An Evaluation of Unstructured Text Mining Software

Micah J. Crowsey, Amanda R. Ramstad, David H. Gutierrez, Gregory W. Paladino, and K. P. White, Jr., *Member, IEEE*

Abstract— Five text mining software tools were evaluated by four undergraduate students inexperienced in the text mining field. The software was run on the Microsoft Windows XP operating system, and employed a variety of techniques to mine unstructured text for information. The considerations used to evaluate the software included cost, ease of learning, functionality, ease of use and effectiveness. Hands on mining of text files also led us to more informative conclusions of the software. Through our evaluation we found that two software products (SAS and SPSS) had qualities that made them more desirable than the others.

INTRODUCTION

Unstructured data exists in two main categories: bitmap objects and textual objects. Bitmap objects are non-language based (e.g. image, audio, or video files) whereas textual objects are "based on written or printed language" and predominantly include text documents [1]. Text mining is the discovery of previously unknown information or concepts from text files by automatically extracting information from several written resources using computer software [15]. In text mining, the files mined are text files which can be in one of two forms. Unstructured text is usually in the form of summaries and user reviews whereas structured text consists of text that is organized usually within spreadsheets. This evaluation focused specifically on mining unstructured text files.

Many industries are relying on the field of text mining to solve specific applications using a variety of software. Currently, there does not exist a comprehensive review or comparison of the top software suites. Our research was performed so users would have an unbiased reference for looking at the different software features and the pros, cons, and methods for how to most efficiently use them. In comparing the software, we looked at the features that the software had as well as the ease of use and learning of the software. We used a test corpus that we developed for this purpose.

The five software suites we reviewed were Leximancer, SAS Enterprise Miner, several products of SPSS, Polyanalyst, and Clarabridge. Most of the software was acquired through an anonymous consulting firm. We also obtained software through the University of Virginia and by requesting online software demos.

The evaluators of the software were four fourth year Systems Engineering students at the University of Virginia who had some previous experience with data mining, but little experience with text mining. This inexperience proved useful in determining the ease of use and learning of the software, though it sometimes posed challenges when attempting to find out how the software operates.

I. TEXT MINING SOFTWARE

A. General Information and Pricing

The chart below shows the company, product, version, and cost of the software. Unfortunately, certain vendors were unwilling to disclose cost information for their products.

Table 1

Company	Product	Version	Cost
SAS	Enterprise Miner	4.3	not provided due to company policy
SPSS	Text mining for Clementine(need Clementine) Text Mining Builder Lexiquest Categorize	10.1 2.0 3.2	$4,534 $18,136 Server (1 CPU) $69,824
Megaputer	Polyanalyst	6.0	Professional Server $80,000 Client $5,000
Leximancer	Leximancer	pro	$2500
Clarabridge	Clarabridge Content Mining Platform	2.1	$75,000

B. Learnability

The criteria used to assess the ease of learning were based on whether software had the following:

- A demo version
- Tutorials
- User's manual

Manuscript received April 5, 2007. This work was supported in part by the University of Virginia under a grant provided by an anonymous consultancy.

M. J. Crowsey is with the University of Virginia, Charlottesville, VA, 22902. (phone: 434-924-5393. e-mail: mjc5s@virginia.edu).

A. R. Ramstad is with the University of Virginia, Charlottesville, VA, 22902. (e-mail: arr9p@virginia.edu).

D. H. Gutierrez is with the University of Virginia, Charlottesville, VA, 22902. (e-mail: dhg9u@virginia.edu)

G. W. Paladino is with the University of Virginia, Charlottesville, VA, 22902. (e-mail: gwp2z@virginia.edu)

K. P. White is with the University of Virginia, Charlottesville, VA, 22902. (e-mail: kpw8h@virginia.edu)

- Sample Solutions
- Online help

The functionality results of the software were recorded using a spots and dots methodology. The ease of learning results can be seen in Table 2.

The project team attempted to learn how to use each of the software suites solely by referencing the help files, tutorials, and other learning tools that were provided along with the software package. These materials provided information on the general process each software suite uses to mine text and on the specific features offered at different steps in the text mining process.

