Text Mining using MLK-speech and predefined positive & negative words

Surabhi Chouhan

December 25, 2016

```
library(tm)
```

```
## Loading required package: NLP
```

```
library(SnowballC)

library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

Read the positive and negative words

```
pos<-read.delim("D:/Surabhi docs/portfolio/positive.txt")

neg<-read.delim("D:/Surabhi docs/portfolio/negative.txt")
```

Clean the data

```
pos <- pos[-1:-33,]
head(pos,33)
```

```
##  [1] abound          abounds         abundance       abundant
##  [5] accessable      accessible      acclaim         acclaimed
##  [9] acclamation     accolade        accolades       accommodative
## [13] accomodative    accomplish      accomplished    accomplishment
## [17] accomplishments accurate        accurately      achievable
## [21] achievement     achievements    achievible      acumen
## [25] adaptable       adaptive        adequate        adjustable
## [29] admirable       admirably       admiration      admire
## [33] admirer
## 2031 Levels: ; ... zippy
```

```
str(pos)
```

```
##  Factor w/ 2031 levels ";",";  ",";          (editors: N. Indurkhya and F. J. Damerau), 2010.",..: 27
```

```
names(pos) <- c("Positive")

head(pos)
```

```
##   Positive       <NA>       <NA>       <NA>       <NA>       <NA>
##     abound    abounds  abundance   abundant accessable accessible
## 2031 Levels: ; ... zippy
```

```
neg <- neg[-1:-33,]
head(neg,33)
```

```
##  [1] 2-faces       abnormal      abolish       abominable    abominably
##  [6] abominate     abomination   abort         aborted       aborts
## [11] abrade        abrasive      abrupt        abruptly      abscond
## [16] absence       absent-minded absentee      absurd        absurdity
## [21] absurdly      absurdness    abuse         abused        abuses
## [26] abusive       abysmal       abysmally     abyss         accidental
## [31] accost        accursed      accusation
## 4808 Levels: ; ... zombie
```

```
str(neg)
```

```
##  Factor w/ 4808 levels ";",";  ",";          (editors: N. Indurkhya and F. J. Damerau), 2010.",..: 27
```

```
names(neg) <- c("Negative")

head(neg)
```

```
##   Negative       <NA>       <NA>       <NA>       <NA>       <NA>
##    2-faces   abnormal    abolish abominable abominably  abominate
## 4808 Levels: ; ... zombie
```

Process in the MLK Speech and perform text mining operations

```
mlk <- read.delim("D:/Surabhi docs/portfolio/MLK-Speech.txt")

str(mlk)
```

```
## 'data.frame':    28 obs. of  1 variable:
##  $ I.am.happy.to.join.with.you.today.in.what.will.go.down.in.history.as.the.greatest.demonstration.f
```

```
tail(mlk)
```

```
##
## 23
## 24
## 25
## 26
## 27
## 28 And when this happens, when we allow freedom to ring, when we let it ring from every village and
```

```
#Create a term matrix

words.vec <- VectorSource(mlk)
words.corpus <- Corpus(words.vec)
words.corpus
```

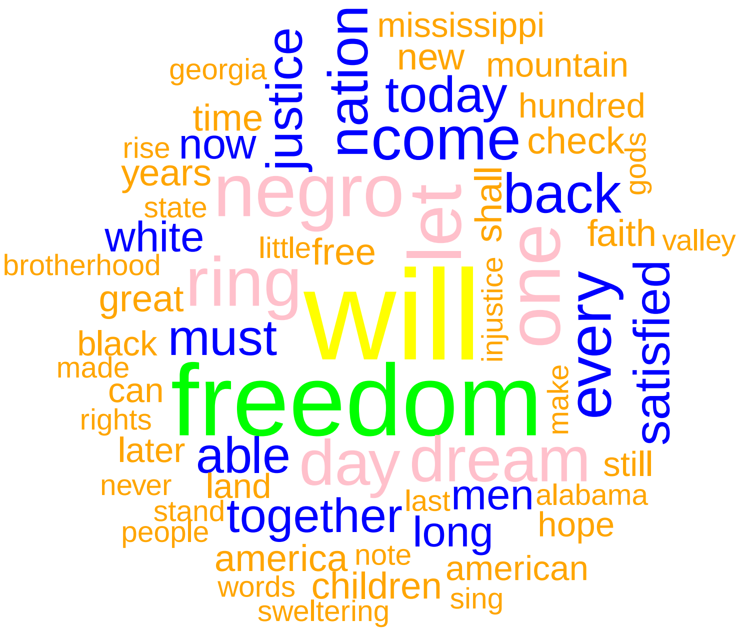
```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 1
```

