ClusterSuite Tutorial

May 2014

Contact Information

- Created by:
  - JJ O'Brien
  - Dr. Stephen Carley (stephen.carley@gmail.com)
  - Dr. Alan Porter (alan.porter@isye.gatech.edu)

- Some of the individual macros and thesauri were not created by the authors of ClusterSuite. Questions pertaining to these macros may take longer to address.

VantagePoint

- ClusterSuite runs as a script within VantagePoint software.
- VantagePoint is a commercial software product designed to be a "powerful text-mining tool for discovering knowledge in search results from patent and literature databases"
- For more information, visit www.thevantagepoint.com

Purpose

- Term clumping macros…
  - Perform dimension reduction on a list, making it more approachable and enabling the user to see the forest for the trees.
  - Minimize noise and maximize prominent topics, making a list easier to work with and enabling the user to (more quickly) extract meaning from large amounts of text.

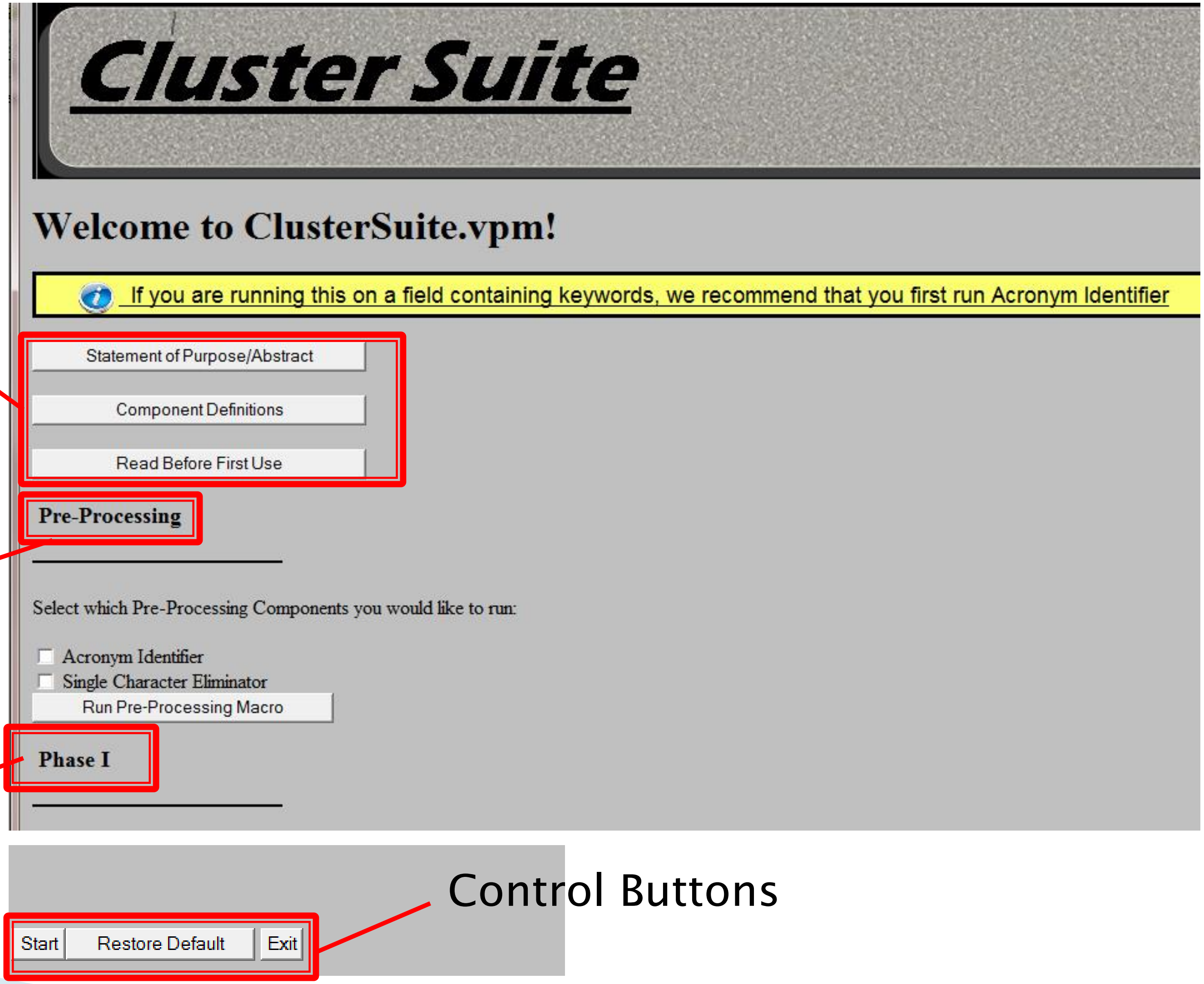
Contents

- A complete ClusterSuite package should contain the following components:
  - ClusterSuite (Folder)
    - Macros (Folder)
      - Resource (Folder)
        - About.html
        - Acronym_Identiifier.xlsm
        - Checklist.html
        - Clustersuite.jpg
        - Info.png
        - finalScreen.html
        - Matrix_to_Columns.xlsm
        - openFile.html
        - Single_Term_Remover.xlsm
        - Tutorial.html
        - Abstract.html
        - Cite.html
      - Acronym_Identifier.vpm
      - Acronym_Identifier_Part2.vpm
      - Clean_List_Iteratively.the
      - CombineAuthorNetworks.the
