Freetext Matching Algorithm Version 15: Guide to Visual Basic code

Anoop Shah, Clinical Epidemiology Group

May 15, 2013

Contents

1 Module MainModule

Module: MainModule – to invoke the program from a command line

1.1 Sub Main

Loads arguments from configuration file and runs the analysis. The command-line argument (Command) is the location of the configuration file.

Arguments: none

Subs and functions called: `fma_gold.do_analysis` subsection 2.2 on page 5
`MainModule.getParameterFromFile` subsection 1.2 on page 4

Called by: none

1.2 Function getParameterFromFile As String

Gets parameter from file, where each line of the file has the format: parameter value (separated by at least one space)

Arguments: `parameterName` – String
`filename` – String

Subs and functions called: none

Called by: `MainModule.Main` subsection 1.1 on page 4

2 Module fma_gold

Module: fma_gold – functions for analysis of free text and output in a format similar to the Clinical Practice Research Datalink 'GOLD' format.

2.1 Global variables and constants

Const `maxtexts` = 200001
Const `delim` = "," (*delimiter*)
`newline` – String (*newline character, will be defined in main_fma_gold*)
Const `maxrows` = 1000
`outdata(maxrows)` – String (*output data staging area*)
`outrows` – Long (*number of rows in output for a single text*)
`pracid(maxtexts)` – Long (*ordered practice identifier*)
`textid(maxtexts)` – Long (*ordered text ID (unique within practice)*)

medcode(maxtexts) – Long (*medcode (may be multiple medcodes for each pracid / textid combination)*)
ntexts – Long (*actual number of texts*)

2.2 Sub do_analysis

FMA gold analysis of free text. medcodefile is the file with medcodes to be appended to the free text to provide the analysis modes. This file is optional. If not provided, medcode is assumed to be zero for all files. If freetext is supplied as an argument to the function, it is analysed (together with origmedcode) and the text and debug output are written in the log file.

Arguments: logfile – String
lookups – String
infile – String (Optional)
medcodefile – String (Optional)
outfile – String (Optional)
freetext – String (Optional)
medcode – String (Optional)
origmedcode – Long (Optional)

Subs and functions called: freetext_core.import_all_lookups subsection 3.4 on page 10
fma_gold.loadMedcodes subsection 2.14 on page 8
fma_gold.gettextid subsection 2.3 on page 5
fma_gold.getpracid subsection 2.4 on page 5
strfunc.dissect subsection 7.9 on page 25
fma_gold.getmedcodes subsection 2.13 on page 7
strfunc.numwords subsection 7.6 on page 24
freetext_core.main_analyse subsection 3.3 on page 9
freetext_core.main_termref subsection 3.2 on page 9
fma_gold.pd_to_fma_gold subsection 2.12 on page 7
terms.std_term subsection 10.14 on page 33

Called by: MainModule.Main subsection 1.1 on page 4

2.3 Function gettextid As Long

Finds textid in a string, at the second position, tab separated In a separate function for error trapping purposes

Arguments: str – String

Subs and functions called: strfunc.dissect subsection 7.9 on page 25

Called by: fma_gold.do_analysis subsection 2.2 on page 5

2.4 Function getpracid As Long

Finds pracid in a string, at the first position, tab separated In a separate function for error trapping purposes

Arguments: str – String

Subs and functions called: strfunc.dissect subsection 7.9 on page 25

Called by: fma_gold.do_analysis subsection 2.2 on page 5

2.5 Function pdYYYYMMDD As Double

Converts a date to YYYYMMDD format

Arguments: `str` – String

Subs and functions called: `strfunc.dissect` subsection 7.9 on page 25

Called by: `fma_gold.pd_to_fma_gold` subsection 2.12 on page 7

2.6 Function pdValue As Double

Returns the value (e.g. medcode, LABS value, duration number)

Arguments: `str` – String

Subs and functions called: `strfunc.dissect` subsection 7.9 on page 25
`strfunc.is_numeric` subsection 7.10 on page 26

Called by: `fma_gold.pd_to_fma_gold` subsection 2.12 on page 7

2.7 Function pdAge As Double

Returns the age in years

Arguments: `str` – String

Subs and functions called: `strfunc.dissect` subsection 7.9 on page 25

Called by: `fma_gold.pd_to_fma_gold` subsection 2.12 on page 7

2.8 Function pdDurUnits As Double

Returns the SUM lookup value for the duration units

Arguments: `str` – String

Subs and functions called: `strfunc.dissect` subsection 7.9 on page 25

Called by: `fma_gold.pd_to_fma_gold` subsection 2.12 on page 7

2.9 Function pdDurValue As Double

Returns the SUM lookup value for the duration units Error trapping in case of type conversion error

Arguments: `str` – String

Subs and functions called: `strfunc.dissect` subsection 7.9 on page 25

Called by: `fma_gold.pd_to_fma_gold` subsection 2.12 on page 7

2.10 Sub addOutputRow

Adds data to the output rows. All arguments are required to be double. Zero values are ignored and considered as missing.

Arguments: medcode_ – Double
enttype_ – Double
data1 – Double (Optional)
data2 – Double (Optional)
data3 – Double (Optional)
data4 – Double (Optional)

Subs and functions called: fma_gold.blankIfZero subsection 2.11 on page 7

Called by: fma_gold.pd_to_fma_gold subsection 2.12 on page 7

2.11 Function blankIfZero As String

Converts a number to a string, returning an empty string if the number is zero.

Arguments: number – Double

Subs and functions called: none

Called by: fma_gold.addOutputRow subsection 2.10 on page 6

2.12 Sub pd_to_fma_gold

Extracts information from pd and converts it to FMA gold format. Stores the extracted information in the outdata array.

Arguments: origmedcode – Long (Optional)

Subs and functions called: pd.max subsection 4.26 on page 19
pd.Attr subsection 4.8 on page 14
fma_gold.addOutputRow subsection 2.10 on page 6
fma_gold.pdValue subsection 2.6 on page 6
pd.mean subsection 4.9 on page 14
fma_gold.pdYYYYMMDD subsection 2.5 on page 6
strfunc.in_set subsection 7.4 on page 23
strfunc.dissect subsection 7.9 on page 25
fma_gold.pdDurValue subsection 2.9 on page 6
fma_gold.pdDurUnits subsection 2.8 on page 6
fma_gold.pdAge subsection 2.7 on page 6
strfunc.is_numeric subsection 7.10 on page 26

Called by: fma_gold.do_analysis subsection 2.2 on page 5

2.13 Function getmedcodes As String

Returns the medcode mapping the given pracid and textid, or 0 if it is not found. The sorting is by pracid then textid. This function uses a binary search, comparing both pracid and textid with the target.

Arguments: targetpracid – Long
targettextid – Long

Subs and functions called: none

Called by: fma_gold.do_analysis subsection 2.2 on page 5

2.14 Function loadMedcodes As String

Loads text id and medcodes from a comma separated text file with optional header: pracid, textid, medcode. If no header, it is assumed that the columns are in this order, otherwise the column names are used and additional columns are allowed. Returns a message stating whether the load was sucessful.

