Week_2_Report

A.Nikitenko

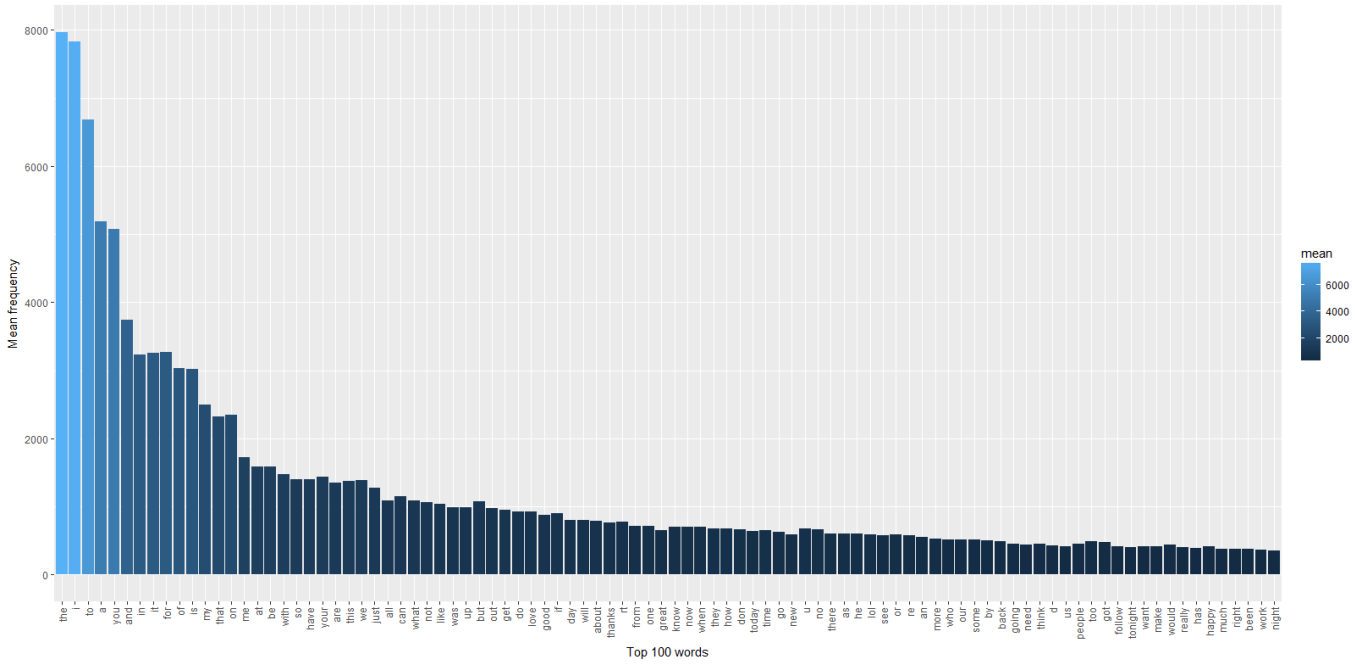
January 8, 2018

Words analysis

This is the week 2 report in Data Science specialization on Coursera, provided by John Hpkins University of Public Health. Report comprises exploratory analysis results of natural language data sets collecting Eanglish texts from Twitter, News services and private blogs. Each of the data set provides a separate text chunks one per line. Data sets are downloaded and unpacked as simple *.txt files. Sice of each file is around 200MB, which is more than a regular PC can handle in reasonable time. Threfore to analyze the content random samples of 20K lines has been taken and after cleaning saved as RDS objects. For the first analysis a simple cleaning has been performed fo each of samples. Cleaning is done in order to exclude unnecessary puntuation, number and other special characters thereby enabling to count words, see their distribution, etc…. Example of top 100 most used words distribution in one of the samples:

```
clean_speech <- readRDS("cleaned_tweet_1_20K.RDS");

freq <- factor(clean_speech$text)
tfreq <- table(freq)
dfreq = as.data.frame(tfreq)
colnames(dfreq) <- c('word','freq')
sorted <-  dfreq[order(-dfreq$freq),]
newhistdata <- sorted[! sorted$word %in% c('s','ll','t','m'),]

histdata <- head(newhistdata,100)

g <- ggplot(histdata, aes(x = reorder(word,-freq), fill = freq)) +
geom_bar(stat="identity", ymin=0, aes(y=freq, ymax=freq), position="dodge")+
theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))

print(g)
```

Word distribution in tweet texts



Unfortunately to reason about large datasets it is not enough to have just a single random sample. Therefore it is necessary to examine all 10 samples in combination.

```
##     word sample1 sample2 sample3 sample4 sample5 sample6 sample7 sample8 sample9 sample10
## 1      a    5312    5205    5125    5046    5170    5295    5233    5104    5181     5262
## 2  about     772     793     773     754     741     753     821     801     804      828
## 3    all    1134    1074    1063    1098    1081    1074    1029    1130    1111     1096
## 5     an     567     569     577     579     564     572     521     529     536      542
## 6    and    3752    3788    3857    3779    3655    3713    3669    3796    3701     3693
## 7    are    1368    1257    1339    1417    1301    1323    1363    1344    1349     1393
## 8     as     623     605     624     542     559     611     587     618     629      593
## 9     at    1601    1601    1606    1630    1588    1634    1566    1590    1520     1562
## 10  back     483     466     520     443     525     495     519     443     502      475
## 11    be    1538    1563    1582    1528    1545    1606    1683    1615    1625     1635
```