      - Matrix_to_Columns.vpm
      - Single_Term_Remover_1.vpm
      - Single_Term_Remove_2.vpm
      - Term Clustering-Cherie.vpm
    - termCluster(Folder)
      - termCluster.jar
      - runTermCluster.xlsm
      - runTermCluster.bat
      - Resource(Folder)
    - Thesaurus(Folder)
      - Acronym_Thesaurus.the
      - Chemical_Compounds.the
      - Common_and_Basic.the
      - NumPunctToSpace.the
      - Scientific_and_Academic.the
      - Single_Terms.the
      - XMLencoding.the
  - Fuzzy (Folder)
    - general-85cutoff-95fuzzywordmatch-1 exact.fuz
  - Metadata (Folder)
    - Documentation components (various formats)
  - TermCluster-Source
    - Java source files
  - ClusterSuite.vpm
  - createEverything.sql
  - Readme.txt

Installation

- Welcome to ClusterSuite.vpm Version 1.0!
- STEP 1: Extract the inner ClusterSuite folder (ClusterSuite\ClusterSuite) to your root VantagePoint folder (usually something like C:\\Program Files\VantagePoint)
- STEP 2: Extract ClusterSuite.vpm to VantagePoint\Macros
- STEP 3: Store contents of the Fuzzy folder(general-85cutoff-95fuzzywordmatch-1 exact.fuz) in VantagePoint\Fuzzy
- STEP 4: Launch VantagePoint and Run ClusterSuite.vpm through Scripts->Run Script
  - Optional: To Add ClusterSuite.vpm to your Scripts menu for quick reference, go to Scripts->Modify Script Menu->Add Script and select ClusterSuite.vpm from your Macros folder

Welcome Screen



Welcome to ClusterSuite.vpm!

If you are running this on a field containing keywords, we recommend that you first run Acronym Identifier

Select which Pre-Processing Components you would like to run:

