

TrueList™

from

DATA DISCOVERY™

Version 2.0
User Documentation

Contents

Overview

Requirements

Using *TrueList*

File Menu

Options Menu

Log Example

ASCII Format

Overview

U.S. and International businesses waste thousands of dollars each year mailing direct marketing materials to duplicate addresses. *TrueList* eliminates this wasteful practice by identifying and removing duplicate addresses from mailing lists. *TrueList* tracks down and weeds out not only exact duplicates, but also entries that contain a high degree of similarity. This process is performed by comparing addresses, calculating the percentage of matching words, and eliminating close matches.

TrueList serves as a filter. It screens mailing lists for duplicates or near duplicates, then generates a mailing list free of redundancies. A dup file can be created to identify duplicate addresses, and a log file contains details about those duplicates.

You have the option of removing duplicates on a case-by-case basis or choosing the automatic removal feature, which makes the decisions for you.

Requirements

TrueList runs on IBMcompatible PCs (80386 or higher) with at least 2 MB RAM and equipped with Microsoft Windows (version 3.0 or higher). A viewing option at the end of *TrueList* requires Windows Notepad or another viewer or editor program.

Using *TrueList*

The following sections describe the primary functions of *TrueList*. The order of the sections generally follow the natural sequence of steps that a user would take.

List Preparation

Starting Program

Input File

ASCII Input Options

dBase Input Options

Output File

Log File

Dup File

Start Processing

Dups Found

Removal Options

View Log

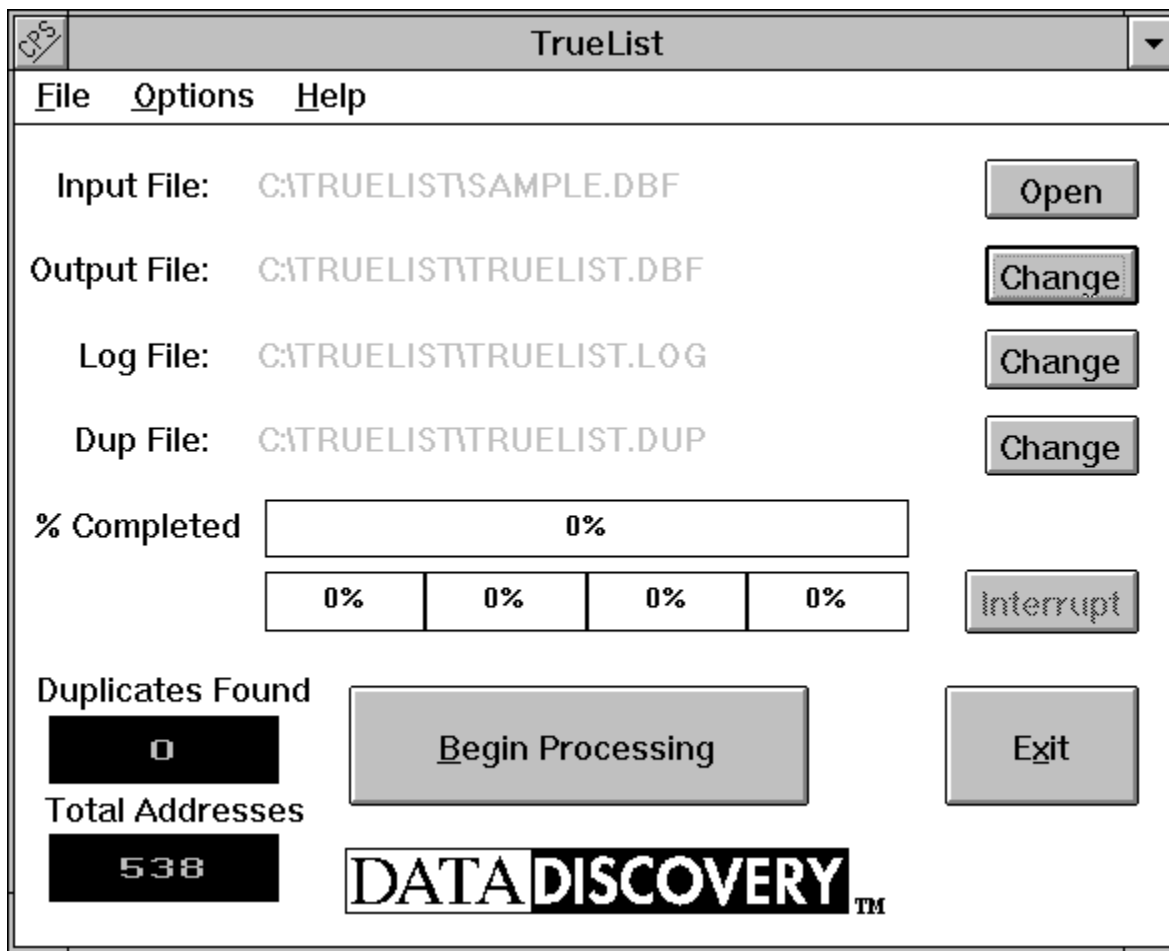
List Preparation

Your mailing list should be in either dBase (.dbf extension) or simple (ASCII) text format. Word processing formats can sometimes be used, but the extra codes used by the word processing system may cause unexpected results. For details on creating an ASCII file, see the **ASCII Format** section below.

ASCII Format

Starting Program

Start the program by doubleclicking the *TrueList* icon in the *TrueList* group. The following dialogue box will appear.

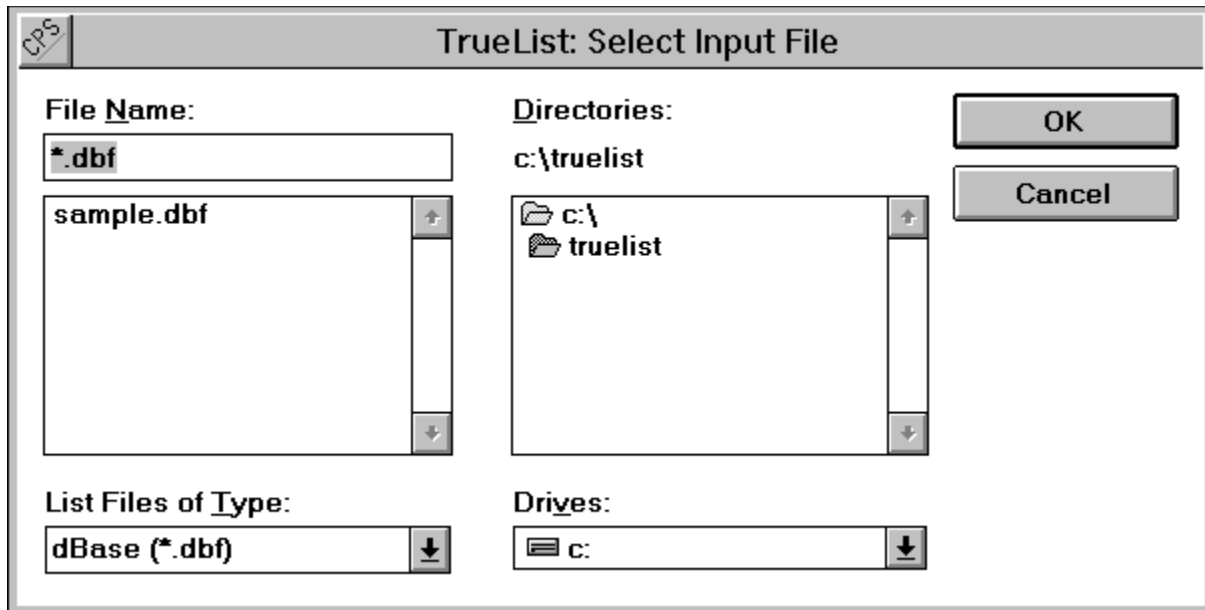


All of the functions of *TrueList* can be accessed from this screen. The menus provide file functions, options, and help. The buttons and other displays provide information about files being used and status. Each of the aspects of this screen are described in sections below.

