A Review of Two Text-Mining Packages: SAS TextMining and WordStat

Angelique DAVI, Dominique HAUGHTON, Nada NASR, Gaurav SHAH, Maria SKALETSKY, and Ruth SPACK

The purpose of this article is to review two text mining packages, namely, WordStat and SAS TextMiner. WordStat is developed by Provalis Research. SAS TextMiner is a product of SAS. We review the features offered by each package on each of the following key steps in analyzing unstructured data: (1) data preparation, including importing and cleaning; (2) performing association analysis; and (3) presenting the findings, including illustrative quotes and graphs. We also evaluate each package on its ability to help researchers extract major themes from a dataset. Both packages offer a variety of features that effectively help researchers run associations and present results. However, in extracting themes from unstructured data, both packages were only marginally helpful. The researcher still needs to read the data and make all the difficult decisions. This finding stems from the fact that the software can search only for specific terms in documents or categorize documents based on common terms. Respondents, however, may use the same term or combination of terms to mean different things. This implies that a text mining approach, which is based on analysis units other than terms, may be more powerful in extracting themes, an idea we touch upon in the conclusion section.

KEY WORDS: Clustering; Correspondence analysis; Theme extraction; Unstructured data.

1. INTRODUCTION

Qualitative data can be highly informative and useful. Analyzing unstructured data—for example, answers to open-ended questions—can provide rich information that might be difficult to obtain using other research measures. However, the analysis of such data can be a very challenging task (Miles 1979). In effect, analyzing qualitative data involves a highly labor-intensive operation. The larger the size of the dataset, the more overwhelming the data analysis is.

To help with the problem of analyzing qualitative data, several software firms have offered text mining packages. The aim of text mining tools is to extract key elements from large unstructured datasets, identify relationships, and summarize the information. Researchers, both academicians and practitioners, are deploying or considering text mining software to deal with their mountains of text. The use of such tools has not been really satisfactory (Robb 2004). Companies, however, continue to improve the tools they offer in an attempt to satisfy researchers. The purpose of this article is to review two text mining packages, namely, WordStat and SAS TextMiner. WordStat is developed by Provalis Research. SAS TextMiner is a product of SAS, a leading statistical package. SAS TextMiner is available as a tool within the SAS Enterprise Miner data mining user interface.

In addition to the introduction, this article consists of five parts. First, we describe the methodology (and dataset) we used to evaluate the two packages. Second, we provide a summary of the requirements needed to set up each package. Third, we review the features offered by each package on each of the following key steps for analyzing unstructured data: (1) data preparation, including importing and cleaning; (2) performing association analysis; and (3) presenting the findings, including illustrative quotes and graphs. Fourth, we evaluate each package on the most challenging task in analyzing qualitative data, namely, extracting major themes from a dataset. Fifth, we provide concluding remarks.

2. METHODOLOGY AND DATA USED IN THE REVIEW

To study the WordStat and SAS TextMining packages, we used each of them to analyze an example dataset. More specifically, we used each software package to extract major themes from the data. Extracting themes from a dataset is the most challenging task in analyzing qualitative data. We also compare the two packages in the areas of descriptive analyses, such as running associations, and drawing graphs and tables. Running associations is commonly used in analyzing large datasets. A common practice in presenting results is the use of graphical illustrations and tabular summaries. The data we used in our analysis were initially collected to help faculty gain insight into the student perspective on fairness/unfairness in grading. A group of 416 undergraduate students enrolled in expository writing courses offered through the English department responded to a questionnaire. The questionnaire consisted mainly of two open-ended questions. The first question stated: "Think about a previous class—in any subject—in which you felt the teacher graded students' work in an unfair way. Discuss how your own work was evaluated and why you felt the grading was unfair." Similarly, students were asked to provide examples of fair grading practices. The participants were also asked for their age, gender, expository writing course they were enrolled in, and semester in college. Answers to the open-ended question on unfair grading practices (stated above) were used in the analysis. This variable is referred to as "UNFAIR" in the rest of the article. Note that answers to open-ended questions represent a typical qualitative dataset that researchers are interested in analyzing. Unfortunately, because the dataset contains named references to specific instructors in some cases, it is not possible to make it publicly available at this point.

Angelique Davi is Assistant Professor, Dominique Haughton is Professor, Nada Nasr is Assistant Professor, Gaurav Shah is Manager, Maria Skaletsky is Research Consultant, and Ruth Spack is Associate Professor, Bentley College, Waltham, MA 02452 (E-mail: DHAUGHTON@bentley.edu).

 DOI: 10.1198/000313005X22987

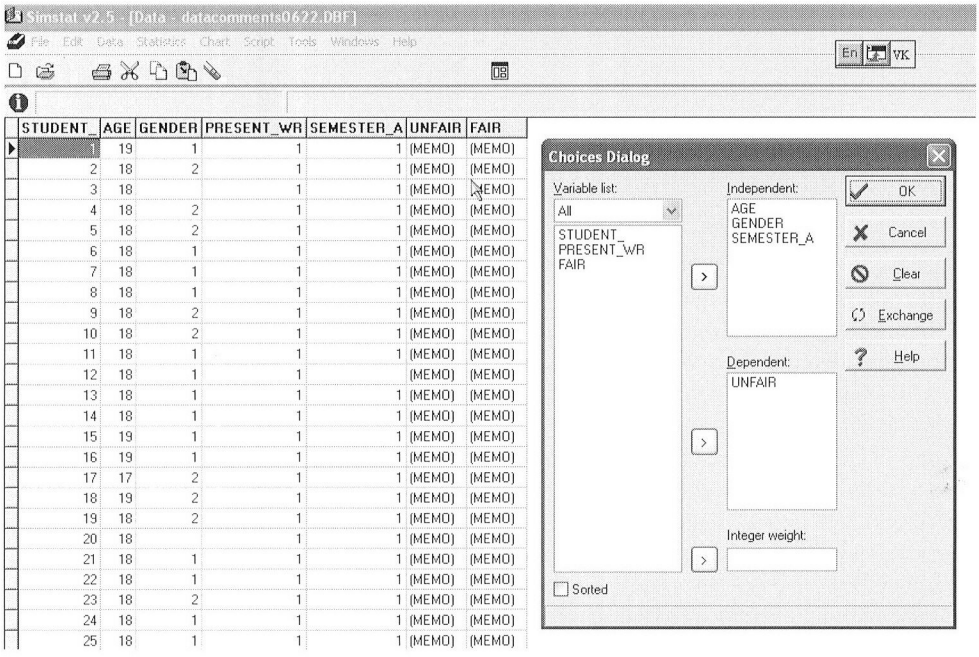


Figure 1. SimStat interface.

3. REQUIREMENTS

Hardware requirements are standard: we ran all analyses on an IBM T40 laptop computer, with an Intel Pentium processor (1500MHz, 598 MHz, and 512MB of RAM). Installations are easy in both cases, although perhaps a bit more awkward in the case of SAS TextMiner. Pricing is as follows: an individual permanent license for WordStat (with Simstat) 2.5 is priced at $525, $375 for academics; the total initial fee for SAS Text Miner would be of $10,551 for academics, and $154,200 for commercial organizations, with an estimated $5,249 renewal fee for 2005 (inclusive of SAS Base and STAT). The bulk of the SAS Text Miner costs are those of adding SAS Enterprise Miner and SAS Text Miner to a typical SAS Base and STAT