Table 2: Learnability

Software				
	Demo version	Tutorial	User's manual	Online help
Clarabridge		X	X	
SAS		X		X
Clementine	X			X
Polyanalyst	X	X	X	X
Leximancer	X	X	X	X

As we were unable to obtain working copies of Clarabridge and Polyanalyst, we were unable to gain experience using these software suites. The evaluation of this software was done by looking at product features, going through live internet demonstrations, and performing research on the software companies' websites.

Overall, the help documentation which accompanied the SAS, SPSS, and Leximancer software was sufficient to learn basic processes and features employed by each suite. SAS and SPSS both offer traditional help files which serve as a good starting point for learning the process that each software suite uses to mine text. These resources provide both an introduction to text mining as well as the process flows that each of the software suites use to accomplish different text mining functions such as text extraction, text link analysis, and categorization. At a more detailed level, the help documentation of both software suites provide information on how a user can manipulate various features at different nodes in the text mining process in order to affect the results yielded. Finally, the documentation also provides information on how to view, edit, and make use of results.

SPSS help documentation presents basic information which provides as user with a quick start to text mining. Example projects also are provided to show how different nodes can work together to accomplish various text mining functions.

SAS documentation presents information on the text mining process and its unique methods and features in much more detail. Step-by-step examples of how to achieve a few text mining functions are also very helpful.

Leximancer presents its help documentation in a slightly different fashion than SAS and SPSS. Leximancer provides several help topics describing how to accomplish certain text mining functions within its architecture. Leximancer's text mining process consists of a preset stream of nodes which a user cannot alter, and information on how a user can manipulate the features of these nodes to achieve different results resides within the nodes themselves.

The fact that Leximancer presents information on the features offered by different nodes in the text mining process within the nodes themselves makes for quick referencing, and the content is very helpful in general. Leximancer also allows the user to adjust the level of complexity of the features it offers.

Although help documentation and tutorials are sufficient for the beginning of the learning curve with these software suites, the group found that more advanced knowledge of how to get the most out of each product is best achieved through an iterative process of hands on experimentation. Manipulating the unique features of each software suite in various ways provides the best knowledge of how to achieve desired results. Also, experimenting with a variety of nodes in sequence allows a user to see how these nodes can interact to achieve more advanced text mining functions.

C. Data Preparation

The importance of data preparation in text mining cannot be stressed enough. Given the importance of proper data preparation to the success of a data mining effort, it is advised that the user perform some "cleaning" on the data to put it into a semi-structured form. Although text mining seeks to find relationships between concepts in unstructured data, we found through our evaluation that mining technology does not eliminate the need for data preparation. If the user wishes to achieve a high level of reliability and extract useful concepts from the data, then structuring the data, even in small ways, is helpful.

To achieve useful information for our evaluation, we ran text files through the software in order to see how they were processed. We also looked at the quality of the results after the mining was completed. For this task, we used HTML documents that were gathered from the University of Virginia *Cavalier Daily* newspaper website, www.cavalierdaily.com. A software program that copies entire websites was used to gather approximately 1200 HTML pages from the website, and these formed the corpus that we ran through the software.

D. Software Pros and Cons

Leximancer is a software suite that focuses on extracting concepts and showing the relationships between those concepts, along with the relationship strength. Although its Java-based user interface is somewhat different from the other software suites evaluated, Leximancer still offers many of the same features that allow a user to manipulate stages in the text mining process. Leximancer's results browser is very effective at presenting several pieces of

information at once and allowing a user to browse extracted concepts and concept relationships. Leximancer's results browser is shown in Figure 1 below.

