Boilerplate

Jinhang Jiang

5/16/2020

Calculation of Boilerplate

```
load("workspaces/CSR_documents_30samples.RData")
```

Tokenize all the sentence

we removed all the numbers here

```
library(koRpus.lang.en)
library(tokenizers)
library(tm)

t<-list(length=nrow(text_stack_sample))
for(row in 1:nrow(text_stack_sample))
{
  print(row)
  if (text_stack_sample[row,1] != "")
  {
    t[[row]] =unlist(tokenize_sentences(removeNumbers(text_stack_sample[row,1])))
  }
}
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
## [1] 17
## [1] 18
## [1] 19
## [1] 20
## [1] 21
## [1] 22
## [1] 23
## [1] 24
## [1] 25
## [1] 26
## [1] 27
## [1] 28
## [1] 29
## [1] 30
```

Get the tetragrams into one list

```r
ngram <- list(length=length(t))
for(i in 1:length(t))
{
  print(i)
  ngram[[i]] = list(length = length(t[[i]]))
  for(j in 1:length(t[[i]]))
  {
    try(
    if(t[[i]][[j]] != "")
    {
      ngram[[i]][[j]] = tokenize_ngrams(t[[i]][[j]],n=4)
    }
    )
  }
}
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
## [1] 17
## [1] 18
## [1] 19
## [1] 20
## [1] 21
## [1] 22
## [1] 23
## [1] 24
## [1] 25
## [1] 26
## [1] 27
## [1] 28
## [1] 29
## [1] 30
```

```
list_tetragrams = list(length(nrow(text_stack_sample)))

for(row in 1:nrow(text_stack_sample))
{
  temp  = unlist(ngram[row])
  temp = as.data.frame(table(temp))
  list_tetragrams[[row]] = temp$temp
}

Fngram<- list(unlist(unlist(list_tetragrams)))
```

Get the tetragrams with frequency between 30% and 75%

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ---------------------------------------------------------------------------------------
## x ggplot2::annotate() masks NLP::annotate()
## x dplyr::filter()     masks stats::filter()
## x dplyr::lag()        masks stats::lag()
## x readr::tokenize()   masks koRpus::tokenize()
```

```
N_table<-as.data.frame(table(Fngram))

N_table2 = N_table%>%
  arrange(desc(Freq))%>%
  mutate(prop=Freq/nrow(text_stack_sample)) %>% filter(prop>0.3 & prop<=0.75)

