detection file. Distributions of time between testing and age to replace small meters in years.
- This analysis creates token, bi-grams, tri-grams, and four-grams of this text field.
- The data can be found http://cowaterefficiency.com/unauthenticated_home (http://cowaterefficiency.com/unauthenticated_home) with permission. Once in the portal, all report years (2013-2017) were selected as well as all water providers.

Recommendations

- Water detection technology could be a useful factor to correlate and predict with water loss. The current input method does not easily allow for this type of analysis.
- Our suggestion is to create a drop-down menu for this particular part of the reporting with the applicable leak detection technologies.

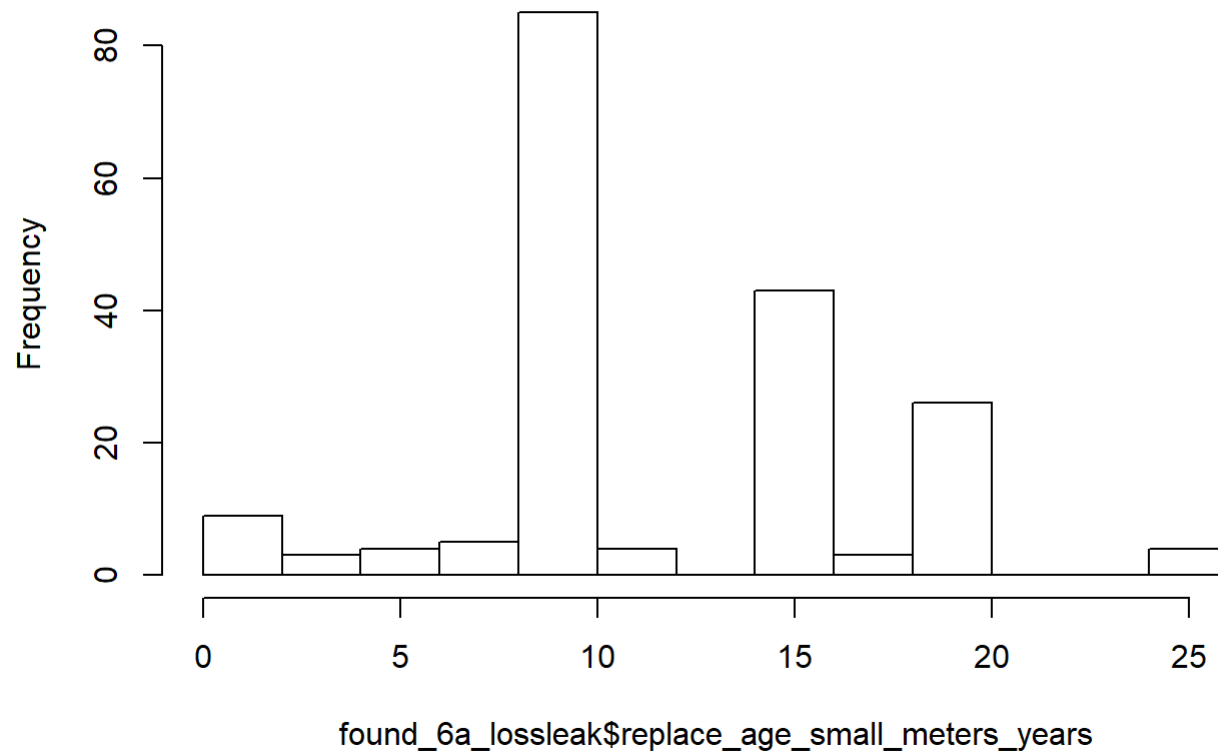
```
# read in data
found_6a_lossleak <- read.csv("EffDataPortal_Output_User690_20181112192716/foundational/foundational_06a_loss_and_leak_det.csv", stringsAsFactors = FALSE)
hist(found_6a_lossleak$time_between_tests_largemeters_years, main="Distribution of Years Between Tests For Large Meters")
```

Distribution of Years Between Tests For Large Meters



```
hist(found_6a_lossleak$replace_age_small_meters_years, main="Distribution of Years Between Tests For Small Meters")
```