Summary of the combined data set

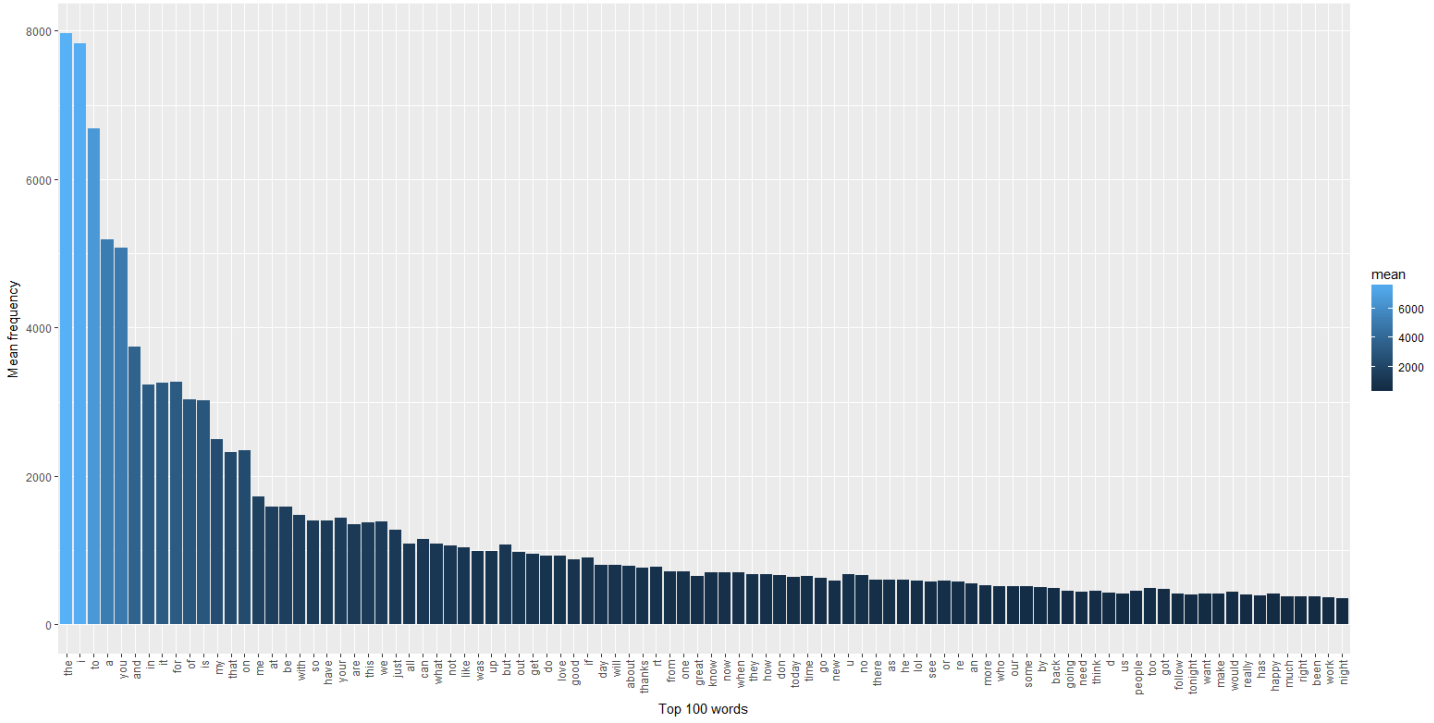
```
summary(finaldataset)
```

```
##       word       sample1        sample2          sample3
##  a      : 1   Min.   : 346   Min.   : 356.0   Min.   : 346
##  about  : 1   1st Qu.: 514   1st Qu.: 522.2   1st Qu.: 514
##  all    : 1   Median : 694   Median : 707.5   Median : 706
##  an     : 1   Mean   :1271   Mean   :1268.7   Mean   :1282
##  and    : 1   3rd Qu.:1321   3rd Qu.:1287.8   3rd Qu.:1320
##  are    : 1   Max.   :7934   Max.   :8091.0   Max.   :8013
##  (Other):88
##     sample4          sample5          sample6          sample7
##  Min.   : 354.0   Min.   : 354.0   Min.   : 342.0   Min.   : 336.0
##  1st Qu.: 534.2   1st Qu.: 512.5   1st Qu.: 518.2   1st Qu.: 519.5
##  Median : 715.0   Median : 704.0   Median : 691.5   Median : 713.0
##  Mean   :1283.4   Mean   :1274.0   Mean   :1285.8   Mean   :1273.2
##  3rd Qu.:1350.5   3rd Qu.:1286.2   3rd Qu.:1310.5   3rd Qu.:1320.8
##  Max.   :8073.0   Max.   :7942.0   Max.   :8090.0   Max.   :7930.0
##
##     sample8          sample9          sample10
##  Min.   : 360.0   Min.   : 344.0   Min.   : 353.0
##  1st Qu.: 514.5   1st Qu.: 492.2   1st Qu.: 523.2
##  Median : 711.0   Median : 693.5   Median : 711.0
##  Mean   :1275.0   Mean   :1271.7   Mean   :1284.1
##  3rd Qu.:1319.2   3rd Qu.:1329.0   3rd Qu.:1343.8
##  Max.   :7931.0   Max.   :7872.0   Max.   :7967.0
##
```

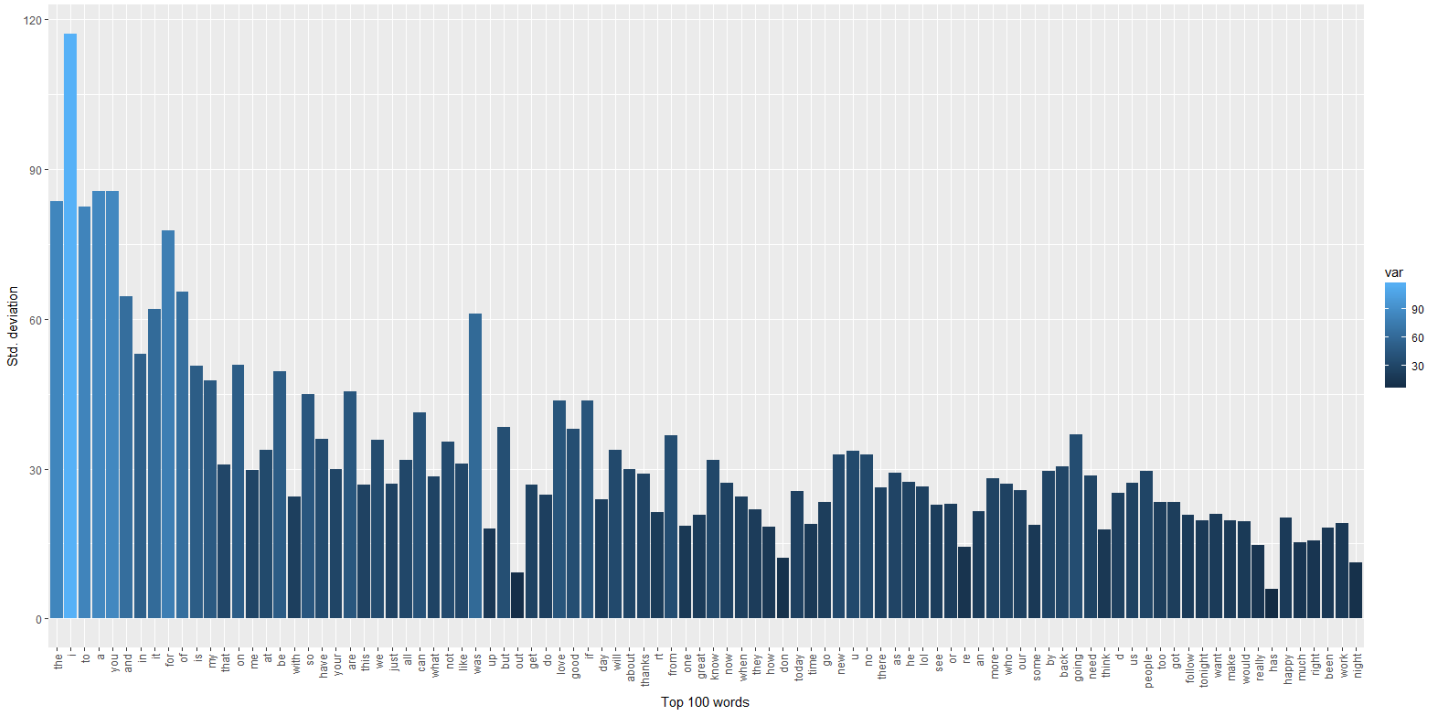
Mean values and variance of word frequencies

In order to determine features of word use it is worth to see mean values of use frequencies and theirstandard deviations.