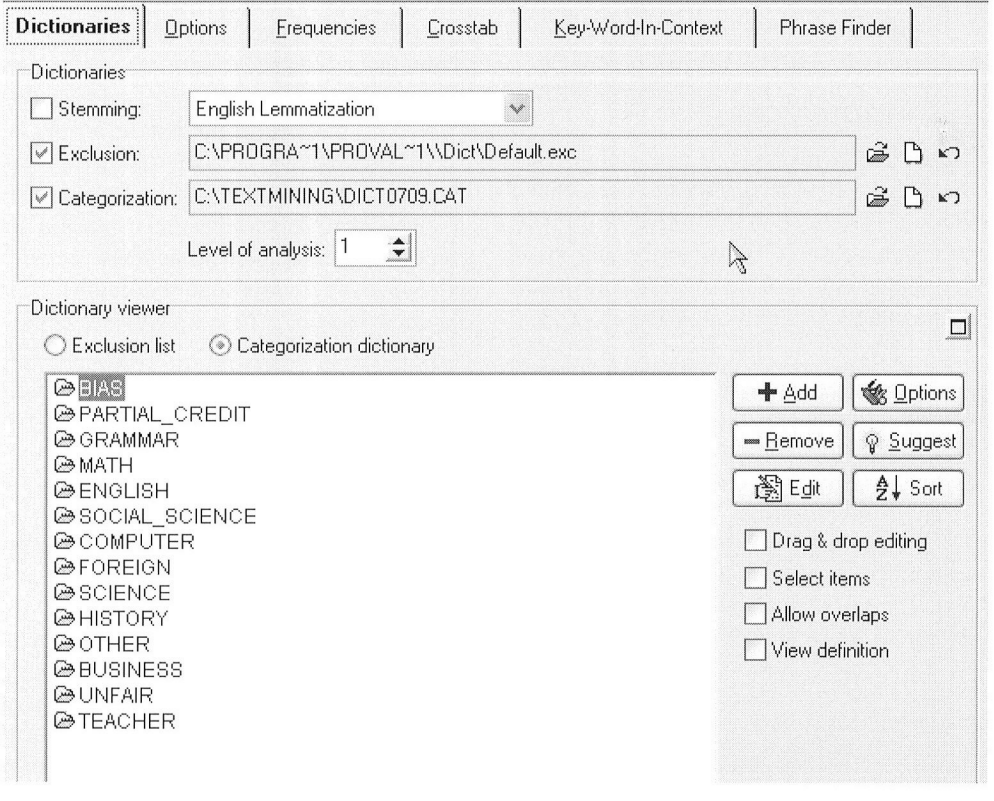
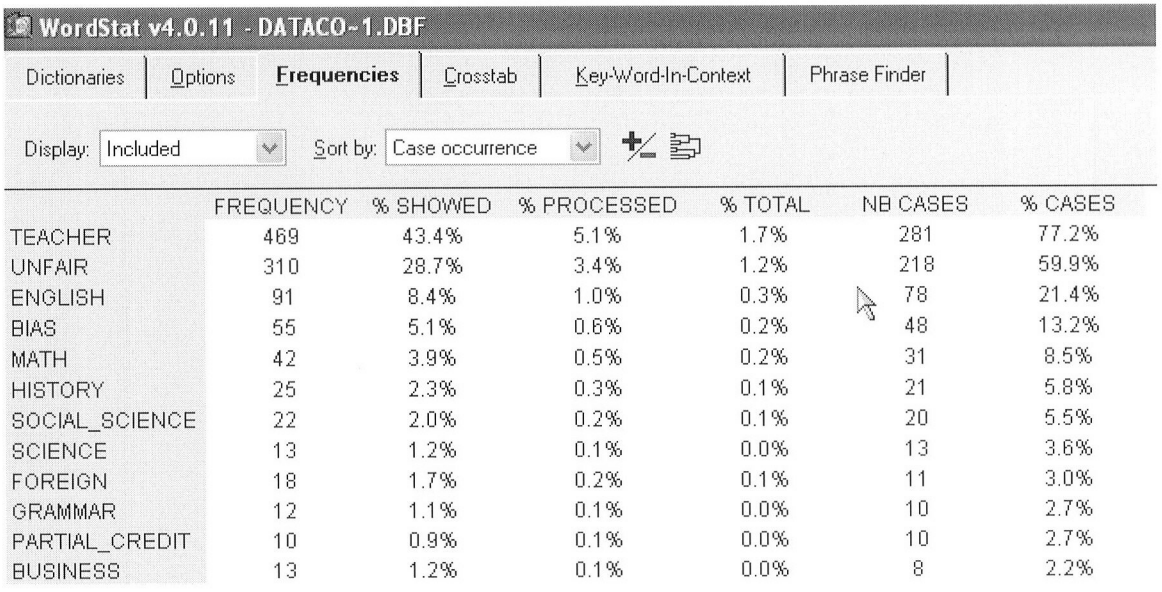


Figure 2. Dictionaries in WordStat.



	FREQUENCY	% SHOWED	% PROCESSED	% TOTAL	NB CASES	% CASES
TEACHER	469	43.4%	5.1%	1.7%	281	77.2%
UNFAIR	310	28.7%	3.4%	1.2%	218	59.9%
ENGLISH	91	8.4%	1.0%	0.3%	78	21.4%
BIAS	55	5.1%	0.6%	0.2%	48	13.2%
MATH	42	3.9%	0.5%	0.2%	31	8.5%
HISTORY	25	2.3%	0.3%	0.1%	21	5.8%
SOCIAL_SCIENCE	22	2.0%	0.2%	0.1%	20	5.5%
SCIENCE	13	1.2%	0.1%	0.0%	13	3.6%
FOREIGN	18	1.7%	0.2%	0.1%	11	3.0%
GRAMMAR	12	1.1%	0.1%	0.0%	10	2.7%
PARTIAL_CREDIT	10	0.9%	0.1%	0.0%	10	2.7%
BUSINESS	13	1.2%	0.1%	0.0%	8	2.2%

Figure 3. Key word frequencies in WordStat.

installation; base SAS and SAS/STAT initially cost $686 and $606, respectively, with an estimated renewal fee for 2005 of $312 and $307.

4. THE WORDSTAT PACKAGE

4.1 Introduction and Data Preparation

The WordStat package is a software product produced by Provalis Research, and needs to be called as an add-on to the statistical package Simstat produced by the same company. It is very easy to read data into Simstat; in our case we imported an Excel file without any problems. Figure 1 shows the Simstat interface with variables such as age, gender, writing course, semester, and the text variables UNFAIR and FAIR. The annotation (MEMO) represents text data. We focus in this review on the variable UNFAIR. In order to run WordStat, UNFAIR needs to be selected as a dependent variable, and some demographics of interest as independent variables through an easy-to-use menu (Figure 1).

Once the WordStat program has been called, a screen such as in Figure 2 appears. WordStat provides by default an exclusion list that consists of words to be excluded from the analysis (words such as "that", "a", etc.). The stemming process (not selected on Figure 2) allows the user to merge words—for example, grade and grades—into one "word".

4.2 Creating Dictionaries

An important and challenging preliminary to most WordStat analyses is the creation of a categorization dictionary. This process requires content knowledge, and involves creating categories and deciding which terms belong to them. For

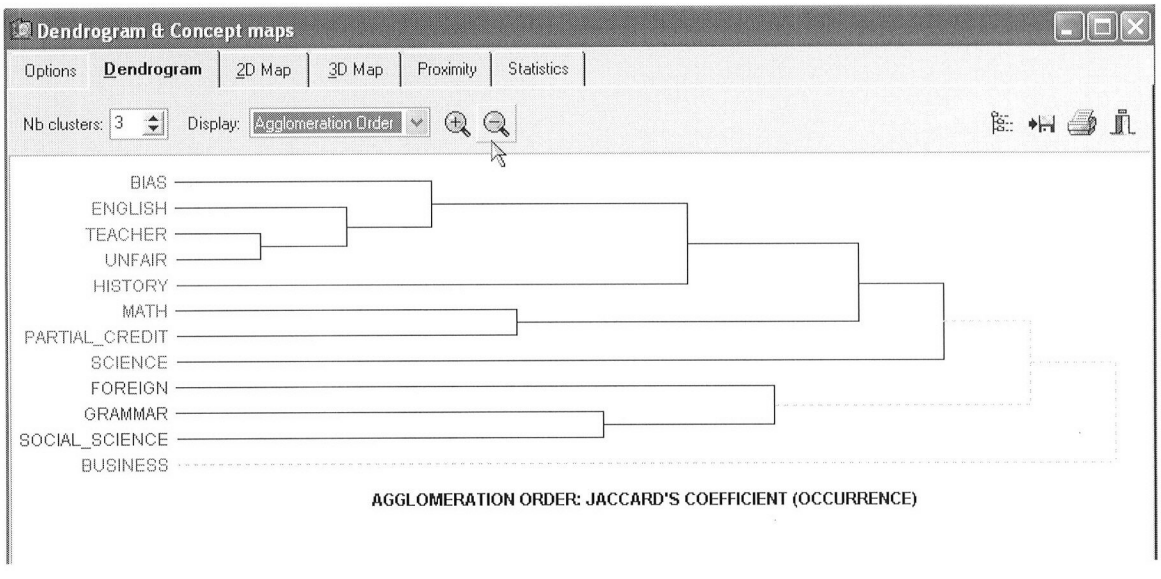


Figure 4. Clustering of key words in WordStat.

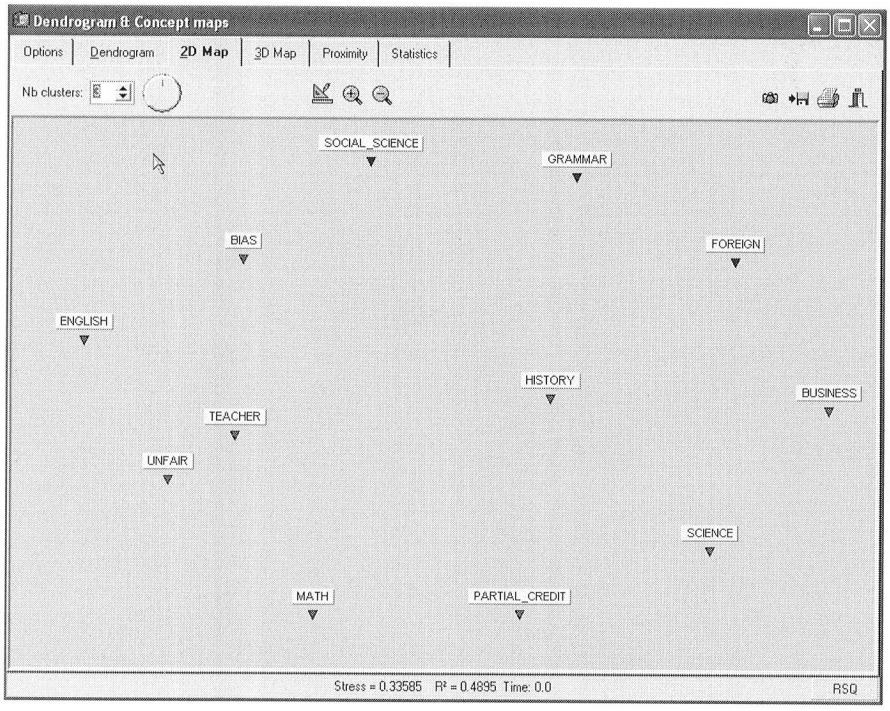


Figure 5. Multidimensional scaling map of key words in WordStat.

instance, the user would need to decide which terms fall under the "MATH" category. For example, the MATH category in our case contains the terms ALGEBRA, MATH, CALCULUS, and STATISTICS. Without a categorization dictionary, WordStat will use all words appearing in the dataset (except excluded words) as categories.