□ Acronym Identifier
□ Single Character Eliminator

Meta-Data

Pre-Processing Macros

Phases

Control Buttons

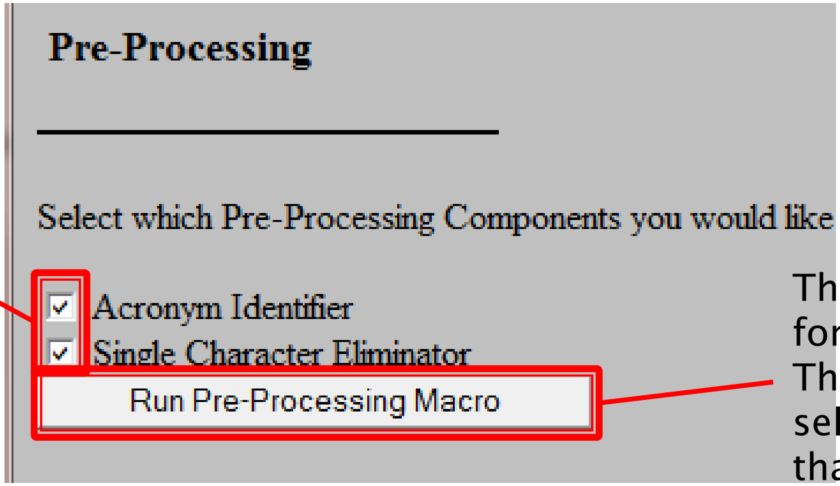
Meta-Data

- Statement of Purpose/Abstract: Self-explanatory
- Component Definitions:
  - Defines each of the scripts, thesauri, and other applications that can be run through ClusterSuite.
  - Grouped according to phase
  - **Read this before deciding which components to run!**
- Read Before First Use:
  - The macro will run more slowly on larger datasets. Some datasets may run very slowly (hours or even days), and some datasets may not run at all. This generally depends on the size of your data and your computer speed.
  - Depending on your data, a TFIDF matrix may be useful. Enter "Ctrl + M" after ClusterSuite runs to create one
  - Combine Author Networks is a very aggressive clumping macro. Cherie's Macro is more precise, but it may cause difficulties on large recordsets

Pre-Processing

- Acronym Identifier: Creates a Thesaurus, which converts acronyms to their full word form when run on a VantagePoint Field
- Single Character Eliminator: This component uses MS Excel to remove items in the VantagePoint field that contain only one character. For example, if a field consisted of three items, "a", "the", and "4". Items "a" and "4" would be removed from the field following the execution of Single Character Eliminator.

Select which components you would like to run



Pre-Processing

Select which Pre-Processing Components you would like

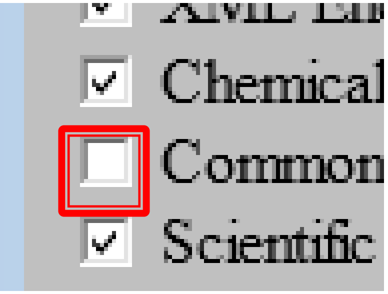
☑ Acronym Identifier

☑ Single Character Eliminator

This button will prompt you for a VantagePoint field. Then, it will run the selected components on that field.

ClusterSuite Runtime Components

- The following slides outline the ClusterSuite components that may be run after the “Start” button is pressed
- If any of these components are undesired, or you notice that your recordset is too dramatically reduced by any given component, simply uncheck that component’s box before pressing “Start”
- Re-arranging components is not supported by ClusterSuite at this time. If you would like to run ClusterSuite out of order, you will have to simulate the desired order by checking and unchecking boxes and running ClusterSuite multiple times.