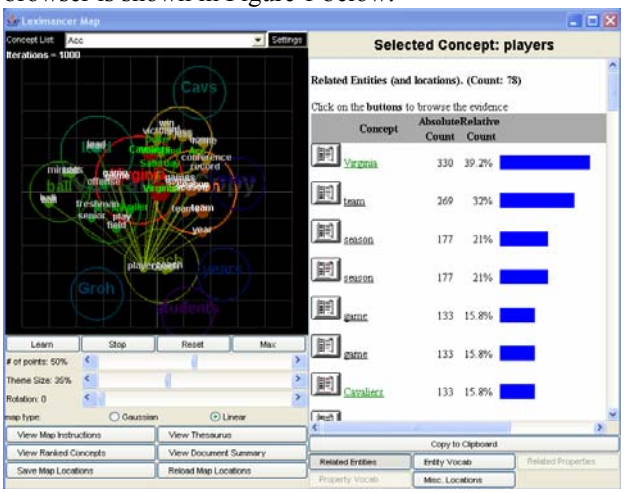

Figure 1: Leximancer concept map and list of linked terms

SAS Enterprise miner uses a unique mining approach called SEMMA (Sampling, Exploration, Modification, Modeling, and Assessment). This package offers features that compare the results of the different types of modeling through statistical analysis and in business terms. The integrated environment allows for the statistical modeling group, business managers and the information technology department to work together. Although SAS supports loading text in a variety of file formats, all externally stored text files must first be converted to a SAS data set via the use of a prewritten macro, adding additional complexity and time consumption to this step in the text mining process.

SAS' user interface is less intuitive than other software, but it still offers many of the same features as other products which affect how text is parsed. In addition to offering these basic features, SAS also offers a user the ability to affect how its algorithm is run which represents parsed documents in a structured, quantitative form. SAS' term extraction generally yields a larger number of terms than other software, but its results browser allows for easy browsing and filtering of these terms. SAS also simplifies the text mining process somewhat by automatically clustering documents and identifying links between terms when a term extraction is executed. Figure 2 below shows SAS' text mining browser for documents, extracted terms, and clusters.

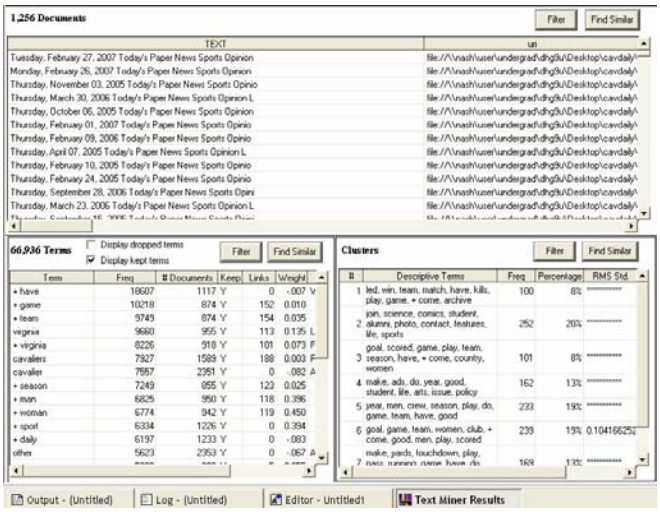

Figure 2: SAS text mining browser

SPSS is flexible in that it supports many text formats, including: plain text, PDF, HTML, Microsoft Office, and XML text files. It has open architecture which allows the program to join together with other text analytics applications including Text Mining Builder, LexiQuest Categorize, and all other SPSS data mining and predictive analytics applications. Clementine's text extraction does well to offer the use several methods for limiting the scope of an extraction and therefore tends to yield the most informative terms and phrases in its results. Figure 3 below shows Clementine's text extraction results browser.

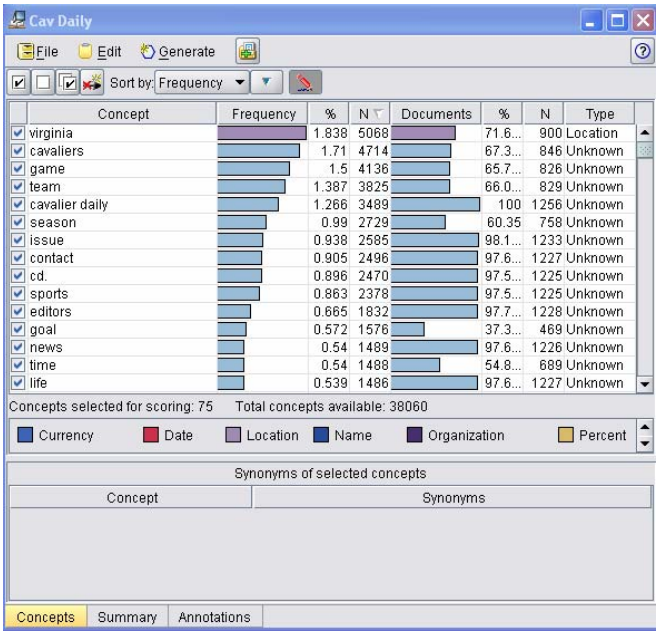

Figure 3: Ranked list of extracted concepts in Clementine

Clementine also includes a text link analysis which has two main functions. The first is that it recognizes and extracts sentiments (i.e. likes and dislikes) from the text with the help of dictionary files created in Text Mining Builder.