N_table2
```

```
##                                                            Fngram Freq      prop
## 1                                           more than million in   22 0.7333333
## 2                                                  at the end of   22 0.7333333
## 3                                          in the united states   21 0.7000000
## 4                                                 by the end of   20 0.6666667
## 5                                           we will continue to   20 0.6666667
## 6                                                 is one of the   19 0.6333333
## 7                                               a member of the   18 0.6000000
## 8                                               one of the most   18 0.6000000
## 9                                                as part of our   18 0.6000000
## 10                                     the communities we serve   18 0.6000000
## 11                                        the united states and   18 0.6000000
## 12                                              a wide range of   17 0.5666667
## 13                                               our goal is to   17 0.5666667
## 14                                   products and services that   17 0.5666667
## 15                                       the board of directors   17 0.5666667
## 16                                               as part of the   17 0.5666667
## 17                                            meet the needs of   17 0.5666667
## 18                                         more than million to   17 0.5666667
## 19                                               as a result of   17 0.5666667
## 20                                      more than million hours   16 0.5333333
## 21                                               the end of the   16 0.5333333
## 22                                      for more information on   15 0.5000000
## 23                                        than million hours of   15 0.5000000
## 24                                             the needs of our   15 0.5000000
## 25                                          to learn more about   15 0.5000000
## 26                                          for more than years   15 0.5000000
## 27                                       girls clubs of america   15 0.5000000
## 28                                     across the united states   14 0.4666667
## 29                                     code of business conduct   14 0.4666667
## 30                                    donated more than million   14 0.4666667
## 31                                           in addition to our   14 0.4666667
## 32                                     a good corporate citizen   14 0.4666667
## 33                                    impact on the environment   14 0.4666667
## 34                                         of our employees and   14 0.4666667
## 35                                    our products and services   14 0.4666667
## 36                                     the communities in which   14 0.4666667
## 37                                  the environmental impact of   14 0.4666667
## 38                                             when it comes to   14 0.4666667
## 39                                             at the same time   14 0.4666667
## 40                                  and chief diversity officer   13 0.4333333
## 41                                         learn more about our   13 0.4333333
## 42                                           more than hours of   13 0.4333333
## 43                                         to more than million   13 0.4333333
## 44                                                as one of the   13 0.4333333
## 45  corporateregister.com limited corporateregister.com limited   13 0.4333333
## 46                                employee resource groups ergs   13 0.4333333
## 47                                         more than percent of   13 0.4333333
## 48                                         of our commitment to   13 0.4333333
## 49                                     the highest standards of   13 0.4333333
## 50                                          to our business and   13 0.4333333
## 51                                             we live and work   13 0.4333333
## 52                                     of products and services   12 0.4000000
## 53                                         an important part of   12 0.4000000
## 54                                  and chief executive officer   12 0.4000000
## 55                                  environmental impact of our   12 0.4000000
## 56                                              in a variety of   12 0.4000000
## 57                                          in our supply chain   12 0.4000000
## 58                                    organizations such as the   12 0.4000000
## 59                                       the next generation of   12 0.4000000
## 60                                          the total number of   12 0.4000000
## 61                                  top companies for diversity   12 0.4000000
## 62                                           in addition to the   12 0.4000000
## 63                                               in the u.s and   12 0.4000000
## 64                                       program is designed to   12 0.4000000
## 65                                          boys girls clubs of   12 0.4000000
## 66                                         and around the world   11 0.3666667
## 67                                               as a leader in   11 0.3666667
## 68                                              as part of this   11 0.3666667
## 69                                               as well as our   11 0.3666667
## 70                                environment health and safety   11 0.3666667
## 71                                           for the first time   11 0.3666667
## 72                                      more information on our   11 0.3666667
## 73                                           one of the largest   11 0.3666667
## 74                                                the end of we   11 0.3666667
## 75                                               to do the same   11 0.3666667
## 76                                             with the goal of   11 0.3666667
## 77                                       and the communities in   11 0.3666667
## 78                                  commitment to diversity and   11 0.3666667
## 79                               communities across the country   11 0.3666667
## 80                                      communities in which we   11 0.3666667
## 81                                contributed more than million   11 0.3666667
## 82                               family online safety institute   11 0.3666667
## 83                                       in more than countries   11 0.3666667
## 84                                             in which we live   11 0.3666667
## 85                                         of more than million   11 0.3666667
## 86                                          of our business and   11 0.3666667
## 87                                        our employees and our   11 0.3666667
## 88                                     the communities where we   11 0.3666667
## 89                              the global reporting initiative   11 0.3666667
## 90                                             the scope of our   11 0.3666667
## 91                                           to make a positive   11 0.3666667
## 92                                               as well as the   11 0.3666667
## 93                                            in the process of   11 0.3666667
## 94                                         to make a difference   11 0.3666667
## 95                                      have the opportunity to   11 0.3666667
## 96                                      the national center for   11 0.3666667
## 97                                                year in a row   11 0.3666667
## 98                                               in the fall of   11 0.3666667
## 99                                            right thing to do   11 0.3666667
## 100                                          the right thing to   11 0.3666667
## 101                                            a broad range of   11 0.3666667
## 102                                          a broader range of   10 0.3333333
## 103                              and career development reviews   10 0.3333333
## 104                                               as of the end   10 0.3333333
## 105                                greenhouse gas ghg emissions   10 0.3333333
## 106                                             in an effort to   10 0.3333333
## 107                                               of the end of   10 0.3333333
## 108                          performance and career development   10 0.3333333
## 109                              regular performance and career   10 0.3333333
## 110                                   senior vice president and   10 0.3333333
## 111                              senior vice president corporate   10 0.3333333
## 112                                             state of the art   10 0.3333333
## 113                                             the power of our   10 0.3333333
## 114                                     vice president and chief   10 0.3333333
## 115                                              we are proud to   10 0.3333333
## 116                                            we are working to   10 0.3333333
## 117                                      with family and friends   10 0.3333333
## 118                                           and we continue to   10 0.3333333
## 119                                               as a member of   10 0.3333333
## 120                                   comply with all applicable   10 0.3333333
## 121                                 for people with disabilities   10 0.3333333
## 122                                      high speed internet and   10 0.3333333
## 123                                             more than of our   10 0.3333333
## 124                                           of our company and   10 0.3333333
## 125                                          of our supply chain   10 0.3333333
## 126                                             on a daily basis   10 0.3333333
## 127                                        our commitment to the   10 0.3333333
## 128                                     products and services to   10 0.3333333
## 129                              reduce the environmental impact   10 0.3333333
## 130                              senior executive vice president   10 0.3333333
## 131                                    the products and services   10 0.3333333
## 132                                         through a variety of   10 0.3333333
## 133                                  to reduce the environmental   10 0.3333333
## 134                                        to the communities we   10 0.3333333
## 135                                            where we live and   10 0.3333333
## 136                                          around the world to   10 0.3333333
## 137                                       committee of the board   10 0.3333333
## 138                                          in which we operate   10 0.3333333
## 139                                              of the board of   10 0.3333333
## 140                                    and in kind contributions   10 0.3333333
## 141                                     in kind contributions to   10 0.3333333
## 142                                           one of the world’s   10 0.3333333
## 143                             organizations across the country   10 0.3333333
## 144                                           the past two years   10 0.3333333
## 145                                 hispanic chamber of commerce   10 0.3333333
## 146                                      employees live and work   10 0.3333333
```

Calculate Number of words and tetragrams in each sentences

