User Guide for the eMERGE-2 CAAD Phenotype NLP System

Version 0.5

July 17, 2014

David Carrell and David Cronkite

Contents

Introduction

Carotid artery atherosclerosis disease (CAAD) is measured quantitatively by Doppler ultrasound and related imaging studies and reported as a percent stenosis. An easy to use, self-installing natural language processing (NLP) software package is provided to all eMERGE-2 sites for extracting these stenosis measures from unstructured clinical text reports. This document explains how to obtain, set up, and use the CAAD NLP system.

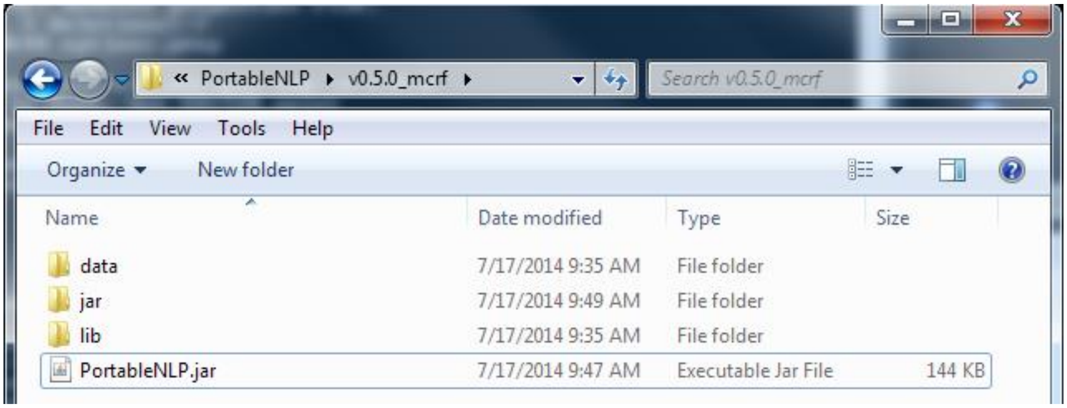
Overview

Software Download

The CAAD NLP software is available for download from PheKB's (www.phekb.org) CAAD phenotype page (http://www.phekb.org/phenotype/caad-carotid-artery-atherosclerosis-disease). Look for the most recent version of a zip archive file that includes "CAAD portable NLP" in its name or label.

Currently, there is an unlabeled "vanilla" version as well as a "Marshfield" version, which assumes that only appropriate laboratory tests are provided as input. If you are unsure of which version to use, select the "vanilla" version.

Download the zip archive (with a name like PortableNLP_v*n.n*.zip) to the local hard drive of the machine from which you want to execute the NLP system. Open the archive and drill into its top-level folder to view its contents:

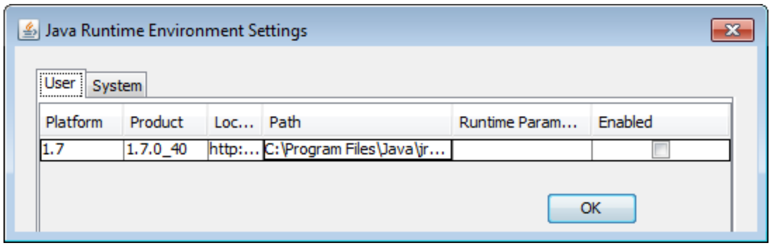


Name	Date modified	Type	Size
data	7/17/2014 9:35 AM	File folder	
jar	7/17/2014 9:49 AM	File folder	
lib	7/17/2014 9:35 AM	File folder	
PortableNLP.jar	7/17/2014 9:47 AM	Executable Jar File	144 KB

Computing Environment Requirements

The following are required to run the CAAD NLP system:

1. Tested on Windows, Linux, and Mac operating system
2. Java version 1.7 or higher. The Windows 7 Control Panel display for a qualifying version of Java looks like this:



Platform	Product	Loc...	Path	Runtime Param...	Enabled
1.7	1.7.0_40	http:...	C:\Program Files\Java\jr...		☐

3. Access, from the local machine, to either a file system or a data base from which the NLP system will read relevant text documents to be processed.