Input File

The input file, shown in the above screen, is the original list (in either ASCII or dBase format). ASCII format files have special requirements which are listed in the **ASCII Format** section below.

Click the **Open** button to the right of the **Input File** prompt to bring up a list of files. The following screen will appear.

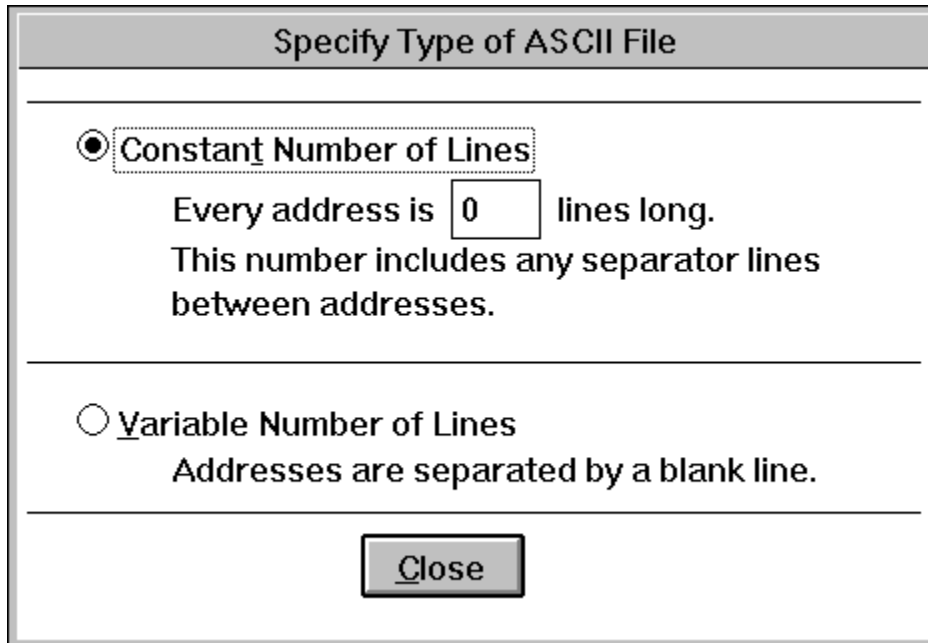


Select the mailing list from the selection box and click **OK**. A sample dBase list (**sample.dbf**) and a sample ASCII list (**sample.asc**) are provided for demonstration purposes.

ASCII Format

ASCII Input Options

If the input file is ASCII text (such as **sample.asc**), the following dialog box will appear after opening the file.



The dialog box is titled "Specify Type of ASCII File". It contains two radio button options. The first option, "Constant Number of Lines", is selected and has a text input field next to it containing the number "0". Below this option is the text: "Every address is 0 lines long. This number includes any separator lines between addresses." The second option, "Variable Number of Lines", is unselected and has the text: "Addresses are separated by a blank line." At the bottom center of the dialog box is a "Close" button.

Select **Constant Number of Lines** if every address contains the same number of text lines. Select **Variable Number of Lines** if the number of lines varies and the addresses are separated by blank lines. For more information on setting the number of lines in an address, see the **ASCII Format** section below.

NOTE: The **sample.asc** file requires **Variable Number of Lines** to be selected.

[Constant Number of Lines](#)

[Variable Number of Lines](#)

dBase Input Options

If the input file is a dBase list (such as **sample.dbf**), the following dialog box will appear after opening the file.

TrueList: Fields to Include

Please select which fields should be considered in the search for duplicates. The fields marked with an X will be included. Try to include those fields which identify an address as unique (such as the person's last name), and exclude fields which are highly repetitive (such as the state in a regional list).

COMPANY

ADDRESS

CITY

STATE

POSTCODE

NAME

Prev Page Next Page Clear All Select All Close

Only selected fields will be used in calculating the percentage of common words between addresses. Thus, you should select fields that are most likely to make a distinction between addresses. In addition, *TrueList* will run faster with fewer fields selected.

If there are more fields than can be shown on the screen, the **Next Page** and **Prev Page** buttons can be used to view more fields. To make wholesale selections, use the **Select All** and **Clear All** buttons. Click the **Close** button to finalize the selection, and the following dialog box will be displayed.

TrueList: Sort Order

Please select up to three fields to sort on. Try to select a sort order that is likely to place probable duplicates near each other.

COMPANY

ADDRESS

CITY(2)

STATE(3)

POSTCODE(1)

NAME

Click the boxes in the order you want to sort the list. The addresses will be sorted on the selected fields, using the first field selected as the primary sort field, the second as the secondary sort field, etc. You may choose up to 3 sort fields, and may also select fields that are not chosen as comparison fields.

The **Next Page** and **Prev Page** buttons are used when there are more fields than can be shown on one screen. If you want to clear the selections and start over, press the **Clear Sort** button. Click the **Close** button to finalize the sort order.

NOTE: *TrueList* detects duplicates within groups of 1,000 addresses (500 if memory is limited). dBase lists can be sorted to bring the addresses within local proximity of each other. To run *TrueList* on multiple sort orders, you can reset the sort order by selecting **Options / DBF Options / Select DBF Sort Order**. Multiple sort orders is useful for finding duplicates that might be missed if they are contained in different groups of addresses.

If your input file is ASCII text, the physical order of the file will determine

which addresses are compared in each group. *TrueList* cannot change this order.

Output File

After running *TrueList*, the output file will contain a clean list of addresses without redundancies. If the input file is ASCII, the output file will always be ASCII.

If the input file is in dBase format, *TrueList* will be able to delete records within the original file. Thus, the output file is usually unnecessary for dBase lists and will be disabled by default. However, an output file can be generated from a dBase list in either dBase or ASCII format.

The default name for the output file is **truelist.dbf** (or `truelist.out` if ASCII), but you can designate another name by clicking the **Change** button to the right and entering the new name. If the output file status is inactive, press the **Open** button to select an output file.

Log File

The log file contains information about removed entries, including retained addresses, words shared by different addresses, and percentage of words in common. You can use the default name, **truelist.log**, or rename the file by clicking the **Change** button and entering the new name.