```
tdm <- TermDocumentMatrix(words.corpus)
```

```
#   Create a list of counts for each word

m <- as.matrix(tdm)
wordCounts <- rowSums(m)
head(wordCounts)
```

```
##     'tis     able    again     ago,   ahead. alabama,
##        1        8        2        1        1        3
```

```
words.corpus <- tm_map(words.corpus, content_transformer(tolower))
words.corpus <- tm_map(words.corpus, removePunctuation)
words.corpus <- tm_map(words.corpus, removeNumbers)
words.corpus <- tm_map(words.corpus, removeWords, stopwords("english"))


##Generate a word cloud
wordcloud(words.corpus, max.words = 100, random.order = FALSE, colors=c("orange","blue","pink","green",
```



```
inspect(tdm)
```

```
## <<TermDocumentMatrix (terms: 561, documents: 1)>>
## Non-/sparse entries: 561/0
## Sparsity           : 0%
## Maximal term length: 15
## Weighting          : term frequency (tf)
##
##                    Docs
## Terms              1
##   'tis             1
##   able             8
##   again            2
##   ago,             1
##   ahead.           1
##   alabama,         3
##   all              7
##   alleghenies      1
##   allow            2
##   almighty,        1
##   alone.           1
##   also             1
##   always           1
##   america          5
##   american         3
##   american,        1
##   and             41
##   architects       1
##   are              8
##   areas            1
##   asking           1
##   autumn           1
##   awakening        1
##   back             8
##   back.            1
##   bad              1
##   bank             1
##   bankrupt.        1
##   basic            1
##   battered         1
##   beacon           1
##   beautiful        1
##   become           1
##   been             2
##   beginning.       1
##   believe          2
##   believes         1
##   bitterness       1
##   black            4
##   blow             1
##   bodies,          1
##   bound            1
##   boys             2
##   bright           1
##   brotherhood.     3
##   brothers,        1
##   brothers.        1
##   brutality.       2
##   business         1
##   but              6
##   california!      1
##   came             2
##   can              4
##   cannot            6
##   capital           1
##   captivity.        1
##   carolina,         1
##   cash              2
##   catholics,        1
##   cells.            1
##   chains            1
##   changed.          1
##   character.        1
##   check             3
##   check,            1
##   check.            1
##   children          3
##   children,         1
##   children.         1
##   cities,           1
##   cities.           1
##   citizens          1
##   citizenship       1
##   city,             1
##   civil             1
##   color             2
##   colorado!         1
##   come             10
##   community         1
##   concerned.        1
##   condition.        1
##   conduct           1
##   constitution      1
##   content           2
##   continue          2
##   cooling           1
##   corners           1
##   country,          1
##   created           1
##   creative          2
##   creed:            1
##   crippled          1
##   crooked           1
##   cup               1
##   curvaceous        1
##   dark              1
##   day               9
##   day,              1
##   day.              1
##   daybreak          1
##   declaration       1
##   decree            1
##   deeds.            1
##   deeply            1
##   defaulted         1
##   degenerate        1
##   demand            1
##   democracy.        1
##   desolate          1
##   despair           1
##   despair.          1
##   destiny           1
##   destiny.          1
##   devotees          1
##   died,             1
##   difficulties      1
##   dignity           2
##   discipline.       1
##   discontent        1
##   discords          1
##   discrimination.   1
##   distrust          1
##   down              3
##   dramatize         1
##   dream             9
##   dream.            2
##   drinking          1
##   dripping          1
##   drug              1
##   emancipation      1
##   emerges.          1
##   end               1
##   end,              1
##   engage            1
##   engulfed          1
##   equal.            1
##   equality.         1
##   even              2
##   every            10
##   evidenced         1
##   exalted,          1
##   exile             1
##   face              1
##   faith             5
##   fall              1
##   fatal             1
##   fathers           1
##   fatigue           1
##   fierce            1
##   finds             1
##   five              1
##   flames            1
##   flesh             1
##   for               8
##   force             1
##   force.            1
##   forever           1
##   former            2
##   foundations       1
##   four              1
##   free              4
##   free.             1
##   freedom          18
##   freedom.          