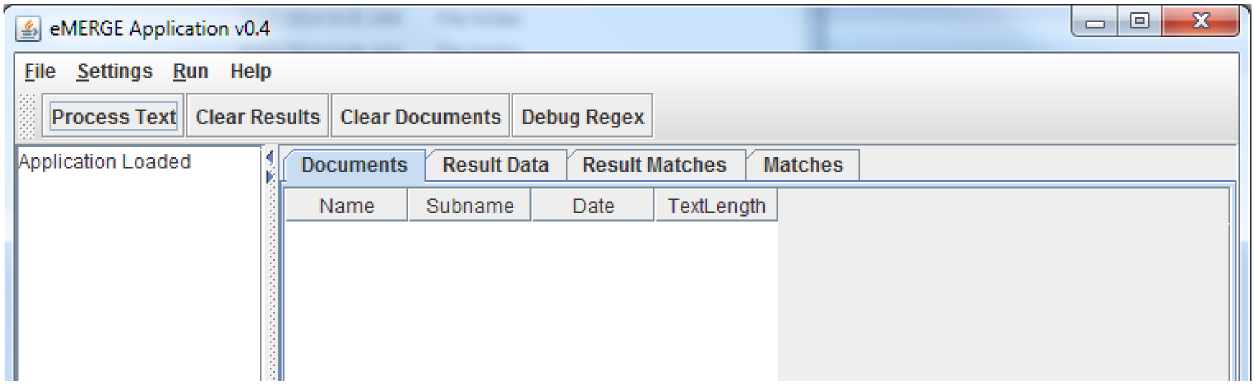
   a. If input files are provided by a database, the database must be MySQL, PostgreSQL, or Microsoft SQL Server, and you must know how to compose an ODBC connection string to link to your database server. You must also know the name of the database table in which the text is stored.
   b. If connectivity to other database systems is desired, please contact David Carrell (carrell.d@ghc.org).

4. Ability to write output files, either to the local file system or to the database (MySQL, PostgreSQL, or Microsoft SQL Server—this does not have to be the same database or type of database from which text is read into the software system).

Launching the NLP System

Start the CAAD NLP system as follows:

5. Open (double-click) the archive file "PortableNLP.zip" then click and drag (or otherwise copy) fold "v0.4" to your local computer. Copying this folder to the desktop or other location on the local machine's hard drive will avoid potential issues that may arise if you try to run this application on a network drive.
6. Open folder "v0.4" and double-click the JAR file named "PortableNLP.jar". This launches the application. Be patient while the application loads—it may take a minute and may appear as though nothing is happening. (On Windows, you can look in the Windows Task Manager's Applications tab to confirm that the application named eMERGE Application v0.5" is opening.) Once the application opens it will display a window similar to the one shown below.
   a. There are menu headings for File, Settings, Run, and Help.
   b. Below the menu are buttons used to Process Text, Clear Results, Clear Documents, and Debug (more on these later).
   c. The tall narrow window on the left displays system messages. After successfully launching the system it will display the message "Application Loaded."
   d. The large window on the right is for viewing information about the documents read into the system (see the Documents tab), and viewing data generated by the NLP system (see the Result Data, Result Matches, and Matches tab). For now, all of these tabs are empty.

