

Explaining the newDxCodes Function

Mark White | markhwhiteii@gmail.com

5/18/2017

This function is very derivative of the newAbbrevReport function, so I will only discuss the code where it differs from the newAbbrevReport function.

```
newDxCodes <- function(channel, month, year, oldcounts) {  
  
  suppressMessages(require(RODBC))  
  suppressMessages(require(dplyr))  
  suppressMessages(require(stringr))  
  suppressMessages(require(stringi))  
  suppressMessages(require(tidytext))  
  
  year_ch <- str_sub(year, start=-2)  
  get_ch <- data.frame(month=c("01", "02", "03", "04", "05", "06",  
                              "07", "08", "09", "10", "11", "12"),  
                      ch=c("jan", "feb", "mar", "apr", "may",  
                           "jun", "jul", "aug", "spt", "oct",  
                           "nov", "dec"))  
  month_ch <- as.character(get_ch[which(get_ch$month %in% month), "ch"])  
  
  newraw <- sqlQuery(channel,  
                    query=paste0("SELECT C_BioSense_ID, Diagnosis_Code  
                                FROM KS_PR_Processed  
                                WHERE MONTH(C_Visit_Date_Time) = ", month,  
                                " AND YEAR(C_Visit_Date_Time) = ", year))  
}
```

All of this is the same up until the SQL query. Instead of pulling down chief complaint and admit reason fields, we are now just pulling down the ID and diagnosis code fields.

```
newraw <- subset(newraw, subset=!is.na(Diagnosis_Code))  
  
newraw <- within(newraw, list(Diagnosis_Code <-  
  str_replace_all(Diagnosis_Code, "[;]", " "),  
                           Diagnosis_Code <-  
  str_replace_all(Diagnosis_Code, "[.]", "")))
```

This code cleans up the data we just pulled down with the query. First, we remove any rows that do not report diagnosis codes using the subset function, asking for all rows where is.na is FALSE. Then we replace all semicolons (what separates each code) with a space. Some facilities report periods in the ICD-10 codes, while others do not. To make this consistent, I'm simply removing all periods.

```

newword <- unnest_tokens(newraw, code, Diagnosis_Code)
newword <- filter(group_by(newword, C_BioSense_ID), !duplicated(code))
newcounts <- count(ungroup(newword), code, sort=TRUE)
colnames(newcounts) <- c("code", paste0("n_", month_ch, year_ch))

oldcounts$new <- factor("not new")
allcounts <- full_join(oldcounts, newcounts, by="code")
allcounts$new <- ifelse(is.na(allcounts$new), "new", "not new")
allcounts[is.na(allcounts)] <- 0

newcodes <- c(allcounts[which(allcounts$new=="new"), "code"])

report <- data_frame()
for (i in 1:length(newcodes)) {
  report <- bind_rows(report, unique(newword[which(newword$code %in%
newcodes[i]), c("C_BioSense_ID", "code")]))
}

report <- report[-grep("^[:digit:]*$", report$code),]

```

All of this is the same, up until the last line. Some facilities are starting to report SNOMED codes. For now, we want to exclude those. What separates these from ICD-10 codes is that they ONLY contain digits. ICD-10 codes always have a letter. So what I do in this last line is get rid of all rows where the code is ONLY digits (i.e., numbers).

```

output <- list(report, subset(allcounts, select=-new))
names(output) <- c("report", "allcounts")
return(output)
}

```

All of this is the same as in the newAbbrvReport function.