**MIMIC-III-CharacterCount.nb.html**

Look at each of the .csv raw data files for unexpected ASCII characters:

ADMISSIONS.csv

CALLOUT.csv

CAREGIVERS.csv

CHARTEVENTS.csv

CPTEVENTS.csv

D\_CPT.csv

D\_ICD\_DIAGNOSES.csv

D\_ICD\_PROCEDURES.csv

D\_ITEMS.csv

D\_LABITEMS.csv

DATETIMEEVENTS.csv

DIAGNOSES\_ICD.csv

DRGCODES.csv

ICUSTAYS.csv

INPUTEVENTS\_CV.csv

INPUTEVENTS\_MV.csv

LABEVENTS.csv

MICROBIOLOGYEVENTS.csv

NOTEEVENTS.csv

OUTPUTEVENTS.csv

PATIENTS.csv

PRESCRIPTIONS.csv

PROCEDUREEVENTS\_MV.csv

PROCEDURES\_ICD.csv

SERVICES.csv

TRANSFERS.csv

Use **charcnt2.exe** command-line program to create frequency count of all characters in each file. (charcnt.exe is another command line program that creates a 16-by-16 matrix of counts. CharCount.exe is a Windows version of charcnt.exe.)

Write results to MIMIC-III-character-counts.csv and “make pretty” manually in **MIMIC-III-character-counts.xlsx** Excel file.

Review of the Excel file shows these possible problems:

* Most problems are in the file NOTEEVENTS.csv
* Line ends predominantly LFs, but why 1030 hex 0Ds?
* Nearly 50,000 tabs are in the file. Can they cause problems?
* What's the purpose for the 677 hex 13 and 49 hex 14 characters?
* Why are two hex 7F characters in the file?

**MIMIC-III-Fix-Bad-characters.html**

Fixes are easier in a Linux environment, but this notebook shows some approaches to fixes in Windows.

**Summary**: The “bad” characters in some of the files may be annoying but probably will not adversely affect most down-stream processing.