Some of these dictionary files are already provided in the software package while others can be developed by the user. The second is that it detects correlations between things such as people/events, or diseases/genes. SPSS also allows a user to manually define a customized taxonomy with the help of LexiQuest Categorize, which uses training data to learn the characteristics of the taxonomy predict the classification of documents.

Polyanalyst, manufactured by Megaputer, can process large scale databases. The software also has a low learning curve and step by step tutorials. The integration of an analysis performed on both structured and unstructured text is available. The results of the analysis can be incorporated into existing business processes.

Clarabridge provides analysis of data and is used in conjunctions with commercially available business intelligence tools which are used to view and interpret the results of this analysis. It also allows parallel processing of large amounts of data. During the processing, entities, relationships, sections, headers, and topics, as well as proximal relationships, tables, and other data are recognized. This information is stored into the capture schema thus maintaining metadata and linking it back to its source. Clarabridge can contain large amounts of data and maintain high throughput. The GUI requires a minimal amount of coding from users and processes can be done without human intervention.

Table 3 shows the different functions that the software have. For some of these functions, such as extraction, all of the software possesses some form of the function. For others, such as clustering, not all of the software have the feature. A table of this form is useful for a quick visual software functionality comparison.

Table 3: Functionality

	SOFTWARE				
FUNCTIONS	SPSS	SAS	Clarabridge	Polyanalyst	Leximancer
Extraction	X	X	X	X	X
Summarization					X
Categorization	X	X	X	X	X
Clustering		X		X	X
Concept Linking	X	X	X	X	X
Decision Trees		X		X	X
Document Linkage		X	X	X	X
Memory based Reasoning	X	X	X	X	
Regression	X	X		X	Exports data for further analysis in other packages
Time Series	X	X	X		Exports data for further analysis in other packages

II. Results

Unfortunately we were unable to run the Cavalier Daily files through Clarabridge because this software package requires additional business intelligence tools in order to achieve readable results.

After running the sample corpus through the other four software suites and comparing the subjective ease of use of the software, two products rose to the top: SAS and SPSS.

The immediate advantage of these pieces of software is that they were developed for large projects, so while processing 1200 documents took a significant amount of time, they were able to display meaningful results. The choice between these products, however, rests on the particular application in which they are used.

Because of the non-intuitive interface and steep learning curve, SAS is best used in situations where the user already has a general understanding of text mining. It is also an excellent choice for processing large amounts of documents, however, it only gives truly meaningful information if the input has been pre-processed and made to be semi-structured.

When running the files through, SPSS proved to be the quickest at mining the files. SPSS also is a good choice for processing large amounts of documents and provides more useful results if the input has been pre-processed and made to be semi-structured. Another benefit that SPSS has over SAS is that SAS extracts a large amount of useful terms.

All of the software products tested primarily extracted sport-related concepts from the given corpus in the explorative analysis that was done. This indicates that in the Cavalier Daily newspaper, sports are the main topics that are reported on. Again, because we were unable to obtain working copies of Clarabridge and Polyanalyst, we were unable to test them using our sample corpus. Further results could be obtained with a deeper analysis, but as we were using our corpus to get only a preliminary idea of the features of the software, we did not pursue a more advanced investigation.

III. Future Work

Future work with text mining software is already underway. While the test corpus that was used to evaluate the software in this report was large, the problem that was attempted to be solved was not well-defined. Therefore, a new problem has been proposed that is well-defined, and work is underway to analyze and solve it.

There is a current project underway in which a group is attempting to extract relationships between the results from social security disability claims in the court systems and the content of the claims that are filed with the courts. This is a problem that is semi-structured and well-defined, and is perfect for further testing of the SAS and SPSS suites.

The data for these cases are being gathered from various state and federal websites that have cases on record having to do with social security disability claims. This data will be collected, parsed, inputted into a database table, and then

processed by SAS and SPSS in order to extract relationships in the data.