Arguments: filename – String

Subs and functions called: fma_gold.isHeader subsection 2.15 on page 8
fma_gold.findColumn subsection 2.16 on page 8
strfunc.dissect subsection 7.9 on page 25

Called by: fma_gold.do_analysis subsection 2.2 on page 5

2.15 Function isHeader As Boolean

Whether str is a possible header in a comma separated file If any of the columns are non-numeric, isHeader is True

Arguments: str – String

Subs and functions called: strfunc.dissect subsection 7.9 on page 25
strfunc.is_numeric subsection 7.10 on page 26

Called by: fma_gold.loadMedcodes subsection 2.14 on page 8

2.16 Function findColumn As Long

Finds out the column number (first position) of colName in allNames, with comma delimiter Returns 0 if column name not found

Arguments: colName – String
allNames – String

Subs and functions called: strfunc.dissect subsection 7.9 on page 25

Called by: fma_gold.loadMedcodes subsection 2.14 on page 8

3 Module freetext_core

Module: freetext_core – core algorithm

3.1 Global variables and constants

Const wordmatchthreshold = 0.73 (*used by readscore*)
debug_string – String (*stores analysis report for an individual text, when running in debug mode*)
death – Boolean (*whether Read term implies death*)
gest – Boolean (*whether Read term refers to weeks gestation*)
spell – Boolean (*whether to use spelling correction*)

3.2 Sub main_termref

Calls main_analyse with appropiate analysis option based on the Read term associated with the text, and depending on the append_term argument it may also append the text to the end of the Read term to appear as it would on the GP's computer.

Arguments: `instring` – String
`Termref` – Long
`spell_` – Boolean (Optional)
`debug_` – Boolean (Optional)
`append_term` – Boolean (Optional) (ByVal)

Subs and functions called: `terms.read_type` subsection 10.13 on page 33
`terms.std_term` subsection 10.14 on page 33
`strfunc.in_set` subsection 7.4 on page 23
`freetext_core.main_analyse` subsection 3.3 on page 9
`pd.mean` subsection 4.9 on page 14
`strfunc.dissect2` subsection 7.9 on page 25
`pd.Attr` subsection 4.8 on page 14
`pd.remove` subsection 4.22 on page 18

Called by: `fma_gold.do_analysis` subsection 2.2 on page 5

3.3 Sub main_analyse

This is the main part of the Freetext Matching Algorithm which calls functions to perform each of the major steps in the analysis of an input text (instring).

Arguments: `instring` – String (ByVal)
`death_` – Boolean (Optional)
`pregnant_` – Boolean (Optional)
`debug_` – Boolean (Optional)
`labtest` – String (Optional)
`spell_` – Boolean (Optional)
`date_only` – Boolean (Optional)
`termstring` – String (Optional)
`append_term` – Boolean (Optional)
`sicknote` – Boolean (Optional)

Subs and functions called: `freetext_core.readscore` subsection 3.9 on page 11
`wordlist.remove_ignore_phrases` subsection 11.11 on page 36
`pd.init_read` subsection 4.19 on page 17
`freetext_core.initial_search` subsection 3.5 on page 10
`attrib.pd_search2` subsection 5.4 on page 20
`pd.show_all_2` subsection 4.6 on page 14
`freetext_core.attrib_search` subsection 3.6 on page 10
`freetext_core.analyse_pd` subsection 3.7 on page 11
`pd.compress` subsection 4.4 on page 13
`checkterms.check_all` subsection 12.3 on page 37
`pd.check_compressed` subsection 4.2 on page 12

Called by: `fma_gold.do_analysis` subsection 2.2 on page 5
`freetext_core.main_termref` subsection 3.2 on page 9

3.4 Function import_all_lookups As String

Imports all lookup tables from text files by calling the appropriate import functions in the modules attrib, checkterms, synonym, terms and wordlist. The text files must have standard names as in the master repository (https://github.com/anoopshah/freetext-matching-algorithm-lookups). Returns a string stating what was imported.

Arguments: `lookupfolder` – String

Subs and functions called: `attrib.import` subsection 5.2 on page 19
`checkterms.import` subsection 12.2 on page 37
`synonym.import` subsection 9.3 on page 28
`wordlist.import_ignore` subsection 11.6 on page 35
`terms.import` subsection 10.5 on page 31
`wordlist.import_wordlist` subsection 11.3 on page 34

Called by: `fma_gold.do_analysis` subsection 2.2 on page 5

3.5 Sub initial_search

Identifies synonyms, words which might be part of a Read term, numbers and dates in the free text, recording the results in the 'meaning' array in the pd module.

Arguments: `debug_` – Boolean (Optional)

Subs and functions called: `pd.max` subsection 4.26 on page 19
`pd.part_nopunc` subsection 4.14 on page 16
`strfunc.get_date` subsection 7.1 on page 22
`pd.part_punc_nospace` subsection 4.15 on page 16
`pd.add_mean` subsection 4.13 on page 15
`synonym.get_search_summary` subsection 9.6 on page 28
`pd.text` subsection 4.23 on page 18
`wordlist.ignorable` subsection 11.10 on page 36
`pd.add_attr` subsection 4.12 on page 15
`strfunc.is_numeric` subsection 7.10 on page 26
`wordlist.wordsearch` subsection 11.9 on page 36
`pd.set_text` subsection 4.24 on page 18

Called by: `freetext_core.main_analyse` subsection 3.3 on page 9

3.6 Sub attrib_search

Extends context attributes found on pattern matching (attrib.pd_search2) to nearby words based on hard-coded patterns.

Arguments: `debug_` – Boolean (Optional)

Subs and functions called: `pd.max` subsection 4.26 on page 19
`pd.Attr` subsection 4.8 on page 14
`strfunc.in_set` subsection 7.4 on page 23
`pd.mean` subsection 4.9 on page 14
`pd.text` subsection 4.23 on page 18
`pd.punct` subsection 4.25 on page 19
`pd.set_attr` subsection 4.10 on page 15

Called by: `freetext_core.main_analyse` subsection 3.3 on page 9

3.7 Sub analyse_pd

Attempts to map sequences of words to Read terms.

Arguments: `debug_` – Boolean (Optional)
`labtest` – String (Optional)

Subs and functions called: `strfunc.in_set` subsection 7.4 on page 23
`pd.mean` subsection 4.9 on page 14
`pd.Attr` subsection 4.8 on page 14
`pd.max` subsection 4.26 on page 19
`pd.punct` subsection 4.25 on page 19
`pd.text` subsection 4.23 on page 18
`list.bestmatch` subsection 6.3 on page 21
`pd.set_mean` subsection 4.11 on page 15
`strfunc.words` subsection 7.3 on page 23
`pd.set_attr` subsection 4.10 on page 15

Called by: `freetext_core.main_analyse` subsection 3.3 on page 9

3.8 Function remove_ignorable As String

Removes ignorable words from a phrase. The argument instring must have one space between words and no punctuation.

Arguments: `instring` – String (ByVal)
`remove_right_left` – Boolean (Optional)

Subs and functions called: `strfunc.numwords` subsection 7.6 on page 24
`strfunc.dissect2` subsection 7.9 on page 25
`wordlist.ignorable` subsection 11.10 on page 36
`strfunc.in_set` subsection 7.4 on page 23

Called by: `list.getlist` subsection 6.6 on page 22
`terms.init_and_sort` subsection 10.6 on page 31

3.9 Function readscore As Single

Returns a score (0 to 100) based on the accuracy and completeness of match between a sequence of words in the free text and a candidate Read term.

Arguments: `pd_start` – Long
`pd_fin` – Long
`Termref` – Long
`debug_` – Boolean (Optional)
`clear_memory` – Boolean (Optional)

Subs and functions called: `terms.std_term` subsection 10.14 on page 33
`strfunc.numwords` subsection 7.6 on page 24
`pd.part_nopunc` subsection 4.14 on page 16
`terms.attrib_str` subsection 10.15 on page 33
`pd.Attr` subsection 4.8 on page 14

strings_Acc97.replace subsection 8.1 on page 27
strfunc.dissect2 subsection 7.9 on page 25
synonym.trylink_2 subsection 9.7 on page 29
strfunc.words subsection 7.3 on page 23
pd.text subsection 4.23 on page 18
pd.true_ subsection 4.7 on page 14
strfunc.in_set subsection 7.4 on page 23
wordlist.ignorable subsection 11.10 on page 36

Called by: freetext_core.main_analyse subsection 3.3 on page 9
list.getlist subsection 6.6 on page 22

3.10 Function fuzzylink As Long

Whether the two words are almost the same (maximum one character difference). Assume the first character is the same and they differ in length by at most 1. Gives a score (letter position of difference, zero if too different).