```r
##  NWoS stands for Number of Words of each Sentence

for (i in 1:nrow(text_stack_sample)){
  text_stack_sample$NWoS[[i]] <- lapply(t[[i]],function(x) str_count(x,'\\w+'))
}


## Num of tetragram in each sentence

sen_list<- list()
```

```r
system.time(
for (i in 1:length(t)){
  print(i)
  sent_tetragram_count_list = list()
  for(sent in 1:length(t[[i]]))
  {
    temp = 0
    for (j in 1:nrow(N_table2)){

      ngrams = na.omit(ngram[[i]][[sent]])

      if(isTRUE(any(unlist(map(ngrams,str_detect,as.character(N_table2[j,1]))))))
      {
        temp = temp + 1
      }

    }
    sent_tetragram_count_list[[sent]] = temp
  }

  sen_list[[i]] = sent_tetragram_count_list

}
)
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
## [1] 17
## [1] 18
## [1] 19
## [1] 20
## [1] 21
## [1] 22
## [1] 23
## [1] 24
## [1] 25
## [1] 26
## [1] 27
## [1] 28
## [1] 29
## [1] 30
```

```
##    user  system elapsed 
##  286.89    0.06  287.39
```

Calculate Boilerplate

```
library(stringr)
## Get the length of each document
text_stack_sample$Length<-str_count(text_stack_sample[,1], '\\w+')

## Final Calculation
for (i in 1:nrow(text_stack_sample)){
  temp = 0
  for (sent in 1:length(text_stack_sample$NWoS[[i]])){
    if (sen_list[[i]][[sent]] != 0){
      temp = temp+text_stack_sample$NWoS[[i]][[sent]]
    }
  }
  text_stack_sample$BoilerPlate[i] = temp / text_stack_sample$Length[i]
}

text_stack_sample$BoilerPlate
```

```
##  [1] 0.1546690 0.1676224 0.2195359 0.1268242 0.1321867 0.1297530 0.2144227
##  [8] 0.2236854 0.2800286 0.1529663 0.2102302 0.1231205 0.1722892 0.1384798
## [15] 0.1124994 0.1502978 0.1231884 0.1075884 0.1494311 0.1500299 0.1468770
## [22] 0.1615247 0.1445208 0.1237633 0.1684940 0.1385408 0.1037732 0.1530027
## [29] 0.1664161 0.1285627
```