1. **Respond to Exercise 3 of Chapter 1. Browse the web for text mining software, and skim through the documentation of a few tools to get a feel for the state-of-the-art. Distinguish between free, free for trial period, and not free software. Include at least one application for each category.**

Free: Apache OpenNLP; high rated, very popular free toolset. Not as well rated as some not free versions, but very popular amongst small businesses. This package seems more focused on information extraction and processing, and less on predication.

Free for Trial Period: RapidMiner Studio; Highly rated. Integrated visualization tools, machine learning workflows/algorithms, database connectivity, data preprocessing.

Not Free: SAS Text Miner; industry standard, high performance fully featured suite of tools. Integrated visualization tools, analysis workflows.

1. **In the VAERS data set create a new variable called SERIOUS (without the double quotes) which is set to Y if any of the DIED, ER\_VISIT, HOSPITAL, or DISABLE variables are equal to Y or set to N if none of these are equal to Y. Also set the VAERS ID to a “key” role rather than an ID role for future work. Try to find some way to show the first half dozen or so rows with the SERIOUS variable.**

See below. Please note, for brevity, only 6 columns are shown (out of 31 in the dataset).

A close up of a piece of paper

Description automatically generated

1. **In any other data set establish a categorical target variable which acts like the dependent variable of a model. It should be a binary/binomial variable with values like Y/N, yes/no, true/false, 1/0, or a multinomial variable with integers values like 1-5 or verbal values like favorable/neutral/unfavorable.  It is the variable that will be predicted by the model built from the other data, text, and/or metadata variables.**

I downloaded weather data from a local weather station through NOAA. One of the existing columns was “HourlyPrecipitation”. This column contained either a null value (no precipitation), ‘T’ for trace precipitation, or a numerical value indication the rate of precipitation. I created a new column (shown in the screenshot below) named ‘precipitation’, which is a categorical True/False, indication whether or not there is precipitation. This column could be predicted using some of the other data in the dataset.

A screenshot of a cell phone

Description automatically generated

1. **Briefly describe the data set and comment on what type of data/text mining data it is as described in Chapter 1 of Weiss et al, and which variables might be suitable for the analyses suggested in Sections 1.5, 1.6, and 1.7.**

Section 1.5 describes clustering. That is, a methodology for grouping similar data together in an effort to learn about potential patterns/trends. A few columns that could be used in this type of analysis, depending on what is being analyzed, might include columns ALLERGIES, AGE\_YEARS, and SEX.

Section 1.6 discusses extracting information out of text. That is, building structured data out of unstructured data. There are a number of unstructured text columns that could be converted to a more structured dataset, such as SYMPTOM\_TEXT, HISTORY, PRIOR\_VAX, and OTHER\_MEDS.

Section 1.7 is about using this dataset to create predictions. Given the data we currently have, we want to be able to predict some outcome/output variable. The columns we would want to use in this analysis include those mentioned in the text above about sections 1.5 and 1.6. There are additional columns that might help prediction as well, such as VAX\_DATE and ONSET\_DATE. An excellent target variable for prediction would be the newly created SERIOUS column.