1
##   fresh             1
##   friends,          1
##   from             18
##   funds             1
##   funds.            1
##   gain              1
##   gaining           1
##   gentiles,         1
##   georgia           1
##   georgia!          1
##   georgia,          1
##   ghetto            1
##   ghettos           1
##   girls             2
##   give              1
##   given             1
##   glory             1
##   god               1
##   god's             3
##   governor          1
##   gradualism.       1
##   granted           1
##   great             5
##   guaranteed        1
##   guilty            1
##   had               1
##   hallowed          1
##   hamlet,           1
##   hampshire.        1
##   hands             2
##   happens,          1
##   happiness.        1
##   has               5
##   hatred.           1
##   have             20
##   having            1
##   heat              2
##   heavy             1
##   heightening       1
##   heights           1
##   heir.             1
##   her               1
##   here              3
##   hew               1
##   high              1
##   highways          1
##   hill              2
##   hills             1
##   hilltops          1
##   himself           1
##   his               3
##   hold              1
##   honoring          1
##   hope              2
##   hope.             2
##   horrors           1
##   hotels            1
##   hundred           4
##   independence,     1
##   inextricably      1
##   injustice         1
##   injustice,        1
##   injustice.        1
##   insofar           1
##   instead           1
##   insufficient      2
##   interposition     1
##   into              4
##   invigorating      1
##   island            1
##   its               3
##   jail              2
##   jangling          1
##   jews              1
##   join              2
##   joyous            1
##   judged            1
##   justice           4
##   justice.          4
##   knowing           2
##   land              3
##   land.             1
##   languishing       1
##   larger            1
##   last!             3
##   later,            4
##   lead              1
##   leads             1
##   left              1
##   legitimate        1
##   let              13
##   liberty,          2
##   life              1
##   life,             1
##   lift              1
##   light             1
##   like              2
##   lips              1
##   little            3
##   live              2
##   lives             1
##   lodging           1
##   lonely            1
##   long              6
##   lookout           1
##   lord              1
##   louisiana,        1
##   low,              1
##   luxury            1
##   made              3
##   magnificent       1
##   majestic          1
##   make              3
##   manacles          1
##   many              1
##   march             1
##   marked            1
##   marvelous         1
##   material          1
##   meaning           1
##   meaning,          1
##   meeting           1
##   men               3
##   men,              3
##   midst             1
##   mighty            2
##   militancy         1
##   millions          1
##   mississippi       1
##   mississippi,      2
##   mississippi.      1
##   mobility          1
##   molehill          1
##   moment.           1
##   momentous         1
##   motels            1
##   mountain          4
##   mountains         1
##   mountainside,     2
##   must              8
##   narrow            1
##   nation            8
##   nation's          1
##   nation.           1
##   needed            1
##   negro            13
##   negro's           2
##   neither           1
##   never             3
##   new               5
##   night             1
##   nineteen          1
##   no,               2
##   nor               1
##   northern          1
##   not              13
##   note              3
##   nothing           1
##   now               5
##   now.              1
##   nullification;    1
##   oasis             1
##   obligation,       1
##   obvious           1
##   ocean             1
##   off               2
##   old               1
##   one              12
##   one.              1
##   only              1
##   only.             1
##   opportunity       1
##   oppression,       1
##   our              16
##   out               3
##   overlook          1
##   own               1
##   owners            1
##   palace            1
##   pass              1
##   path              1
##   pennsylvania!     1
##   people            2
##   people,           1
##   persecution       1
##   physical          2
##   pilgrim's         1
##   place             1
##   places            2
##   plain,            1
##   plane             1
##   pledge            1
##   police            2
##   poverty           1
##   pray              1
##   presence          1
##   pride,            1
##   process           1
##   proclamation.     