The General Inquirer User's Guide

Vanja Buvač and Philip Stone
(buvac,pjstone)@fas.harvard.edu

April 2, 2001

Abstract

The General Inquirer is a content analysis program described in [2] and [1]. This document explains how to use the Java language implementation of the General Inquirer (version j1.0). Familiarity with the web version of the General Inquirer (`http://inquirer.wjh.harvard.edu/GI`) is a prerequisite to this document.

1 Starting the Program

The General Inquirer is distributed as a zip archive. The contents of the archive together with links to tools for unpacking it on various platform is described in Appendix A. Uncompress the archive in a directory of your choice.

The current version of the General Inquirer is written in Java which makes it usable on any operating systems with a Java Virtual Machine (JVM). In case that a JVM is not pre-installed on your computer, instructions on obtaining and installing a JVM are available at Sun's website (`http://www.java.sun.com`). The General Inquirer has been successfully used on the following operating systems: Linux, Windows9x, MacOs, Solaris, MkLinux, and a wide variety of other Unix based systems.

1.1 Starting the program from the command prompt

Once you have successfully installed the JVM on your computer, you need to start the `giGui` class. If a command prompt is available on your operating system you can start this class with the following command.

```
java  -mx128m giGui
```

The option `-mx128m` asks the JVM to allocate 128 megabytes to the execution of General Inquirer. Depending on the size of your text files and computer memory you might want to change this number.

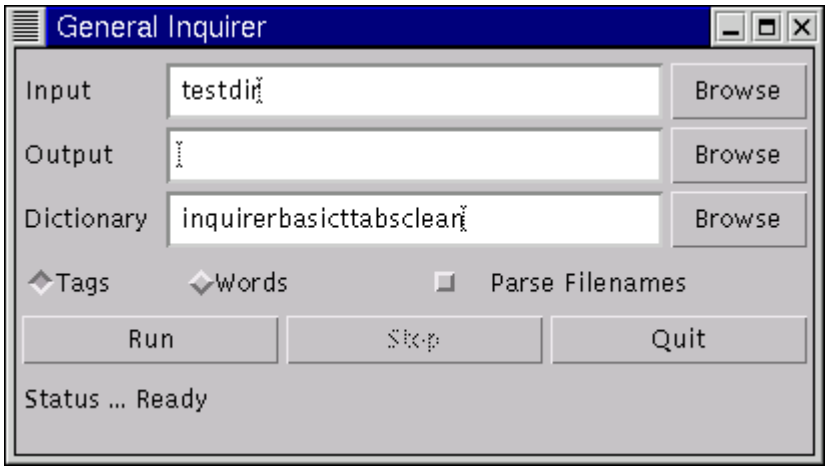

Figure 1: The General Inquirer main window.

1.2 Starting the program under Windows

On Windows the General Inquirer can be started by executing the batch file `clickme.bat`. In case the JVM is not pre-installed you will need to follow the instructions in the previous subsection (1.1).

1.3 Starting the program under MacOs

On the MacOs the General Inquirer can be started by executing the batch file `clickmemac`. In case the JVM is not pre-installed or you get some other error please consult Apple's web site for alternative ways of starting the program (`http://developer.apple.com/java/`).

2 Using the Program

Once you have successfully started the program you will see a window similar to the one shown in Figure 1 displayed on your screen.

2.1 Test run

To test the program simply click on the `Run` button without changing any arguments. This will execute the General Inquirer on the directory `testdir` which is included in the distribution, the results will be printed to the standard output which is most probably your command prompt or shell window, and the dictionary used will be the `inquirerbasicttabsclean`. First you will have to wait for the dictionary to load. The status bar will show the message "Loading dictionary ... please wait." In a minute or two, once the dictionary is loaded, the files in the directory will be processed one by one and the output will probably

race by your command prompt window. the status bar will tell you what the General Inquirer is doing at every point.