☑ XML En

☑ Chemical

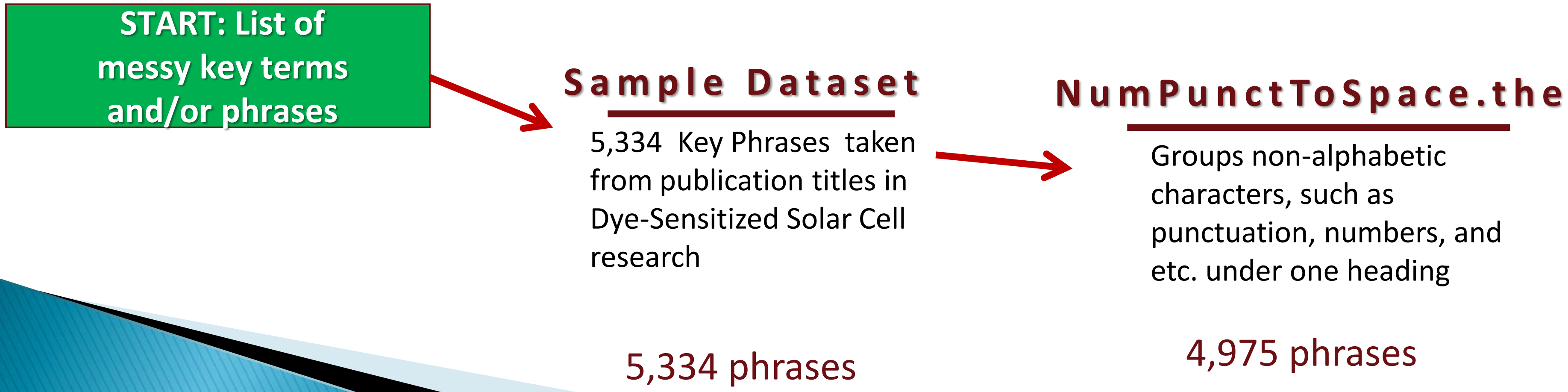
□ Common

☑ Scientific

Phase I: NumPunctToSpace

- NumPunctToSpace: Removes keywords beginning with non-alpha numeric characters

Sample Trial



START: List of messy key terms and/or phrases

Sample Dataset

5,334 Key Phrases taken from publication titles in Dye-Sensitized Solar Cell research

5,334 phrases

NumPunctToSpace.the

Groups non-alphabetic characters, such as punctuation, numbers, and etc. under one heading

4,975 phrases

Phase I: XML Encoding

- XML Encoding: Removes XML tags

Sample Trial

NumPunctToSpace.the

Groups non-alphabetic characters, such as punctuation, numbers, and etc. under one heading

4,975 phrases

→

XMLEncoding.the

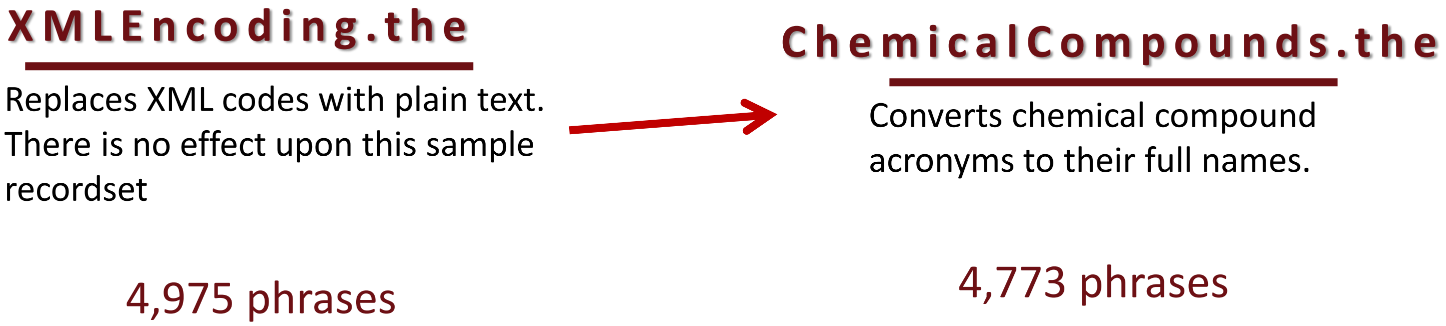
Replaces XML codes with plain text. There is no effect upon this sample recordset

4,975 phrases

Phase I: Chemical Compounds

- Chemical Compounds: Consolidates chemical compounds and their abbreviations

Sample Trial



XMLEncoding.the

Replaces XML codes with plain text. There is no effect upon this sample recordset

4,975 phrases

ChemicalCompounds.the

Converts chemical compound acronyms to their full names.

4,773 phrases

Phase I: Common and Basic

- Common and Basic: Consolidates multiple common keywords into a single header

Sample Trial

ChemicalCompounds.the

Converts chemical compound acronyms to their full names.