Arguments: ref_word – String
test_word – String

Subs and functions called: none

Called by: wordlist.wordsearch subsection 11.9 on page 36

4 Module pd

Module: pd – arrays for holding individual words of the text being analysed (limit of 1000 words), and functions for pattern matching

4.1 Global variables and constants

Const maxpartdata = 1000
partdata_used – Long (*number of words in the input text*)
partdata(maxpartdata) – String (*array containing individual words in the input free text*)
punc(maxpartdata) – String (*punctuation*)
attrib(maxpartdata) – String (*attribute e.g. negative, family etc.*)
meaning(maxpartdata) – String (*interpreted meaning e.g. Read code or date*)

4.2 Sub check_compressed

Checks that attributes and values are consistent. This function must be run after sub compress. It also converts gestational ages into a 'LABS' output data type, checks that there is only one gestational age and checks that systolic blood pressure is greater than diastolic. It also checks that dateprev, datenext etc. refer to clinical events.

Arguments: maybe_pregnant – Boolean (Optional)
labtest – String (Optional)

Subs and functions called: strfunc.words subsection 7.3 on page 23
pd.Attr subsection 4.8 on page 14
pd.remove subsection 4.22 on page 18

pd.set_attr subsection 4.10 on page 15
terms.true_term subsection 10.11 on page 32
strfunc.dissect2 subsection 7.9 on page 25
pd.set_mean subsection 4.11 on page 15
strfunc.in_set subsection 7.4 on page 23
pd.mean subsection 4.9 on page 14
pd.remove_from_compressed subsection 4.3 on page 13
terms.linkto subsection 10.16 on page 33

Called by: freetext_core.main_analyse subsection 3.3 on page 9

4.3 Sub remove_from_compressed

Removes all entries with a certain attribute from the pd arrays if there is a risk it might be wrong.

Arguments: attr_to_remove – String (Optional) (ByVal)
type_to_remove – String (Optional) (ByVal)

Subs and functions called: pd.remove subsection 4.22 on page 18
strfunc.dissect2 subsection 7.9 on page 25
pd.mean subsection 4.9 on page 14

Called by: pd.check_compressed subsection 4.2 on page 12

4.4 Sub compress

Converts the pd arrays from a list of words from the original text (i.e. one entry per text) to a list of interpreted results (i.e. one entry per output value). The original text and punctuation are removed. This is used as an intermediate stage in the construction of the final output.

Arguments: none

Subs and functions called: pd.Attr subsection 4.8 on page 14
strfunc.in_set subsection 7.4 on page 23
pd.mean subsection 4.9 on page 14
pd.set_mean subsection 4.11 on page 15
pd.correct_attr subsection 4.5 on page 13
pd.set_attr subsection 4.10 on page 15
pd.remove subsection 4.22 on page 18

Called by: freetext_core.main_analyse subsection 3.3 on page 9

4.5 Function correct_attr As Boolean

Returns True if the attribute is appropriate for the extracted data type

Arguments: pos – Long

Subs and functions called: strfunc.dissect2 subsection 7.9 on page 25
pd.mean subsection 4.9 on page 14
strfunc.in_set subsection 7.4 on page 23
pd.Attr subsection 4.8 on page 14

Called by: pd.compress subsection 4.4 on page 13

4.6 Sub show_all_2

Adds the whole of the pd arrays to the debug string, for use when analysing a single text in debug mode.

Arguments: none

Subs and functions called: none

Called by: `freetext_core.main_analyse` subsection 3.3 on page 9

4.7 Function true_ As Boolean

Returns True if the attribute at this position is not 'negative'.

Arguments: `pos` – Long

Subs and functions called: `pd.Attr` subsection 4.8 on page 14

Called by: `freetext_core.readscore` subsection 3.9 on page 11

4.8 Function Attr As String

Returns the attribute at this position.

Arguments: `pos` – Long

Subs and functions called: none

Called by: `fma_gold.pd_to_fma_gold` subsection 2.12 on page 7
`freetext_core.main_termref` subsection 3.2 on page 9
`freetext_core.attrib_search` subsection 3.6 on page 10
`freetext_core.analyse_pd` subsection 3.7 on page 11
`freetext_core.readscore` subsection 3.9 on page 11
`pd.check_compressed` subsection 4.2 on page 12
`pd.compress` subsection 4.4 on page 13
`pd.correct_attr` subsection 4.5 on page 13
`pd.true_` subsection 4.7 on page 14
`synonym.trylink_2` subsection 9.7 on page 29
`checkterms.check_all` subsection 12.3 on page 37

4.9 Function mean As String

Returns the interpreted meaning at this position.

Arguments: `pos` – Long

Subs and functions called: none

Called by: `fma_gold.pd_to_fma_gold` subsection 2.12 on page 7
`freetext_core.main_termref` subsection 3.2 on page 9
`freetext_core.attrib_search` subsection 3.6 on page 10
`freetext_core.analyse_pd` subsection 3.7 on page 11
`pd.check_compressed` subsection 4.2 on page 12
`pd.remove_from_compressed` subsection 4.3 on page 13
`pd.compress` subsection 4.4 on page 13

pd.correct_attr subsection 4.5 on page 13
checkterms.check_all subsection 12.3 on page 37

4.10 Sub set_attr

Sets the attribute at this position to a specific value.

Arguments: new_attribute – String
pos – Long

Subs and functions called: none

Called by: freetext_core.attrib_search subsection 3.6 on page 10
freetext_core.analyse_pd subsection 3.7 on page 11
pd.check_compressed subsection 4.2 on page 12
pd.compress subsection 4.4 on page 13
attrib.pd_search2 subsection 5.4 on page 20
checkterms.check_all subsection 12.3 on page 37

4.11 Sub set_mean

Sets the interpreted meaning at this position to a specific value.

Arguments: new_meaning – String
pos – Long

Subs and functions called: none

Called by: freetext_core.analyse_pd subsection 3.7 on page 11
pd.check_compressed subsection 4.2 on page 12
pd.compress subsection 4.4 on page 13
attrib.pd_search2 subsection 5.4 on page 20

4.12 Sub add_attr

Sets the attribute for a range of positions to a specific value.

Arguments: new_attribute – String
pos_start – Long
pos_fin – Long (Optional)
ignore_if_already – Boolean (Optional)

Subs and functions called: none

Called by: freetext_core.initial_search subsection 3.5 on page 10

4.13 Sub add_mean

Sets the interpreted meaning for a range of positions to a specific value.

Arguments: new_meaning – String
pos_start – Long
pos_fin – Long (Optional)
ignore_if_already – Boolean (Optional)

Subs and functions called: none

Called by: freetext_core.initial_search subsection 3.5 on page 10

4.14 Function part_nopunc As String

Returns a string containing a defined set of words from the text with no punctuation.

Arguments: start – Long (Optional)
fin – Long (Optional) (ByVal)

Subs and functions called: pd.max subsection 4.26 on page 19

Called by: freetext_core.initial_search subsection 3.5 on page 10
freetext_core.readscore subsection 3.9 on page 11
list.bestmatch subsection 6.3 on page 21
synonym.trylink_2 subsection 9.7 on page 29

4.15 Function part_punc_nospace As String

Returns a string containing a defined set of words and punctuation but without spaces either side of punctuation.