Mean values



Standard deviation around mean



Blog texts

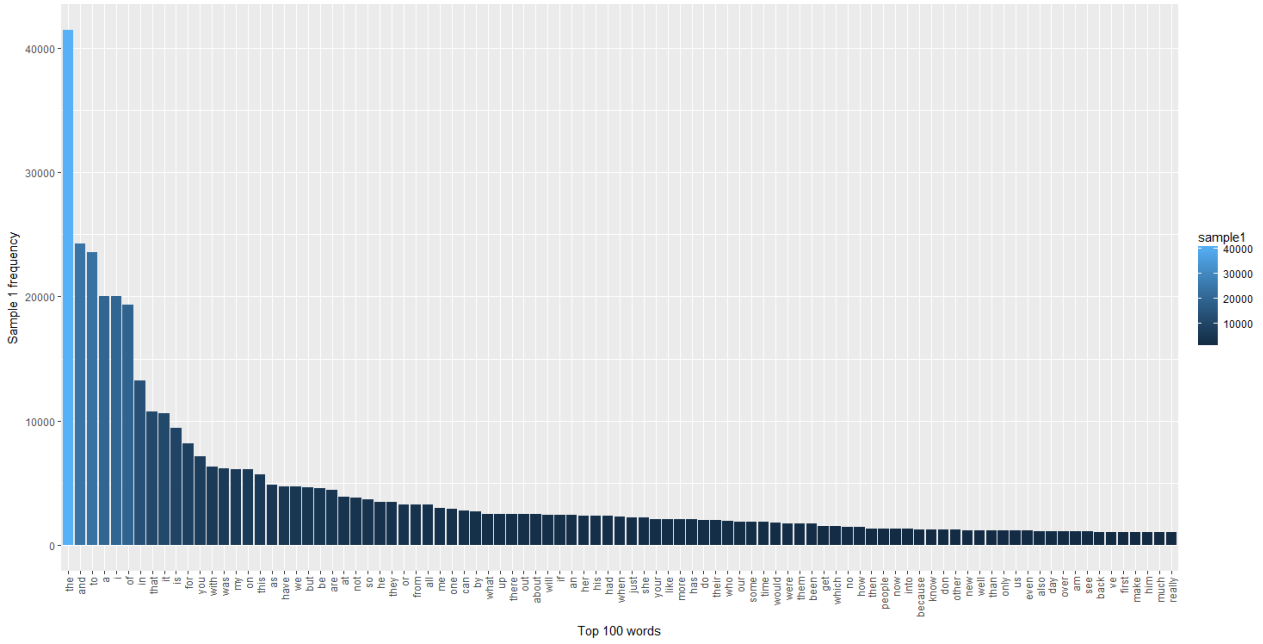
Tha same analysis has been done for Blog texts, using 10 random samples in combination. Only 10 words are presented in the following table.

```
##      word sample1 sample2 sample3 sample4 sample5 sample6 sample7 sample8  sample9 sample10
## 1       a   20064   19864   20181   19765   19957   20645   20157   20357    20452    20342
## 2   about    2490    2628    2654    2532    2515    2605    2554    2482     2627     2595
## 3     all    3248    3236    3280    3312    3266    3229    3216    3281     3281     3323
## 4    also    1168    1246    1217    1203    1187    1276    1211    1141     1195     1233
## 5      am    1136    1123    1136    1106    1086    1144    1119    1153     1143     1100
## 6      an    2416    2448    2474    2442    2309    2460    2418    2441     2408     2366
## 7     and   24245   24843   24533   24122   24082   25024   24131   24502    24593    24314
## 8     are    4489    4349    4319    4188    4110    4295    4084    4278     4229     4363
## 9      as    4854    5020    4944    4873    4973    5214    4950    5023     5029     4978
## 10     at    3913    3761    3820    3831    3814    3937    3779    3855     3851     3825
```

Summary of the combined blog data set

```
summary(finaldataset)

##       word       sample1         sample2         sample3
##  a      : 1   Min.   : 1043   Min.   : 1094   Min.   : 1084
##  about  : 1   1st Qu.: 1339   1st Qu.: 1346   1st Qu.: 1383
##  all    : 1   Median : 2297   Median : 2322   Median : 2314
##  also   : 1   Mean   : 4349   Mean   : 4384   Mean   : 4393
##  am     : 1   3rd Qu.: 3913   3rd Qu.: 3927   3rd Qu.: 3976
##  an     : 1   Max.   :41444   Max.   :41922   Max.   :41383
##  (Other):87
##     sample4         sample5         sample6         sample7
##  Min.   : 1025   Min.   : 1016   Min.   : 1088   Min.   : 1043
##  1st Qu.: 1366   1st Qu.: 1361   1st Qu.: 1411   1st Qu.: 1384
##  Median : 2295   Median : 2264   Median : 2383   Median : 2255
##  Mean   : 4299   Mean   : 4305   Mean   : 4463   Mean   : 4362
##  3rd Qu.: 3854   3rd Qu.: 3904   3rd Qu.: 3937   3rd Qu.: 3974
##  Max.   :40741   Max.   :41326   Max.   :41837   Max.   :41542
##
##     sample8         sample9         sample10
##  Min.   : 1046   Min.   : 1069   Min.   : 1080
##  1st Qu.: 1340   1st Qu.: 1391   1st Qu.: 1355
##  Median : 2299   Median : 2326   Median : 2220
##  Mean   : 4373   Mean   : 4394   Mean   : 4365
##  3rd Qu.: 3899   3rd Qu.: 3927   3rd Qu.: 3825
##  Max.   :41909   Max.   :42011   Max.   :41412
##
```