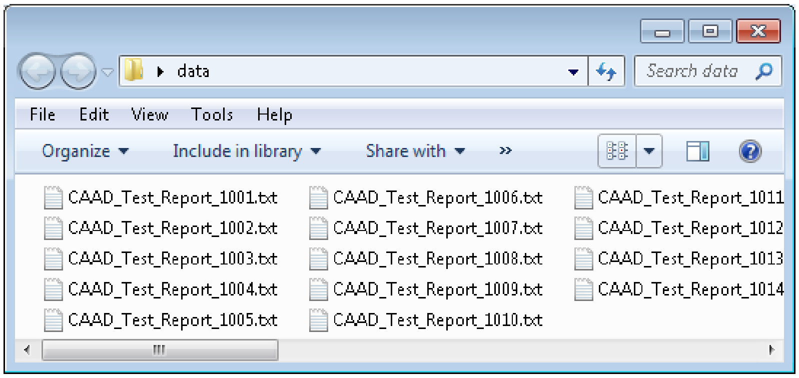


Application Loaded

Processing the Test Data

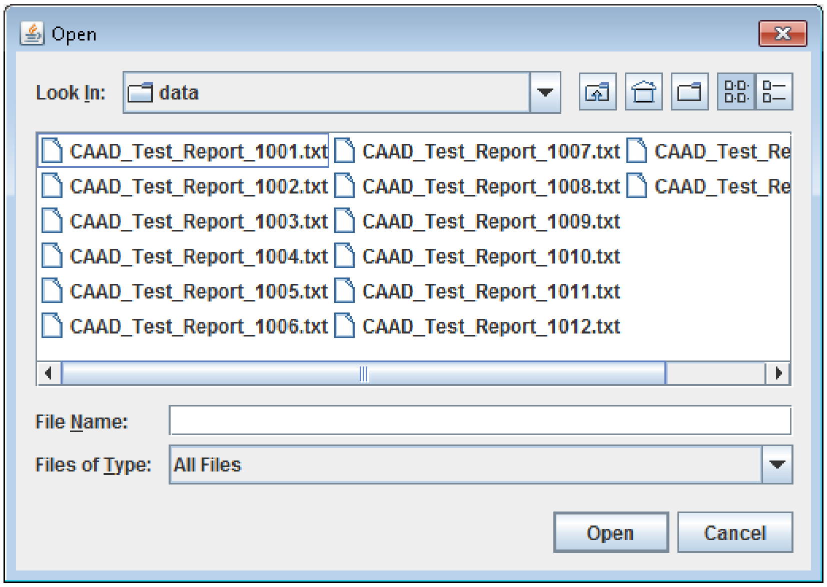
The application ships with a small set of de-identified/fictitious text documents which can be processed to test the NLP system. The test data consist of a set of individual text documents in a folder named "data" that is part of the archive. You can now process the test data as follows.

7. Inside the folder named "v0.5" find the folder named "data." It contains 14 plain files with names like CAAD_Test_Report_1001.txt, as shown below. You may open these reports in a text editor if you want to see what they look like.



CAAD_Test_Report_1001.txt
CAAD_Test_Report_1002.txt
CAAD_Test_Report_1003.txt
CAAD_Test_Report_1004.txt
CAAD_Test_Report_1005.txt
CAAD_Test_Report_1006.txt
CAAD_Test_Report_1007.txt
CAAD_Test_Report_1008.txt
CAAD_Test_Report_1009.txt
CAAD_Test_Report_1010.txt
CAAD_Test_Report_1011
CAAD_Test_Report_1012
CAAD_Test_Report_1013
CAAD_Test_Report_1014

8. Launch the NLP system BY double-clicking the JAR file (as in Step 5 above).
9. When the application opens, confirm that "Application Loaded" is displayed in the left pane. There may be other messages displayed regarding driver setup, but you can ignore these for purposes of processing the test data.
10. From the application's **File** menu select **Load Data → From File ….** This will display an Open file dialog. Use the Open file dialog to navigate to the data folder containing the 14 test notes (described above). Once you have drilled into the "data" folder the open dialog should look similar to this:



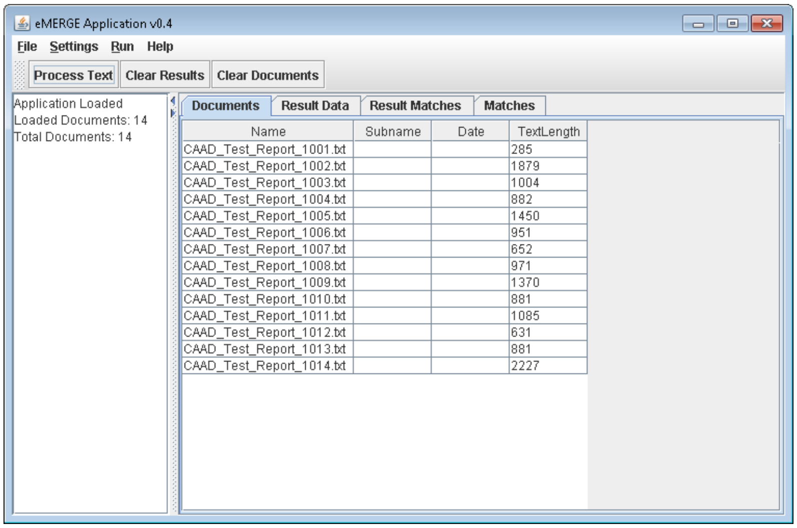
Look In: data

CAAD_Test_Report_1001.txt
CAAD_Test_Report_1002.txt
CAAD_Test_Report_1003.txt
CAAD_Test_Report_1004.txt
CAAD_Test_Report_1005.txt
CAAD_Test_Report_1006.txt
CAAD_Test_Report_1007.txt
CAAD_Test_Report_1008.txt
CAAD_Test_Report_1009.txt
CAAD_Test_Report_1010.txt
CAAD_Test_Report_1011.txt
CAAD_Test_Report_1012.txt
CAAD_Test_Re
CAAD_Test_Re