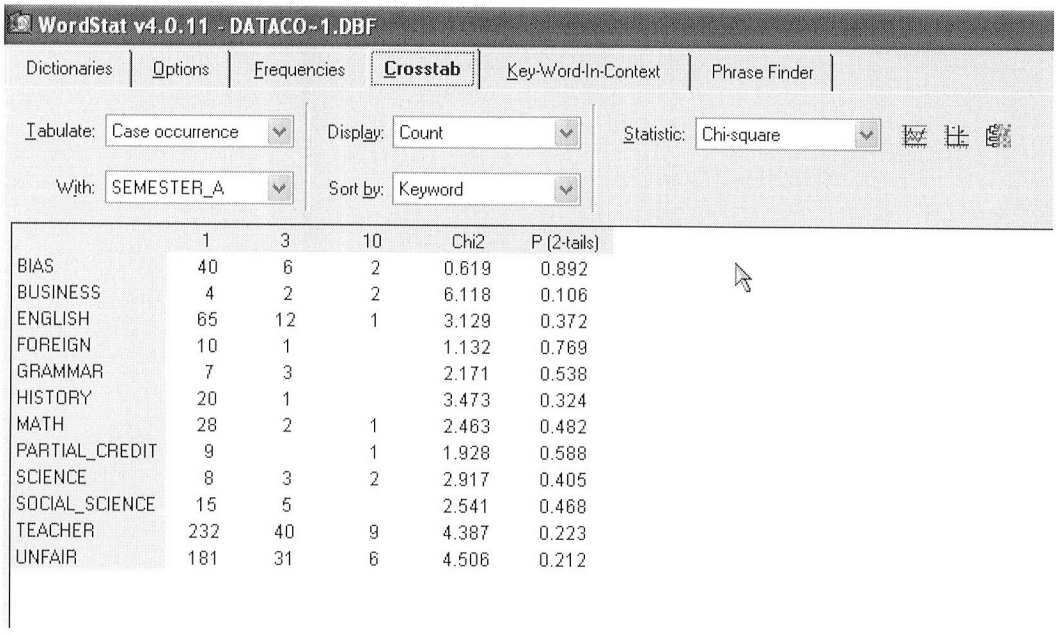
For purposes of this review, we set up a dictionary, which is given in the Appendix, with an objective of identifying which topics (math, English, social science, etc.) tended to be associated with various key words such as bias, partial credit, unfair, and so on.

4.3 Reporting Results

Once the categories are created, a WordStat analysis consists of a number of tabulations and cross-tabulations of the categories (against some categories of the independent variables) to be described below, and then of a statistical treatment of the resulting tables (such as correspondence analysis).

Frequencies of key words are given in Figure 3; for example, the key word UNFAIR appears 310 times and in 218 of the 416 documents.

Once the frequencies are calculated, clustering can be performed on the key words. WordStat performs a hierarchical clus-



	1	3	10	Chi2	P (2-tails)
BIAS	40	6	2	0.619	0.892
BUSINESS	4	2	2	6.118	0.106
ENGLISH	65	12	1	3.129	0.372
FOREIGN	10	1		1.132	0.769
GRAMMAR	7	3		2.171	0.538
HISTORY	20	1		3.473	0.324
MATH	28	2	1	2.463	0.482
PARTIAL_CREDIT	9		1	1.928	0.588
SCIENCE	8	3	2	2.917	0.405
SOCIAL_SCIENCE	15	5		2.541	0.468
TEACHER	232	40	9	4.387	0.223
UNFAIR	181	31	6	4.506	0.212

Figure 6. Cross-tabulation of key words with semester in WordStat.

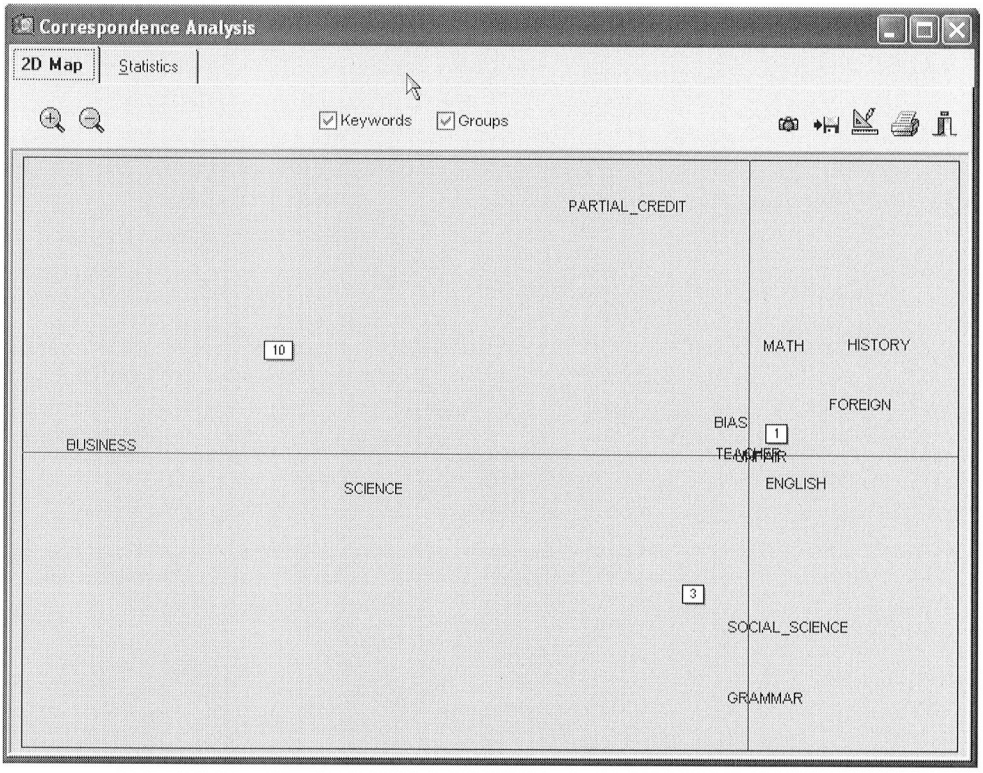


Figure 7. Correspondence analysis of key words with semester in WordStat.

tering of key words, using as a distance by default the Jaccard coefficient $(a/(a+b+c))$, where a is the number of documents where both key words occur, and b (relatively c) is the number of documents where one key word occurs but not the other. For example, it appears that MATH and PARTIAL CREDIT are clustered together, as well as BIAS, ENGLISH, TEACHER and UNFAIR. The dendrogram from this clustering is given in Figure 4.

From this clustering, it is quite easy to obtain a multidimensional scaling map, presented in Figure 5, where the clusters can be visualized.

WordStat also offers the possibility of cross-tabulating key words with independent variables, such as gender, age, semester in which the student is enrolled, and so on. In Figure 6, we present the result of cross-tabulating the key words with SEMESTER. Because most respondents were enrolled in semesters 1 and 3 (first semester freshman and sophomore, respectively) we coded all remaining semesters with 10.

A chi-square analysis is provided for each key word individually cross-tabulated with SEMESTER. For instance, the individual table for BIAS would consist of six entries, corresponding to the number of occurrences and nonoccurrences of BIAS terms across the semesters. Note that none of the chi-square analyses are significant, but the overall cross-tabulation gives rise to row and column profiles that can be used to perform a correspondence analysis. The results of this correspondence analysis are given in Figures 7 and 8.

Figure 8 provides the data which underlie the graph in Figure 7. Note that "Axes 1" ("Axes 2") in Figure 8 refers to the vertical (horizontal) axis in Figure 7. As is known in correspondence analysis settings, and is clearly explained in the WordStat manual, the closeness of key words to each other can be judged by how close they appear on the graph, and similarly for the semester categories 1, 3, and 10. However, how close a key word is to a semester category can be judged only by looking at the angle between the segments joining the origin and the key word, and the origin and the category.

Moreover, the further a key word (or a semester category) is from the origin, the less typical its row profile (column profile) is. We can thus see that key words such as BIAS, TEACHER, and ENGLISH have row profiles (across semesters) which are close to the row profile of the whole set of key words. That means that the use of the key words BIAS, TEACHER, and ENGLISH by students enrolled in different semesters is typical of the use of key words in general by students enrolled in different semesters. Locations far from the origin represent outlying behavior, as is the case for BUSINESS, SCIENCE, or semester category 10. That means that key words such as BUSINESS and SCIENCE tend to differ from other key words in how students enrolled in different semesters use them, and that semester category 10 differs from other semester categories in how different key words are used by students.

Semester category 1 is close (using the angle mentioned above) to ENGLISH, HISTORY, and MATH while semester category 3 is close to GRAMMAR or SOCIAL SCIENCE. That means, as one might expect, that first semester students were more likely to comment on perceptions of unfair treatment in classes related to English, history, and math (probably because these subjects tend to enroll a lot of first semester students).

4.4 Additional Features

We mention two more features of WordStat which are quite helpful. Figure 9 gives a listing of most frequent sentences (con-

EIGENVALUES

Eigenvalues	Percentages	Cumul. Percent
0.033	71.482	71.482
0.013	28.518	100.000

VARIABLES COORDINATES

Item.	Axes 1	Axes 2	Axes 3
1	0.306	0.341	
3	-0.605	-2.388	
10	-5.214	1.740	

WORDS/CATEGORIES COORDINATES

Item.	Axes 1	Axes 2	Axes 3
BIAS	-0.210	0.510	
BUSINESS	-7.166	0.076	
ENGLISH	0.522	-0.527	
FOREIGN	1.227	0.813	
GRAMMAR	0.179	-4.160	
HISTORY	1.444	1.843	
MATH	0.379	1.834	
PARTIAL_CREDIT	-1.355	4.194	
SCIENCE	-4.148	-0.638	
SOCIAL_SCIENCE	0.430	-2.970	
TEACHER	-0.004	-0.020	
UNFAIR	0.134	-0.071	

Figure 8. Correspondence analysis (key words with semester) statistics in WordStat.

sisting of between 3 and 6 words), and Figure 10 gives a listing of key words in context. For instance "GRADING WAS UNFAIR" appeared 19 times in 18 different documents.