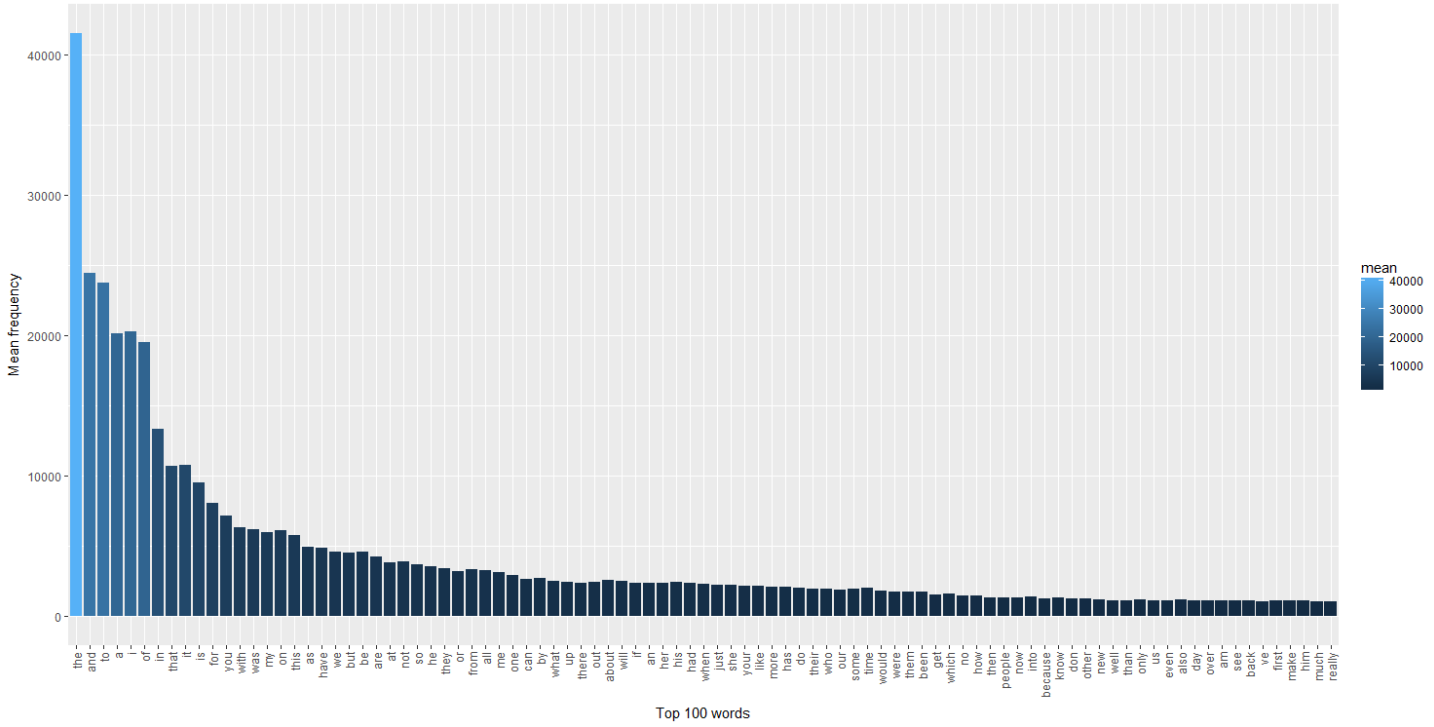
Word distribution in blog texts



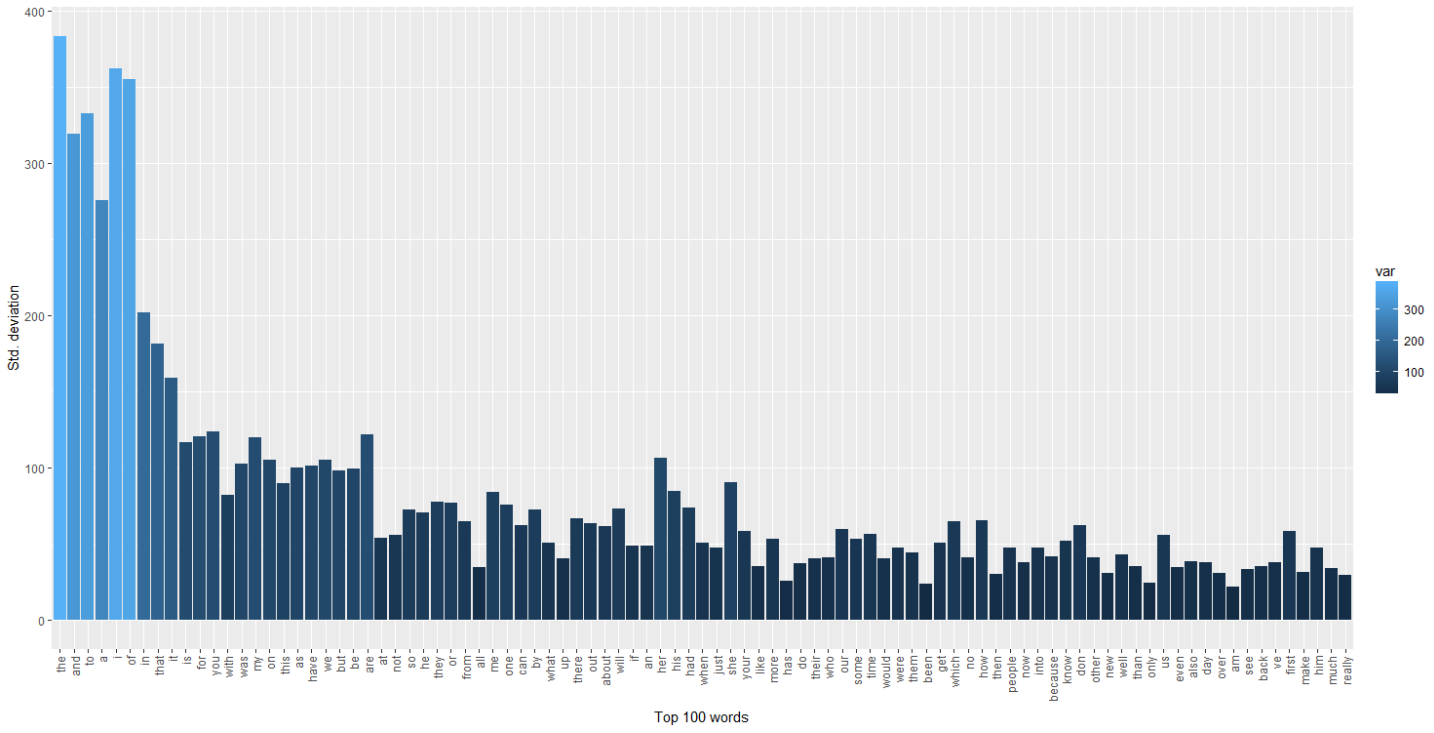
Mean values and variance of word frequencies

In order to determine features of word use it is worth to see mean values of use frequencies and theirstandard deviations.