4,773 phrases

→

Common and Basic.the

Groups "stopwords" and other basic, simple, and common words under a single heading

3,236 phrases

Phase I: Scientific and Academic

- Scientific and Academic: Consolidates multiple keywords common to scientific and academic publications into one header

Sample Trial

Common and Basic.the

Groups "stowords" and other basic, simple, and common words under a single heading

3,236 phrases

Scientific and Academic.the

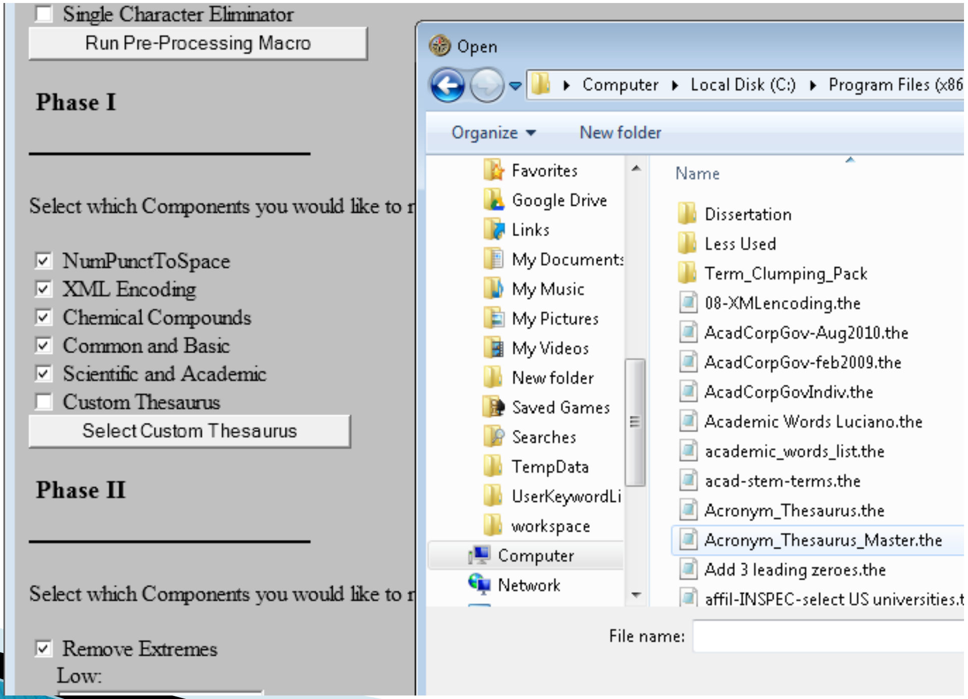
Groups over 3,000 common scientific and academic terms into a single heading

3,080 phrases

Phase I: Custom Thesaurus

The "Custom Thesaurus" option allows the user to run a thesaurus not included in the five default ClusterSuite thesauri.

To use a custom thesaurus, check the "Custom Thesaurus" checkbox. Then, use the "Select Custom Thesaurus" button to navigate to the location of your thesaurus on your computer. The thesaurus will now run at the end of Phase I.



□ Single Character Eliminator

Phase I

Select which Components you would like to r

☑ NumPunctToSpace
☑ XML Encoding
☑ Chemical Compounds
☑ Common and Basic
☑ Scientific and Academic
□ Custom Thesaurus

Phase II

Select which Components you would like to r

☑ Remove Extremes
Low:

Phase II: Remove Extremes

This macro is useful in trimming off outlier terms with unusually high or low record counts.

Both the Low and High Extremes options may be run with either "Raw Numbers" or "Percentages“.

- "Raw Numbers": Use this option if you would like to trim off items that contain at least as many records as your input under "High" or as few records as your input under "Low“.
- "Percentage": Use this option if you would like to remove all items that contain a set percent of the total records in the field. For example, imagine you had four items in a field with 10 total records: "A (5 records)“, "B (2 records)", "C (2 records)", and "D (1 record)". If you wanted to remove "A" and "D", you would enter 50 under "High" and 10 under "Low". Select the "%" radio button for each. This will leave you with a field containing only "B" and "C" because "A” is greater than or equal to 50% and "D" is less than or equal to 10%