In Figure 10, we can see the different ways the key word UNFAIR was used in responses.

5. THE SAS TEXT MINER PACKAGE

5.1 Introduction and Data Preparation

The SAS Text Miner package is available as an add-on to SAS Enterprise Miner. So both a standard installation of SAS and Enterprise Miner are necessary before the installation of SAS Text Miner can be contemplated. Here we are using here versions 9.1 of all packages.

Interestingly, the SAS Text Miner approach, at least as far as statistical analyses are concerned, is quite different from that of WordStat, with pros and cons to be discussed as we go along. For instance, SAS Text Miner clusters documents, not key words, and then lists the most frequent terms that appear in the doc-

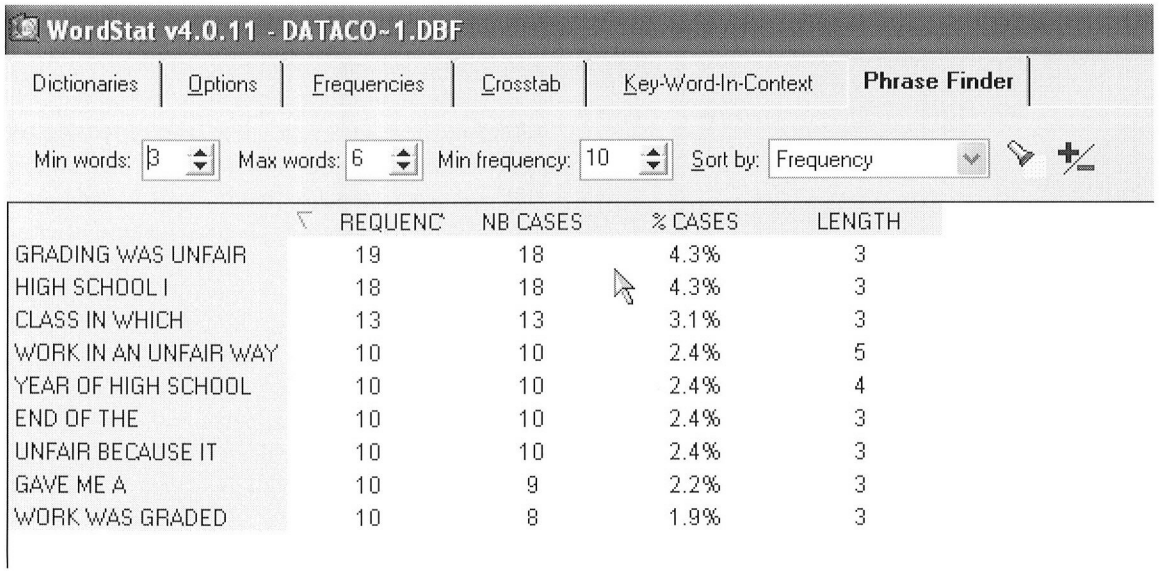


Figure 9. Listing of most frequent sentences in WordStat.

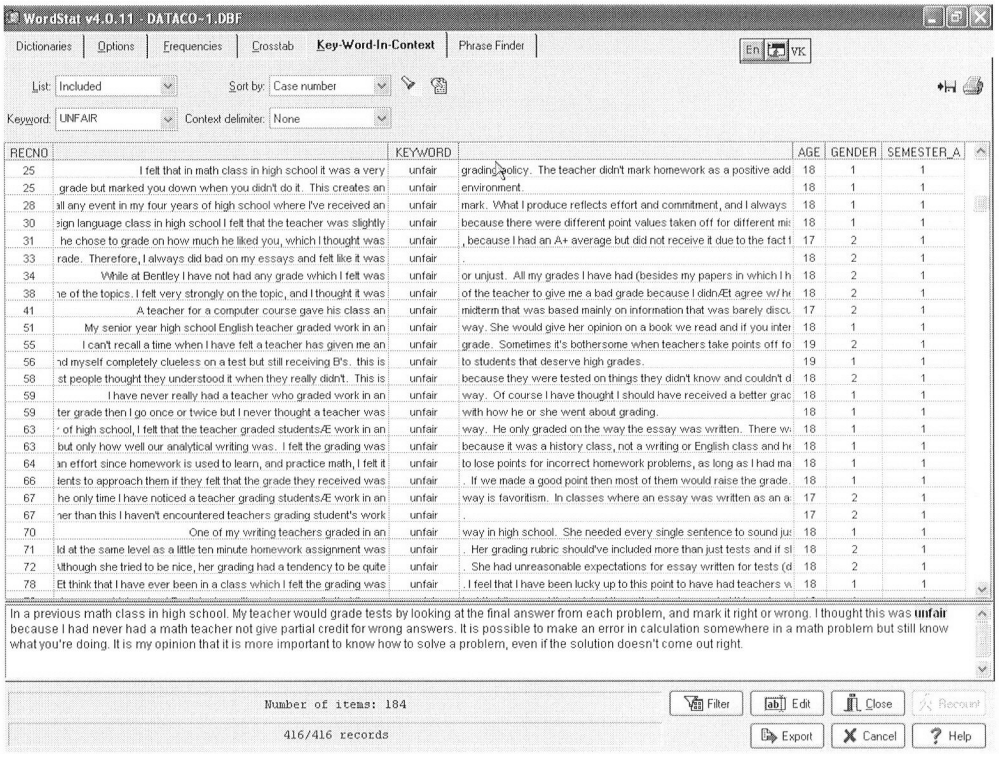


Figure 10. Listing of key word in context in WordStat.

uments in each cluster. SAS Text Miner begins an analysis by building a matrix with terms as rows and documents as columns, with cells equal to the number of occurrences of a term in a document. Because such a matrix is very large, SAS Text Miner uses SVD (singular value decomposition) to reduce the dimensionality of the matrix, so that documents can be represented with smaller vectors and terms as smaller rows.

Figure 11 gives a sense of the Text Miner user interface, which is that of Enterprise Miner, with Text Miner as a node in the diagram.

By default, SAS Text Miner creates groups of equivalent words, which will be treated as terms in the analysis. For instance, "+ grade" consists of "graded", "grading", and "grades". Note that the two different occurrences of "+ grade" in the table

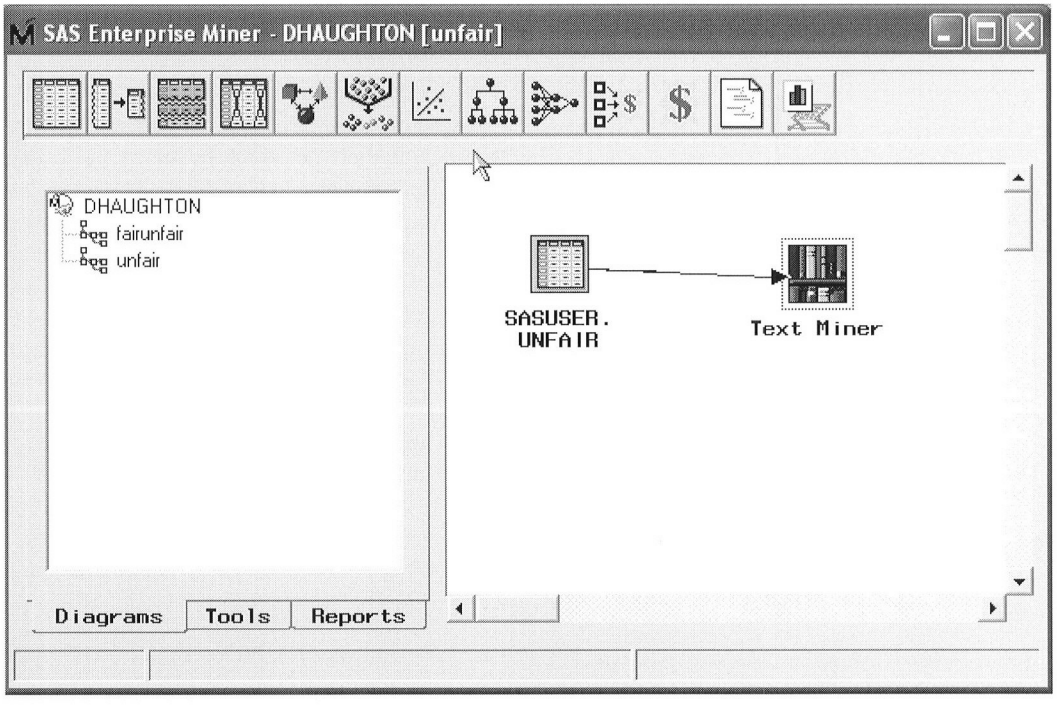


Figure 11. SAS Enterprise Miner Text Miner interface.

in Figure 12 correspond to roles of the words as a noun and as a verb.

5.2 Creating Dictionaries

In SAS Text Miner, creating dictionaries as was done in WordStat is not needed per se. Themes arise out of analysis results, for example on the clustering of documents (see Figure 15).

