Informatics Applied to Qualitative Data: Part-I (MPH Course: Health Informatics and Decision Making)

Zoie Shui-Yee Wong; Tshewang Gyeltshen

Contents

1. Loading free text documents and creating text Corpus

We will use the `tm` package in R to create a text corpus and perform initial data preprocessing.

```r
rm(list = ls())
setwd("C:/Users/zoies/Desktop/StLuke_MicrosoftPC/Teaching/HI-2023/W8Handson")
txt.csv <- read.csv("GPSAsummary_class.csv")
#load tm package
library(tm)
txt.csv$GPSA.summary <-  iconv(txt.csv$GPSA.summary, to ="utf-8")
txt<-Corpus(VectorSource(as.vector(txt.csv$GPSA.summary)))

# Created a corpus, you may inspect the first 3 documents using the following code
inspect(txt[1:3])
```

<> Metadata: corpus specific: 1, document level (indexed): 0 Content: documents: 3

[1] This article is a summary of an alert from the Pennsylvania Patient Safety Authority addressing the role of storage of medications in medication patient safety incidents. Actions to reduce risk for pharmacies are provided. [2] This Risk Alert discusses two medication patient safety incidents involving two high alert drugs - warfarin and heparin - where the wrong dose was dispensed and/or administered.
[3] This Risk Alert discusses five medication patient safety incidents with high alert medications.

```
# or you wish to inspect a particular doc
writeLines(as.character(txt[[20]]))
```

This short prevention strategy describes resources that should be readily available to deal with local anesthetic toxicity. The focus is on keeping a lipid emulsion for treatment and processes to ensure effective mitigation of the toxicity.

2. Preprocessing / Data preparation - Text transformation

Let use `getTransformations()` function to see all available text transformations currently available within `tm`

```
getTransformations()
```

[1] “removeNumbers” “removePunctuation” “removeWords”
[4] “stemDocument” “stripWhitespace”

2.1 tolower

```
txt <- tm_map(txt, content_transformer(tolower))
inspect(txt[1:3])
```

<> Metadata: corpus specific: 1, document level (indexed): 0 Content: documents: 3

[1] this article is a summary of an alert from the pennsylvania patient safety authority addressing the role of storage of medications in medication patient safety incidents. actions to reduce risk for pharmacies are provided. [2] this risk alert discusses two medication patient safety incidents involving two high alert drugs - warfarin and heparin - where the wrong dose was dispensed and/or administered.
[3] this risk alert discusses five medication patient safety incidents with high alert medications.

2.2 removeNumbers

```
txt <- tm_map(txt, removeNumbers)
inspect(txt[1:3])
```

<> Metadata: corpus specific: 1, document level (indexed): 0 Content: documents: 3

[1] this article is a summary of an alert from the pennsylvania patient safety authority addressing the role of storage of medications in medication patient safety incidents. actions to reduce risk for pharmacies are provided. [2] this risk alert discusses two medication patient safety incidents involving two high alert drugs - warfarin and heparin - where the wrong dose was dispensed and/or administered.
[3] this risk alert discusses five medication patient safety incidents with high alert medications.

2.3 removePuncuation

```
txt <- tm_map(txt, removePunctuation)
inspect(txt[1:3])
```

<> Metadata: corpus specific: 1, document level (indexed): 0 Content: documents: 3

[1] this article is a summary of an alert from the pennsylvania patient safety authority addressing the role of storage of medications in medication patient safety incidents actions to reduce risk for pharmacies are provided [2] this risk alert discusses two medication patient safety incidents involving two high alert drugs warfarin and heparin where the wrong dose was dispensed andor administered
[3] this risk alert discusses five medication patient safety incidents with high alert medications

2.4 removeWords

```
txt <- tm_map(txt, removeWords, stopwords("english"))
inspect(txt[1:3])
```

<> Metadata: corpus specific: 1, document level (indexed): 0 Content: documents: 3

[1] article summary alert pennsylvania patient safety authority addressing role storage medications medication patient safety incidents actions reduce risk pharmacies provided [2] risk alert discusses two medication patient safety incidents involving two high alert drugs warfarin heparin wrong dose dispensed andor administered
[3] risk alert discusses five medication patient safety incidents high alert medications

2.5 stripWhitespace

```
txt <- tm_map(txt, stripWhitespace)
inspect(txt[1:3])
```

<> Metadata: corpus specific: 1, document level (indexed): 0 Content: documents: 3

[1] article summary alert pennsylvania patient safety authority addressing role storage medications medication patient safety incidents actions reduce risk pharmacies provided [2] risk alert discusses two medication patient safety incidents involving two high alert drugs warfarin heparin wrong dose dispensed andor administered
[3] risk alert discusses five medication patient safety incidents high alert medications

2.6 stemDocument

```
txt <- tm_map(txt, stemDocument, language = "english")
inspect(txt[1:3])
```

<> Metadata: corpus specific: 1, document level (indexed): 0 Content: documents: 3

[1] articl summari alert pennsylvania patient safeti author address role storag medic medic patient safeti incid action reduc risk pharmaci provid [2] risk alert discuss two medic patient safeti incid involv two high alert drug warfarin heparin wrong dose dispens andor administ
[3] risk alert discuss five medic patient safeti incid high alert medic

```r
txtAll <- TermDocumentMatrix(txt)

inspect(txtAll[1:8, 1:3])
```

«TermDocumentMatrix (terms: 8, documents: 3)» Non-/sparse entries: 16/8 Sparsity : 33% Maximal term length: 7 Weighting : term frequency (tf) Sample : Docs Terms 1 2 3 action 1 0 0 address 1 0 0 alert 1 2 2 articl 1 0 0 author 1 0 0 incid 1 1 1 medic 2 1 2 patient 2 1 1

```r
#display the first 8 terms, from the first 3 documents
```

3. Saving Structured Data

```r
txt.matrix <- as.matrix(txtAll) ## turn doc matrix to matrix
txtdata<- data.frame(txt.matrix) ## turn matrix to data frame


## Transpose dataframe - from column to row
txtdata <- t(txtdata)
txtdata <- as.data.frame(txtdata)

## indexing row
rownames(txtdata) <- 1:nrow(txtdata)

# add the LASA class into the dataframe
txtdata$LASAcases<-txt.csv$LASAcases

## write csv file
write.csv(txtdata, "DatasetLASA.csv")
```