Mean values



Standard deviation around mean



News texts

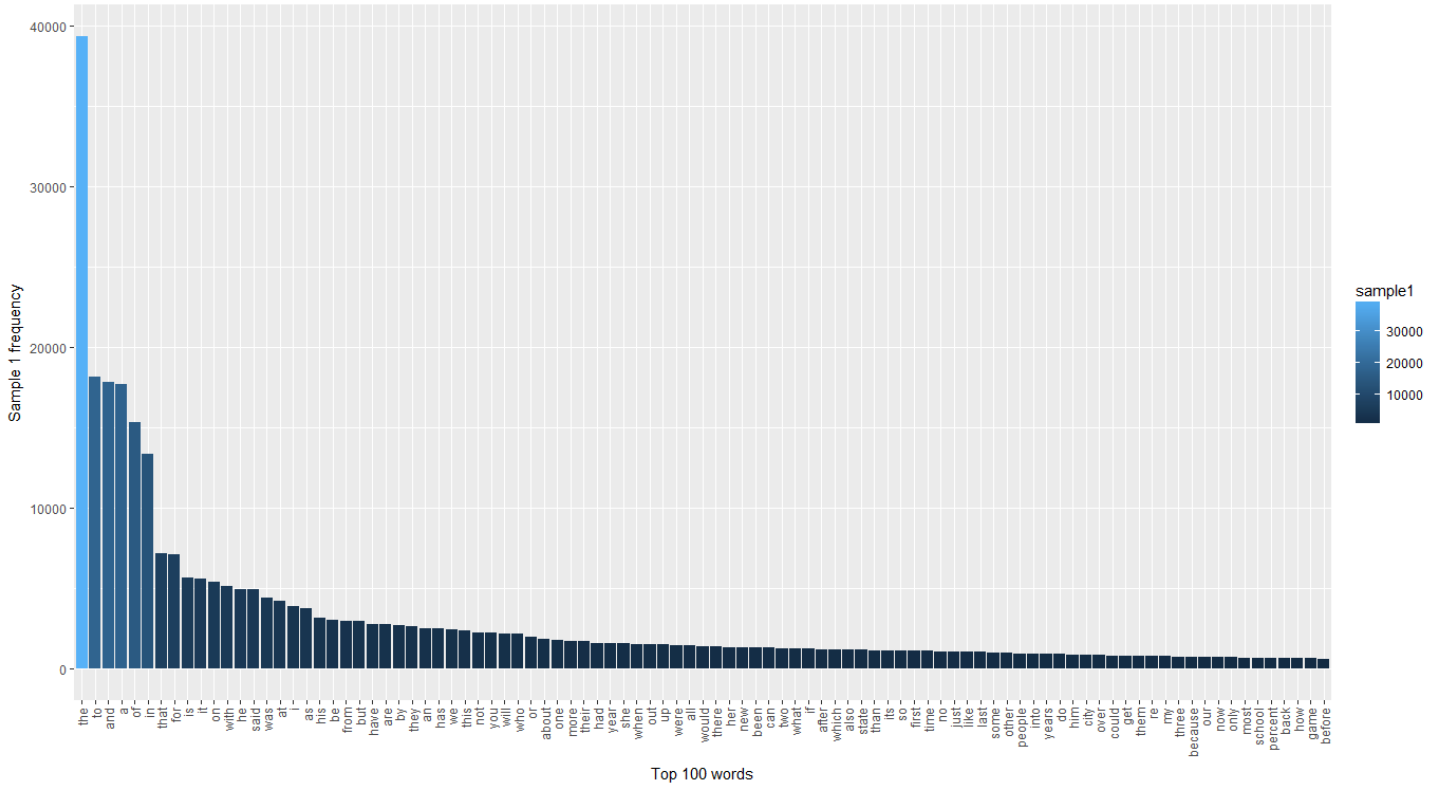
Tha same analysis has been done for news texts as well, using 10 random samples in combination.

```
##     word sample1 sample2 sample3 sample4 sample5 sample6 sample7 sample8  sample9 sample10
## 1      a   17713   17581   17793   17559   17685   17709   17891   17867    17641    17839
## 2  about    1845    1750    1798    1763    1805    1853    1769    1735     1790     1862
## 3  after    1236    1199    1220    1215    1278    1243    1187    1223     1220     1236
## 4    all    1450    1405    1439    1446    1465    1487    1469    1491     1450     1466
## 5   also    1192    1174    1218    1184    1134    1142    1191    1208     1179     1158
## 6     an    2502    2344    2380    2363    2495    2468    2373    2396     2502     2471
## 7    and   17843   17681   17661   17912   17624   18030   17782   17777    17605    17868
## 8    are    2792    2803    2833    2757    2712    2727    2734    2858     2778     2747
## 9     as    3794    3691    3760    3811    3723    3775    3770    3917     3714     3801
## 10    at    4241    4197    4378    4227    4299    4356    4312    4381     4361     4489
```

Distribution of words use in news texts

```
summary(finaldataset)
```

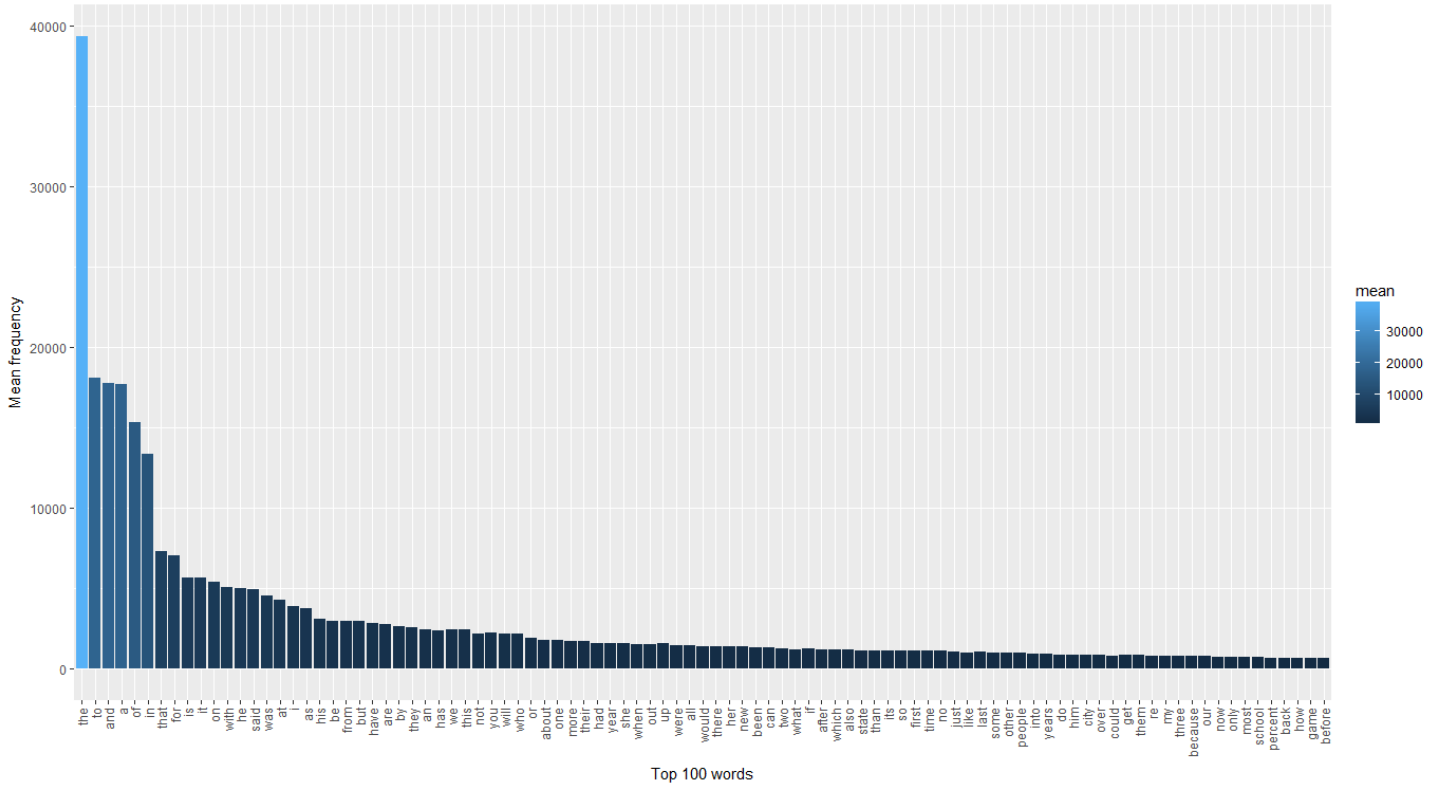
```
##       word       sample1           sample2         sample3
##  a      : 1   Min.   :  649.0   Min.   :  649   Min.   :  634
##  about  : 1   1st Qu.:  980.5   1st Qu.:  984   1st Qu.: 1020
##  after  : 1   Median : 1419.0   Median : 1435   Median : 1427
##  all    : 1   Mean   : 3109.0   Mean   : 3096   Mean   : 3126
##  also   : 1   3rd Qu.: 2743.0   3rd Qu.: 2753   3rd Qu.: 2764
##  an     : 1   Max.   :39331.0   Max.   :39487   Max.   :39549
##  (Other):89
##     sample4         sample5         sample6         sample7
##  Min.   :  657   Min.   :  637   Min.   :  649   Min.   :  667.0
##  1st Qu.:  996   1st Qu.: 1000   1st Qu.: 1004   1st Qu.:  992.5
##  Median : 1437   Median : 1446   Median : 1465   Median : 1415.0
##  Mean   : 3115   Mean   : 3102   Mean   : 3115   Mean   : 3120.5
##  3rd Qu.: 2716   3rd Qu.: 2706   3rd Qu.: 2734   3rd Qu.: 2728.5
##  Max.   :39436   Max.   :39259   Max.   :39292   Max.   :39229.0
##
##     sample8           sample9         sample10
##  Min.   :  650.0   Min.   :  659   Min.   :  684
##  1st Qu.:  995.5   1st Qu.:  975   1st Qu.: 1002
##  Median : 1462.0   Median : 1450   Median : 1466
##  Mean   : 3121.0   Mean   : 3088   Mean   : 3146
##  3rd Qu.: 2755.0   3rd Qu.: 2697   3rd Qu.: 2684
##  Max.   :39340.0   Max.   :38946   Max.   :39963
##
```