5.3 Reporting Results

The term frequencies given in Figure 12 are very similar to those given in WordStat with, for example, "+ teacher" appearing 470 times in 283 documents.

SAS Text Miner offers two possible clustering options, one with the EM (expectation-maximization) algorithm, and a hierarchical clustering. The results of a clustering depend on the preliminary SVD decomposition, which is performed each time the Text Miner node is run. The number of dimensions can be chosen by the user, or decided by SAS Text Miner according to a choice (made by the user) of low, medium, or high resolution. The user also selects the number of descriptive terms to be listed as most frequently used in the documents in each cluster (five by default), as well as a maximum (or exact) number of clusters.

Figure 13 presents the results of an EM clustering, with medium SVD resolution, five descriptive terms, and the default maximum number of clusters (40). Figure 14 presents the same

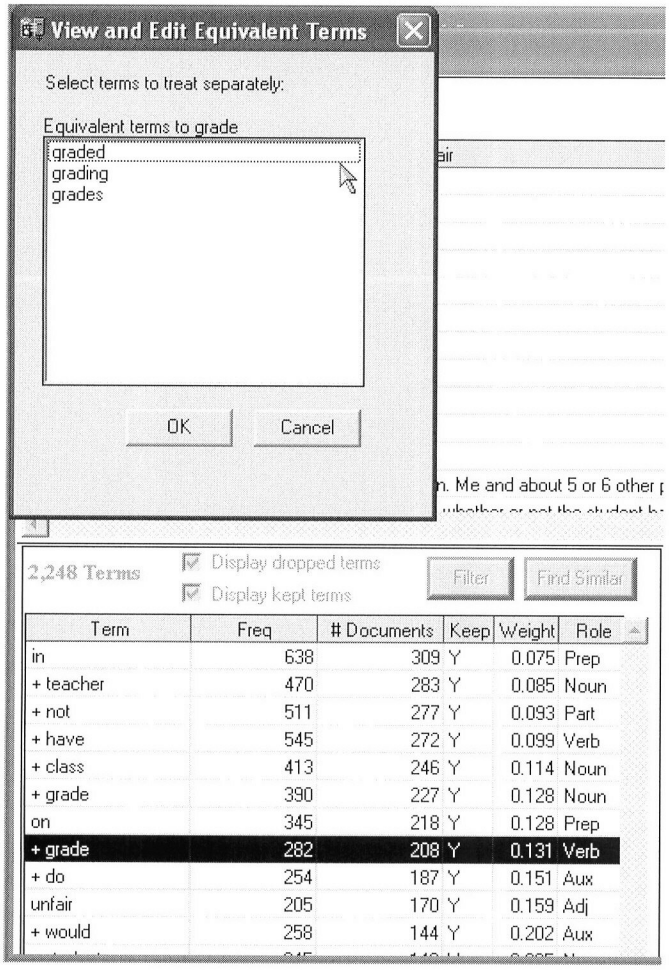


Figure 12. SAS Enterprise Miner Text Miner listing of equivalent terms.

analysis, but this time with full text visible for the documents. Eight clusters were identified by SAS Text Miner. In order to interpret the clusters we can select, for instance, 20 descriptive terms; the result is given in Figure 15.

Carefully thinking about the terms in each cluster, the researcher might be able to, in a preliminary fashion, come up with possible themes. For instance, using the descriptive terms in Figure 15, we can identify the following themes arising from the responses: perceptions of unfairness related to essay writing (points taken off) in English class (cluster 1), perceptions of bias (related for example to different opinions regarding style) in English classes (or classes which require English writing; cluster 2), perceptions of unfairness in relation to tests and quizzes (cluster 3), blank responses with one additional document containing quite a few extra blanks in it (cluster 4), perceptions of unfairness in grading writing assignments in high school (cluster 5), perceptions of bias related to some students being "liked" more than others (cluster 6), perception of unfair grading in the context of group work (cluster 7), perceptions of unfair amount of partial credit received, in the context of multiple choice exams, notably in math courses, and notably in high school (cluster 8).

Interestingly, SAS Text Miner identified some links also identified by WordStat; for example, the link between the terms math and partial credit. One significant advantage of the SAS Text Miner approach is that it does not require the creation of a dictionary of key words.

Figure 16 gives a sense of what a hierarchical cluster solution would look like; we used the same settings as for the EM solution, and five descriptive words for the clusters. The first split occurs between the cluster with all (but one) blank documents and the rest of the documents, and further splits follow. SAS Text Miner provides a tree representation for the hierarchical cluster solution (Figures 17 and 18). By positioning a cursor over a node, the 5 descriptive terms can be viewed. The partial credit cluster appears near the bottom of Figure 18.

5.4 Additional Features

We finally mention two additional features of SAS Text Miner: a search for "similar" terms, documents, or clusters. Terms are "similar" if their positions in the space spanned by the SVD dimensions are close to each other. The results of a search for ten terms "similar" to "unfair" are given in Figure 19. "Similarity" in this case refers to similar distribution of terms across the documents. Figures 20 and 21 provide a SAS Text Miner concept link search for the term "+ student". Figure 20 shows that "+ student" appeared in 143 documents, and is linked to other terms such as "unfair way", "+ know", or "other students". In order for SAS Text Miner to create a link from term 1 to term 2, term 2 must appear when term 1 does at least 1% of the time (5% by default), and the relationship between term 1 and terms 2 must be extremely, highly or somewhat significant (we used the "somewhat" setting here). Figure 21 reveals that "+ student" and "+ know" occur together in 32 documents, and that "+ know" occurred in 58 documents.

6. EXTRACTING THEMES

The most challenging task in analyzing qualitative data is the extraction of major themes from the data. For instance, in

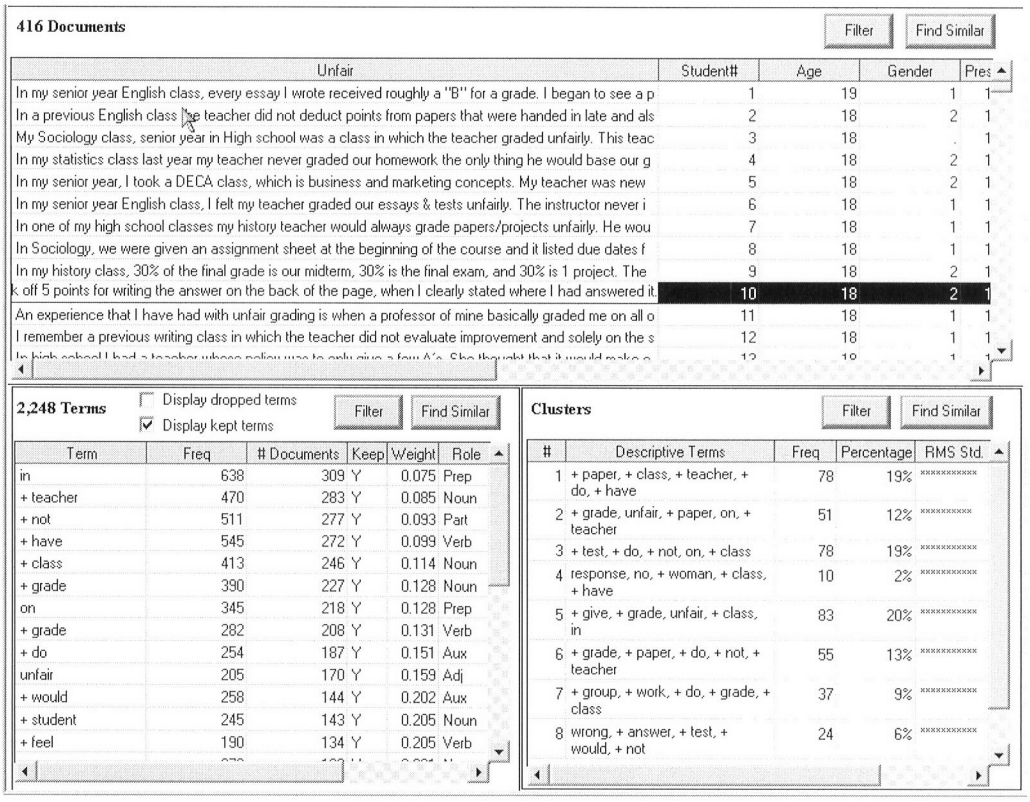


Figure 13. Result window from SAS Text Miner, with EM clustering (five descriptive terms, medium SVD resolution).

the data we used for this article, the main issue was to identify grading practices that students perceived as unfair. Ideally, the software would reduce the need for the researcher to work manually on the data. Usually, researchers read through the data several times looking for emerging themes. Once the themes are identified, researchers use those themes to classify the answers under different categories. This classification is needed to quantify any results. It also helps with locating quotes under a specific

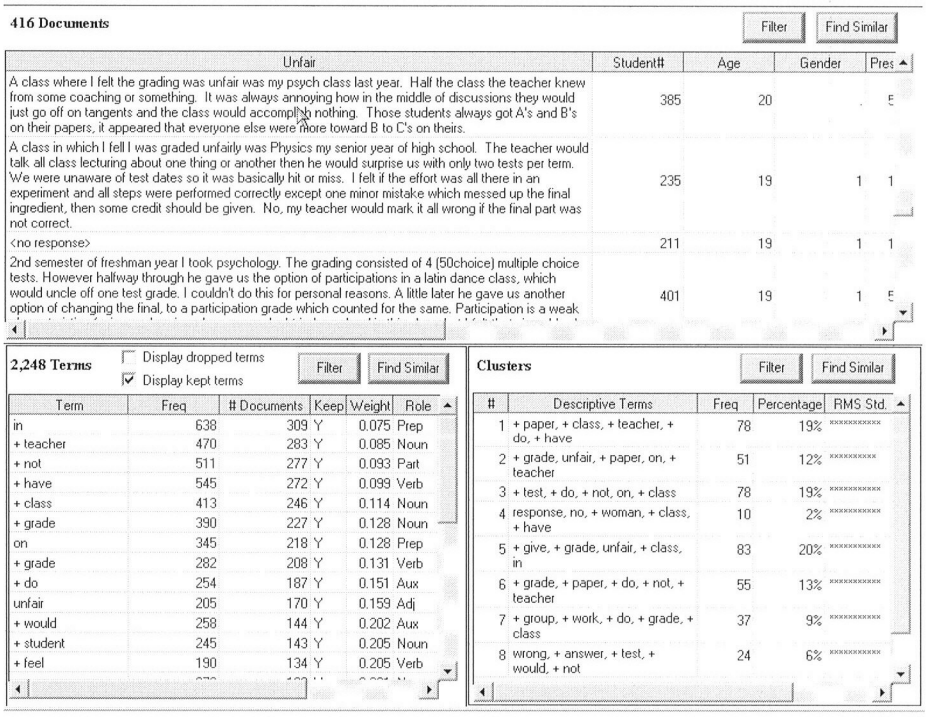


Figure 14. Result window from SAS Text Miner, with EM clustering, and full text view.

