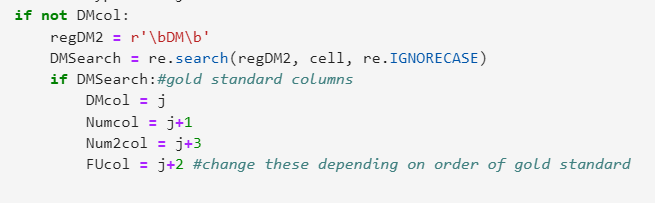
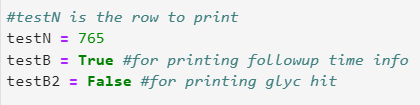
**Instructions for running:**

To run a specific excel sheet, make sure it is in the Input folder in the Follow-up folder and that the file path is changed.  


Then, if running on a file with a gold standard, make sure to change the relative order of the columns (in particular Num2col and FUcol.

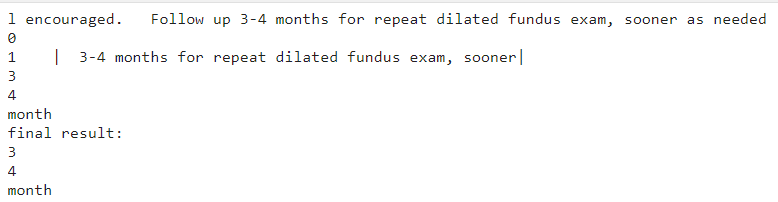
If the number columns are right next to each other e.g. 2|3|weeks, then set Num2col = j +2 and FUcol = j+3

If you want to debug a specific row, you can do so here:



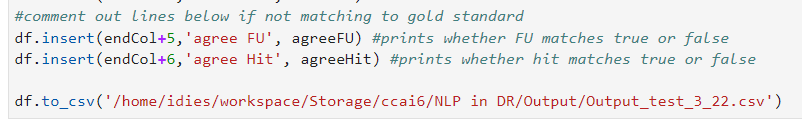
This will print the output of that row with either the follow-up info, or hit info selected

An example follow-up debugging output is:



The first line is the last 85 characters of the note. Each of the subsequent lines is the number of the priority and the surrounding text of any hits for that priority. E.g. follow-up is the second priority so it is listed under 1. The final date of followup is then printed for that cell.

Finally the last thing to consider if the notes contain gold standard annotations or not. If not, then the following can be commented out:

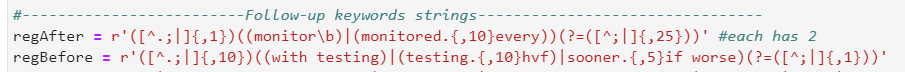


**Summary**:

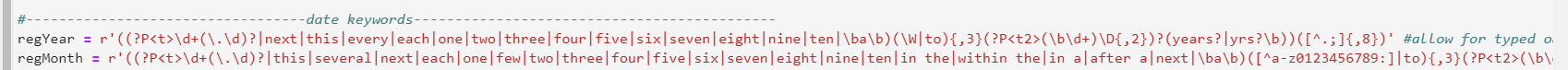
Cell 1: package imports

Cell 2: regex search strings

* First section contains the follow-up search keywords grouped and each group has its own capture sensitivity before and after the keyword e.g. keywords like “monitor” are usually only followed by the time of follow-up instead of preceded by the time and thus only the characters after the hit would be capture for further searching



The first and last parenthesized groups are the 2 surrounding texts capture. The carrot indicates that it will stop capturing if it encounters a semicolon etc. The “?=” in the end group makes it capture the group using positive lookahead, allowing for overlapping captures if there happen to be nearby keywords