Dup File

The dup file contains addresses that were removed from the mailing list. The dup file is written in ASCII format if the input is ASCII. If the input file is dBase, the output will normally be written in dBase format, unless the user chooses ASCII format.

Start Processing

To start the search process, click the **Begin Processing** button. The top progress bar shows the percent of all addresses that have been processed. The bottom progress bar is used for each group of addresses and fills with different colors to show the processing phases:

Building Index, Filling Matrix, Finding Duplicates, and Writing Output. If you decide to stop the program during execution, click the **Interrupt** button.

Dups Found

When *TrueList* finds a duplicate, it invokes a window similar to that shown below.

The dialog box is titled "TrueList: Duplicate Found [54% Match]". It contains two columns: "Address 1 (#1)" and "Address 2 (#2)".

Address 1 (#1)	Address 2 (#2)
John Q. Public XYZ Corporation 100 North Main Street Anytown, USA 12345	XYZ Corp. 100 N. Main, Suite 10 Anywhere, USA 12345 C/O Jonathan Public

54% Match on Words: PUBLIC XYZ 100 MAIN USA 12345

View Options

Full Records
 Selected Fields

Keep First **Keep Both** **Process Automatically**
Keep Second **Keep Neither** **Interrupt Search**

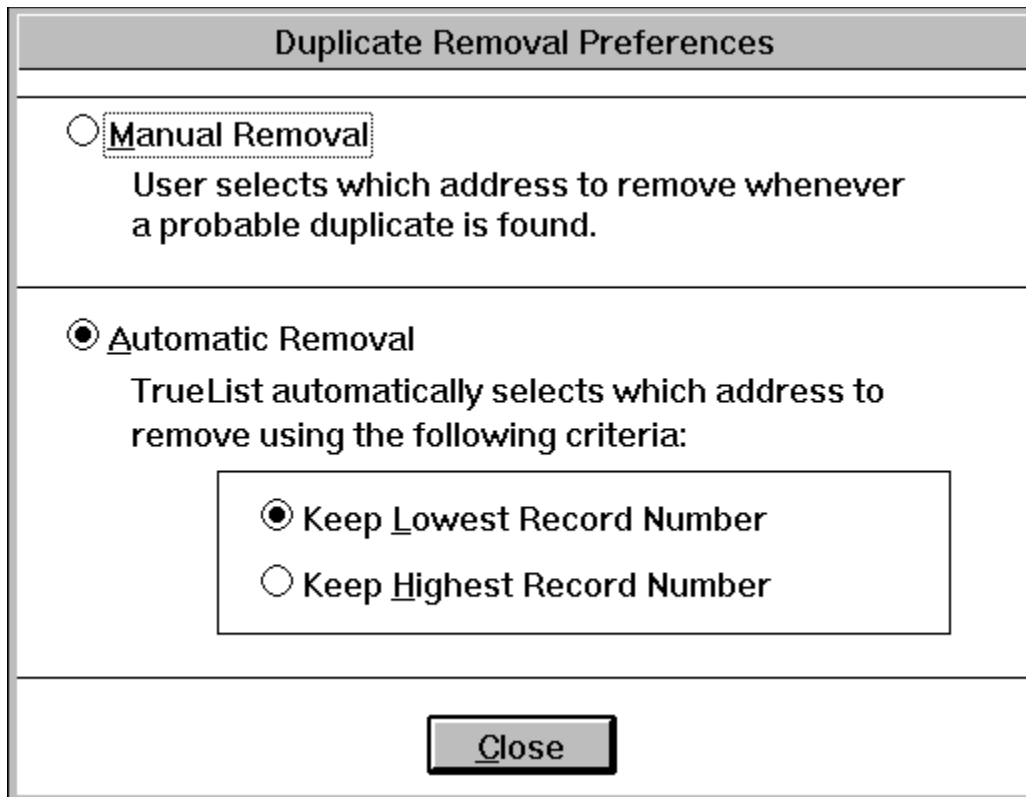
When the input file is dBase, the user may choose to either view all fields in the record or only those fields that were chosen for comparison. If the input file is ASCII, these view options will not be available, and will always show the complete address.

Click the **Keep First** button to retain the first address and remove the second from the mailing list. Select **Keep Second** to retain the second and remove the first. To retain both addresses, click **Keep Both**. To delete both, click **Keep Neither**.

To halt the comparison process, click **Interrupt**. If you want *TrueList* to make the decisions for you, choose **Process Automatically**.

Removal Options

If the user selects **Process Automatically** from the duplicates window or selects **Options / DBF Options / Duplicate Removal Preferences**, the following dialog will appear.



The image shows a dialog box titled "Duplicate Removal Preferences". It contains two main sections. The first section is "Manual Removal", which is currently unselected (radio button is empty). Below it is the text: "User selects which address to remove whenever a probable duplicate is found." The second section is "Automatic Removal", which is selected (radio button has a dot). Below it is the text: "TrueList automatically selects which address to remove using the following criteria:". This section contains a sub-dialog box with two options: "Keep Lowest Record Number" (selected) and "Keep Highest Record Number" (unselected). At the bottom of the main dialog box is a "Close" button.

If the **Manual Removal** option is chosen, *TrueList* displays each pair of duplicate addresses one pair at a time.

The **Automatic Removal** option allows automatically removing either the lowest numbered or highest numbered record in all cases without asking the user for verification. The numbering of the addresses is based on the physical order of the list and is useful for keeping either the oldest or newest record.

NOTE: If there are more than two addresses that are duplicates of each other, *TrueList* will display each pair of addresses.

View Log

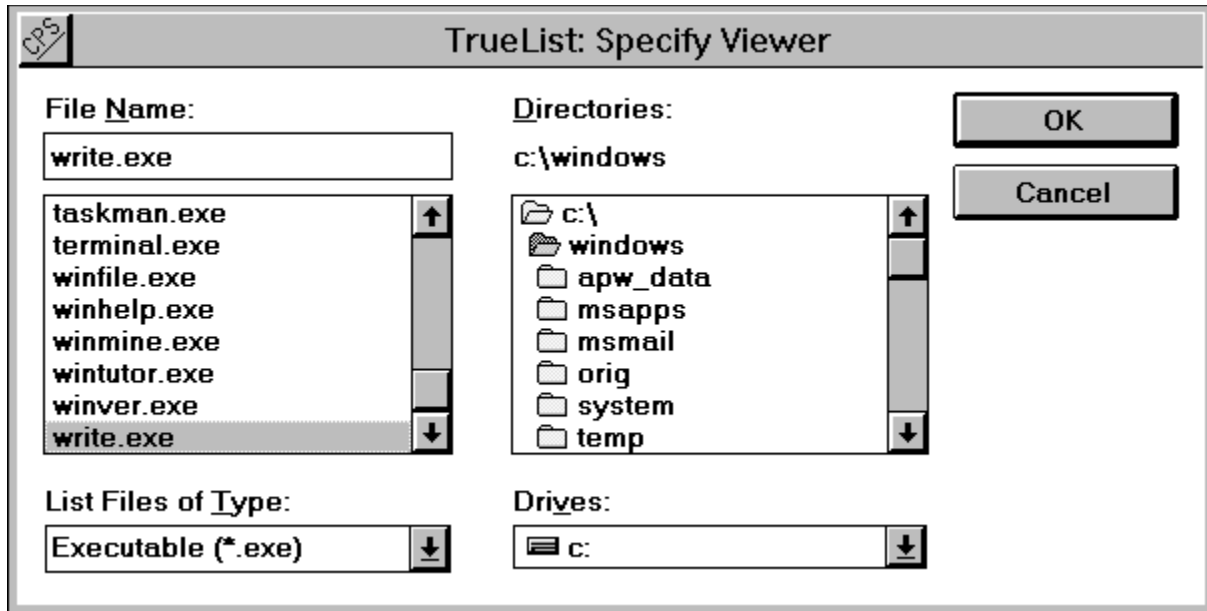
If the log file option was turned on, this screen will appear when *TrueList* finishes processing.