1
##   prodigious        1
##   promise           1
##   promises          1
##   promissory        2
##   prosperity.       1
##   protest           1
##   protestants       1
##   pursuit           1
##   quest             1
##   quick             1
##   racial            2
##   racists,          1
##   real              1
##   reality           1
##   realize           2
##   red               1
##   redemptive.       1
##   refuse            2
##   remind            1
##   republic          1
##   rest              1
##   returns           1
##   revealed,         1
##   revolt            1
##   riches            1
##   right             1
##   righteousness     1
##   rightful          1
##   rights            1
##   rights,           1
##   rights.           1
##   ring              9
##   ring,             1
##   ring.             2
##   rise              3
##   robbed            1
##   rock              1
##   rockies           1
##   rolls             1
##   rooted            1
##   rough             1
##   rude              1
##   sacred            1
##   sadly             1
##   sands             1
##   satisfied         5
##   satisfied,        2
##   satisfied?        1
##   satisfy           1
##   say               2
##   score             1
##   seared            1
##   security          1
##   see               1
##   seek              1
##   segregation       2
##   self-evident:     1
##   selfhood          1
##   sense             1
##   shadow            1
##   shake             1
##   shall             5
##   shameful          1
##   signed            1
##   signing           1
##   signs             1
##   sing              2
##   sing.             1
##   sisters           1
##   sit               1
##   situation         1
##   sixty-three       1
##   skin              1
##   slave             1
##   slaves            2
##   slopes            1
##   slums             1
##   smaller           1
##   snowcapped        1
##   society           1
##   solid             1
##   some              3
##   somehow           1
##   something         1
##   sons              2
##   soul              1
##   south             2
##   speed             1
##   spiritual,        1
##   spot              1
##   staggered         1
##   stand             3
##   state             3
##   stating           1
##   steam             1
##   still             4
##   stone             2
##   storms            1
##   straight,         1
##   stream.           1
##   stripped          1
##   struggle          2
##   suffering         1
##   suffering.        1
##   summer            1
##   sunlit            1
##   sweet             1
##   sweltering        3
##   symbolic          1
##   symphony          1
##   table             1
##   take              1
##   tennessee!        1
##   thank             1
##   that             23
##   that;             1
##   the             100
##   thee              1
##   thee,             1
##   their             8
##   there             6
##   these             1
##   they              3
##   thirst            1
##   this             19
##   those             2
##   though            1
##   threshold         1
##   tied              1
##   time              5
##   today             3
##   today,            3
##   today.            2
##   together          1
##   together,         5
##   together.         1
##   tomorrow,         1
##   tranquility       1
##   tranquilizing     1
##   transform         1
##   transformed       1
##   travel,           1
##   trials            1
##   tribulations.     1
##   true              1
##   true.             1
##   truths            1
##   turn              1
##   unalienable       1
##   unearned          1
##   unmindful         1
##   unspeakable       1
##   until             4
##   upon              1
##   urgency           2
##   usual.            1
##   valley            3
##   vast              1
##   vaults            1
##   veterans          1
##   vicious           1
##   victim            1
##   village           1
##   violence.         1
##   vote              1
##   vote.             1
##   walk              1
##   walk,             1
##   wallow            1
##   warm              1
##   was               2
##   waters            1
##   well              1
##   were              1
##   when              7
##   where             3
##   which              5
##   whirlwinds         1
##   white              6
##   whites             1
##   who                4
##   whose              1
##   will              25
##   winds              1
##   with              14
##   with.              1
##   withering          1
##   words              3
##   work               2
##   would              2
##   wrongful           1
##   wrote              1
##   years              5
##   yes,               1
##   york               1
##   york.              1
##   you                7
##   your               1
```