Arguments: start – Long
fin – Long

Subs and functions called: pd.max subsection 4.26 on page 19

Called by: freetext_core.initial_search subsection 3.5 on page 10
attrib.pd_search2 subsection 5.4 on page 20

4.16 Function matchpattern As Boolean

Returns True if the set of up to 5 words or meanings (w1-w5) with punctuation (p1-p5) match a set of entries in partdata

Arguments: partdata_pos – Long
w1 – String
p1 – String
w2 – String
p2 – String
w3 – String
p3 – String
w4 – String
p4 – String
w5 – String
p5 – String

Subs and functions called: pd.matchposition subsection 4.17 on page 17

Called by: attrib.pd_search2 subsection 5.4 on page 20

4.17 Function matchposition As Boolean

Returns True if there is a match between the search word and text. The argument 'word' can represent either text or meaning (if enclosed in []).

Arguments: partdata_pos – Long
word – String (ByVal)
punct – String (ByVal)

Subs and functions called: strfunc.dissect2 subsection 7.9 on page 25
pd.matchoption subsection 4.18 on page 17

Called by: pd.matchpattern subsection 4.16 on page 16

4.18 Function matchoption As Boolean

Match the free text and single position match meaning / words

Arguments: partdata_pos – Long
word – String (ByVal)
punct – String (ByVal)

Subs and functions called: strfunc.dissect subsection 7.9 on page 25
strfunc.words subsection 7.3 on page 23
pd.text subsection 4.23 on page 18

Called by: pd.matchposition subsection 4.17 on page 17

4.19 Sub init_read

Initialises the 'partdata' and 'punc' arrarys in the pd module with words and punctuation from the free text, by parsing the raw free text string. Also converts symbols '+' and '' to the word 'and', and '' (used for CPRD anonymised words) to the word 'anonymised', to avoid it being recognised as part of a Read term.

Arguments: instring – String

Subs and functions called: pd.clear subsection 4.21 on page 18
pd.st_type subsection 4.20 on page 17
strfunc.is_numeric subsection 7.10 on page 26

Called by: freetext_core.main_analyse subsection 3.3 on page 9

4.20 Function st_type As Long

Returns the type of a text string; 0 if it is a single space, 1 if it is part of a word, 2 if it is a number, and 3 if it does not fit into any of the other categories (i.e. if it is punctuation).

Arguments: instring – String

Subs and functions called: strfunc.is_text subsection 7.5 on page 24
strfunc.is_numeric subsection 7.10 on page 26

Called by: pd.init_read subsection 4.19 on page 17

4.21 Sub clear

Clears the 'partdata', 'punc', 'attrib' and 'meaning' arrays in the pd module.

Arguments: none

Subs and functions called: none

Called by: `pd.init_read` subsection 4.19 on page 17

4.22 Sub remove

Removes data from the arrays in the pd module between the specified positions

Arguments: `pos1` – Long
`pos2` – Long (Optional)

Subs and functions called: none

Called by: `freetext_core.main_termref` subsection 3.2 on page 9
`pd.check_compressed` subsection 4.2 on page 12
`pd.remove_from_compressed` subsection 4.3 on page 13
`pd.compress` subsection 4.4 on page 13
`checkterms.check_all` subsection 12.3 on page 37

4.23 Function text As String

Returns the word at a particular position (from the 'partdata' array).

Arguments: `position` – Long

Subs and functions called: none

Called by: `freetext_core.initial_search` subsection 3.5 on page 10
`freetext_core.attrib_search` subsection 3.6 on page 10
`freetext_core.analyse_pd` subsection 3.7 on page 11
`freetext_core.readscore` subsection 3.9 on page 11
`pd.matchoption` subsection 4.18 on page 17
`attrib.pd_search2` subsection 5.4 on page 20
`list.bestmatch` subsection 6.3 on page 21

4.24 Sub set_text

Replaces the word at a particular position (in the 'partdata' array).

Arguments: `new_text` – String
`position` – Long

Subs and functions called: none

Called by: `freetext_core.initial_search` subsection 3.5 on page 10

4.25 Function punct As String

Returns the punctuation at a particular position (from the 'punc' array).

Arguments: `position` – Long

Subs and functions called: none

Called by: `freetext_core.attrib_search` subsection 3.6 on page 10
`freetext_core.analyse_pd` subsection 3.7 on page 11

4.26 Function max As Long

Returns the total number of words in the input text

Arguments: none

Subs and functions called: none

Called by: `fma_gold.pd_to_fma_gold` subsection 2.12 on page 7
`freetext_core.initial_search` subsection 3.5 on page 10
`freetext_core.attrib_search` subsection 3.6 on page 10
`freetext_core.analyse_pd` subsection 3.7 on page 11
`pd.part_nopunc` subsection 4.14 on page 16
`pd.part_punc_nospace` subsection 4.15 on page 16
`attrib.pd_search2` subsection 5.4 on page 20
`checkterms.check_all` subsection 12.3 on page 37

5 Module attrib

Module: attrib – code related to the attributes table. The table is loaded from a text file by the 'import' function

5.1 Global variables and constants

Const `maxattrib` = 400
`w(5, maxattrib)` – String (*pattern of up to 5 words*)
`p(5, maxattrib)` – String (*options for punctuation associated with each word*)
`a(5, maxattrib)` – String (*attribute associated with each word*)
`death_only(maxattrib)` – Boolean (*whether this attribute pattern is only applicable in 'death' mode*)
`numwd(maxattrib)` – Long (*number of words (1 to 5) in this pattern*)
`order(maxattrib)` – Double (*order of this row in the lookup table; not used in the actual algorithm but loaded for debug purposes.*)
`num` – Integer

5.2 Function import As String

Imports attributes lookup table and returns a string stating what was imported. The table must be already be sorted in order; this is checked but not corrected.

Arguments: `filename` – String

Subs and functions called: strfunc.dissect subsection 7.9 on page 25
synonym.import subsection 9.3 on page 28
attrib.dissect2_options subsection 5.3 on page 20

Called by: freetext_core.import_all_lookups subsection 3.4 on page 10
synonym.import subsection 9.3 on page 28
terms.import subsection 10.5 on page 31
checkterms.import subsection 12.2 on page 37

5.3 Function dissect2_options As String

Counts the number of options in a string and puts it at the front of the string, for future use by the dissect2 function. e.g. 'word—another—option' is converted to '3—word—another—option'.

Arguments: instring – String

Subs and functions called: none

Called by: attrib.import subsection 5.2 on page 19

5.4 Sub pd_search2

Tries each attribute pattern in turn to see whether it applies to the free text being analysed. Results are added to attribute fields in the arrays in module pd.

Arguments: debug_ – Boolean (Optional)
death – Boolean (Optional)

Subs and functions called: pd.max subsection 4.26 on page 19
pd.matchpattern subsection 4.16 on page 16
pd.set_attr subsection 4.10 on page 15
pd.set_mean subsection 4.11 on page 15
strfunc.dissect2 subsection 7.9 on page 25
pd.text subsection 4.23 on page 18
pd.part_punc_nospace subsection 4.15 on page 16

Called by: freetext_core.main_analyse subsection 3.3 on page 9

6 Module list

Freetext Matching Algorithm: natural language analysis system for clinical text Copyright: Anoop Dinesh Shah, 2012, 2013 Email: anoop@doctors.org.uk

6.1 User-defined data types

termlist
Data elements:
Termref(maxtermlist) – Long (*medcode that the term maps to*)
words(maxtermlist) – String (*phrase variant which maps to this medcode*)
score(maxtermlist) – Single (*readscore*)
num – Long (*number of terms in termlist*)

6.2 Global variables and constants

Const maxtermlist = 50 (*maximum number of terms to consider*)
Const threshold_high = 91 ((*for readscore - don't analyse further*))
Const threshold = 87 ((*for readscore - minimum*))

6.3 Function bestmatch As String

Returns a string containing the medcode and readscore for the best possible Read term match for a portion of the text. The match may be to an 'alternate' Read term, which is converted to the linked preferred term in the final output.