On its default settings, the macro removes all items that constitute 80% or more of the total records in a field for Remove High Extremes. However, Remove Low Extremes is turned off by default in this spot. If you are using data that is particularly large and not easily processed by Combine Lists Iteratively, you may want to use Remove Low Extremes to make your data more manageable

Sample Trial

Scientific and Academic.the

Groups over 3,000 common scientific and academic terms into a single heading →

3,080 phrases

Remove Extremes

- These two macros are useful in removing items with unusually high or low record counts.
- Useful in removing clumped term headers, such as “Common and Basic”

Remove High

1,054 phrases →

Remove Low

1,049 phrases

Phase III: Clean List Iteratively

- This macro runs the (aggressive) general.fuz list cleanup algorithm on a given field until it cannot be run anymore. This script is intended to be run on keywords fields. Fields, such as Countries, will not yield meaningful results. This macro does not need to be applied to fields that contain too few list items. For example, a field with only 10 list items will likely only need to run one iteration of the .fuz file. In these cases directly running a single instance of general–85cutoff–95fuzzywordmatch–1 exact.fuz should suffice.
    - Running this macro on a field with a particularly large number of list items is not recommended, as it may take the macro up to several days to finish running on some large fields.

Sample Trial

Remove Extremes

- These two macros are useful in removing items with unusually high or low record counts.
- Useful in removing clumped term headers, such as “Common and Basic”

1,049 phrases

→

Combine Lists Iteratively

Iteratively runs a fuzzy-search phrase- cleaning script to clump similar keywords together

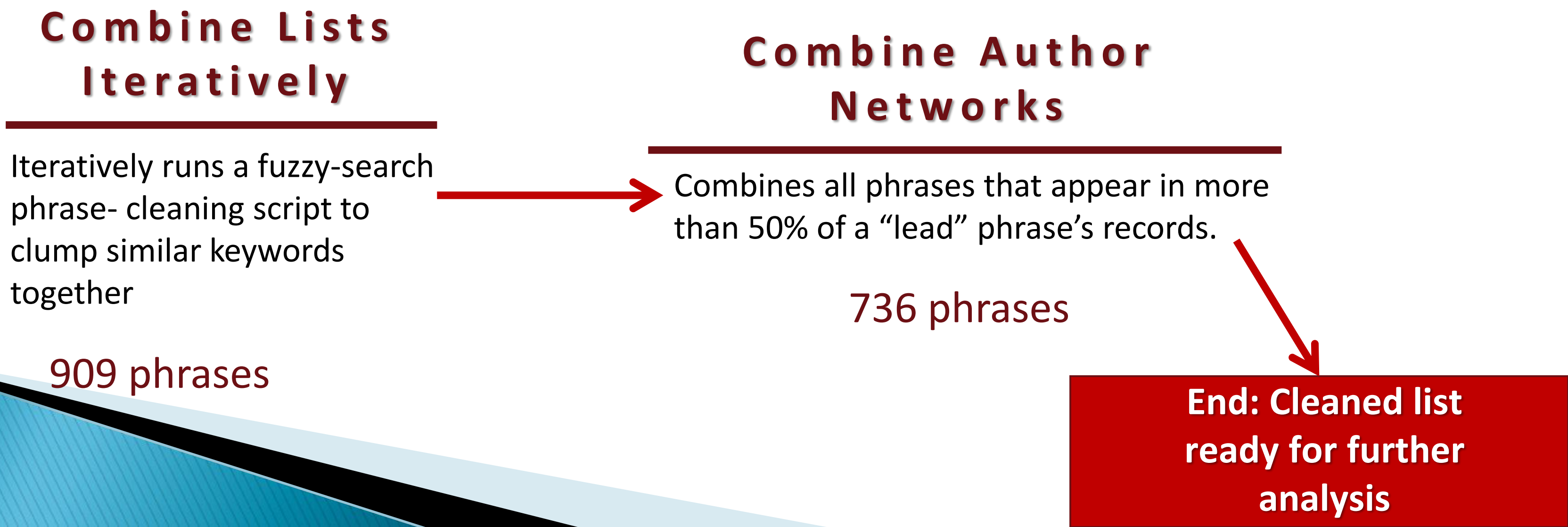
909 phrases

Phase III: Combine Author Network

The original version of this macro was designed to create and run a thesaurus to combine all authors who appear in more than 50% of a lead author's publication. For our purposes, this macro would be better titled "Combine Term Networks". We use it to clump terms that are found with 50% or more occurrences of the "lead term".