```
m = as.matrix(tdm)
wordCounts = rowSums(m)


wordCounts <- sort(wordCounts, decreasing=TRUE)
head(wordCounts)
```

```
##  the  and will that have this
##  100   41   25   23   20   19
```

```
str(wordCounts)
```

```
##  Named num [1:561] 100 41 25 23 20 19 18 18 16 14 ...
##  - attr(*, "names")= chr [1:561] "the" "and" "will" "that" ...
```

```
total <- sum(wordCounts)
words <- names(wordCounts)
str(words)
```

```
##  chr [1:561] "the" "and" "will" "that" "have" "this" ...
```

```
#Step 3: Determine how many positive words were in the speech

countPos <- match(words,pos,nomatch=0)
countPos
```

```
##   [1]    0    0    0    0    0    0  765    0    0    0    0    0    0    0
##  [15]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##  [29]    0    0    0    0    0    0    0  672  856    0    0    0 1565    0
##  [43]    0    0    0    0    0    0  763    0    0    0    0    0    0    0
##  [57]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##  [71]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##  [85]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##  [99]    0    0  378  438    0    0    0    0    0    0    0    0    0    0
## [113]    0    0    0    0 1087    0 1178    0    0    0    0    0    0    0
## [127]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [141]    0    0    0    0    0    0 1986    0    0    0    0    0    0    0
## [155]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [169]    0  173    0    0    0    0    0    0    0  234    0    0    0    0
## [183]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [197]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [211]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [225]    0    0    0    0    0    0  430    0    0    0    0    0    0    0
## [239]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [253]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [267]    0    0    0    0    0    0  767    0    0    0  782  786    0    0
## [281]    0    0    0    0    0    0  821    0    0    0    0    0    0    0
## [295]  873    0    0    0    0    0    0    0    0    0    0    0    0    0
## [309]    0    0    0    0    0  922    0    0    0    0    0    0    0    0
## [323]    0    0 1026    0    0    0 1045    0    0    0    0 1072 1074    0
## [337]    0    0    0    0    0    0    0    0    0    0    0    0    0 1127
## [351] 1134 1136    0    0    0    0 1144    0    0    0    0    0    0    0
## [365]    0    0    0    0    0 1188    0    0    0    0    0    0    0    0
## [379]    0    0    0    0    0    0 1216    0    0    0    0    0    0    0
## [393]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [407]    0    0    0    0    0    0    0    0 1355 1371 1373    0    0    0
## [421]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [435]    0 1532 1536 1537    0    0    0    0    0    0    0    0    0    0
## [449]    0    0    0    0    0 1567    0    0    0    0    0    0    0    0
## [463]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [477]    0    0    0    0 1642    0    0    0    0    0    0    0    0    0
## [491]    0    0    0    0    0    0    0    0 1768    0    0    0    0    0
## [505] 1791    0    0    0    0    0    0    0    0    0    0    0 1829    0
## [519]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [533]    0    0    0    0    0    0    0    0    0    0    0    0    0 1928
## [547]    0 1935    0    0    0    0    0    0    0    0    0    0    0    0
## [561]    0
```

```r
matchCounts <- wordCounts[which(countPos    !=  0)]
length(matchCounts) #42 positive words
```