2.2 Explaining the options

The basic function of the General Inquirer is to generate a count of words falling into a given semantic category. There are a few options that which enable you to tailor this process to your particular needs. All the options are available through the main window. Here is the explanation of all the items on General Inquirer's control window.

Input is the name of the text file that you want content analyzed. If the name refers to a directory, all the files in the directory are analyzed. The output in this case is a matrix where files are represented by row entries. If the name refers to a single file then only one file is analyzed, and the output is a single row. The input files should be in plain text format; no special formating is necessary.

Output is the name of the file where you want the output to go. If this field is left blank, the output goes to standard output which is probably the command prompt.

Dictionary is the name of the file containing the dictionary. If you are not interested in changing the tags and entries in the dictionary, you won't need to change this text field.

Browse The browse buttons located to the right of the text areas let you browse through the your files to choose to one you want.

Tags If this option is selected the output of the analysis will be a summary of the tags assigned to the file(s) processed.

Words With this option you can get a count of the words appearing in the processed file(s). In this output format the rows in the output matrix contain the words, and the columns contain the file(s). In other words, the output matrix is a "transposed" version of the tags matrix.

Parse Filenames When processing multiple files it might be useful to encode some variable information directly into the filename. Selecting this option will parse the filenames and include the variable names it finds in the output matrix. Details of how the filenames are parsed are covered in section 2.5.

Run This button is used to start executing the General Inquirer.

Stop This button halts the execution.

Quit This button exits the General Inquirer.

file1	file2	file3	file4	format	wordcount	leftovers	Positiv	...
gore	speech	announce	1	r	1239	52	96	...
gore	speech	announce	1	s	1239	4.196933	7.748184	...
gore	speech	announce	2	r	1505	41	112	...
gore	speech	announce	2	s	1505	2.7242525	7.4418607	...
bush	speech	announce	1	r	2036	53	187	...
bush	speech	announce	1	s	2036	2.6031435	9.184676	...
bush	speech	announce	2	r	1048	13	92	...
bush	speech	announce	2	s	1048	1.240458	8.7786255	...
gore	speech	education	1	r	4009	149	316	...
gore	speech	education	1	s	4009	3.7166376	7.882265	...
gore	speech	education	2	r	1689	63	166	...
gore	speech	education	2	s	1689	3.7300177	9.8283	...

Table 1: Tag output matrix for the files in *testdir*.

Status This area is used to tell you what the current status of the General Inquirer is. Specifically, it will tell you when it is loading the dictionary, processing a given file, generating output, or ready for the next task.

2.3 Tag Output format

When the tag option is selected the General Inquirer outputs a matrix containing the counts and percentages of words tagged with a given semantic category. Table 1 shows the results from the texts in the *testdir* directory with the filenames parsed for variables. The first four columns of the matrix contain the variables extracted from the filenames of the processed files. The fifth, *format*, column contains the "format" of the numbers contained in the given row. Each file contains two rows. The first is the raw count output which is simply the count of words in a given semantic category. This format is labeled with a *r* in the format column. The second is a scaled count which represents the percentage of words appearing in a given category. This number is computed as $100 * \frac{\text{rawcount}}{\text{wordcount}}$. This format is labeled with a *s* in the format column.

The sixth, *wordcount*, column shows the total number of words present in the given document. The seventh, *leftovers*, column shows the number of words that were not fount in the dictionary.

The remaining columns contain the counts (or percentages) of words in a given tag. For a full description of the tags used in the current dictionary see `http://www.wjh.harvard.edu/~inquirer`.

