1. Load data in R

*Mydata<-read.table(“nameofdataset”,sep=”|”,header=T)*

2. Run descriptive summaries on the data- sample commands include

*names(Mydata)* – to get the rows

*summary(Mydata$ColumnNo)* -- to get a descriptive summary of a specific row

in case of numbers use

*summary(as factor(Mydata$ColumnNoOrName))*

We need to get the fields to use(as we can see we have too many fields already) To do this , we use the field metrics demonstrated ;last time

Summary of commands is : to install a package in R you need internet

library(entropy)

install.packages("entropy")

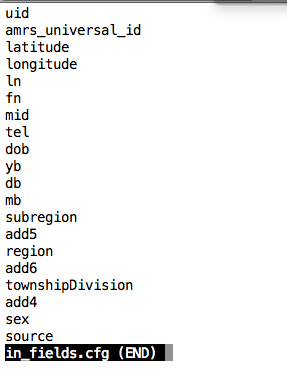
source("is\_good\_match\_field.R")

hct.fm<-field\_metrics(hct)

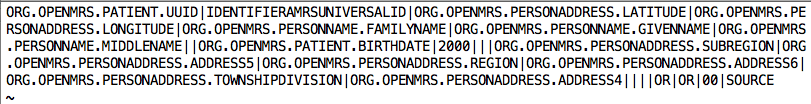
3. Clean data and transform it to the format for matching

We use the preprocessor script (preprocess.pl) – after doing the steps below

3.a Format your header to ensure it rhymes with the in.fields.cfg file (ensure that the values in this file correspond to those in your header



here is an example of a header



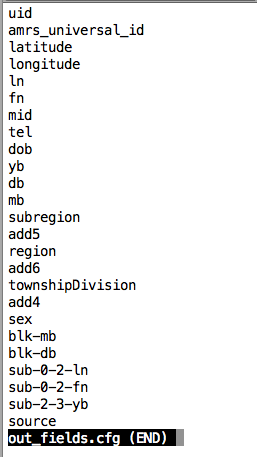
you should edit the header to tally with the in.fields file names

This will vary with the dataset that you are working on

The preprocessor and in.fields and out.fields files should all be in the same folder…because they reference to each other

I use emacs to change the header (you can use any text editor to change these files)

For the out.fields file, use the same fields as the in.fields files with a few modifications (see below)



In the out\_fields.cfg you add additional fields that hold the nysiis (phonetic transformation of names)

These fields include (refer to the screen shot above)

Blk-mb

Blk-db

Sub\*