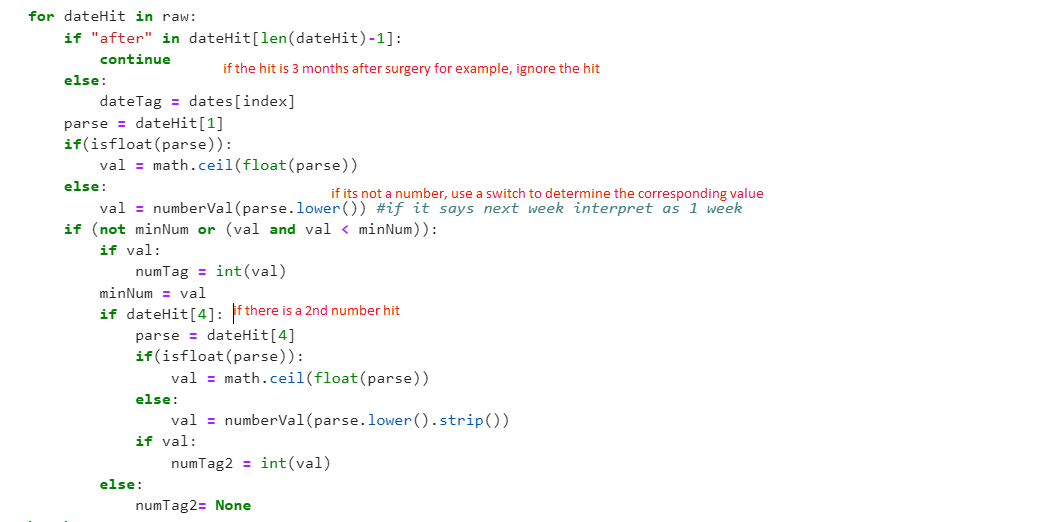
* regSearches is the order of priority of the keyword groups for problem lists in terms of which group is more likely to contain the actual follow-up time e.g. the text around “return to clinic” is more likely to contain the correct follow-up time than that of “evaluate”.
* The next section is used for searching the first row and identifying which columns contain which relevant information
* Date keywords contains each of the numeric and date-type regex keywords that are detected for the 4 date types (year, month week, and day). The order these types are searched are set in regDate. There are also a couple miscellaneous keywords used in the regex Extra strings

In this example : regYear = r'((?P<t>\d+(\.\d)?|next|this|every|each|one|two|three|four|five|six|seven|eight|nine|ten|\ba\b)(\W|to){,3}(?P<t2>(\b\d+)\D{,2})?(years?|yrs?\b))([^.;]{,8})' -> (?P<t>\d+(\.\d)?|next|this|every|each|one|two|three|four|five|six|seven|eight|nine|ten|\ba\b) captures the date either as a number or a typed out letter and (\W|to) {,3} is the part scanning in the space or hyphens between two part numbers e.g. 3 **to** 4 months. The (?P<t2>(\b\d+)\D{,2})? Captures the 2nd number if there is one and (years?|yrs?\b) matches the date-type

* DR hits section contain the keywords used to determine if a cells note is related to diabetic retinopathy

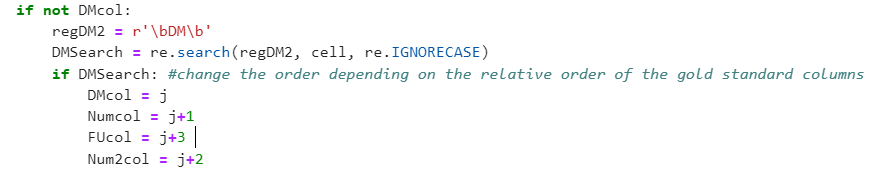
Cell 3: function definitions

* Most are self explanatory; the main one is the findDate function which is passed in the search string and iterates through the 4 date-type regex strings, capturing all hits. It then compares to see which of the hits has the smallest number e.g. 3 days is saved over 7 days. Finally, if there were no hits, the text is searched for any miscellaneous hits, such as “annual” or “monthly”

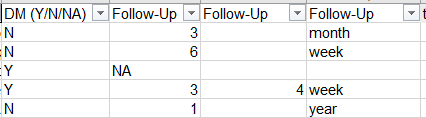


Cell 4: The setup cell

* The input file path is set, all the loop variables are reset and the excel input is converted to a dataframe and each of the columns of the first row are looped through to determine which column contains which information. We also determine which columns to output our follow-up information to.

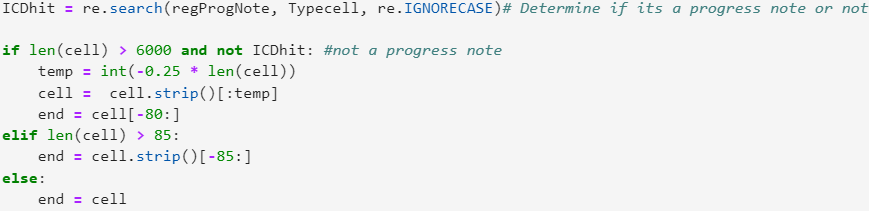


In this example, the gold standard of the DM hit column is searched for and the column number, j, is saved. The number 1, number 2 (if applicable), and the follow-up date type columns are manually filled in depending on the file format:

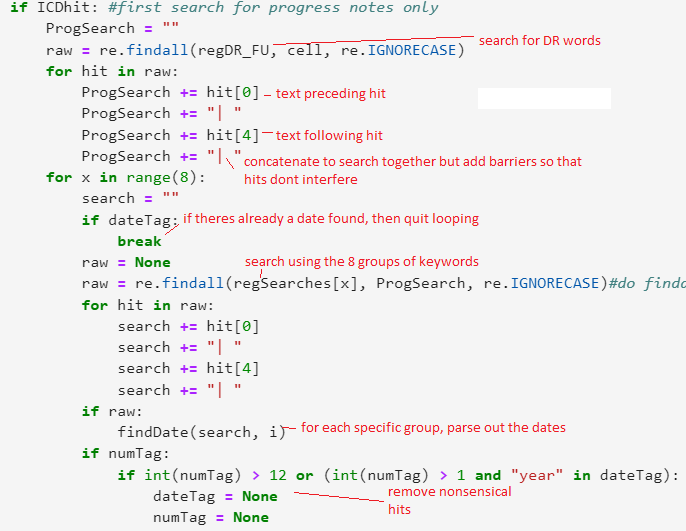


Cell 5: The data

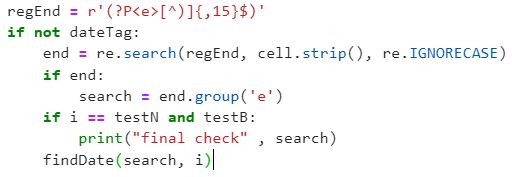
* Starts off resetting all loop variables and then loops through the data frame row by row, saving cells of interest as strings. E.g. the cell containing the note text. In the note type cell, search if the note is a progress note or not. For especially long problem list notes, strip off anything over the 6000th character since they mostly contain background information not necessarily relevant to DR. Next save the last 85 characters of each note so that the end can be searched first.



1. The first search is only for progress notes, the longer notes that often have background information and it searches first for DR related terms, saving the text a couple characters before and a couple after each hit. These texts are then searched for follow-up keywords in the priority listed above and the text before and after those hits are saved and run in the find\_date function described earlier.



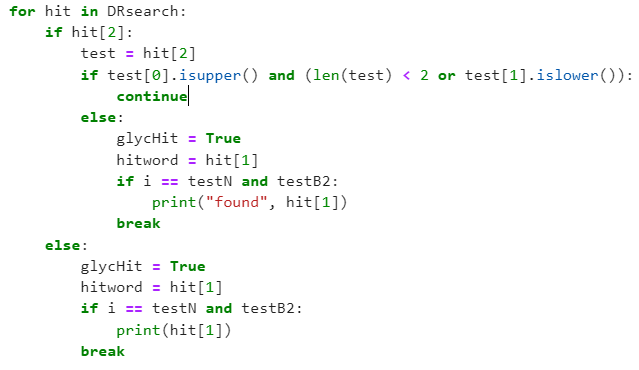
1. If there is no hit in the first search, then the end bit of the note (last 85 characters) is searched using the priorities and the same process of saving the surrounding text and running the date function on it is used.
2. If there is no hit in the first two searches, then the entire note is searched using the priorities and the same process of saving the surrounding text and running the date function on it is used.
3. If none of those searches yielded a hit, then the final search simply takes the last 15 characters and regardless if there is a follow-up keyword, searches them for a number and a date-type using the date function.



Searches the last 15 characters for a date regardless of keywords

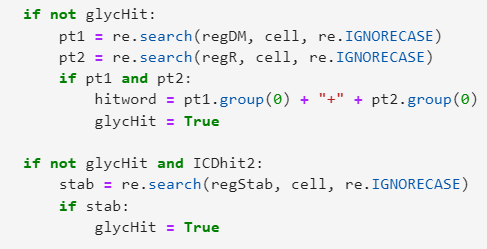
At each stage, unreasonable results are filtered out, such as followups that are over 2 years and any number greater than 12 (anything above 12 months is rarely used)

* Next the ICD tag is searched for problem lists to determine if the note is already tagged to be DR related. Using the same DR keywords as before, the note is searched to determine if it is a note relevant to DR. The hit “DR” is ignored if it is in reference to a doctor’s name e.g. Dr James and this is done by heuristic (if the following word starts with an upper case letter).



Keep looping on hits until it is not a doctor’s name hit and then break since it’s a confirmed DR hit

* + If there are no hits, then two categories of less specific hits (Diabetes Mellitus and retinopathy )are individually searched and if both are found, then the note is counted as a DR hit
  + Finally if there are still no hits and the note is ICD tagged to be DR related, then even more general search terms, such as “stable” are used to confirm if the note is a hit



The 2 special cases if the first yields no hit

* Finally the results are compared to the golden standard and the false/true positives/negatives are calculated and printed out

Cell 6:

* The results are added as columns to the dataframe and it is exported to another excel file