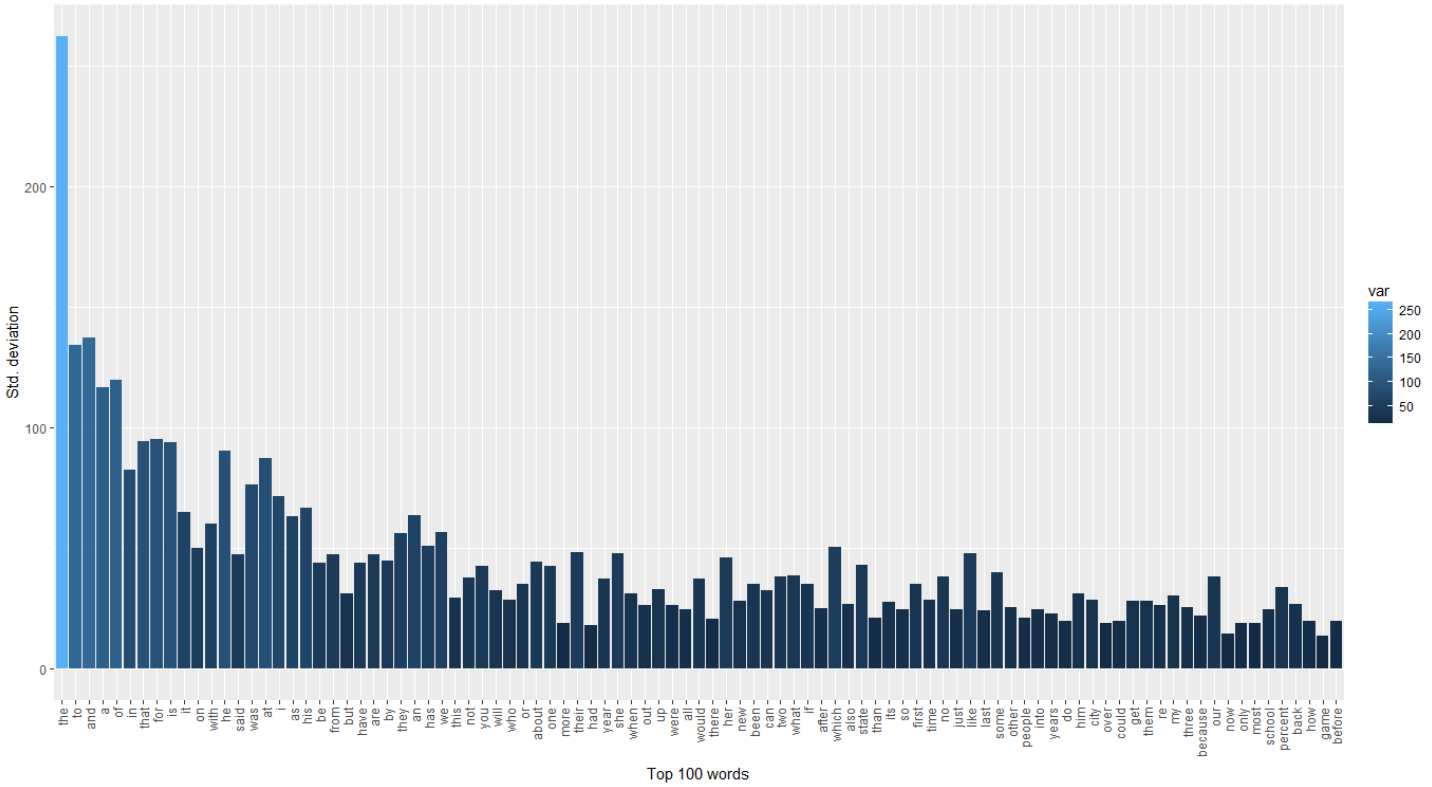
Mean values and variance of word frequencies

In order to determine features of word use it is worth to see mean values of use frequencies and theirstandard deviations.

Mean values



Standard deviation around mean



N-grams analysis

In the same way as done with use of single words, n-grams analysis is done. In particular 2-grams analysis for every data set. In the same way as for single word analysis the data sets has been sampled preprocessed using the following code:

```
on <- file("en_US.twitter.txt", "r")
lines <- readLines(con);

for(i in 1 : 10){
     text <- sample(lines, 20000)
     oz.str = paste(text, collapse = " ")

     # removing punctuation and unneccessary characters
     oz.corpus = Corpus(VectorSource(oz.str))
     oz.corpus = tm_map(oz.corpus, tolower)
     oz.corpus = tm_map(oz.corpus, removePunctuation, preserve_intra_word_dashes = FALSE)

     str(oz.corpus)
     filename <- paste0("corpus_twitter_",i,"_20K.RDS");
     saveRDS(oz.corpus,filename);
}
close(con)

con <- file("en_US.blogs.txt", "r")
lines <- readLines(con);

for(i in 1 : 10){
     text <- sample(lines, 20000)
     oz.str = paste(text, collapse = " ")

     # removing punctuation and unneccessary characters
     oz.corpus = Corpus(VectorSource(oz.str))
```

```
     oz.corpus = tm_map(oz.corpus, tolower)
     oz.corpus = tm_map(oz.corpus, removePunctuation, preserve_intra_word_dashes = FALSE)

     str(oz.corpus)
     filename <- paste0("corpus_blogs_",i,"_20K.RDS");
     saveRDS(oz.corpus,filename);
}
close(con)

con <- file("en_US.news.txt", "r")
lines <- readLines(con);

for(i in 1 : 10){
     text <- sample(lines, 20000)
     oz.str = paste(text, collapse = " ")

     # removing punctuation and unneccessary characters
     oz.corpus = Corpus(VectorSource(oz.str))
     oz.corpus = tm_map(oz.corpus, tolower)
     oz.corpus = tm_map(oz.corpus, removePunctuation, preserve_intra_word_dashes = FALSE)

     str(oz.corpus)
     filename <- paste0("corpus_news_",i,"_20K.RDS");
     saveRDS(oz.corpus,filename);
}
close(con)
```