Distribution of Years Between Tests For Small Meters



Text Analysis of Leak Detection Type with N-grams

```
leaks <- found_6a_lossleak  
head(leaks[order(leaks$ce_annual_ndx),], n=3)
```

```

## ce_annual_ndx meter_test_program awwa_policy_adherence
## 1          1912          YES          YES
## 2          1913          YES          NO
## 3          1916          YES          NO
##
## non_awwa_test_procedures
## 1
## 2          Test as needed to address non-reading or customer requests
## 3 We replace about 20-30 meters per year based on age and largely on reactionary measures due to leakage.
##
## astest_largemeters_describe
## 1 Yes, we field test using AWWA recommendations for yearly intervals between testing. Go to meter pit, test port, hose t
o field test unit which is calibrated to shop tanks. 1-1.5 are changing out not tested
## 2
## 3          yes- by contract on concurrent system - Ute Mountain Tribe
##
## No
## time_between_tests_largemeters_years replace_age_small_meters_years
## 1          3.5          15
## 2          2.0          3
## 3          0.0          10
##
## water_loss_comments
## 1 4-12" meters are tested every year, 3" every 2 years, 2" every 4 years, for total average of 3.5 years
## 2
## 3
##
## leak_detection_field_technology
## 1 Leak correlators, loggers, and listening devices. Correlators to locate known leaks/breaks, loggers for undiscovered l
eaks, and microphones on hydrants. Exceptions reporting utilized (attached).
## 2
## 3          visual, water loss-increased flows from water plant
##
## Sonic, hydrostatic, visual
## pctannual_leak_inspection pctannual_pipe_replaces
## 1          35          0.5
## 2          70          2.0
## 3          20          5.0
##
## leak_detection_comments
## 1 Very proactive (survey over 300 miles of pipe per year listening for leaks).
## 2
## 3

```

```
leaks_short <- leaks[,c(1,9 )]
head(leaks_short, n=3)
```

```
##    ce_annual_ndx
## 1          1912
## 2          1913
## 3          1916
##
##                                leak_detection_field_technology
## 1 Leak correlators, loggers, and listening devices. Correlators to locate known leaks/breaks, loggers for undiscovered leaks, and microphones on hydrants. Exceptions reporting utilized (attached).
## 2
##                                visual, water loss-increased flows from water plant
## 3
##                                Sonic, hydrostatic, visual
```

```
library(dplyr)
library(tidytext)
leaks_short2<- leaks_short %>% unnest_tokens(word, leak_detection_field_technology)

# get rid of stop words
tidy_leaks <- leaks_short2 %>% anti_join(stop_words)
```

Most used Words

- These are the top 20 most common words typed into the leak detection technology field.

```
head(tidy_leaks %>% count(word, sort=TRUE), n=20)
```

```
## # A tibble: 20 x 2
##   word      n
##   <chr>    <int>
## 1 leak      96
## 2 detection  50
## 3 water     32
## 4 correlators 31
## 5 visual    29
## 6 leaks     28
## 7 loggers   28
## 8 data      27
## 9 surface   22
## 10 listening 21
## 11 technology 21
## 12 noise     20
## 13 system    20
## 14 equipment  19
## 15 acoustic  18
## 16 sonic     17
## 17 field     16
## 18 devices   14
## 19 meter     14
## 20 subsurface 13
```