Arguments: pd_start – Long
pd_fin – Long
debug_ – Boolean (Optional)

Subs and functions called: terms.exact_read_termref subsection 10.12 on page 33
pd.part_nopunc subsection 4.14 on page 16
pd.text subsection 4.23 on page 18
list.getlist subsection 6.6 on page 22
list.display subsection 6.7 on page 22
list.expand subsection 6.4 on page 21

Called by: freetext_core.analyse_pd subsection 3.7 on page 11

6.4 Function expand As termlist

Returns a termlist which expands the input termlist by generating variants of the text fragment using the synonym table. The function searches for up to 5 words at a time, starting with longer possible matches.

Arguments: in_list – termlist
pd_start – Long (Optional)
pd_fin – Long (Optional)
leeway – Long (Optional)

Subs and functions called: checkterms.in_list subsection 12.4 on page 37
strfunc.numwords subsection 7.6 on page 24
strfunc.words subsection 7.3 on page 23
synonym.s1_pos subsection 9.9 on page 29
synonym.s1_priority subsection 9.12 on page 30
synonym.s2 subsection 9.11 on page 30
list.getlist subsection 6.6 on page 22
list.add_termlists subsection 6.5 on page 21
synonym.s1 subsection 9.11 on page 30

Called by: list.bestmatch subsection 6.3 on page 21

6.5 Function add_termlists As termlist

Appends one termlist to another, and returns the combined termlist.

Arguments: t1 – termlist
t2 – termlist

Subs and functions called: none

Called by: `list.expand` subsection 6.4 on page 21

6.6 Function getlist As termlist

Creates a list of potential Read term matches to a text phrase, returning the result as a termlist object. Calculates the readscore for each match. If no matches are found, the function removes the words 'left' and 'right' from the text and tries again (by recursion). The leeway argument is currently not used, but it may be possible in the future to alter this to allow the function to attempt to match to terms with a different number of non-ignorable words.

Arguments: `words` – String (ByVal)
`pd_start` – Long (Optional)
`pd_fin` – Long (Optional)
`leeway` – Long (Optional)

Subs and functions called: `freetext_core.remove_ignorable` subsection 3.8 on page 11
`strfunc.numwords` subsection 7.6 on page 24
`strfunc.bag_of_words` subsection 7.11 on page 26
`terms.pos_bagofwords` subsection 10.9 on page 32
`freetext_core.readscore` subsection 3.9 on page 11
`terms.termref_bagofwords` subsection 10.10 on page 32

Called by: `list.bestmatch` subsection 6.3 on page 21
`list.expand` subsection 6.4 on page 21

6.7 Sub display

Adds the contents of termlist to the debug string. This is used when running the program in test mode, to produce an analysis report for a single text.

Arguments: `in_list` – termlist

Subs and functions called: `checkterms.in_list` subsection 12.4 on page 37
`terms.std_term` subsection 10.14 on page 33

Called by: `list.bestmatch` subsection 6.3 on page 21

7 Module strfunc

Module: strfunc – functions for manipulating strings

7.1 Function get_date As String

Attempts to identify dates and durations in almost any format, returning a string stating the date or duration in a standardised format. The first 9 characters are the date/duration type, followed by a single space, followed by a number or date. The possible types are: DURA_gest (gestational age), DURA_days (duration in weeks), DURA_wks_ (duration in weeks), DURA_mths (duration in weeks), DURA_yrs_ (duration in years), DATE_time (time, e.g. 12:15), DATE_full (full date, e.g. 9-May-2013), DATE_year (year only)

Arguments: s – String
get_time – Boolean (Optional)

Subs and functions called: strfunc.in_set subsection 7.4 on page 23
strings_Acc97.replace subsection 8.1 on page 27
strings_Acc97.monthname subsection 8.3 on page 27
strfunc.dissect2 subsection 7.9 on page 25
strfunc.get_date_average subsection 7.2 on page 23

Called by: freetext_core.initial_search subsection 3.5 on page 10

7.2 Function get_date_average As String

Provides a replacement for the first number (s1) from phrases such as 2-3 weeks, 5-6 days etc. The average duration is used, rounded up (no fractions in the result).

Arguments: s1 – String
s2 – String

Subs and functions called: none

Called by: strfunc.get_date subsection 7.1 on page 22

7.3 Function words As String

Extracts individual words from a string, assuming one space between words and no spaces at the beginning of the string.

Arguments: phrase – String (ByVal)
start – Long
numwd – Long (Optional)
finish – Long (Optional)

Subs and functions called: strfunc.dissect2 subsection 7.9 on page 25
strfunc.numwords subsection 7.6 on page 24

Called by: freetext_core.analyse_pd subsection 3.7 on page 11
freetext_core.readscore subsection 3.9 on page 11
pd.check_compressed subsection 4.2 on page 12
pd.matchoption subsection 4.18 on page 17
list.expand subsection 6.4 on page 21
strfunc.bag_of_words subsection 7.11 on page 26
synonym.trylink_2 subsection 9.7 on page 29
wordlist.add_to_wordlist subsection 11.2 on page 34

7.4 Function in_set As Boolean

Whether target is one of a, b, c, d, e etc. The function does not consider any entries after the first empty string.

Arguments: Target – String
a – String
b – String
c – String (Optional)
d – String (Optional)

e – String (Optional)
f – String (Optional)
g – String (Optional)
h – String (Optional)
i – String (Optional)
j – String (Optional)
k – String (Optional)
l – String (Optional)
m – String (Optional)
n – String (Optional)
o – String (Optional)

Subs and functions called: none

Called by: fma_gold.pd_to_fma_gold subsection 2.12 on page 7
freetext_core.main_termref subsection 3.2 on page 9
freetext_core.attrib_search subsection 3.6 on page 10
freetext_core.analyse_pd subsection 3.7 on page 11
freetext_core.remove_ignorable subsection 3.8 on page 11
freetext_core.readscore subsection 3.9 on page 11
pd.check_compressed subsection 4.2 on page 12
pd.compress subsection 4.4 on page 13
pd.correct_attr subsection 4.5 on page 13
strfunc.get_date subsection 7.1 on page 22
strfunc.is_numeric subsection 7.10 on page 26
terms.import subsection 10.5 on page 31

7.5 Function is_text As Boolean

Whether a string consists entirely of lower case text.

Arguments: instring – String

Subs and functions called: none

Called by: pd.st_type subsection 4.20 on page 17

7.6 Function numwords As Long

Returns the number of words in a string, assuming exactly one space between adjacent words.

Arguments: instring – String (ByVal)

Subs and functions called: none

Called by: fma_gold.do_analysis subsection 2.2 on page 5
freetext_core.remove_ignorable subsection 3.8 on page 11
freetext_core.readscore subsection 3.9 on page 11
list.expand subsection 6.4 on page 21
list.getlist subsection 6.6 on page 22
strfunc.words subsection 7.3 on page 23
strfunc.bag_of_words subsection 7.11 on page 26
synonym.import subsection 9.3 on page 28
synonym.trylink_2 subsection 9.7 on page 29
wordlist.add_to_wordlist subsection 11.2 on page 34

7.7 Function num_diff_char As Long

Counts the number of characters which are different between str1 and str2. Ignores any differences beyond the length of the shorter string. If there are more than 3 differences, num_diff_char returns '4' and the exact number of differences is not counted.

Arguments: `str1` – String
`str2` – String

Subs and functions called: none

Called by: none

7.8 Function dissect As String

Extracts part of a string between two delimiters. Uses the VBA.split function via 'dissect2'. The functions dissect and dissect2 are identical apart from the order of the arguments.