The hope is that this processing will lead to discoveries about what types of claims are most often approved or rejected by the courts, if there is such a relationship. For example, it might be the case that if a person mentions "Lou Gehrig's disease" in their claims, that they are almost always approved for their claim. If such a relationship were true, then text mining software like SAS and SPSS should be able to extract it through predictive capabilities.

IV. Conclusion

The following goals were achieved by the conclusion of this project:

- Identified the common needs of users of text mining tools through researching the industries that use text mining and the applications for which text mining is employed.
- Addressed the current state in text mining through background research in the field and hands on experience.
- Evaluated and compared text mining software. This goal can be improved upon in future projects by considering an expanded set of evaluation criteria.

Appendix

This Appendix provides a glossary of terms commonly used in discussions of text mining software.

KDD-knowledge discovery and data mining

Queries-a common way of extracting information from databases

Tuples-finite sequence or ordered list of object

Ease of Learning-how easy or hard it is to learn how to use the software

Ease of Use-once the software is learned, how easy or hard it is to use the software

Clustering-Clustering algorithms find groups of items that are similar. For example, clustering could be used by an insurance company to group customers according to income, age, types of policies purchased and prior claims experience.

Decision tree-A tree-like way of representing a collection of hierarchical rules that lead to a class or value.

Regression tree-A decision tree that predicts values of continuous variables.

Time series model-A model that forecasts future values of a time series based on past values.

Extraction-locating specific pieces of data and extracting it from the document

Summarization- summarization extracts the most relevant phrases or even sentences from a document.

Concept Linking- Usually comes in the form of some web-like visualization in which the links between extracted concepts are shown based on their co-occurrence and proximity within documents.

Document Linkage – The ability to view in the results where in the documents the concept occurs. Results link back to input documents.

Categorization- Organization of documents into predefined categories based on existence of specified indicator concepts within the documents.

Memory-based Reasoning- MBR uses training records to train a neural network to learn to predict certain characteristics of new documents

Acknowledgment

The Capstone team would like to thank their technical advisor, Professor K. Preston White for guiding them through the capstone project. Also, the team would like to acknowledge Elder Research Incorporated for allowing them to participate in such a rewarding project through funding it. The team would like to also thank Debbie and Jordan for everything they have done for the team.

References

[1] Weglarz, G. (2004). Two worlds of data – unstructured and structured. *DM Review*, 14(9), 19-22.

[2] J. Elder et al., *An Evaluation of High-end Data Mining Tools for Fraud Detection*. Available: http://www.datamininglab.com/Portals/0/tool eval articles/ smc98 abbot mat eld.pdf.

[3] Megaputer, (2002), Polyanalyst Powerpoint.

[4] S. Grimes, *The Word on Text Mining*. Available: http://altaplana.com.

[5] *Saving Lives, Cutting Patient Costs: Louisville Hospitals Advance with SAS Text Miner*, SAS, 2006. Available: http://www.sas.com/success/louisville.html.

[6] R. Rao, *From Unstructured Data to Actionable Intelligence*, IT Pro, 9202(03), 2003, pp. 1-7.

[7] P. Fule, J. Roddick, *Detecting Privacy and Ethical Sensitivity in Data Mining Results*, School of Informatics and Engineering, Flinders University, South Australia.

[8] *Text Mining Terms*, SPSS White Paper, Sept. 2006.

[9] K. Michel, J. Elder, *Evaluation of Fourteen Desktop Data Mining Tools*, 1998.

[10] M. Goebel, L. Gruenwald, *A Survey of Data Mining and Knowledge Discovery Software Tools*, SIGKDD Explorations, pp. 20-33, 1999.

[11] Apte, C. (2002). Business applications of data mining. *Communications of the ACM, 45*(8), 49-53.

[12] Blumberg, R. (2003). The problem with unstructured data. *DM Review*, *13*(2), 42-4.

[13] Grimes, S. (2005). *The Developing Text Mining Market*. [White paper, electronic version]. Retrieved October 12, 2006 from www.textminingnews.com.
[14] Hand, D., Mannila, H., Smyth, P. (2001). *Principles of Data Mining*. MIT Press: Cambridge, MA.
[15] Marti Hearst. "What is Text Mining?" 17 October 2003. http://www.ischool.berkeley.edu/~hearst/text-mining.html (29 October 2006).