File Name:

Files of Type: All Files

11. You may open one file by clicking on it, then clicking the Open button. You may open several files by using the Shift or Ctrl keys in conjunction with mouse clicks (e.g., clicking the first file in the list, then holding down the Shift key while clicking the last file in the list selects all 14 documents). After selecting the files you want to process, click Open.
12. The application's Documents tab will now display the list of files just opened. Notice that the log window reports the number of documents loaded (14 in this example), and the total number of documents loaded since launching the application (also 14 in this example).

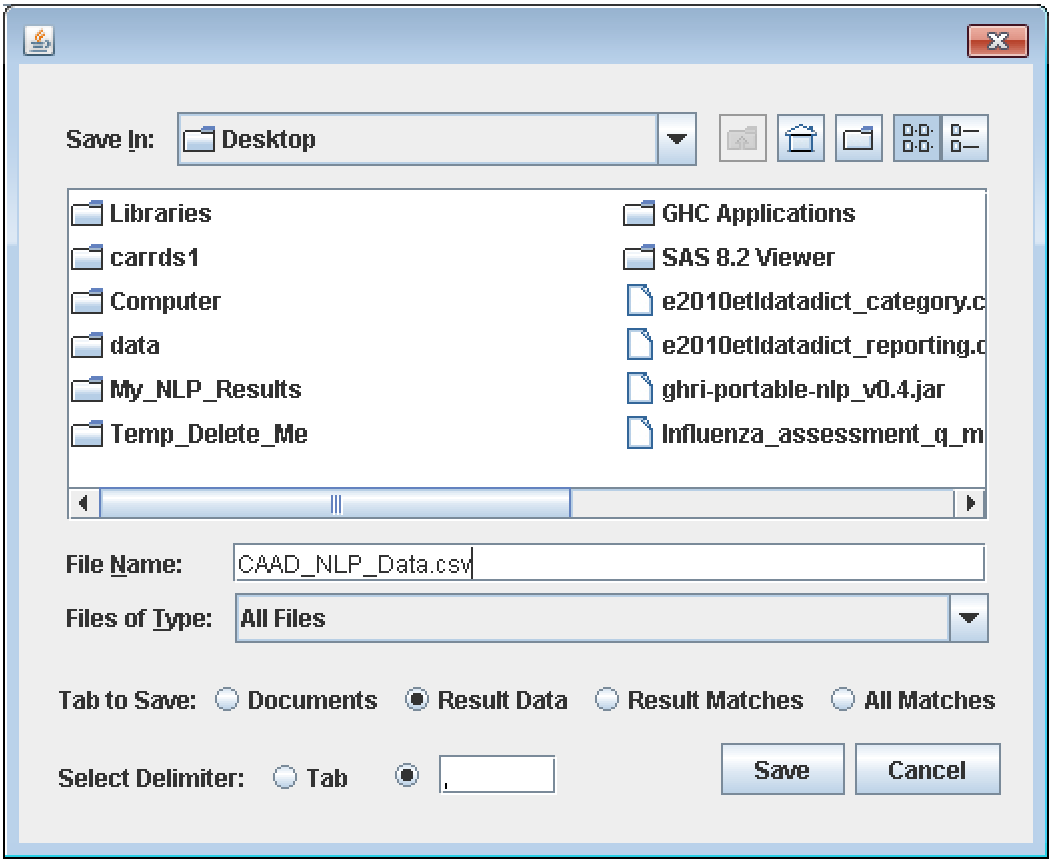


```
Application Loaded
Loaded Documents: 14
Total Documents: 14
```

Name	Subname	Date	TextLength
CAAD_Test_Report_1001.txt			285
CAAD_Test_Report_1002.txt			1879
CAAD_Test_Report_1003.txt			1004
CAAD_Test_Report_1004.txt			882
CAAD_Test_Report_1005.txt			1450
CAAD_Test_Report_1006.txt			951
CAAD_Test_Report_1007.txt			652
CAAD_Test_Report_1008.txt			971
CAAD_Test_Report_1009.txt			1370
CAAD_Test_Report_1010.txt			881
CAAD_Test_Report_1011.txt			1085
CAAD_Test_Report_1012.txt			631
CAAD_Test_Report_1013.txt			881
CAAD_Test_Report_1014.txt			2227

13. Process the loaded documents with the CAAD NLP system by clicking the **Process Text** button. The documents will be processed, usually quickly. Notice the message in the log window that reports processing time in milliseconds. It generally takes only a few dozen milliseconds to process the 14 test documents.