Arguments: `in_string` – String
`number` – Long
`delimiter` – String (Optional)

Subs and functions called: `strfunc.dissect2` subsection 7.9 on page 25

Called by: `fma_gold.do_analysis` subsection 2.2 on page 5
`fma_gold.gettextid` subsection 2.3 on page 5
`fma_gold.getpracid` subsection 2.4 on page 5
`fma_gold.pdYYYYMMDD` subsection 2.5 on page 6
`fma_gold.pdValue` subsection 2.6 on page 6
`fma_gold.pdAge` subsection 2.7 on page 6
`fma_gold.pdDurUnits` subsection 2.8 on page 6
`fma_gold.pdDurValue` subsection 2.9 on page 6
`fma_gold.pd_to_fma_gold` subsection 2.12 on page 7
`fma_gold.loadMedcodes` subsection 2.14 on page 8
`fma_gold.isHeader` subsection 2.15 on page 8
`fma_gold.findColumn` subsection 2.16 on page 8
`pd.matchoption` subsection 4.18 on page 17
`attrib.import` subsection 5.2 on page 19
`synonym.import` subsection 9.3 on page 28
`terms.import` subsection 10.5 on page 31
`checkterms.import` subsection 12.2 on page 37

7.9 Function dissect2 As String

Extracts part of a string between two delimiters. Uses the VBA.split function via 'dissect2', with a fallback to the dissect3 function in the strings_Acc97 module if this function is not found. The functions dissect and dissect2 are identical apart from the order of the arguments.

Arguments: `in_string` – String
`delimiter` – String (Optional)
`number` – Long (Optional)

Subs and functions called: `strings_Acc97.dissect3` subsection 8.2 on page 27

Called by: freetext_core.main_termref subsection 3.2 on page 9
freetext_core.remove_ignorable subsection 3.8 on page 11
freetext_core.readscore subsection 3.9 on page 11
pd.check_compressed subsection 4.2 on page 12
pd.remove_from_compressed subsection 4.3 on page 13
pd.correct_attr subsection 4.5 on page 13
pd.matchposition subsection 4.17 on page 17
attrib.pd_search2 subsection 5.4 on page 20
strfunc.get_date subsection 7.1 on page 22
strfunc.words subsection 7.3 on page 23
strfunc.dissect subsection 7.9 on page 25
synonym.s1_priority subsection 9.12 on page 30
checkterms.check_all subsection 12.3 on page 37
checkterms.if_qualify subsection 12.5 on page 38
checkterms.if_dequalify subsection 12.6 on page 38

7.10 Function is_numeric As Boolean

Determines whether a string contains only a single number or part of a single number. If lab_results_mode is TRUE, words like 'normal', 'abnormal' etc. are considered to be numbers.

Arguments: instring – String
lab_results_mode – Boolean (Optional)
dont_ignore_large_numbers – Boolean (Optional)

Subs and functions called: strfunc.in_set subsection 7.4 on page 23

Called by: fma_gold.pdValue subsection 2.6 on page 6
fma_gold.pd_to_fma_gold subsection 2.12 on page 7
fma_gold.isHeader subsection 2.15 on page 8
freetext_core.initial_search subsection 3.5 on page 10
pd.init_read subsection 4.19 on page 17
pd.st_type subsection 4.20 on page 17

7.11 Function bag_of_words As String

Creates a bag-of-words representation of a string: all words in alphabetical order, no duplicates, one space between words. This function can only handle up to 10 words; any additional words are ignored.

Arguments: instring – String

Subs and functions called: strfunc.numwords subsection 7.6 on page 24
strfunc.words subsection 7.3 on page 23
wordlist.quicksort subsection 11.5 on page 35

Called by: list.getlist subsection 6.6 on page 22
terms.init_and_sort subsection 10.6 on page 31

8 Module strings_Acc97

Module: strings_Acc97 – functions for manipulating strings that are provided in VBA for Access 2003 but not in Access 97

8.1 Function replace As String

Returns bigstring with every instance of lookstring replaced with replacestring

Arguments: bigstring – String
lookstring – String
replacestring – String

Subs and functions called: none

Called by: freetext_core.readscore subsection 3.9 on page 11
strfunc.get_date subsection 7.1 on page 22
wordlist.remove_ignore_phrases subsection 11.11 on page 36
wordlist.initial_process subsection 11.12 on page 36

8.2 Function dissect3 As String

Equivalent to the VBA.Split() function in Access 2003, so this program can run in Access 97.

Arguments: in_string – String
delimiter – String (Optional)
number – Long (Optional)

Subs and functions called: none

Called by: strfunc.dissect2 subsection 7.9 on page 25

8.3 Function monthname As String

Name of the month (either full name or short name).

Arguments: number – Integer
short – Boolean

Subs and functions called: none

Called by: strfunc.get_date subsection 7.1 on page 22

9 Module synonym

Module: synonym – code for handling synonyms

9.1 Global variables and constants

Const maxsynonym = 20000
s_used – Long
s1_sorted(maxsynonym) – String (*sorted text word/phrase (duplicates are allowed)*)
s1_result(maxsynonym) – String (*priority and numwords, used for get_search_summary*)
s1_s2(maxsynonym) – String (*Read word/phrase*)
s2_sorted(maxsynonym) – String (*sorted Read word/phrase (duplicates are allowed)*)
s2_s2num(maxsynonym) – Long (*number of words in Read word/phrase*)
s2_s1num(maxsynonym) – Long (*number of words in text word/phrase*)
s2_priority(maxsynonym) – Long (*priority of synonym pair*)
s2_s1(maxsynonym) – String (*text word/phrase*)

9.2 Function numrows As Long

Returns the number of synonyms (s_used) for use by external functions.

Arguments: none

Subs and functions called: none

Called by: `wordlist.import_wordlist` subsection 11.3 on page 34

9.3 Function import As String

Imports the synonym table from the text lookup file. Returns a string stating whether the table was imported successfully.

Arguments: `filename` – String

Subs and functions called: `strfunc.dissect` subsection 7.9 on page 25
`attrib.import` subsection 5.2 on page 19
`synonym.heap_s2` subsection 9.5 on page 28
`strfunc.numwords` subsection 7.6 on page 24
`synonym.heap_s1` subsection 9.5 on page 28

Called by: `freetext_core.import_all_lookups` subsection 3.4 on page 10
`attrib.import` subsection 5.2 on page 19

9.4 Sub heap_s2

Heap helper function for sorting the synonym table by Read word.

Arguments: `i` – Long (ByVal)
`iMin` – Long
`iMax` – Long

Subs and functions called: none

Called by: `synonym.import` subsection 9.3 on page 28

9.5 Sub heap_s1

Heap helper function for sorting the synonym table by text word.

Arguments: `i` – Long (ByVal)
`iMin` – Long
`iMax` – Long

Subs and functions called: none

Called by: `synonym.import` subsection 9.3 on page 28

9.6 Function get_search_summary As String

Returns the s1_result for an entry in the s1_sorted column (text word/phrase). Uses a binary search algorithm.

Arguments: `instring` – String

Subs and functions called: none

Called by: `freetext_core.initial_search` subsection 3.5 on page 10

9.7 Function trylink_2 As String

Tries to match a Read term segment to pd (the text being analysed between pd_start and pd_fin). The algorithm starts from the beginning of the Read term segment, trying to match the whole of pd between pd_start and pd_fin, then tries to get the largest possible match. If not possible, it tries smaller segments of the Read term but always starting from the beginning. The output is a string with the following values (space separated): priority position_within_pd_start position_within_pd_fin read_fin. If the Read term segment is identical to the text (pd), the output has priority 6.