NOTE: This macro has been known to produce dramatic drops in item-count.
For best results, "Remove Extremes" before running.
There is now also a Remove Low Extremes Option at this stage (not shown in example)

Sample Trial



Combine Lists Iteratively

Iteratively runs a fuzzy-search phrase- cleaning script to clump similar keywords together

909 phrases

Combine Author Networks

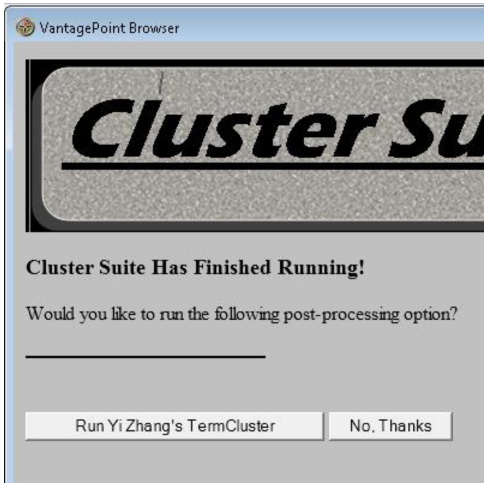
Combines all phrases that appear in more than 50% of a “lead” phrase’s records.

736 phrases

End: Cleaned list ready for further analysis

Yi Zhang's Term Cluster

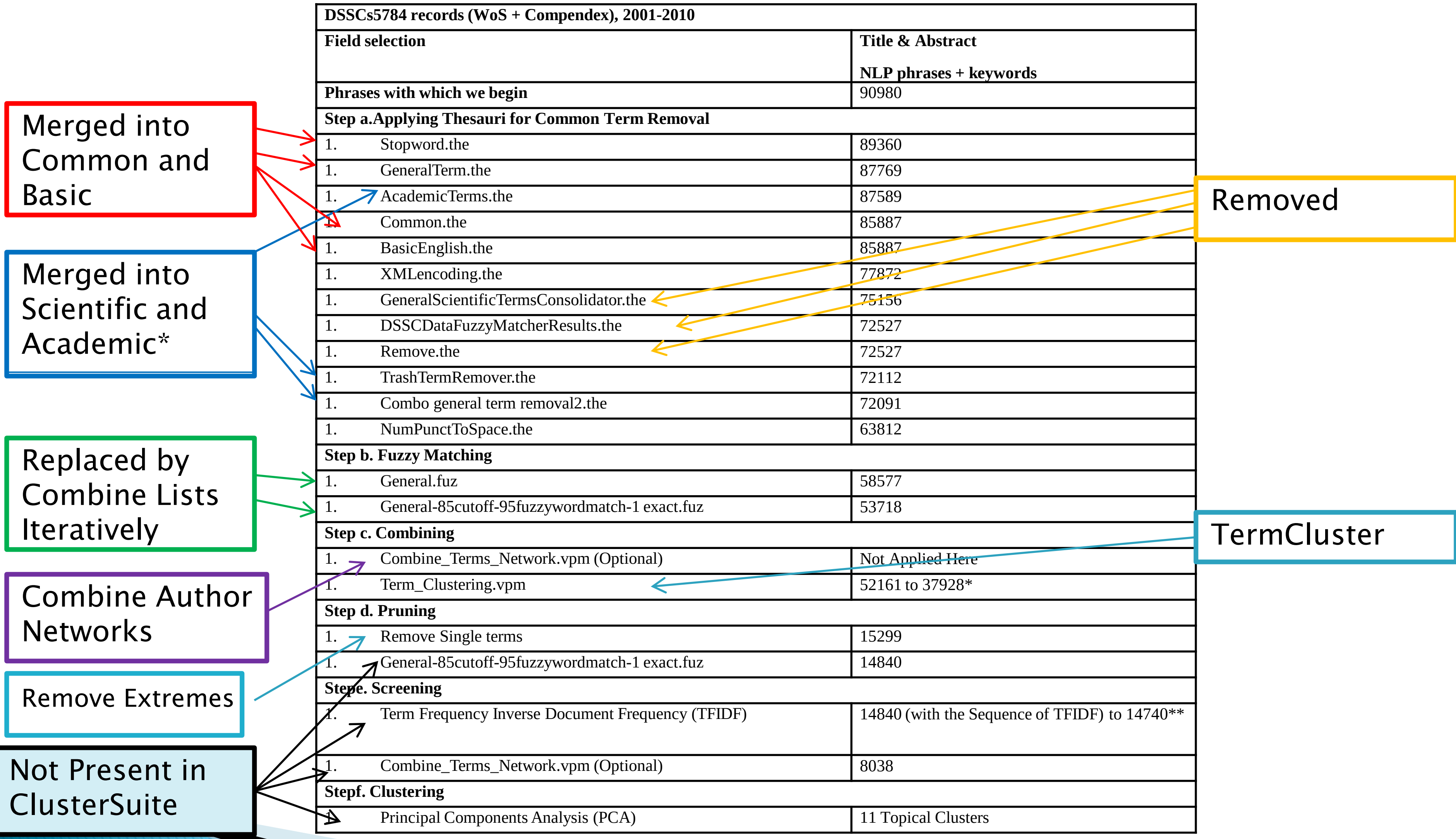
- TermCluster is used here to group (cluster) the items produced in the final, cleaned ClusterSuite output.
- After ClusterSuite finishes running, the user will be prompted to launch TermCluster.jar
- TermCluster requires a MySQL server to run. Refer to supplementary documentation for more information