#	Descriptive Terms	Freq	%	RMS Std.
1	off, + point, + writing, + essay, + homework, english class, + make, + time, + assignment, + answer, + paper, + feel, + good, + write, + give, should, no, + class, + teacher, unfair	78	19%	0.1066333394
2	other students, english teacher, english, own, + style, english, + opinion, + work, + writing, + receive, + write, grading, other, with, + grade, unfair, + paper, + base, high school, + like	51	12%	0.103921704
3	+ test, quizzes, + test, + question, + fail, + thing, can, well, + do, + not, but, + hard, on, when, all, fair, should, + good, + like, + give	78	19%	0.1065503221
4	response, no, + woman, + class, + have	10	2%	0.0512804746
5	+ assignment, only, + year, + make, same, grading, + know, could, more, + give, + grade, other, as, + bad, + receive, grading, + essay, never, unfair, with	83	20%	0.1035279399
6	always, unfairly, fair, + look, never, + grade, + like, 's, into, + work, like, + paper, + feel, + do, + not, + student, + bad, + would, more, + year	55	13%	0.1023508415
7	+ group, + project, + professor, + work, different, people, out, + evaluate, + do, + opinion, when, + time, unfairly, + base, should, other, as, + work, + receive, same	37	9%	0.1047414592
8	wrong, partial, partial credit, choice, multiple choice, wrong answers, + answer, credit, multiple, math, + test, bentley, + problem, but, + would, + thing, high school, + give, no, + feel	24	6%	0.0970475959

Figure 15. Results of clustering (EM) with SAS Text Miner, with 20 descriptive terms.

theme, a common and highly desirable activity in the analysis of qualitative data.

6.1 WordStat

Although WordStat provides a variety of tools to help with finding themes, most of the sophisticated work has to be done by the researcher. One of the tools that WordStat offers is a dictionary that includes all the words in the dataset being analyzed. The dictionary excludes words that do not add meaning to a sentence such as pronouns and prepositions. Moreover, the software can run frequencies on words, which allows the researcher to know the number of times a term occurred. The frequency

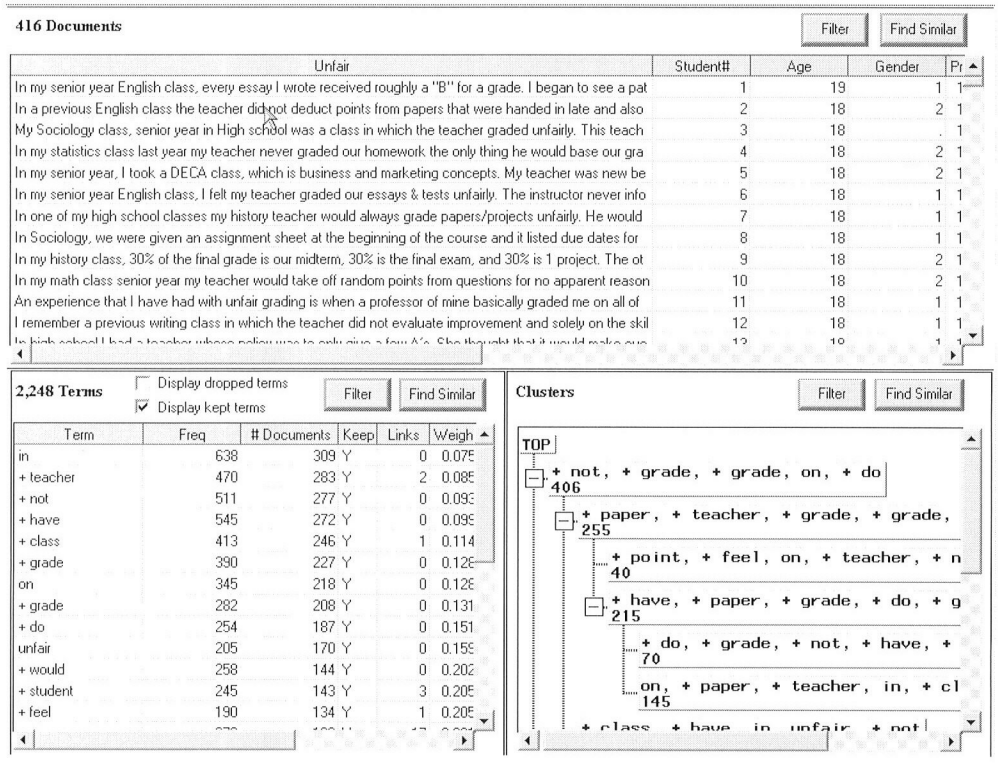


Figure 16. Results of hierarchical clustering with SAS Text Miner; interface.

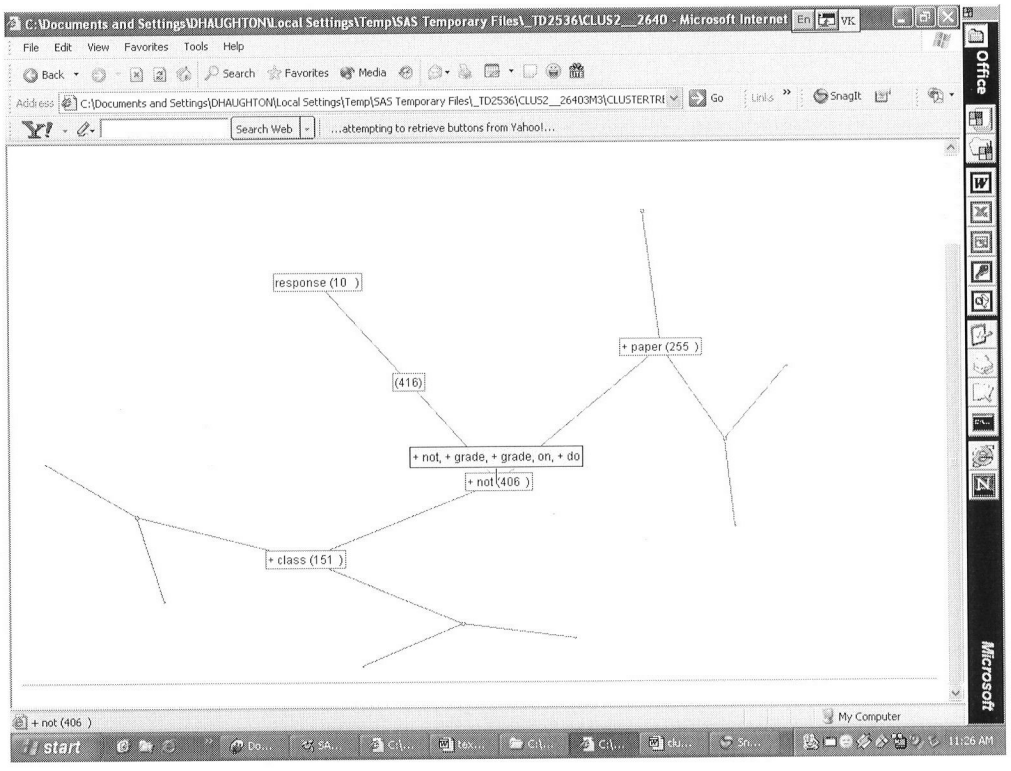


Figure 17. Results of hierarchical clustering with SAS Text Miner; tree view 1.

feature and the availability of a dictionary are intended to help the researcher identify major categories in the data. However, we did not find that searching by word was as helpful as we had anticipated initially. Respondents used the same word in different contexts that meant completely different things. For instance, the software points to the existence of the word fair in the two phrases "this is fair" and "this is not at all fair" when the meanings are totally different. To help overcome this problem, the software allows the researcher to search by clauses. Although