14. After documents have been processed you may inspect results on the three results tabs—Result Data, Result Matches, and Matches. Tab Matches provides the most granular data. The Result Matches tab provides data slightly less granular. The Result Data tab sometimes (but not always) provides a higher level of summary data than the Result matches tab.
15. To save the NLP-generated results, use the menus to select **File** → **Export Data** … → **Save to File**. This opens a dialog as shown below.
    a. Give the file a name in the **File Name** field.
    b. Specify the **Tab to Save** (you may choose between Documents, Result Data, Result Matches, and All Matches). The Result Data tab contains the data the NLP system was indended to extract.
    c. Specify the delimiter of your choice and a file name, then click Save, as shown below.



Save In: Desktop

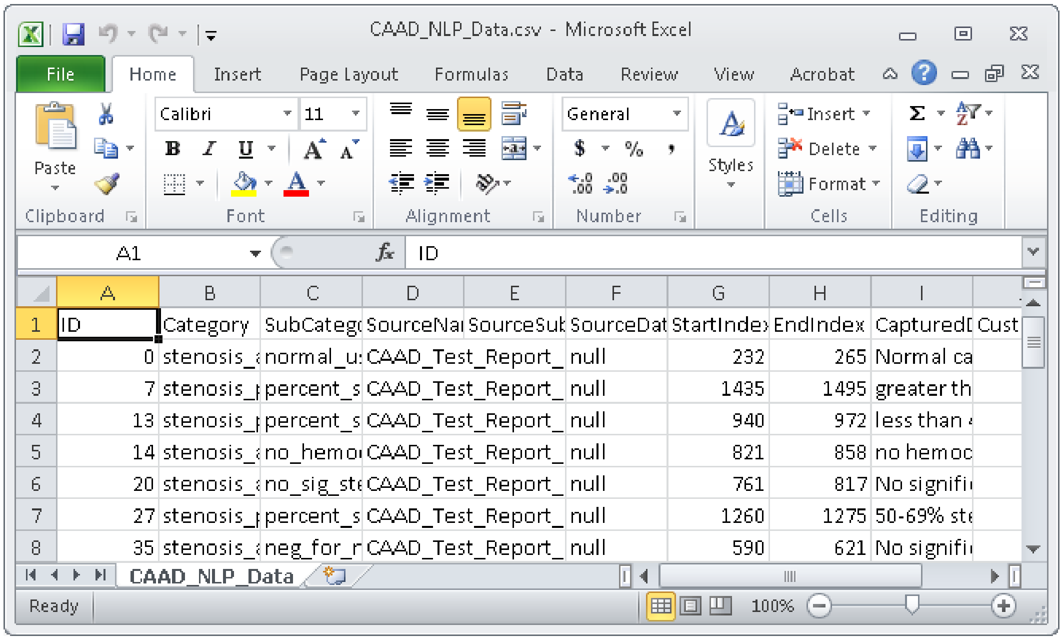
File Name: CAAD_NLP_Data.csv

Files of Type: All Files

Tab to Save: ○ Documents ● Result Data ○ Result Matches ○ All Matches

Select Delimiter: ○ Tab ● ,

16. The save data can be opened in other applications, such as Excel, shown below:



A1 ID

	A	B	C	D	E	F	G	H	I	
1	ID	Category	SubCateg	SourceNa	SourceSu	SourceDat	StartInde	EndIndex	Captured	Cust
2	0	stenosis_	normal_u	CAAD_Test_Report_		null	232	265	Normal ca	
3	7	stenosis_	percent_s	CAAD_Test_Report_		null	1435	1495	greater th	
4	13	stenosis_	percent_s	CAAD_Test_Report_		null	940	972	less than	
5	14	stenosis_	no_hemo	CAAD_Test_Report_		null	821	858	no hemoc	
6	20	stenosis_	no_sig_ste	CAAD_Test_Report_		null	761	817	No signifi	
7	27	stenosis_	percent_s	CAAD_Test_Report_		null	1260	1275	50-69% ste	
8	35	stenosis_	neg_for_r	CAAD_Test_Report_		null	590	621	No signifi	

17. Exit the program by selecting from the **File** menu the **Exit** command.

Exiting the NLP System

To exit the system:

18. Select from the menu **File → Exit**.

Debugging

Overview

While the NLP system is designed to be as robust as possible, variation in free text reports at different sites will likely cause the system some problems.

