Cleaning Text Data Using R

I have a data frame having more than 100 columns and 1 million rows. One column is the text data. The text data column contains huge sentences. I have written a code to clean the data but it's not cleaning. I want to remove all stop words, "the", "you", "like" "for" so on.

```
scorel= function(sentences, pos.words, .progress='none')
{
  require(plyr)
 require(stringr)


scores = laply(sentences, function(sentence, pos.words)
{

# clean up sentences with R's regex-driven global substitute, gsub():
sentence = gsub('[[:punct:]]', '', sentence)
sentence = gsub('[[:cntrl:]]', '', sentence)
sentence = gsub('\\d+', '', sentence)
sentence = gsub("@\\w+ *", "", sentence)
# and convert to lower case:
sentence = tolower(sentence)

# split into words. str_split is in the stringr package
word.list = str_split(sentence, '\\s+')
words = unlist(word.list)
# compare our words to the dictionaries of positive & negative terms
pos.matches = match(words, pos.words)
# match() returns the position of the matched term or NA
# we just want a TRUE/FALSE:
#     pos.matches = !is.na(pos.matches)

pos.matches=!is.na(pos.matches)

# and conveniently enough, TRUE/FALSE will be treated as 1/0 by sum():
#score = sum(pos.matches)
score = sum(pos.matches)
return(score)
 }, #pos.words, neg.words, .progress=.progress )
  pos.words, .progress=.progress )

 scores.df = data.frame(score=scores, text=sentences)
 return(scores.df)
```

```
}
 Data <- read.csv("location", stringsAsFactors=FALSE)
 Data<-Data[!duplicated(Data), ]
 Text <- data.frame(as.factor(Data$speech))
 names(Text)<-"Conversation"
 textf<-Text$Conversation
 textf<- unique(textf)
Text <- as.factor(textf)

score<- scorel(Text, disgust, .progress='text')


r
```

asked 2 hours ago

Olay
18 4

3 Are you aware of `tm` package? Check this link – Sotos 2 hours ago

In this link we have to convert to Corpus and then remove stop words... Is there any way we can convert back to data frame and pass it to function above. – Olay 1 hour ago

@Olay Kindly check the answer that I have provided. You can directly copy the suggested code into your code that mentioned in the question. – Saurabh13 1 hour ago

1 Answer

You can try `tm` package as follow:

```
corpus <- Corpus(VectorSource(sentence)) # Convert input data to corpus
corpus <- tm_map(corpus, removeWords, stopwords('english')) # Remove stop word using tm
package
dataframe<-data.frame(text=unlist(sapply(corpus, `[`, "content")),
                      stringsAsFactors=F) # Convert data back to data frame from corpus
sentence<-as.character(dataframe)
```

R console output is as follow:

```
> sentence=c('this is an best example','A person is nice')
> sentence
[1] "this is an best example" "A person is nice"
> corpus <- Corpus(VectorSource(sentence))
> corpus <- tm_map(corpus, removeWords, stopwords('english'))
> dataframe<-data.frame(text=unlist(sapply(corpus, `[`, "content")),
+                       stringsAsFactors=F)
> sentence<-as.character(dataframe)
> sentence
[1] "c(\"   best example\", \"A person  nice\")"
```

Hope it works for you

answered 1 hour ago

Saurabh13

176 11