2.4 Word Output format

When the word option is selected the General Inquirer outputs a matrix containing the counts of the words contained in a give document. Table 2 shows the word output matrix for the texts found in the *testdir* directory. In contrast

word	gore speech announce 1	gore speech announce 2	bush speech announce 1	bush speech announce 2	gore speech education 1	gore speech education 2
TOTAL	1239	1505	2036	1048	4009	1689
A	22	39	55	28	82	29
ABILITY	0	1	0	0	1	1
ABLE	0	0	0	1	1	0
ABOUT	3	3	3	2	9	8
ABOUT#1	3	3	3	2	9	7
ABOUT#2	0	0	0	0	0	1
ABUNDANT	1	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
appliance	0	0	0	0	0	1
up-to-the-minute	0	0	0	0	0	1
$60	0	0	0	0	0	1
website	0	0	0	0	0	2
providers	0	0	0	0	0	1
enrollees	0	0	0	0	0	1
labor-management	0	0	0	0	0	1

Table 2: Word output matrix for the files in *testdir*.

to the tag output format, the columns of the matrix represent the files, and the rows contain the given words. In this example the parse filenames option was selected, so the variables extracted from the filenames are found in the first four rows. The fifth, *TOTAL* row contains the total count of words in the document. The following rows contain the wordcounts for the words in the dictionary. Once this list is exhausted, a count of the words not in the dictionary (leftover words) is displayed.

2.5 Parsing variables out of the filenames

When the *Parse Filenames* option is selected the General Inquirer creates creates columns from this file name identification information in the output spreadsheet according to the following procedure:

1. All the characters in a file name starting with the first period are removed. For example, `.txt` and `.doc` will be removed.

2. Each word in a file name (separated by spaces) is given a separate ID field.

3. The ID fields are labeled ID1, ID2, etc. It may be helpful to rename these columns with more descriptive labels later on the statistical spreadsheet.

4. For the last word in the file name (which may be the only word if there is but one): The computer tests to see if it begins with a character. If

it does, it then looks for a digit in the word. If a digit is found, then all characters up to the digit are made into one ID field and the characters starting with the digit are made a second ID field.

Here are some examples of how variables are parsed out of the filenames.

- The filename `bush speech defense1` is made into 4 ID fields for the candidate name, the type of document, the topic, and the serial number within that group: (1) bush, (2) speech, (3) defense, (4) 1.
- The filename `UMIN 0225.txt` will have the `.txt` removed and be made into two fields, one for Univ. of Minnesota, the second for the newspaper date (February 25). The date field may be further recoded into groupings by the statistical software.
- The filename `DH134.TXT` will have the `.TXT` removed and separated into two fields, `DH` for a high performer and `134` for the respondent's ID number.
- The filename `C87` will similarly be two fields, with the `C` for conservative party and `87` indicating the year of the party manifesto.

3 License

This version of the General Inquirer is made available exclusively for educational and research purposes. In publications please reference this document and [2]. Harvard University and The Gallup Organization have supported the development of this version of the General Inquirer; please consider acknowledging their support. Please do not distribute the General Inquirer on your own. We are more than happy to give copies to other researchers, but we would like to know who is using it.

This program is provided "as is" and carries no warranties of any kind. Comments and questions are always welcome.

A Description of the distribution

The General Inquirer is currently distributed as a zip archive containing the following files and directories.

filename	size	description
clickme.bat		Startup file for Windows9x.
clickmemac		Startup file for MacOs.
manual.pdf		This document.
inquirerbasicttabsclean	2911237	The dictionary.
DisambigRules	277167	This disambiguation rules.
GeneralInquirer.class	61520	Main Java class.
giGui.class	7943	Java class for the user interface.
giThread.class	6187	Java class for concurency.
testdir	directory	Contains a few text files.

To uncompress this archive you will need to have a zip archiving program available on your computer. On Unix based computers the `unzip` command will uncompress a zip archive into the current directory. On the Windows platforms the winzip utility (`http://www.winzip.com`) is widely used for this purpose. The free evaluation copy is likely to be sufficient. On the MacOs platforms the maczip (`http://www.sitec.net/maczip/` utility is will do the job.

References

[1] Edward F. Kelly and Philip J. Stone. *Computer Recognition of English Word Senses.* North-Holland Publishing, 1975.

[2] Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, Daniel M. Ogilvie, and associates. *The General Inquirer: A Computer Approach to Content Analysis.* MIT Press, 1966.