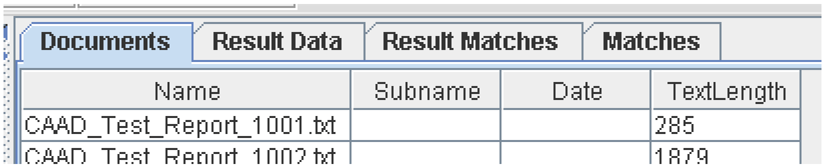
Some common problems include:

- No output in the “Result Data” tab

Reporting Problems

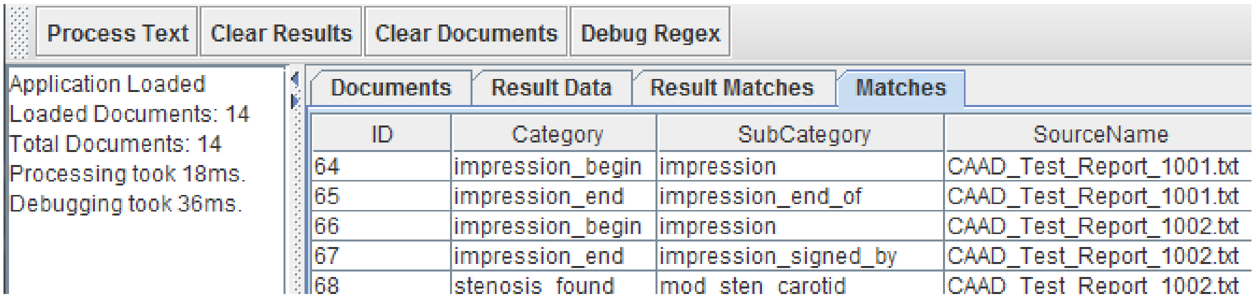
If you encounter problems, please follow the guide below to help us figure out what any issues are:

1. Check the “TextLength” column in the “Documents” tab to make sure the values are not getting truncated. You can also check the actual text that’s being loading by going to **Settings -> Columns…** and selecting the check mark next to “Preview of Text” to look at the text that’s being loaded.



Name	Subname	Date	TextLength
CAAD_Test_Report_1001.txt			285
CAAD_Test_Report_1002.txt			1879

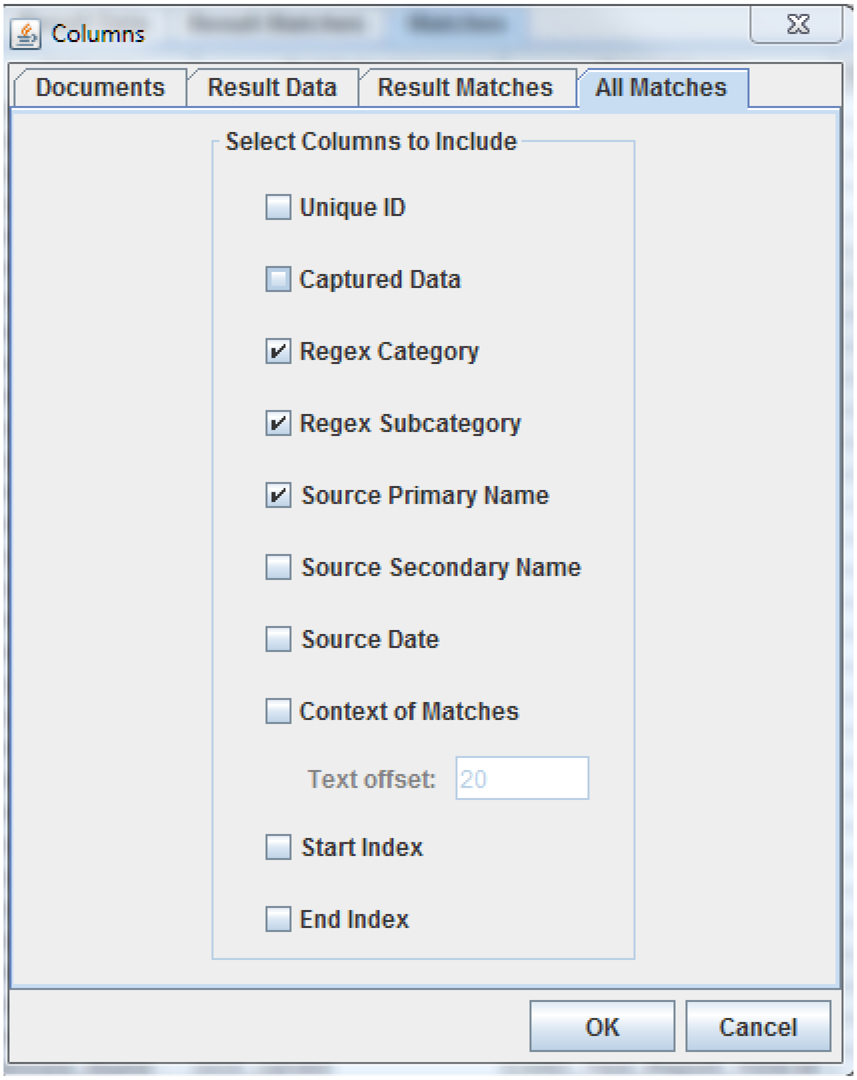
2. Clear any existing results by clicking the “Clear Results” button, and then select the “Debug Regex” button. The “Matches” tab should now be populated.