If you want to see information about the removed entries, click **Yes**. The following dialog box will appear.



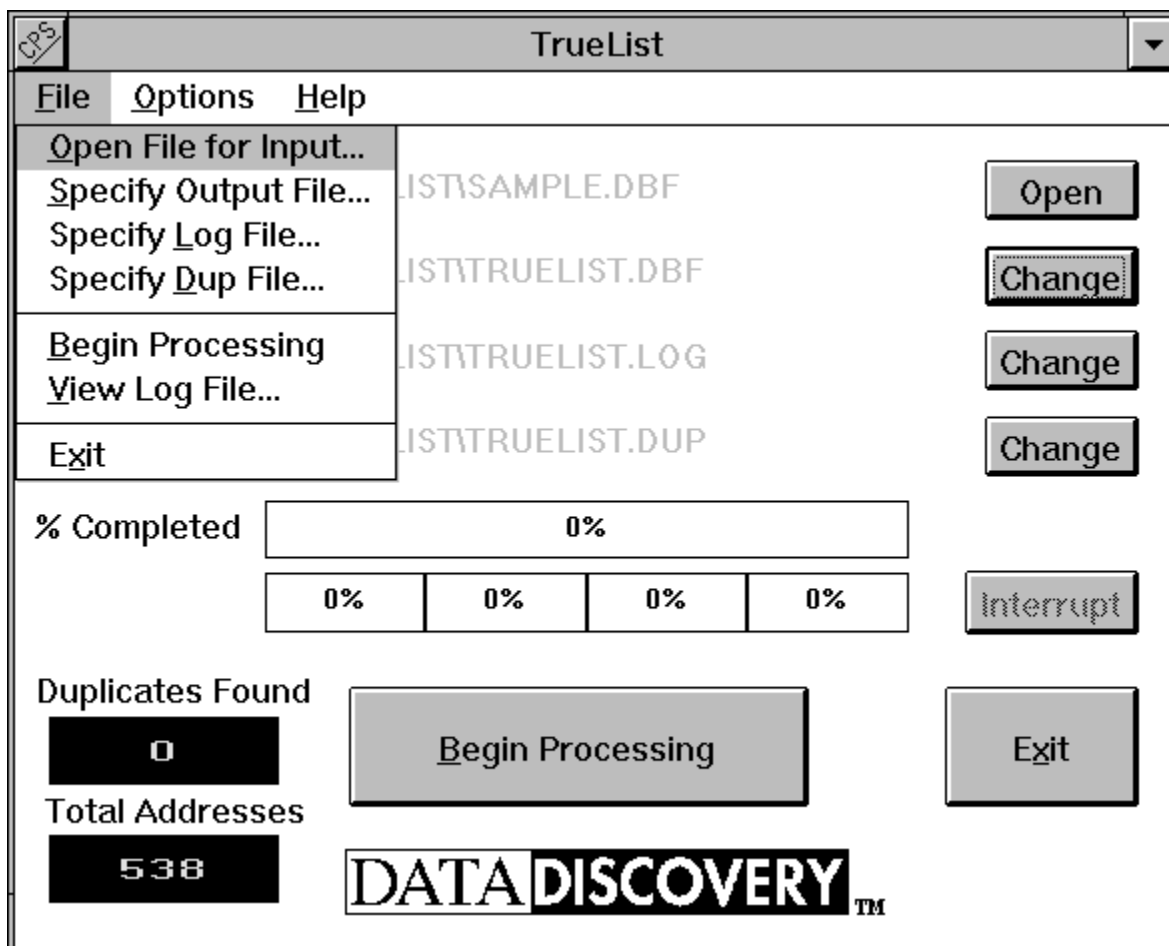
To view the log file using NotePad, click the **Accept** button. The **Browse** button allows you to select another viewing program. Clicking it invokes the following dialog box.



Use this dialog box to select the program with which to view the log file. Once selected, press the **OK** button on this screen and the **Accept** button on the previous screen.

File Menu

The following screen shows the **File** menu.