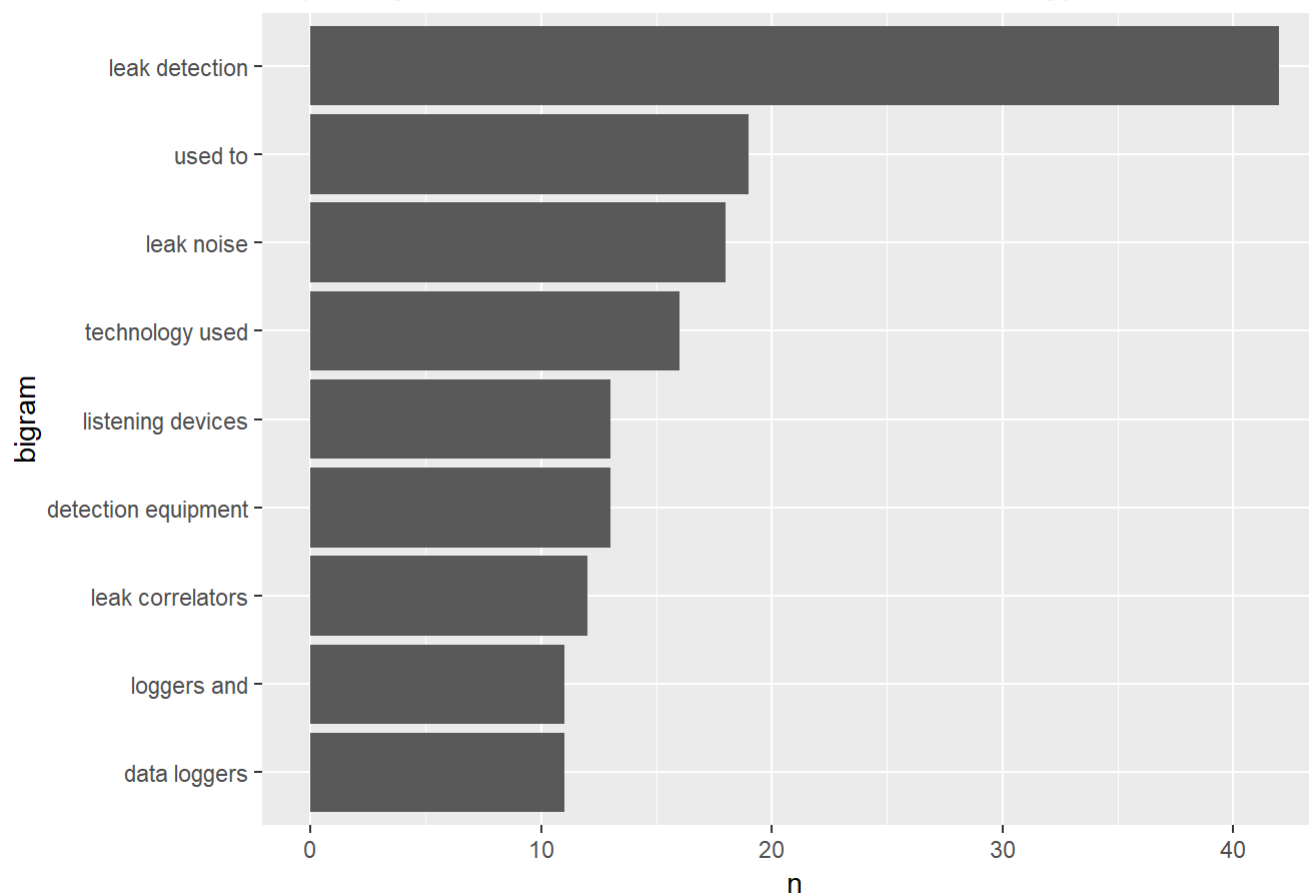
Finding n-grams

- These are the top 10 bi, tri, and 4-grams for the leak detection technology column.

```
#2-gram
tidy_bigram <- leaks_short %>% unnest_tokens(bigram, leak_detection_field_technology, token="ngrams", n=2)
bill10 <- head(tidy_bigram %>% count(bigram, sort=TRUE), n=10)
# remove NA ( reorder)
bill10 <- bill10[!is.na(bill10$bigram),]
bill10 <- bill10 %>% mutate(bigram=reorder(bigram, n))

ggplot(bill10, aes(bigram, n))+
  geom_col()+
  coord_flip()+
  ggtitle("Top 8 Bigrams From Leak Detection Field Technology")
```