```
## [1] 42
```

```r
#Step 4: Determine how many negative words were in the speech

countNeg <- match(words,neg,nomatch=0)
countNeg
```

```
##   [1]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##  [15]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##  [29]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##  [43]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##  [57]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##  [71]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##  [85]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
##  [99]    0    0    0    0    0    0    0    0    0    0    0    0    0 2476
## [113]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [127]    0    0    0    0    0 3483    0    0    0    0    0 3857    0    0
## [141]    0 4081    0    0    0    0    0    0    0    0    0    0    0    0
## [155]    0    0    0    0    0    0    0    0    0  245    0    0    0  273
## [169]    0    0    0    0    0  339  376    0    0    0    0    0    0    0
## [183]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [197]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [211]    0  759  773    0    0  825    0    0    0    0    0    0    0    0
## [225]  903    0    0  989  992    0    0    0    0    0 1070    0 1146    0
## [239]    0 1322    0    0 1394    0    0    0    0    0    0    0    0    0
## [253]    0    0 1530    0 1574 1612    0 1621 1664    0    0    0    0    0
## [267]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [281]    0    0 1841    0    0    0    0    0    0    0    0    0 1921    0
## [295]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [309]    0    0    0    0    0    0    0    0    0 2380 2431    0    0    0
## [323]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [337]    0    0    0    0    0    0    0    0 2738    0    0    0    0    0
## [351]    0    0    0    0    0    0    0    0    0    0    0    0 2857    0
## [365]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [379]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [393]    0    0 3166    0    0    0    0    0    0    0 3263    0    0    0
## [407]    0    0 3341    0    0    0    0    0    0    0    0    0 3393    0
## [421]    0    0    0    0    0    0    0    0    0    0    0    0    0 3579
## [435]    0    0    0    0    0    0    0    0    0    0    0    0    0 3615
## [449] 3618    0 3648    0    0    0    0    0    0    0    0    0    0    0
## [463]    0 3768 3774    0    0    0    0    0    0    0    0    0 3856    0
## [477]    0    0    0    0    0    0    0    0    0    0    0    0    0    0
## [491]    0    0    0    0 4145    0    0    0    0    0    0    0    0    0
## [505]    0    0    0    0    0 4239    0    0    0    0    0    0    0    0
## [519]    0    0    0    0    0    0    0    0    0    0    0    0 4541    0
## [533]    0    0    0    0 4631    0    0    0    0    0    0    0 4671    0
## [547]    0    0    0    0    0    0    0    0    0 4772    0    0    0    0
## [561]    0
```

```r
matchCounts <- wordCounts[which(countNeg    !=  0)]
length(matchCounts) #46 negative words
```

```
## [1] 46
```

```r
#Step 5: Redo the 'positive' and 'negative' calculations for each 25% of the speech

#dividing speech into 4 equal parts
newdataframes = split(words, sample(rep(1:2:3:4, 459)))
```

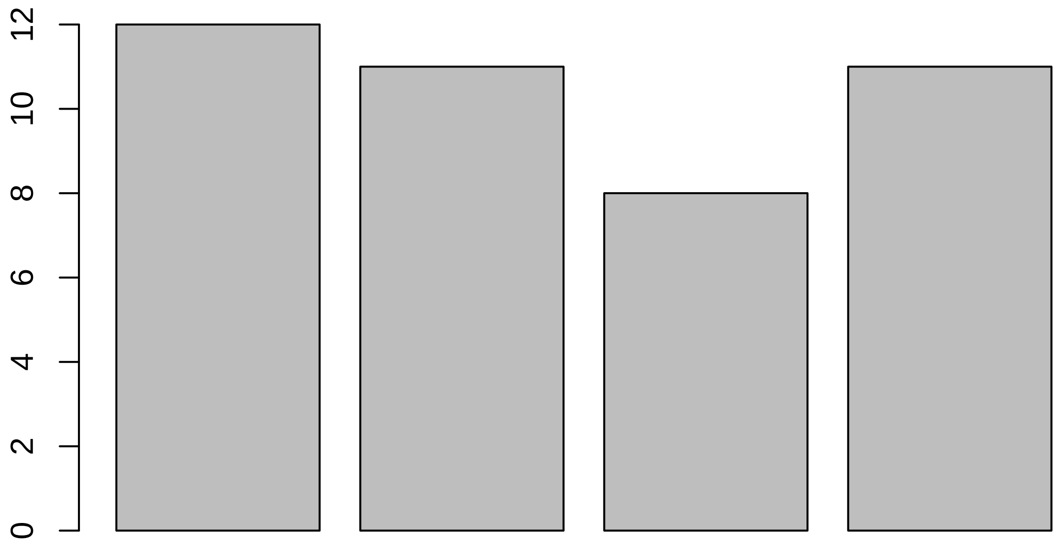
```
## Warning in 1:2:3: numerical expression has 2 elements: only the first used
```