Cluster Suite Has Finished Running!

Would you like to run the following post-processing option?

ClusterSuite compared to Term Clumping steps outlined in paper by Zhang, Porter, et al. (2012)[1]



DSSCs5784 records (WoS + Compendex), 2001-2010	
Field selection	**Title & Abstract** **NLP phrases + keywords**
Phrases with which we begin	90980
Step a.Applying Thesauri for Common Term Removal	
1. Stopword.the	89360
1. GeneralTerm.the	87769
1. AcademicTerms.the	87589
1. Common.the	85887
1. BasicEnglish.the	85887
1. XMLencoding.the	77872
1. GeneralScientificTermsConsolidator.the	75156
1. DSSCDataFuzzyMatcherResults.the	72527
1. Remove.the	72527
1. TrashTermRemover.the	72112
1. Combo general term removal2.the	72091
1. NumPunctToSpace.the	63812
Step b. Fuzzy Matching	
1. General.fuz	58577
1. General-85cutoff-95fuzzywordmatch-1 exact.fuz	53718
Step c. Combining	
1. Combine_Terms_Network.vpm (Optional)	Not Applied Here
1. Term_Clustering.vpm	52161 to 37928*
Step d. Pruning	
1. Remove Single terms	15299
1. General-85cutoff-95fuzzywordmatch-1 exact.fuz	14840
Stepe. Screening	
1. Term Frequency Inverse Document Frequency (TFIDF)	14840 (with the Sequence of TFIDF) to 14740**
1. Combine_Terms_Network.vpm (Optional)	8038
Stepf. Clustering	
1. Principal Components Analysis (PCA)	11 Topical Clusters

Merged into Common and Basic

Merged into Scientific and Academic*

Replaced by Combine Lists Iteratively

Combine Author Networks

Remove Extremes

Not Present in ClusterSuite

Removed

TermCluster

*Also contains an additional thesaurus not listed here

[1] Zhang Yi, Alan L. Porter, Zhengyin Hu, Ying Guo, and Nils C. Newman. (2012). "Term Clumping" for Technical Intelligence: A Case Study on Dye-Sensitized Solar Cells. *The International Conference On Innovation Management And Policy.*