Top 8 Bigrams From Leak Detection Field Technology

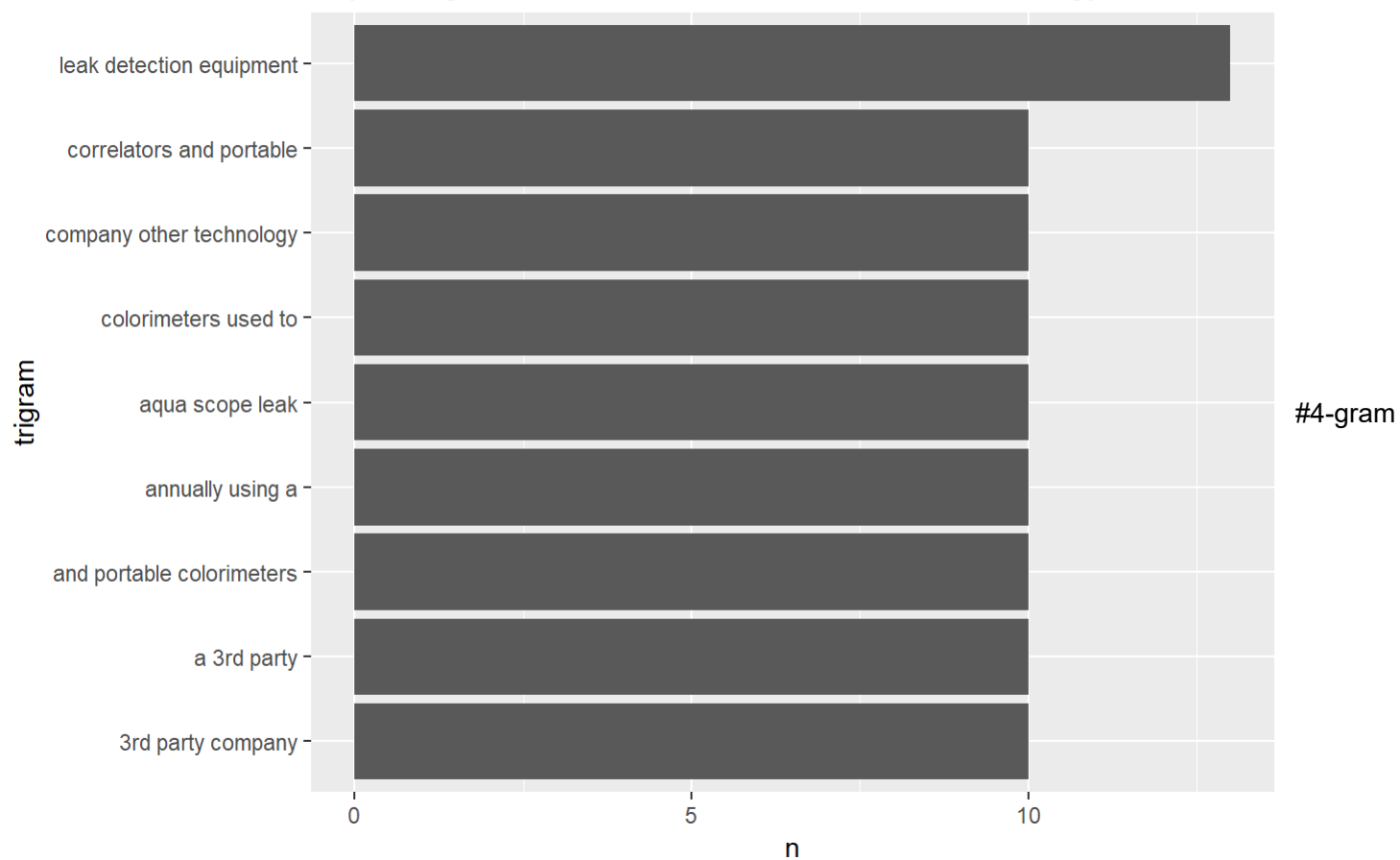


```
#3-gram
tidy_trigram <- leaks_short %>% unnest_tokens(trigram,leak_detection_field_technology, token="ngrams", n=3)
tri10 <- head(tidy_trigram %>% count(trigram, sort=TRUE), n=10)

# remove NA ( reorder)
tri10 <- tri10[!is.na(tri10$trigram),]
tri10 <- tri10 %>% mutate(trigram=reorder(trigram, n))

ggplot(tri10, aes(trigram, n))+
  geom_col()+
  coord_flip()+
  ggtitle("Top 9 Trigrams From Leak Detection Field Technology")
```

Top 9 Trigrams From Leak Detection Field Technology



```

tidy_fourgram <- leaks_short %>% unnest_tokens(fourgram,leak_detection_field_technology, token="ngrams", n=4)
four10 <- head(tidy_fourgram %>% count(fourgram, sort=TRUE), n=10)

# remove NA ( reorder)
four10 <- four10[!is.na(four10$fourgram),]
four10 <- four10 %>% mutate(fourgram=reorder(fourgram, n))

ggplot(four10, aes(fourgram, n))+
  geom_col()+
  coord_flip()+
  ggtitle("Top 9 Fourgrams From Leak Detection Field Technology")

```