```
## Warning in 1:2:3:4: numerical expression has 3 elements: only the first
## used
```

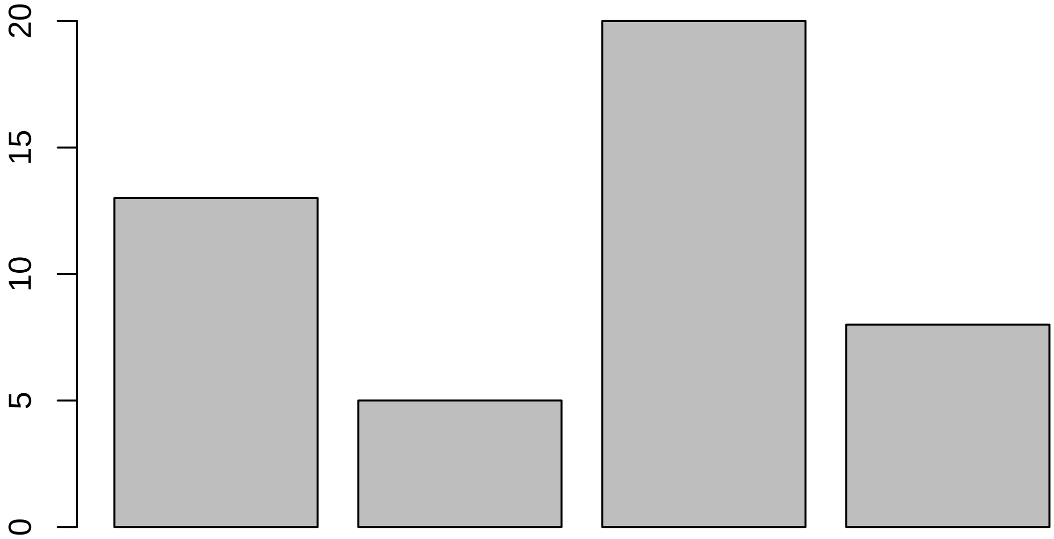
```
## Warning in split.default(words, sample(rep(1:2:3:4, 459))): data length is
## not a multiple of split variable
```

```r
#now matching to get the positive words
#first 25 percent
matchedpos1 = match(newdataframes$`1`,pos,nomatch = 0)

#count of match words
matchcountspos1 = wordCounts[which(matchedpos1  !=  0)]
pos1 = length(matchcountspos1)
```

```
#second 25 percent
matchedpos2 = match(newdataframes$`2`,pos,nomatch = 0)

#count of match words
matchcountspos2 = wordCounts[which(matchedpos2  !=  0)]
pos2 = length(matchcountspos2)

#third 25 percent
matchedpos3 = match(newdataframes$`3`,pos,nomatch = 0)

#count of match words
matchcountspos3 = wordCounts[which(matchedpos3  !=  0)]
pos3 = length(matchcountspos3)

#fourth 25 percent
matchedpos4 = match(newdataframes$`4`,pos,nomatch = 0)

#count of match words
matchcountspos4 = wordCounts[which(matchedpos4  !=  0)]
pos4 = length(matchcountspos4)

#barchart for comparing the result
h = c(pos1,pos2,pos3,pos4)
barplot(h)
```



```r
#similarly for the negative words

#first 25 percent
matchedneg1 = match(newdataframes$`1`,neg,nomatch = 0)

#count of match words
matchcountsneg1 = wordCounts[which(matchedneg1  !=  0)]
neg1 = length(matchcountsneg1)

#second 25 percent
matchedneg2 = match(newdataframes$`2`,neg,nomatch = 0)

#count of match words
matchcountsneg2 = wordCounts[which(matchedneg2  !=  0)]
neg2 = length(matchcountsneg2)

#third 25 percent
matchedneg3 = match(newdataframes$`3`,neg,nomatch = 0)

#count of match words
matchcountsneg3 = wordCounts[which(matchedneg3  !=  0)]
neg3 = length(matchcountsneg3)

#fourth 25 percent
matchedneg4 = match(newdataframes$`4`,neg,nomatch = 0)

#count of match words
matchcountsneg4 = wordCounts[which(matchedneg4  !=  0)]
neg4 = length(matchcountsneg4)

#barchart for seeing the result
J = c(neg1,neg2,neg3,neg4)
barplot(J)
```


0
5
10
15
20