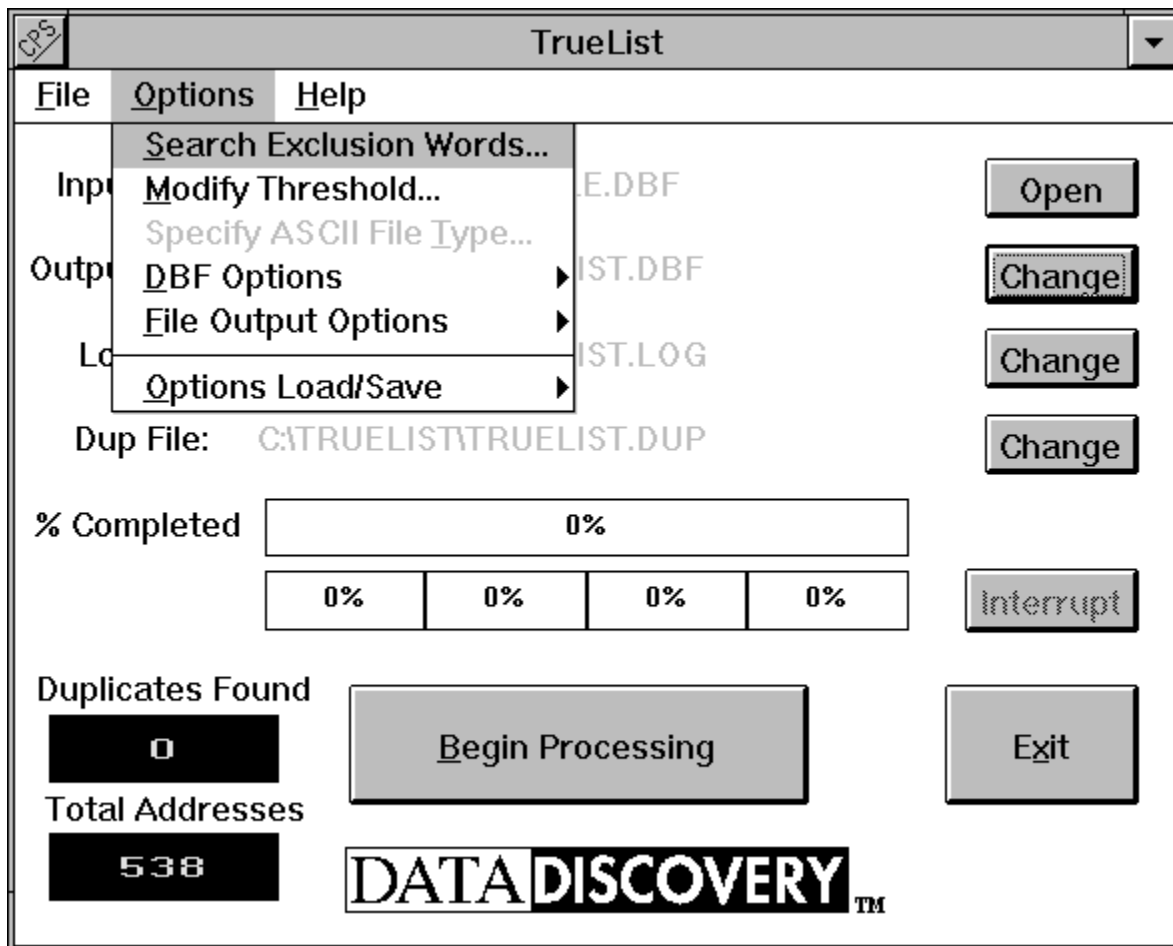


Most of these menu options are alternatives to the buttons on the main *TrueList* screen. They perform exactly the same functions as the buttons. These functions are described in the **Using TrueList** sections above. There is an additional function that allows viewing the log file, which is normally done at the end of processing.

[Using TrueList](#)

Options Menu

The following screen illustrates the **Options** menu .



These options are described in the sections below.

[Exclusion Words](#)

[Threshold](#)

[ASCII File type](#)

[DBF Options](#)

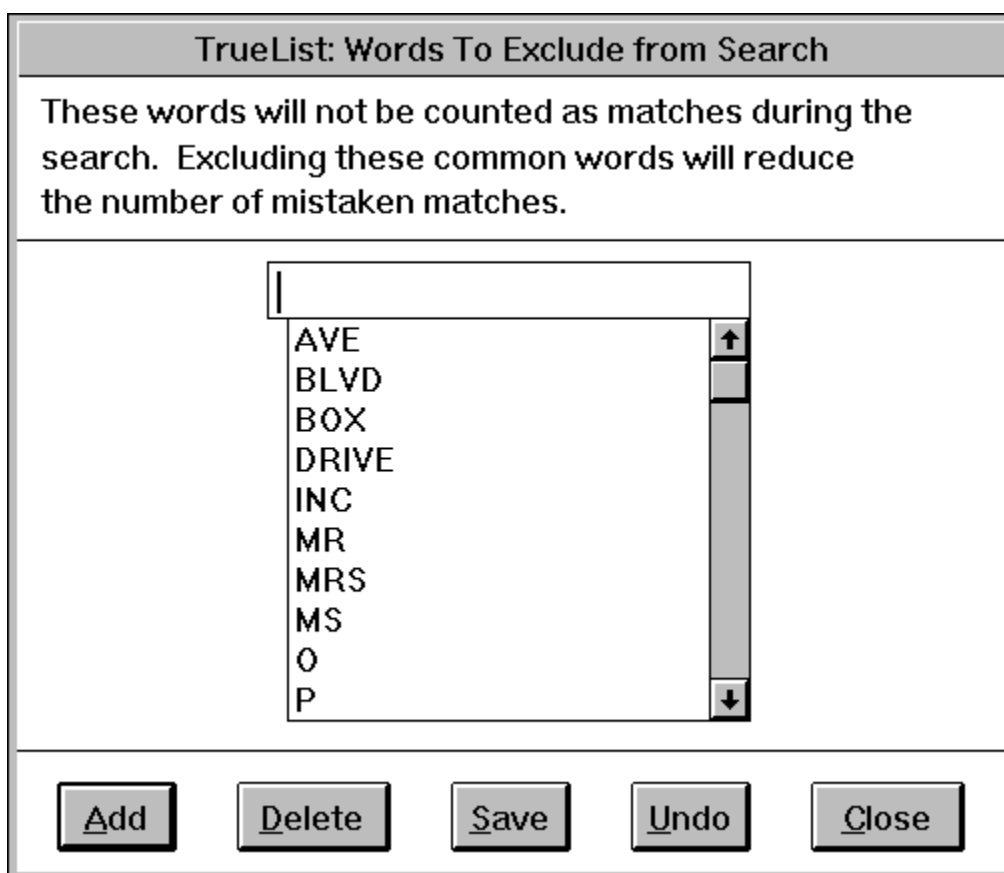
[File Output Options](#)

[Options Load/Save](#)

Exclusion Words

To enhance the accuracy of the matching system, *TrueList* maintains a list of "exclusion words." These are words, such as "street," that appear frequently in addresses. *TrueList* does not consider exclusion words in determining the likelihood of a match. This technique reduces the number of incorrect matches, or "false positives."

TrueList provides a default list of exclusion words. These are the words used most often in mailing addresses. To view this list, click **Options** in the initial *TrueList* screen and select **Search Exclusion Words**. The following box will be displayed.



You can modify the exclusion words list by adding new words or deleting existing words.

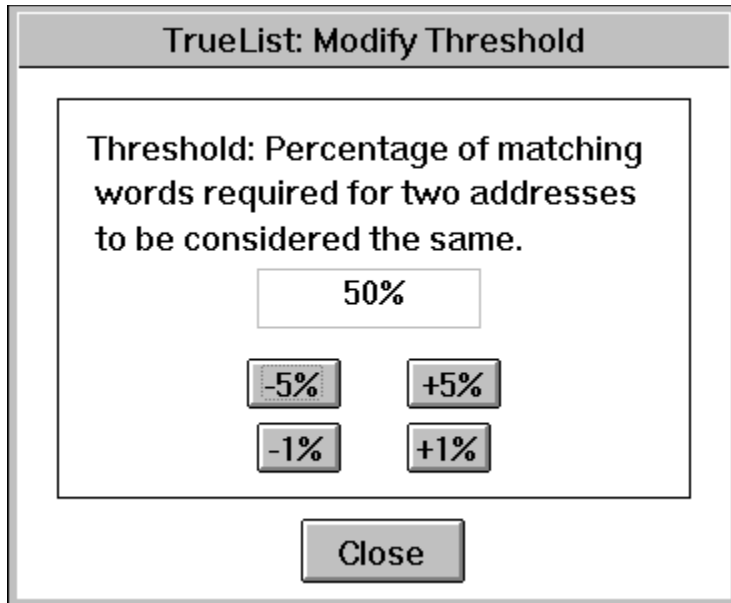
- o To add a word, type it into the edit box (above AVE) and click the **Add** button. The new word will appear below.
- o To delete a word, click that word and it will move to the edit box. Then click the **Delete** button.