Application Loaded
Loaded Documents: 14
Total Documents: 14
Processing took 18ms.
Debugging took 36ms.

ID	Category	SubCategory	SourceName
64	impression_begin	impression	CAAD_Test_Report_1001.txt
65	impression_end	impression_end_of	CAAD_Test_Report_1001.txt
66	impression_begin	impression	CAAD_Test_Report_1002.txt
67	impression_end	impression_signed_by	CAAD_Test_Report_1002.txt
68	stenosis_found	mod_sten_carotid	CAAD_Test_Report_1002.txt

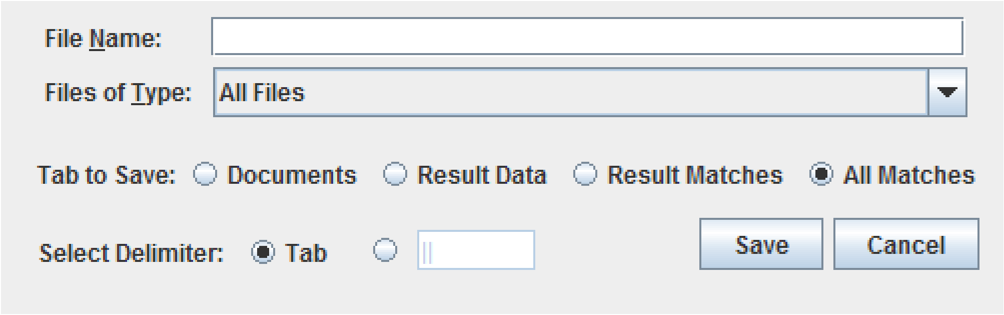
3. Next, go to **Settings -> Columns…**, navigate to the “All Matches” tab and make sure only “Regex Category”, “Regex Subcategory”, and “Source Primary Name” (only if the SourceName column does not contain unique identifiers!—we just want to know which regular expressions have matched in which reports) are checked.



Select Columns to Include

☐ Unique ID
☐ Captured Data
☑ Regex Category
☑ Regex Subcategory
☑ Source Primary Name
☐ Source Secondary Name
☐ Source Date
☐ Context of Matches
Text offset: 20
☐ Start Index
☐ End Index

4. Finally, output the “Matches” tab to a text file (**File -> Export Data -> Save to File…**) and send it to David Carrell (carrell.d@ghc.org) for review. Make sure the “All Matches” radio button is selected.



File Name:
Files of Type: All Files
Tab to Save: ○ Documents ○ Result Data ○ Result Matches ● All Matches
Select Delimiter: ● Tab ○ ||

5. If possible, also send some example **snippets of text** from an actual carotid ultrasound report. Snippets of interest would be those which describe the **type of exam** (examples from our reports are “*CAROTID DUPLEX ULTRASOUND*” and “*EXAMINATION: Bilateral carotid ultrasound*”) and **the findings** (examples from our reports are “*narrowing of 60% to 70%*”, “*50-75% stenosis*

of the right internal carotid artery" and "*Negative for any significant stenosis in the left carotid artery*"). Send these snippets of text to David Carrell (carrell.d@ghc.org).