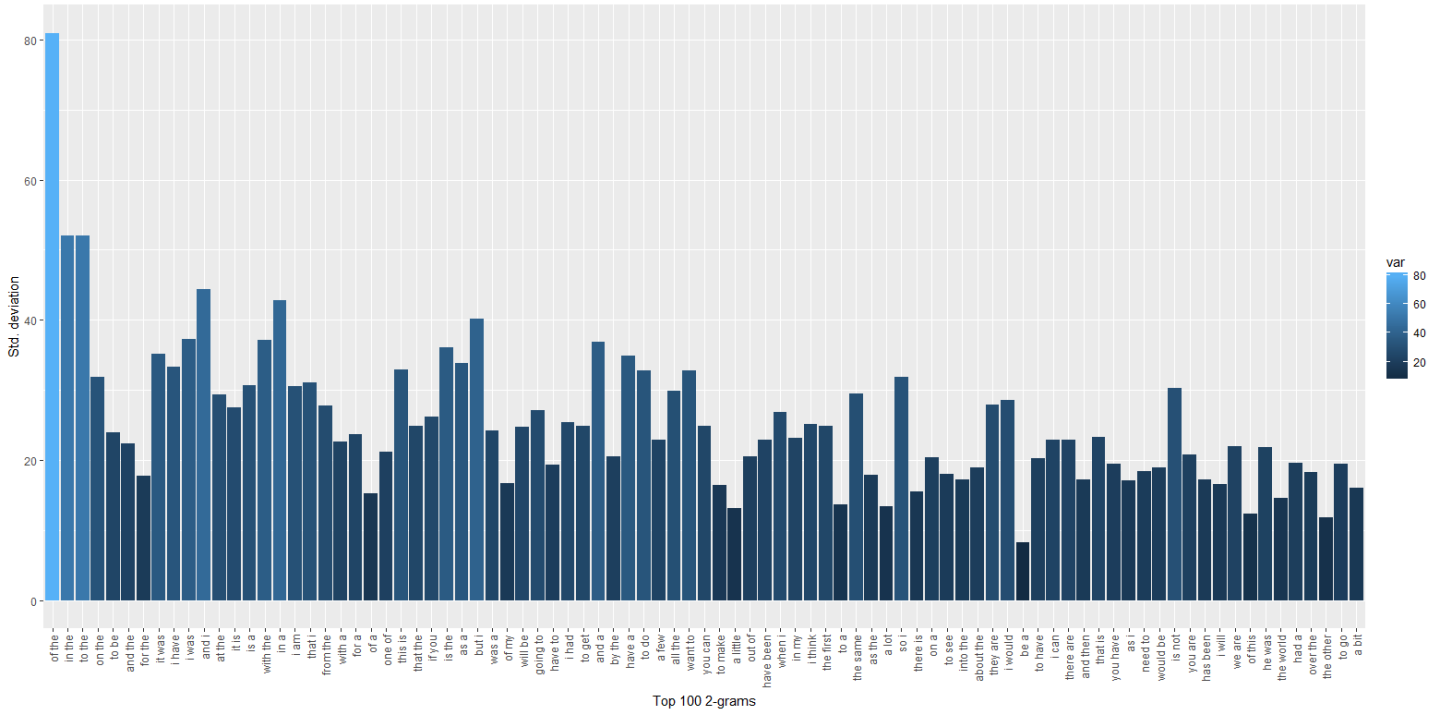
The code is prepared using guidance from: http://www.katrinerk.com/courses/words-in-a-haystack-an-introductory-statistics-course/schedule-words-in-a-haystack/r-code-the-text-mining-package

After the data is prepared, n-gram analysis can be done, using the following code, which calculates the basic statistics for top100 2-grams:

```
mergedData <- data.frame(gram = as.character(),
                         freq = as.numeric());

for(i in 1 : 10){
     filename <- paste0("corpus_blogs_",i,"_20K.RDS");
     oz.corpus <- readRDS(filename)

     cleaned.oz.str = as.character(oz.corpus)[1]
     oz.words = strsplit(cleaned.oz.str, " ", fixed = T)[[1]]
     oz.bigrams = vapply(ngrams(oz.words, 2), paste, "", collapse = " ")
     oz.bigram.counts = as.data.frame(xtabs(~oz.bigrams))

     new.oz.bigram.counts <- oz.bigram.counts[order(-oz.bigram.counts$Freq),]
     colnames(new.oz.bigram.counts) <- c('gram','freq')
     histdata <- head(new.oz.bigram.counts,100)

     if(nrow(mergedData)>0){
          m1 <- merge(mergedData, histdata, by.x = "gram", by.y = "gram",all.x=TRUE)
          mergedData <- m1
     }
     else{
          mergedData <- histdata;
     }
     print(paste0("Iteracija  ",i))
}

colnames(mergedData) <-
c('gram','sample1','sample2','sample3','sample4','sample5','sample6','sample7','sample8','sample9','sample10'
)
finaldataset <- mergedData[complete.cases(mergedData), ]

finaldataset$mean <- apply(finaldataset[,2:11],1,mean)
finaldataset$var <- apply(finaldataset[,2:11],1,var)
finaldataset$var <- sqrt(finaldataset$var)
```