- o To reverse your last addition or deletion, click the **Undo** button.
- o To save your changes to disk, click **Save**. The new list will be used in future *TrueList* operations.
- o To exit the **Exclusion Words** window, click the **Close** button. If you have made changes that were not saved, you will have the option of saving the word list at this time.

NOTE: If you would like to use the changes in the current execution of *TrueList*, but do not want to make the changes permanent, click **No**.

Threshold

The "threshold" determines the minimum percentage of matching words required for any two addresses to be classified as duplicates. Selecting **Modify Threshold** from the **Options** menu summons this dialogue box.



The dialog box has a title bar that reads "TrueList: Modify Threshold". Inside, there is a text area with the following text: "Threshold: Percentage of matching words required for two addresses to be considered the same." Below this text is a text input field containing "50%". Underneath the input field are four buttons arranged in two rows: the top row contains "-5%" and "+5%", and the bottom row contains "-1%" and "+1%". At the bottom center of the dialog box is a "Close" button.

To modify the threshold level, click the increase (+5%, +1%) or decrease (5%, 1%) buttons.

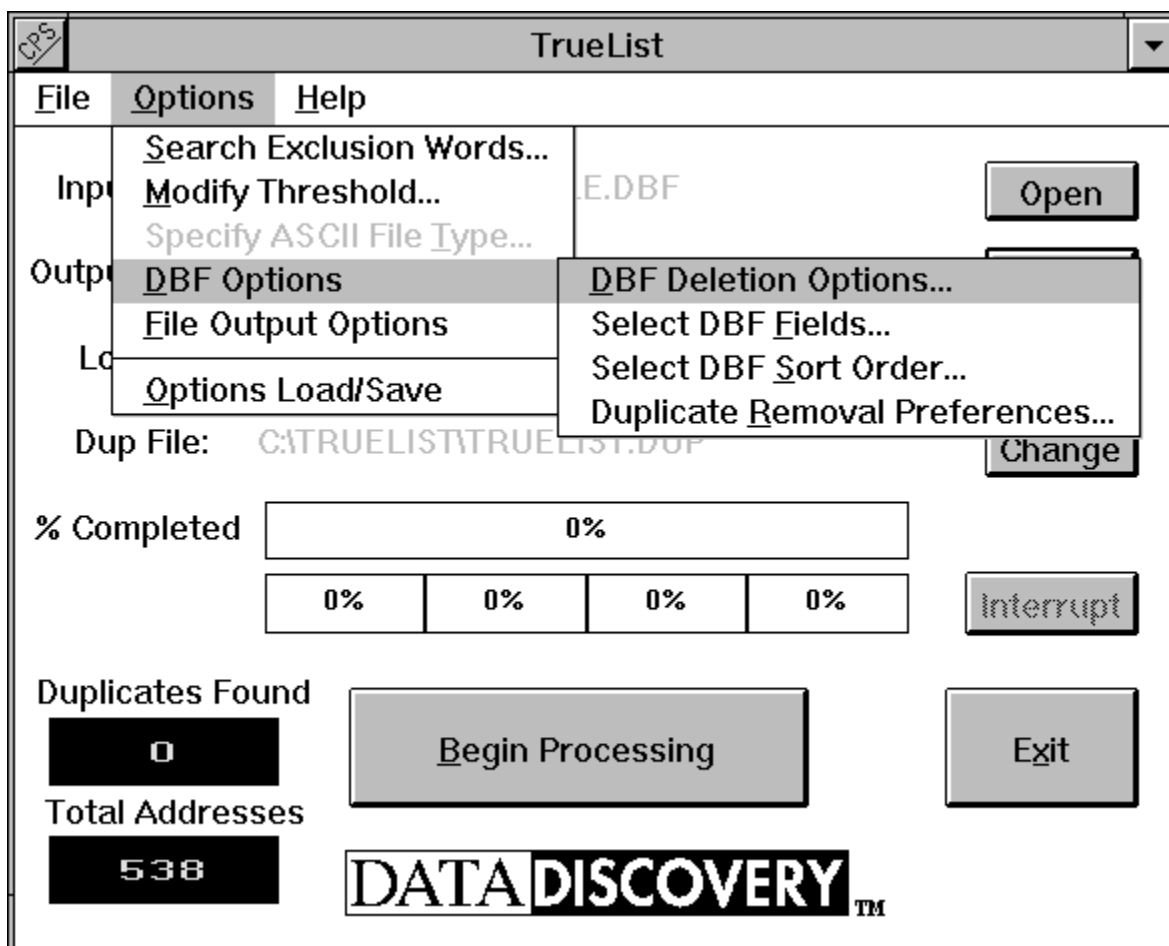
ASCII File Type

This option defines the input data structure for ASCII text files. It is normally set when an ASCII file is opened for input. See the **ASCII Input Options** section above.

ASCII Input Options

DBF Options

The following screen shows the DBF Options sub-menu.



The DBF options are described below.

DBF Deletion Options

Select DBF Fields

Select DBF Sort Order

Duplicate Removal Preferences

DBF Deletion Options

Choosing **DBF Deletion Options** invokes the following dialog box.

DBF Deletion Options

Delete duplicates from DBF file

Removed records will be marked for deletion by setting the DELETE flag in the DBF File. Unless you pack the file, deleted records may still be recovered at a later time.

Allow the use of deleted records

Records marked for deletion will be included in the search for duplicate records.

Undelete Records

When you choose to keep a deleted record, the record's DELETE flag will be cleared.

Close

Each of these options is activated by checking the corresponding box.

If you check **Delete duplicates from the DBF file**, removed records will be identified by marking the delete flag for each deleted record. The records can be physically removed through a database system's pack function.

If you check **Allow the use of deleted records**, comparisons for duplicate addresses will include records that have their delete flag marked.

If you check **Undelete records**, and you later select to keep an address that was previously deleted, its delete flag will be cleared.