Arguments: `read_term_segment` – String (ByVal)
`pd_start` – Long
`pd_fin` – Long
`cur_true` – Boolean

Subs and functions called: `pd.part_nopunc` subsection 4.14 on page 16
`strfunc.numwords` subsection 7.6 on page 24
`strfunc.words` subsection 7.3 on page 23
`pd.Attr` subsection 4.8 on page 14
`synonym.s2_pos` subsection 9.9 on page 29

Called by: `freetext_core.readscore` subsection 3.9 on page 11

9.8 Function s2_pos As Long

Returns the topmost position of s2 (partial Read term) text in the s2 sorted list

Arguments: `s2_text` – String

Subs and functions called: none

Called by: `synonym.trylink_2` subsection 9.7 on page 29

9.9 Function s1_pos As Long

Returns the topmost position of s1 text in the s1 sorted list.

Arguments: `s1_text` – String

Subs and functions called: none

Called by: `list.expand` subsection 6.4 on page 21

9.10 Function s2 As String

Returns the part Read term (s2) at a particular position in the s1 table.

Arguments: `s1_pos` – Long

Subs and functions called: none

Called by: `list.expand` subsection 6.4 on page 21

9.11 Function s1 As String

Returns the part text (s1) at a particular position in the s1 table.

Arguments: `s1_pos` – Long

Subs and functions called: none

Called by: `list.expand` subsection 6.4 on page 21
`wordlist.import_wordlist` subsection 11.3 on page 34

9.12 Function s1_priority As Long

Returns the priority at a particular position in the s1 table.

Arguments: `s1_pos` – Long

Subs and functions called: `strfunc.dissect2` subsection 7.9 on page 25

Called by: `list.expand` subsection 6.4 on page 21

10 Module terms

Module: terms – Read terms as used by the program

10.1 Global variables and constants

Const `max_usedterms` = 100000
Const `max_allterms` = 150000
`a_std_term(max_usedterms)` – String (*array of std_term (sorted) to get termref*)
`a_termref(max_usedterms)` – Long (*termref (medcode) linked to a_std_term*)
`a_terms_used` – Long (*number of entries in a_std_term and a_termref*)
`b_termref(max_allterms)` – Long (*all terms (native, virtual or alternate), sorted by termref*)
`b_std_term(max_allterms)` – String (*standardised Read term*)
`b_attrib_str(max_allterms)` – String (*attribute string*)
`b_type(max_allterms)` – String (*data type of Read term (pregnancy, labtest, death etc.)*)
`b_linkto(max_allterms)` – Long (*the actual medcode in the output*)
`b_terms_used` – Long (*number of records in the 'b' arrays*)
`c_bagofwords(max_usedterms)` – String (*sorted array of 'bag of non-ignorable words'*)
`c_termref(max_usedterms)` – Long (*termref (medcode) for each bag of words*)
Const `headerNative` = "medcode" (*headings for nativeterms lookup file*)
Const `headerVirtual` = "medcode" (*headings for virtualterms lookup file*)
Const `headerAlternate` = "medcode" (*headings for alternateterms lookup file*)

10.2 Function numrows_a_c As Long

Returns a_terms_used for use by external functions. Tables a and c contain only terms used to match to

Arguments: none

Subs and functions called: none

Called by: `wordlist.import_wordlist` subsection 11.3 on page 34

10.3 Function numrows_b As Long

Returns b_terms_used for use by external functions. Table b contains all Read terms, including those that may be associated with text but are not used for matching.

Arguments: none

Subs and functions called: none

Called by: none

10.4 Function get_bagofwords As String

Returns the value of c_bagofwords for a particular position, for use by external functions.

Arguments: `pos` – Long

Subs and functions called: none

Called by: `wordlist.import_wordlist` subsection 11.3 on page 34

10.5 Function import As String

Imports the text files with native Read terms, Virtual Read terms for coding and alternate terms (variants of native or virtual terms which have identical meaning). Not all the native terms may be coded to; only those with include=TRUE. These term files are stored on the GitHub repository. Argument: termsection = native, virtual or alternate. They must be loaded in that order, because the medcodes must be in order.

Arguments: `filename` – String
`termsection` – String

Subs and functions called: `strfunc.in_set` subsection 7.4 on page 23
`attrib.import` subsection 5.2 on page 19
`strfunc.dissect` subsection 7.9 on page 25
`terms.init_and_sort` subsection 10.6 on page 31

Called by: `freetext_core.import_all_lookups` subsection 3.4 on page 10

10.6 Sub init_and_sort

Initialises c using table a, and sorts tables a and c. This must be run after tables a and b have been filled by the import function.

Arguments: none

Subs and functions called: `strfunc.bag_of_words` subsection 7.11 on page 26
`freetext_core.remove_ignorable` subsection 3.8 on page 11
`terms.heap_bagofwords` subsection 10.7 on page 32
`terms.heap_std_term` subsection 10.8 on page 32

Called by: `terms.import` subsection 10.5 on page 31

10.7 Sub heap_bagofwords

Heap helper function for sorting the bag of words vectors (table c).

Arguments: `i` – Long (ByVal)
`iMin` – Long
`iMax` – Long

Subs and functions called: none

Called by: `terms.init_and_sort` subsection 10.6 on page 31

10.8 Sub heap_std_term

Heap helper function for sorting by std_term (table a).

Arguments: `i` – Long (ByVal)
`iMin` – Long
`iMax` – Long

Subs and functions called: none

Called by: `terms.init_and_sort` subsection 10.6 on page 31

10.9 Function pos_bagofwords As Long

Returns the position of first or last termref for which the bag of words (c_bagofwords) matches instring. If there is no match, zero is returned. Specify search_top = True to return the topmost match, or search_top = False for the last match.

Arguments: `search_top` – Boolean
`instring` – String

Subs and functions called: none

Called by: `list.getlist` subsection 6.6 on page 22

10.10 Function termref_bagofwords As Long

Returns the termref from the bag of words table (c_termref) in a particular position.

Arguments: `position` – Long

Subs and functions called: none

Called by: `list.getlist` subsection 6.6 on page 22

10.11 Function true_term As Boolean

Whether a term contains any true parts.

Arguments: `Termref` – Long

Subs and functions called: `terms.attrib_str` subsection 10.15 on page 33

Called by: `pd.check_compressed` subsection 4.2 on page 12

10.12 Function exact_read_termref As Long

Attempts to find an exact match to Read std_term, and returns the medcode (termref) of the match. Binary search of a_std_term.

Arguments: `search_term` – String (ByVal)

Subs and functions called: none

Called by: `list.bestmatch` subsection 6.3 on page 21

10.13 Function read_type As String

Returns the type code of the Read Term (whether pregnancy, death, labtest etc.) by binary search on table b.

Arguments: `Termref` – Long

Subs and functions called: none

Called by: `freetext_core.main_termref` subsection 3.2 on page 9

10.14 Function std_term As String

Returns the standardised term for a termref, by a binary search on table b.

Arguments: `Termref` – Long

Subs and functions called: none

Called by: `fma_gold.do_analysis` subsection 2.2 on page 5
`freetext_core.main_termref` subsection 3.2 on page 9
`freetext_core.readscore` subsection 3.9 on page 11
`list.display` subsection 6.7 on page 22

10.15 Function attrib_str As String

Returns the attribute string for a termref, by a binary search on table b.

Arguments: `Termref` – Long

Subs and functions called: none

Called by: `freetext_core.readscore` subsection 3.9 on page 11
`terms.true_term` subsection 10.11 on page 32

10.16 Function linkto As String

Returns the linked termref (e.g. for alternate Read terms) by binary search on table b.