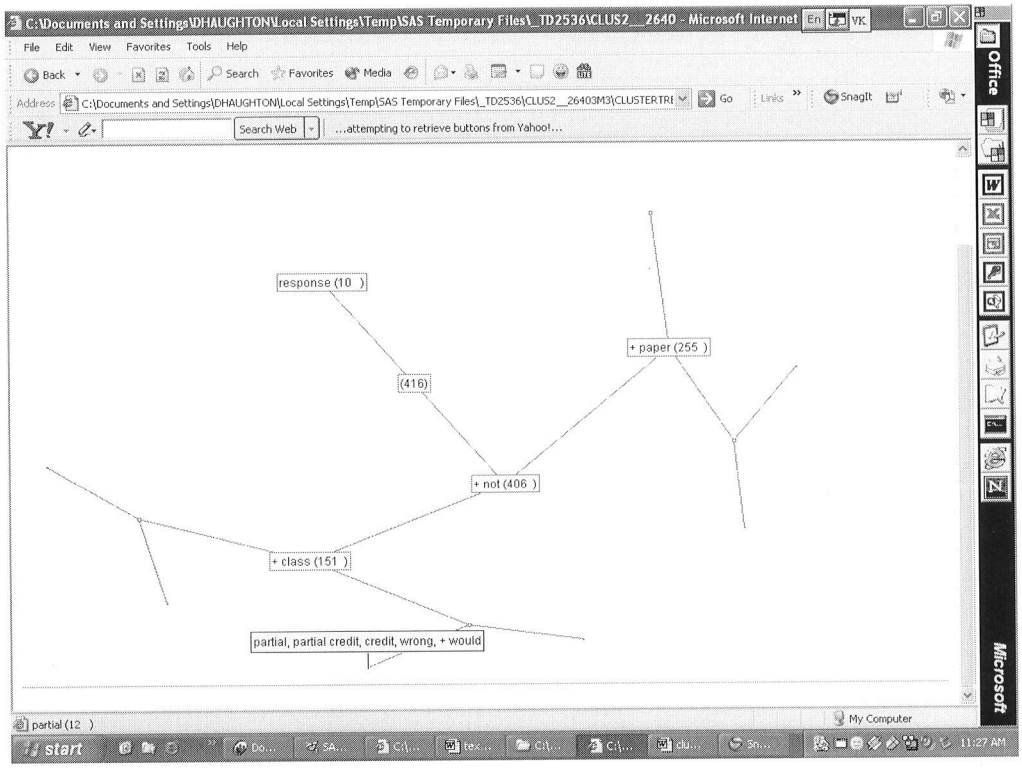


Figure 18. Results of hierarchical clustering with SAS Text Miner; tree view 2.

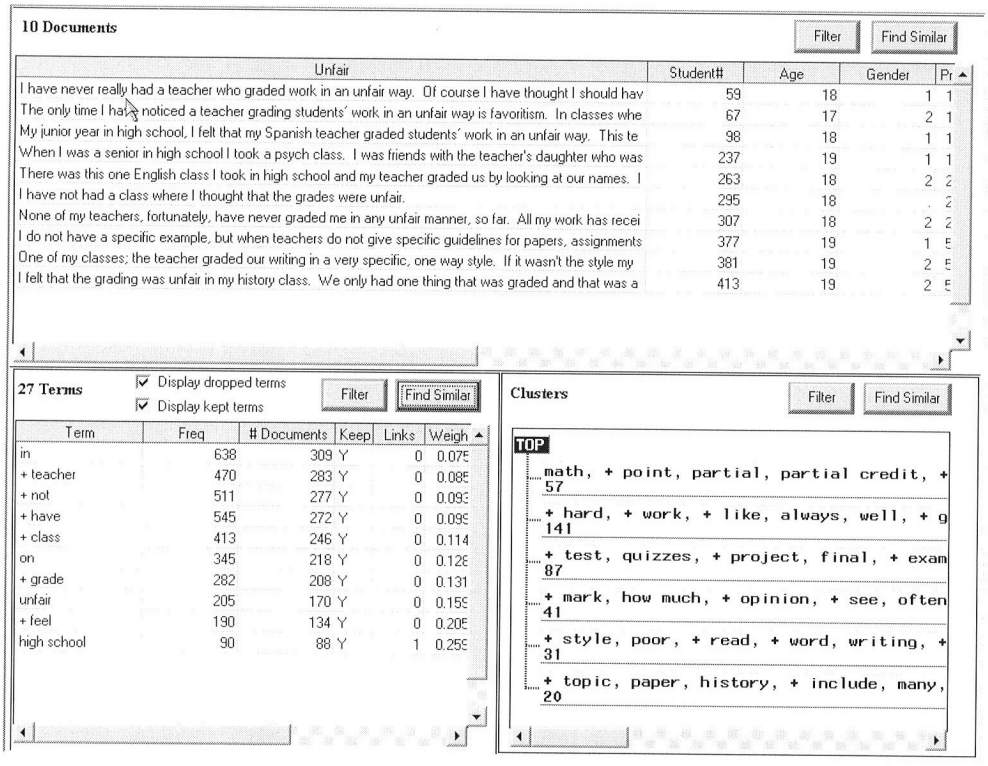


Figure 19. SAS TextMiner listing of 10 most similar documents to "unfair."

helpful in some cases, this feature does not solve the problem already described because the researcher cannot look for chunks of words unless they occur consecutively in a sentence. Hence, searching for "not fair" will not capture something like "not at all fair."

Searching for a word (or phrase) in context is helpful for two main reasons. First, once the identification of major categories is done, searching for words in context allows the researcher to examine whether cases using a specific word really fall under the related category. For example, let us consider a case where the

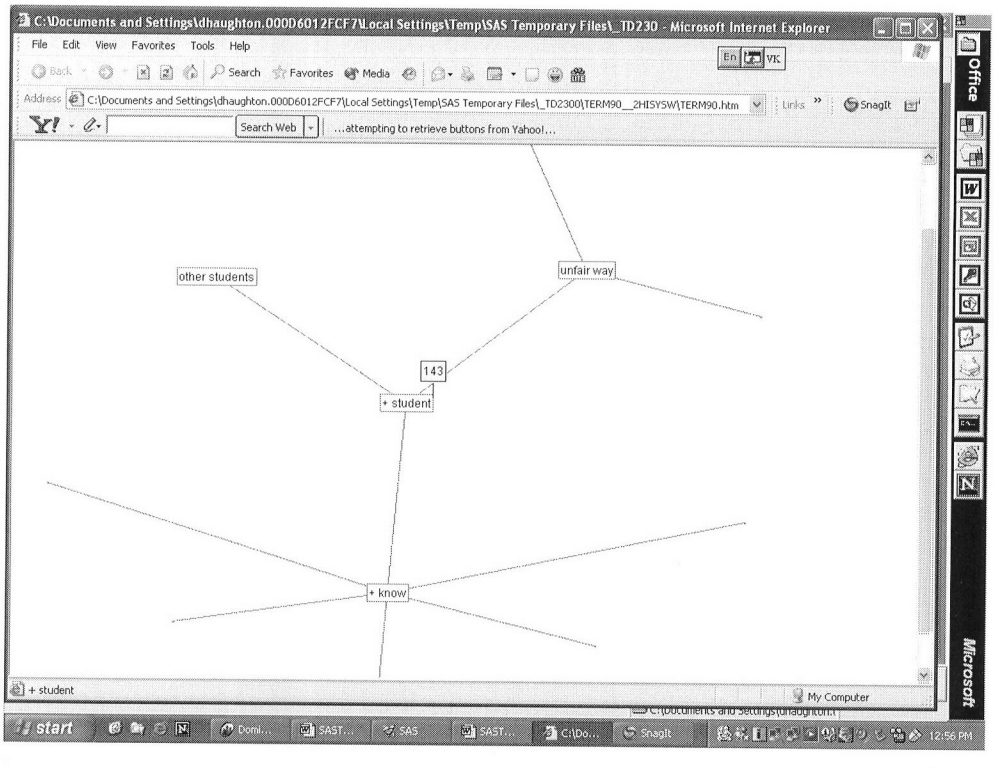


Figure 20. SAS TextMiner concept link tree for "student"; view 1.

researcher had in mind a category entitled "unclear expectations" for grading unfairness with the word "clear" belonging to this category. Searching for the word in context resulted in those two statements: (1) "The instructor never made it clear to us what we should do to get a good grade," and (2) "Although the steps I followed to solve the problem were very clear, the instructor gave me partial credit because he wanted us to use his solution." In this example, reading the word clear in the two contexts shows that the former statement belongs to the category of unclear expectations and the latter does not. Second, searching for word in context makes it easy for the researcher to locate illustrative quotes under a specific theme, a common practice in presenting findings from qualitative data. Note that the researcher needs to review all the contexts to decide on the fit (or lack of) under a certain category.

In addition to the difficulties described above, we faced a few more problems in our attempt to extract themes from the data. The software does not allow the use of the same word in two categories. Although the researcher may find it useful to use the word clear in each of the two categories of "unclear expectations" and "bias," he or she will not be able to do so because the software allows for using the term "clear" under one category only. Moreover, the software does not allow for studying more than one text variable in a given analysis. Hence, the links between answers to two open-ended questions cannot be investigated. For instance, in our case, the researchers could not use the software to examine whether respondents' perceptions of unfair grading practices (as stated in their answers to the first open-ended question) paralleled their perceptions of fair practices (revealed in their answers to the second open-ended question).

In brief, using WordStat was marginally helpful in extracting themes from the data. As explained above, the researcher needs to do most of the work manually. Although the software helps in locating and counting words and organizing responses under preset categories, it does so in a mechanically "blind" way. Sophisticated categorization is done by the researcher after actually reading through the data. Additionally, the researcher needs to carefully check every single sorting done by the software.

6.2 SAS TextMiner

Although SAS TextMiner provides a variety of tools to help with finding preliminary themes, the researcher needs to synthesize, confirm (or disconfirm), refine, and finalize those themes. The key characteristic of SAS TextMiner is its ability to cluster documents and not just terms inside documents. Grouping documents together and identifying the set of common terms that led to this grouping allows researchers the possibility of identifying, at a preliminary level, possible themes. This process, however, involves two key loopholes. First, the set of common terms (see Figure 15 for examples) might not be meaningful, that is, extracting a theme from the given terms is not always straightforward. Second, once a preliminary theme is identified, the researcher still needs to read all of the related documents to ensure that they indeed fit the posited theme.