Select DBF Fields

The **Select DBF Fields** option provides the same selections as were described in the **dBase Input Options** section above.

dBase Input Options

Select DBF Sort Order

The **Select DBF Sort Order** option provides the same selections as were described in the **dBase Input Options** section above.

dBase Input Options

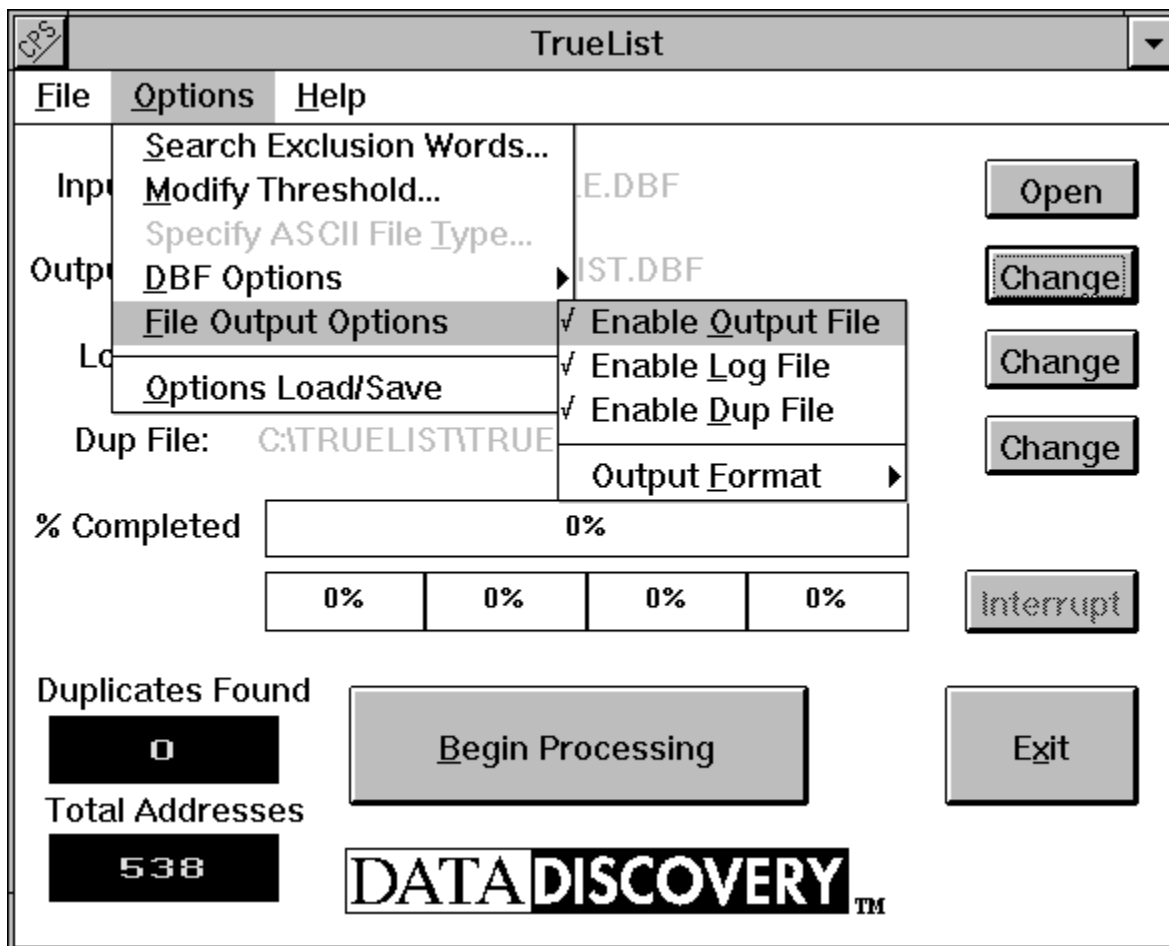
Duplicate Removal Preferences

The **Duplicate Removal Preferences** option allows you to set the processing options shown in the **Dups Found** section above.

Dups Found

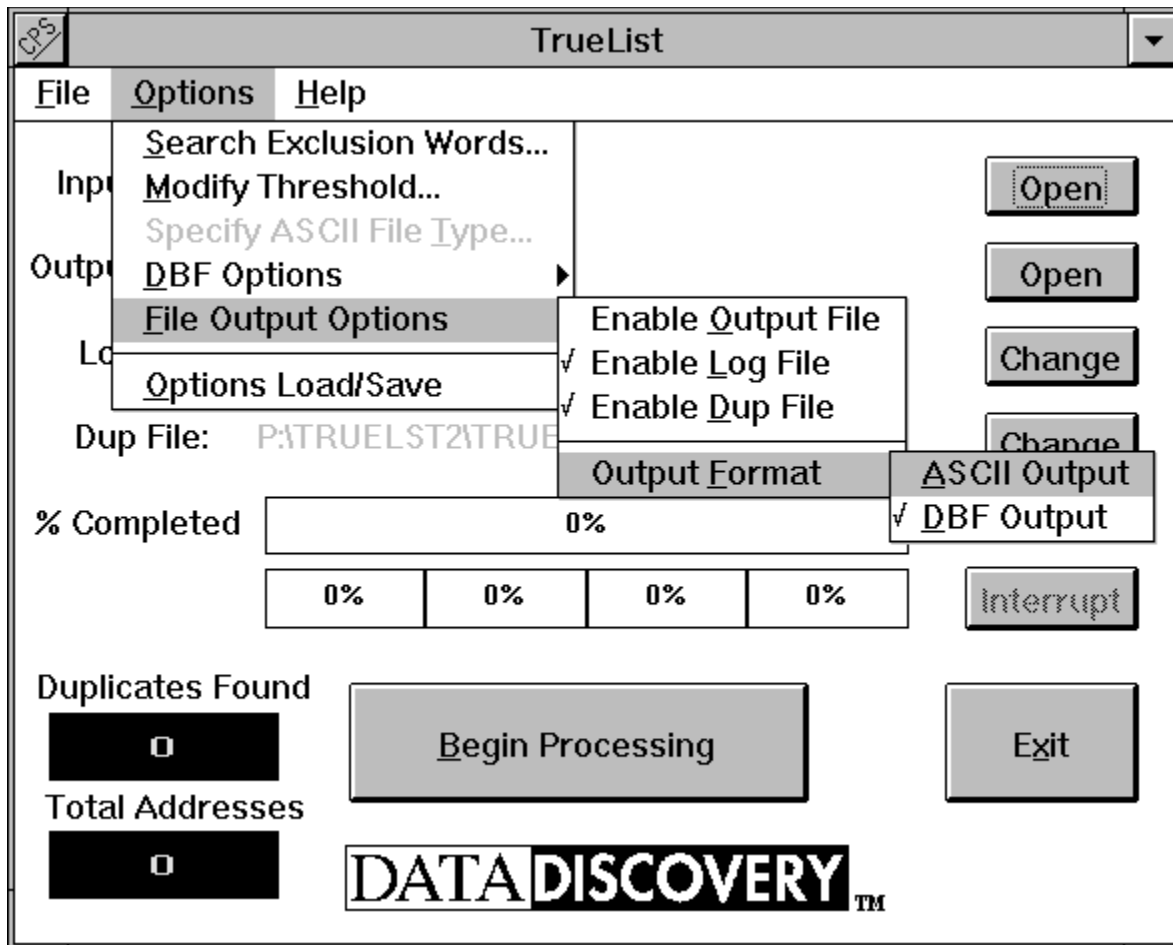
File Output Options

The following screen illustrates the File Output options.



Selecting the first three of these options (with a check mark) enables the output, log and dup files, respectively. Clicking on either option toggles the option on or off.