Arguments: `Termref` – Long

Subs and functions called: none

Called by: `pd.check_compressed` subsection 4.2 on page 12

11 Module wordlist

Module: wordlist – clinical and non-clinical words for spelling correction

11.1 Global variables and constants

Const `maxwords` = 100000 (*Maximum number of entries in the 'w' arrays (list of all words)*)
Const `maxignore` = 100 (*Number of entries in the 'ignore' table*)
Const `maxletters` = 30 (*Number of letters per word*)
`w_words(maxwords)` – String (*array of clinical and non-clinical words (no duplicates)*)
`w_clinical(maxwords)` – Boolean (*whether the word is possibly part of a clinical term*)
`w_top(maxletters)` – Long (*start position for words of different lengths*)
`w_max` – Long (*total number of words*)
`ignorelist(maxignore)` – String (*words which can be ignored e.g. if, and, of, the*)
`ignorelistnum` – Long (*number of words in ignorable list*)
`ignorephrase(maxignore)` – String (*words which can be ignored e.g. if, and, of, the*)
`ignorephrasenum` – Long (*number of phrases in ignorable phrases list*)

11.2 Sub add_to_wordlist

Adds a word or words to wordlist, and automatically sorts and compresses the wordlist when necessary

Arguments: `words_to_add` – String

Subs and functions called: `strfunc.numwords` subsection 7.6 on page 24
`wordlist.sort_and_compress_wordlist` subsection 11.4 on page 34
`strfunc.words` subsection 7.3 on page 23

Called by: `wordlist.import_wordlist` subsection 11.3 on page 34

11.3 Function import_wordlist As String

Creates a list of words in clinical terms and other English words. Gets text words from the synonyms table. Returns a string stating what was imported.

Arguments: `wordlistfile` – String

Subs and functions called: `synonym.numrows` subsection 9.2 on page 28
`wordlist.add_to_wordlist` subsection 11.2 on page 34
`synonym.s1` subsection 9.11 on page 30
`terms.numrows_a_c` subsection 10.2 on page 30
`terms.get_bagofwords` subsection 10.4 on page 31
`wordlist.sort_and_compress_wordlist` subsection 11.4 on page 34

Called by: `freetext_core.import_all_lookups` subsection 3.4 on page 10

11.4 Sub sort_and_compress_wordlist

Sorts the wordlist and removes duplicates. All words are preceded by the number of letters so they are sorted by number of letters then the text (alphabetically)

Arguments: none

Subs and functions called: wordlist.quicksort subsection 11.5 on page 35

Called by: wordlist.add_to_wordlist subsection 11.2 on page 34
wordlist.import_wordlist subsection 11.3 on page 34

11.5 Sub quicksort

Sorts a vector of strings

Arguments: tosort – Variant
start – Long (ByVal)
finish – Long (ByVal)

Subs and functions called: none

Called by: strfunc.bag_of_words subsection 7.11 on page 26
wordlist.sort_and_compress_wordlist subsection 11.4 on page 34
wordlist.import_ignore subsection 11.6 on page 35

11.6 Function import_ignore As String

Imports ignore.txt and ignorephrase.txt. ignore.txt should be sorted alphabetically but is re-sorted to ensure that the string comparison order is identical to that which will be used for binary searching. ignorephrase.txt contains semi-structured phrases which might be found in the raw text and should be ignored. Neither file has a header row.

Arguments: ignorefile – String
ignorephrase_file – String

Subs and functions called: wordlist.quicksort subsection 11.5 on page 35

Called by: freetext_core.import_all_lookups subsection 3.4 on page 10

11.7 Function in_wordlist As String

Returns CLIN for clinical words, WORD for non-clinical words and blank for words not found in the wordlist.

Arguments: instring – String

Subs and functions called: none

Called by: wordlist.wordsearch subsection 11.9 on page 36

11.8 Function approx_wordlist As Long

Approximate position of a word in the wordlist (sorted by wordlength, then word)

Arguments: instring – String

Subs and functions called: none

Called by: wordlist.wordsearch subsection 11.9 on page 36

11.9 Function wordsearch As String

Tries to convert a word into a standard form (or without spelling mistakes) which is in wordlist. Returns CLIN (for a clinical word) or WORD (for any other word) followed by the correctly spelled word, or blank if the spelling cannot be corrected

Arguments: `word` – String (ByVal)
`do_spellcheck` – Boolean (Optional)

Subs and functions called: `wordlist.in_wordlist` subsection 11.7 on page 35
`wordlist.approx_wordlist` subsection 11.8 on page 35
`freetext_core.fuzzylink` subsection 3.10 on page 12

Called by: `freetext_core.initial_search` subsection 3.5 on page 10

11.10 Function ignorable As Boolean

Returns True if a word is in the ignorable list for Read matching. Uses a binary search algorithm. The ignorable list must be sorted.

Arguments: `instring` – String

Subs and functions called: none

Called by: `freetext_core.initial_search` subsection 3.5 on page 10
`freetext_core.remove_ignorable` subsection 3.8 on page 11
`freetext_core.readscore` subsection 3.9 on page 11

11.11 Function remove_ignore_phrases As String

Returns instring with phrases found in 'ignorephrase' list removed. The function tries each phrase to remove in turn in the order they appear in the table.

Arguments: `instring` – String

Subs and functions called: `strings_Acc97.replace` subsection 8.1 on page 27
`wordlist.initial_process` subsection 11.12 on page 36

Called by: `freetext_core.main_analyse` subsection 3.3 on page 9

11.12 Function initial_process As String

Pre-processor to remove semi-structured Vision text in specific formats.

Arguments: `instring` – String

Subs and functions called: `strings_Acc97.replace` subsection 8.1 on page 27

Called by: `wordlist.remove_ignore_phrases` subsection 11.11 on page 36

12 Module checkterms

Module: checkterms – checks for occurence (or not) of words in the text to validate or invalidate some termrefs (medcodes)

12.1 Global variables and constants

Const maxcheckterms = 100
Termref(maxcheckterms) – Long (*medcode of output term*)
Qualify(maxcheckterms) – String (*word fragments which must be present in the text for the medcode to be returned*)
Dequalify(maxcheckterms) – String (*word fragments which must not be present in the text for the medcode to be returned' verbatim words or phrases*)
used – Long (*number of entries*)

12.2 Function import As String

Imports the checkterms table from text file. Returns a text statement stating whether it was successful. This text can be displayed on screen or added to a log file.

Arguments: filename – String

Subs and functions called: strfunc.dissect subsection 7.9 on page 25
attrib.import subsection 5.2 on page 19

Called by: freetext_core.import_all_lookups subsection 3.4 on page 10

12.3 Sub check_all

Carries out the actual checking

Arguments: checkstring – String
debug_ – Boolean (Optional)
sicknote – Boolean (Optional)
death – Boolean (Optional)
date_only – Boolean (Optional)

Subs and functions called: pd.max subsection 4.26 on page 19
pd.mean subsection 4.9 on page 14
pd.set_attr subsection 4.10 on page 15
pd.Attr subsection 4.8 on page 14
pd.remove subsection 4.22 on page 18
strfunc.dissect2 subsection 7.9 on page 25
checkterms.in_list subsection 12.4 on page 37
checkterms.if_qualify subsection 12.5 on page 38
checkterms.if_dequalify subsection 12.6 on page 38

Called by: freetext_core.main_analyse subsection 3.3 on page 9

12.4 Function in_list As Long

Returns the row number of the termref (medcode) in the checkterms table

Arguments: in_termref – Long

Subs and functions called: none

Called by: list.expand subsection 6.4 on page 21
list.display subsection 6.7 on page 22
checkterms.check_all subsection 12.3 on page 37

12.5 Function if_qualify As Boolean

Returns TRUE if one of the qualifying phrases is present in the text, FALSE otherwise.

Arguments: `pos` – Long
`checkstring` – String

Subs and functions called: `strfunc.dissect2` subsection 7.9 on page 25

Called by: `checkterms.check_all` subsection 12.3 on page 37

12.6 Function if_dequalify As Boolean

whether one of the dequalifying terms is present in the text

Arguments: `pos` – Long
`checkstring` – String

Subs and functions called: `strfunc.dissect2` subsection 7.9 on page 25

Called by: `checkterms.check_all` subsection 12.3 on page 37