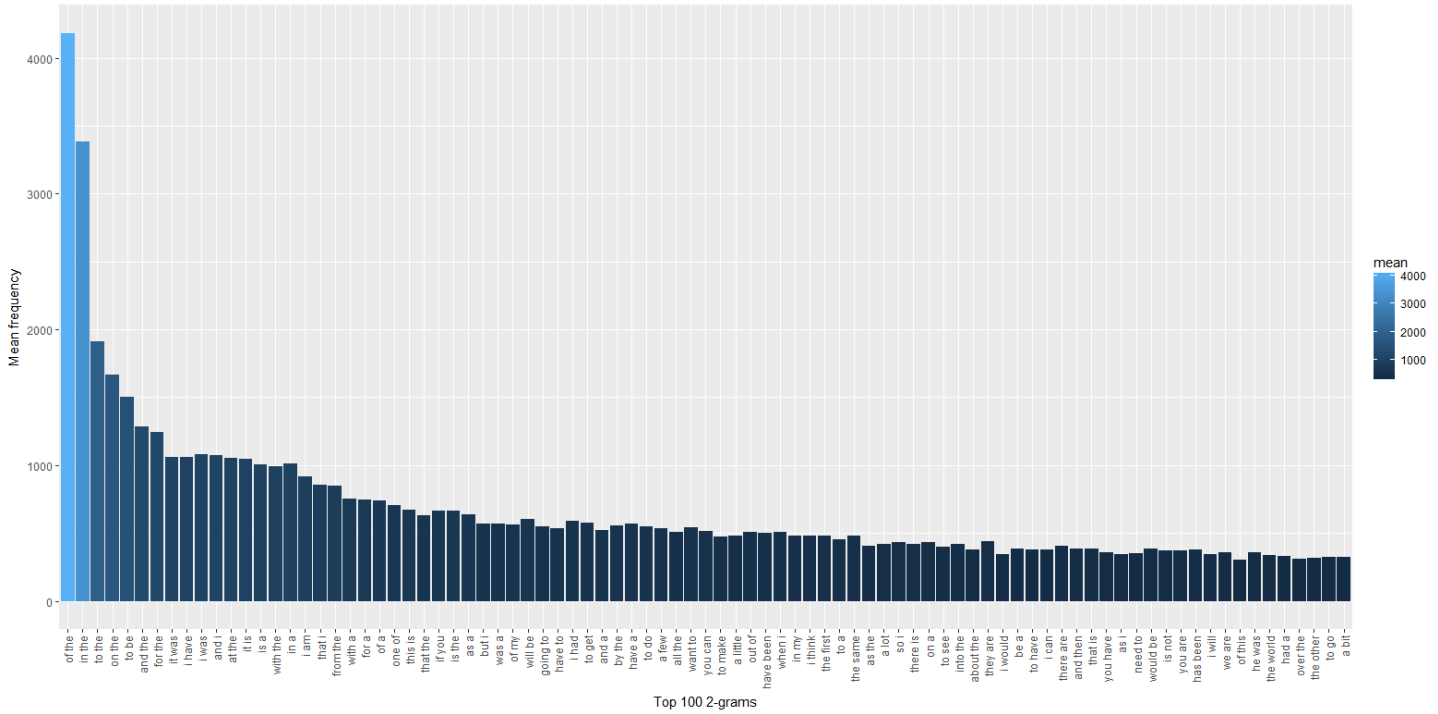
```
#preprocessed data is stored in RDS file for later use
saveRDS(finaldataset,"n_grams_blogs.RDS")

g <- ggplot(finaldataset, aes(x = reorder(gram,-sample1), fill = var)) +
     geom_bar(stat="identity", ymin=0, aes(y=var, ymax=var), position="dodge")+
     theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))+
     labs(x = "Top 100 2-grams", y = "Std. deviation")
g
```

Standard deviation of Blog 2-grams:



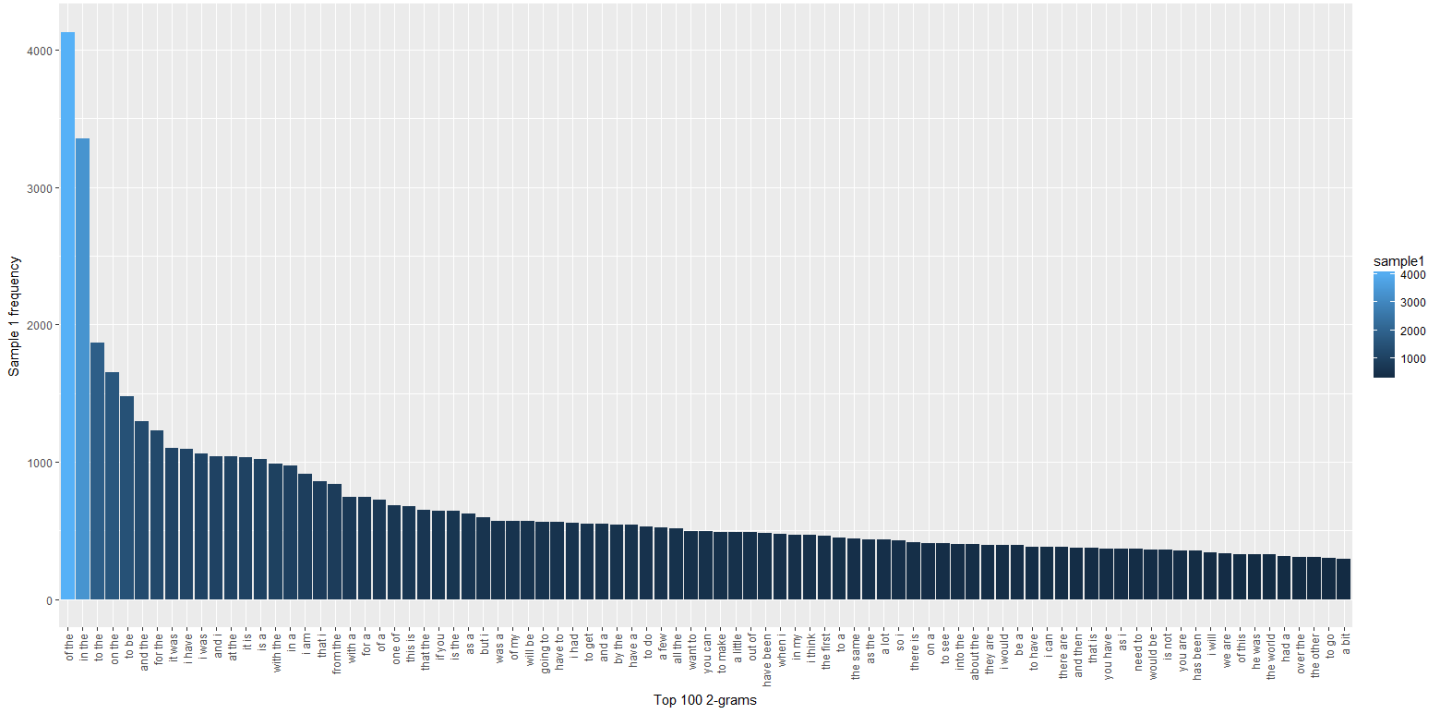

```
g <- ggplot(finaldataset, aes(x = reorder(gram,-sample1), fill = mean)) +
     geom_bar(stat="identity", ymin=0, aes(y=mean, ymax=mean), position="dodge")+
     theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))+
     labs(x = "Top 100 2-grams", y = "Mean frequency")
g
```

Mean values of Blog 2-grams:




```
g <- ggplot(finaldataset, aes(x = reorder(gram,-sample1), fill = sample1)) +
     geom_bar(stat="identity", ymin=0, aes(y=sample1, ymax=sample1), position="dodge")+
     theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
     labs(x = "Top 100 2-grams", y = "Sample 1 frequency")
g
```

Sample (Sample 1) values of Blog 2-grams:



Conclusions

Single word analysis shows that word use in general is similar from one data source to another, however it slightly differs and reflects the nature of the texts and their purpose. Twitter texts are short and reflects regular chat with limited diversity of the used words while other data sources are more related to more complex written thoughts. Thereby the use of words is different.

While none of the whole texts has not been studied it is done using 10 random sample of relatively small simple size 20K lines out of 2M (on average). In combination the samples allow to draw general conclusions and make assumptions of particular word use, which will be the key for building predictive models. This is justified by relatively small standard deviation value of word use frequencies.

N-gram analysis was more focused phrases with length 2 (words). The analysis was done in the same manner as single word analysis allowing to see actual use of consecutive words. In Blogs case the most phrases are something like “of the”, “I am”, and alike, which to large extent corresponds to English grammar constructions and do not propose any surprises.

The further analysis should be focused on modelling the English constructions.