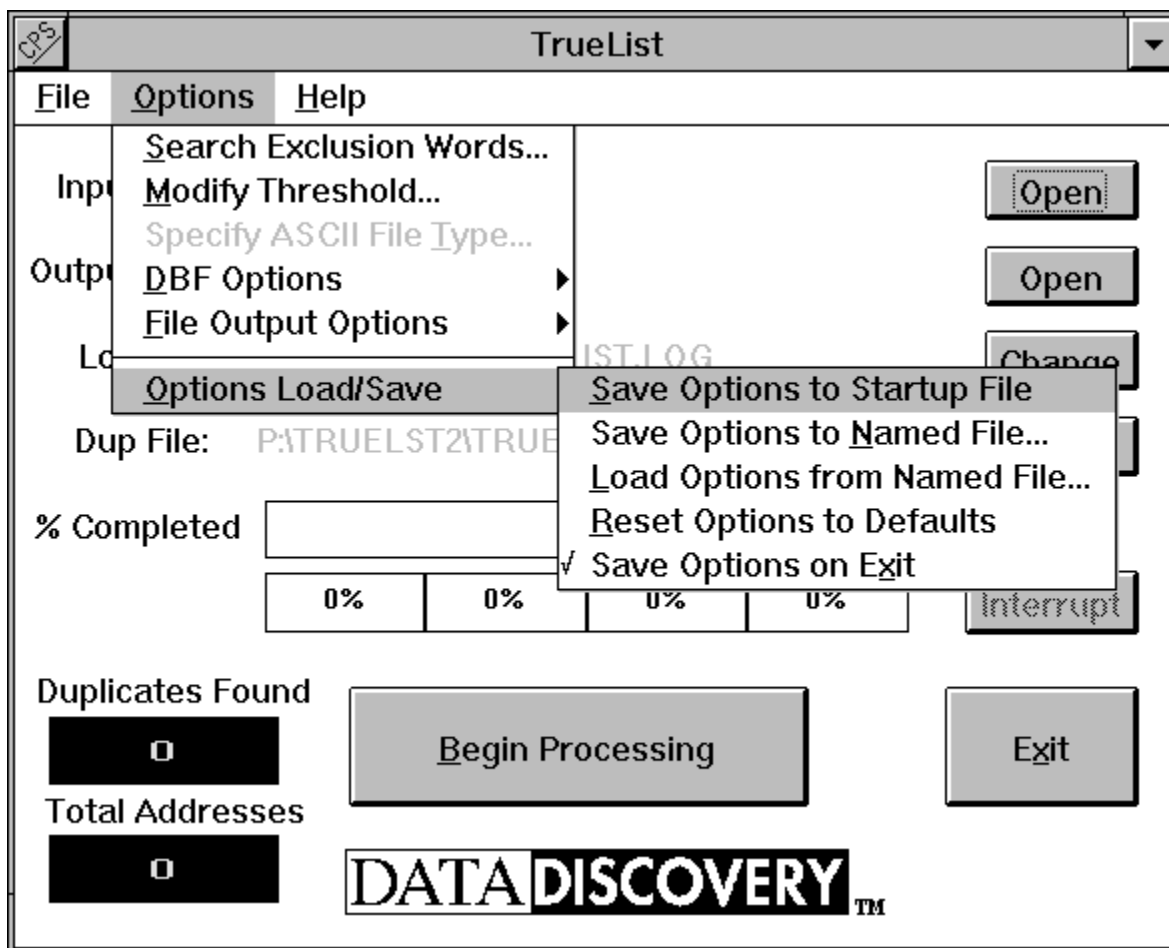
When the **Output Format** option is chosen, the following menu appears.



If the input file is dBase, the user is given the option of creating the output file in either ASCII or dBase format. If the input file is ASCII, the output must also be ASCII.

Options Load/Save

The *TrueList* option values (such as threshold) can be saved and loaded between executions of the program. On start-up, *TrueList* loads the option values from a configuration file, **truelist.cfg**, and uses those values during execution. The following dialog illustrates the ways in which the option values can be saved and loaded.



Selecting **Save Options to Startup File** will save the current settings of the options values to the configuration file so they will be used automatically next time *TrueList* is executed.

If **Save Options to Named File** is chosen, the user is given the ability to save the option values to a specific file. They can later be loaded into *TrueList* by choosing **Load Options from Named File**. Saving option values to named files allows you to maintain sets of options for different situations.

If you wish to set all options back to the "factory settings", you may do so by

selecting **Reset Options to Defaults**. And finally, if **Save Options on Exit** is chosen, *TrueList* will automatically save the options values to the start-up file, thereby continuously maintaining the values.

Log Example

This example shows how information about removed addresses appears in the log file.

TrueList log for file: C:\TRUELIST\JOHNQ.LST

DATE: 03/15/1994 TIME: 13:38

=====

Removed

XYZ CORP.

100 N. MAIN, SUITE 10

ANYWHERE, USA 12345

C/O JONATHAN PUBLIC

Kept

JOHN Q. PUBLIC

XYZ CORPORATION

100 NORTH MAIN STREET

ANYTOWN, USA 12345

54% Match On Words: PUBLIC, XYZ, 100, MAIN, USA, 12345

The log file can be used to analyze the mailing list to determine the best settings for the threshold and other parameters.

ASCII Format

If your mailing list is in dBase (.dbf) format, *TrueList* can work directly on your mailing list. Otherwise, the list will need to be converted to a simple (ASCII) text file for input into *TrueList*. For example, a word processing merge list should be saved as ASCII and page breaks replaced with blank lines.

No special format is required for items within individual addresses. The words (i.e. strings of letters or digits) in the address can be in any order. For example, the person's name can be on the first line, or it can be on the last line. The format of the addresses is only for clarity of display when selecting which address to remove. When creating an ASCII file, you should use spaces between words on the same line, and a carriage return and line feed at the end of each line.

TrueList does, however, need to know how many lines there are in each address. As shown in the **ASCII Input Options** section, there are two options to specifying the number of lines in an address: **Constant number of lines** and **Variable number of lines**.

The number of lines in an address is set initially after the **Begin Processing** button is clicked. It can be reset by using the **Options** menu, as described in the **ASCII Input Options** section.

ASCII Input Options

Constant number of lines

Variable number of lines

Constant Number

Addresses with a constant number of lines have the following criteria:

- o All addresses contain the same number of lines
- o The number must include any separation lines between addresses

For the addresses below, the number of address lines should be set to 4 (three content lines plus one for the blank line between addresses).

John Q. Public
100 North Main Street
Anytown USA 12345

Susan B. Anthony
1776 Independence Ave
Freedom USA 56789

Variable Number

Addresses with a variable number of lines have the following criteria:

- o Addresses can contain a different number of lines
- o Addresses must be separated by one or more blank lines

When there are a variable number of lines in addresses, a blank line indicates the end of one address and the beginning of the next address, as shown below.

John Q. Public
XYZ Corp.
100 North Main Street
Anytown USA 12345

Susan B. Anthony
1776 Independence Ave
Freedom USA 56789

NOTE: If there are no blank lines to separate addresses when using a variable number of lines, *TrueList* will not be able to find the end of the addresses. Thus, it will view the entire file as one large address.

On the other hand, if blank lines appear in the middle of an address, *TrueList* will inadvertently conclude that the text between those blank lines are complete addresses.