Similar to WordStat, SAS TextMiner offers many features intended to help the researcher mine the data. The software can run frequencies on words, which allows the researcher to know the number of times a term occurred. It also allows the researcher to search by clauses. Although helpful in some cases, those features are not as helpful as one would think for reasons discussed earlier under WordStat. Searching for words in context is possi-

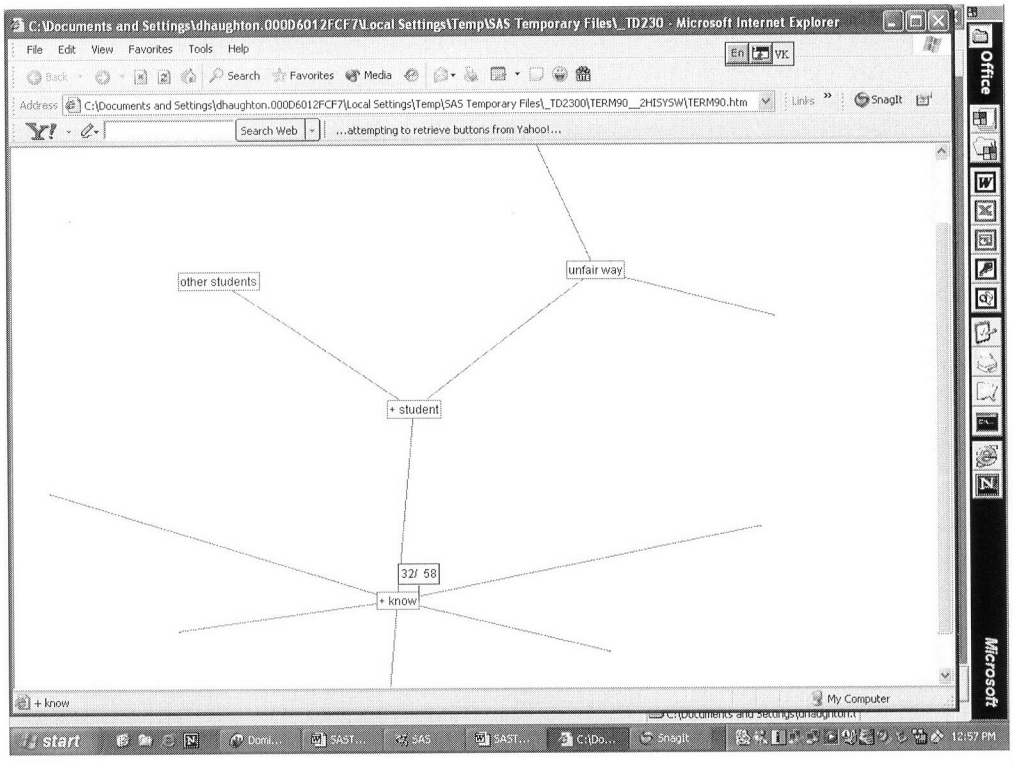


Figure 21. SAS TextMiner concept link tree for "student"; view 2.

ble with SAS TextMiner; this allows for checking whether a case fits under a particular theme. It also helps in locating illustrative quotes for results presentation.

In brief, using SAS TextMiner was helpful, but in a limited way, in extracting themes from the data. The fact that the software groups documents, that is, responses, based on a set of common terms allows for the possibility of identifying preliminary themes. The researcher, however, needs first to make sense of the set of common terms in each group or cluster of responses. The researcher then needs to read all the responses and decides on the fit of each response under a specific theme. Hence, while the software helps in locating and counting words and clustering responses, the researcher needs to do most, if not all, of the time-consuming work.

7. CONCLUDING REMARKS

In conclusion, both packages offer a variety of features that help researchers mine qualitative data, run associations (e.g., with demographic variables), and present results. In extracting themes from unstructured data, SAS TextMiner has the potential of providing a greater assistance to the researcher than WordStat through helping in the identification of preliminary themes. This advantage is due mainly to the fact that SAS TextMiner clusters similar responses while WordStat follows created dictionaries to locate terms or clauses. It is worth noting that SAS TextMiner is significantly more expensive than WordStat. It is likely that few facilities, or individual users, will be able to afford the investment.

Unfortunately, both packages fall short of saving researchers the time and effort needed to actually read the data. This finding stems from the fact that the software can search for specific terms in documents or categorize documents based on common terms. Respondents, however, may use the same term or combination of terms to mean different things.

Note that the data we used are fully unstructured. Based on varied experiences, respondents were providing an example of unfair grading. Hence, not only was the question open-ended, but the answers could be too varied to anticipate any possible categories of answers. This fact might have added to the difficulties we faced in our attempt to analyze the data. A more structured dataset where possible answers could be more easily categorized would have been easier to analyze using either software.

The limitations that emerged in both packages are inherent to a text analysis where the unit of analysis is the key word. We mention here briefly a new direction in text mining where the unit of analysis is the "event". Events are statements, such as "Organization X is hostile to Person Y." Analyzing events involves a knowledge of entities as well as semantic and syntactic rules, and an understanding of grammatical relationships. This approach has been implemented by the Insightful Corporation, in the product InFact (we refer the reader to the InFact white paper listed in the references). At the present time, its cost is likely to be prohibitive to academic users, and its main users are pharmaceutical and intelligence organizations.

APPENDIX: WORDSTAT DICTIONARY USED IN THIS STUDY

BIAS

- BIAS (1)
- BIASED (1)
- BIASES (1)
- DIDN'T_LOOK_AT_OUR_NAME (1)
- DISLIKE_YOU (1)
- DISLIKED_CERTAIN_STUDENTS (1)
- DISLIKED_ME (1)
- DISLIKED_THE_STUDENT (1)
- DISLIKED_US (1)
- FAVOR (1)
- FAVORED (1)
- FAVORING (1)
- FAVORITE_STUDENTS (1)
- FAVORITES (1)
- FAVORITISM (1)
- FIRST_IMPRESSIONS (1)
- HATED (1)
- IMAGE_OF_THE_PERFECT (1)
- JUDGE (1)
- JUDGMENT (1)
- JUDGMENT (1)
- LIKED (1)
- LIKES (1)
- NAME_ON_THE_PAPER (1)
- NO_MATTER_HOW (1)
- NO_MATTER_WHAT (1)
- PARTIALITY (1)
- PERSONAL_AGENDA (1)
- PERSONAL_BELIEFS (1)
- PERSONAL_OPINION (1)
- PERSONAL_VIEWS (1)
- SAME_ANSWER (1)
- SAME_GRADE (1)
- THEIR_FAVORITE (1)
- SEE_A_PATTERN (1)

PARTIAL_CREDIT

- PARTIAL_CREDIT (1)

GRAMMAR

- GRAMMAR (1)
- GRAMMATICAL (1)

MATH

- ALGEBRA (1)
- MATH (1)
- CALCULUS (1)
- STATISTICS (1)

ENGLISH

- ENGLISH (1)
- EXP_101 (1)

EXP_201 (1)
EXP101 (1)
EXP201 (1)
LITERATURE (1)
PHILOSOPHY (1)
POETRY (1)
EXPOS (1)

SOCIAL_SCIENCE

AMERICAN_STUDIES (1)
ECON (1)
ECONOMICS (1)
GOVERNMENT (1)
PSYCH (1)
PSYCHOLOGY (1)
SOCIAL_STUDIES (1)
SOCIOLOGY (1)

COMPUTER

COMPUTER (1)
BNC (1)
IT (1)
IT_101 (1)
IT101 (1)

FOREIGN

FOREIGN (1)
FRENCH (1)
SPANISH (1)
LATIN (1)

SCIENCE

ASTRONOMY (1)
BIOLOGY (1)
CHEMISTRY (1)
PHYSICS (1)
SCIENCE (1)

HISTORY

HISTORY (1)

OTHER

CHRISTIAN (1)
COMMUNITY (1)
FILM (1)
HEALTH (1)
LEGAL (1)
THEOLOGY (1)
LAW (1)
WOOD (1)

BUSINESS

ACCOUNTING (1)
BUSINESS (1)
FI_310 (1)
FI310 (1)
FINANCE (1)
GB_102 (1)
GB_103 (1)
GB_201 (1)
GB102 (1)
GB103 (1)
GB201 (1)
GB_203 (1)
GB203 (1)
MARKETING (1)
MANAGEMENT (1)

UNFAIR

HARD (1)
UNFAIR (1)
UNFAIRLY (1)

TEACHER

TEACHER (1)
TEACHERS (1)
INSTRUCTORS (1)
INSTRUCTOR (1)
PROFESSOR (1)
PROF (1)

REFERENCES

Insightful Corporation [cited 1 November 2004], Infact White Paper. Available at http://www.insightful.com/DocumentsLive/22/31/InFact_White_Paper_2.9.pdf.

Miles, M. B. (1979), "Qualitative Data as an Attractive Nuisance: The Problem of Analysis," *Administrative Science Quarterly*, 24, 590.

Provalis Research [cited 1 November 2004], WordStat. Available at http://www.simstat.com/wordstat.htm.

Robb, D. (2004), "Text Mining Tools Take on Unstructured Data," *Computerworld*, June 21.

SAS [cited 1 November 2004], SAS Text Miner. Available at http://www.sas.com/technologies/analytics/datamining/textminer/.