DRAFT

Text Mining:

A Survey Application

Energy Efficiency

Text Mining: A Survey Application
By Sara Saulcy, March 31, 2011
Draft 7

Questionnaires using closed-ended questions are in many ways ideal for both respondents and researchers because they simplify the task of providing and compiling information, respectively. However, they limit the information provided by respondents. Open-ended questions give respondents the opportunity to expand on questionnaire topics and introduce new topics. These provide researchers with richer information than otherwise obtained from closed-ended questions. In this article R&P explores employer responses to an open-ended question about skills needed by newly-hired employees.

In the spring of 2010, Research & Planning (R&P), along with several other state Labor Market Information offices, received research funding to study the workforce and, as a subset, jobs in the workforce that spend time on activities increasing energy efficiency, utilizing or developing renewable energy resources, or preserving and/or restoring the environment, hereafter referred to as *environmental jobs*. R&P designed a mail questionnaire that not only contained a question intended to measure the degree to which a job was involved in any of these environmental activities, but was also designed to capture and assess the types of skills needed for jobs in Wyoming (see cover letter and questionnaire in the Appendix). Skills were assessed using *open-ended* and *closed-ended* questions (see page ? for further discussion). Ultimately, the goal of this project is to determine what types of jobs are available in the state, which ones are likely to be environmental jobs, the skills associated with these jobs, and if there is any significant differences between environmental jobs and jobs that are not environmental. This article

explores a subset of jobs using a process known as *text mining* (Knapp, forthcoming). Please see the text box for definitions for words in italics.

[This section for a text box]

Definitions

Both hire and exit (both): where the employee started a job and worked for a firm only within a quarter.

Categories: a grouping of words that form a common theme.

Closed-ended questions: closed-ended questions limit the number of responses possible that a respondent can provide. Questions that ask respondents to choose a number on a scale are examples of closed-ended questions.

Concepts: words or phrases that are extracted from the text data. Concepts may be included in libraries to aid extraction.

Environmental jobs: those that involve activities and duties related to increasing energy efficiency, utilizing or developing renewable energy resources, or preserving and/or restoring the environment some or all of the time.

Extract, extraction: computer processing of text when data are imported into TAS.

Extraction results: key words and phrases that are extracted from text responses when data are imported into TAS.

Forced in: manually placing concepts into categories.

Forced out: manually removing concepts from categories.

Hire: a person is hired for a job in a specified quarter and is still employed by that employer in the next quarter.

New-hire: someone who was hired by a firm that they had not worked for in at least the last 20 years (the time frame for which Research & Planning has Unemployment Insurance wage records).

Open-ended questions: in open-ended questions, response options are not limited. “Why did you choose Wyoming as a place to live” is an example of an open-ended question.

Library: a set of words or phrases consisting of preexisting concepts and user-identified concepts which aid in extracting data for use in categories.

Rules: statements that can be created to automatically classify records into a category based on a logical expression.

TAS: Text Analytics for Surveys, a software package designed to extract meaning from open-ended questions and other similar data.

Text mining: the process of examining text for themes that can then be quantified.

With the research about text mining, R&P is seeking to accomplish three goals:

- Demonstrate the application of text mining software to survey research;
- Identify skills needed by today’s job seekers in general, and specifically as skills relate to environmental jobs; and
- Research which informs employees, educators, policy makers, and training providers about the skills necessary to get a job in today’s labor market.

For the purpose of this study, R&P was interested in sampling from only those employees that were designated *hires* excluding those that fell in the *both* category. Specifically, we only included employees that were considered a *new-hire* during the quarter of interest. Rehires were

excluded to control for the confounding effects of seasonal re-hiring and to eliminate circumstances where employers and employees based hiring decisions on prior joint human capital and business investment. Finally, R&P was most interested in including new-hires that were retained by the same employer for at least two quarters. These jobs were probably more likely to require a training or educational investment by the employer. This was so R&P could also track what kinds of jobs employers were hiring for and the skills required for those positions (see Knapp, forthcoming, for additional discussion of the survey and sampling methodology).

The questionnaire was comprised of two types of questions: closed-ended questions and open-ended questions. Closed-ended questions limit the number of responses possible that a respondent can provide. These types of questions force respondents to choose from a limited number of responses. An example of a closed-ended question is “What time is it where you live?” In open-ended questions, response options are not limited. “What is your opinion of Wyoming’s economy?” is an example of an open-ended question. To assess employer skills needs, R&P first asked employers closed-ended questions about five types of skills: service orientation; critical thinking; reading comprehension; technology design; and operation and control (see Table 2 and page ? for definitions). These were selected after cognitive interviews were conducted by the Wyoming Survey & Analysis Center (WYSAC, 2010). Cognitive testing helps to determine if a questionnaire is serving its intended purpose. The five skills on the original questionnaire were chosen from the most frequently occurring skills for environmental jobs. Cognitive testing revealed that the five skills were measuring the same concept. The questionnaire was revised to include skills that contrasted with one another such that the five skills would be measuring the importance of different skills.

R&P then asked employers to answer an open-ended question about which skill they considered most important for the job. The question indicated that it could be one of the five skills previously mentioned or another skill. Although the questionnaire asked employers to provide a single skill, several employers reported two or more skills having equal importance.

In order to evaluate respondents' answers to the open-ended question, R&P used a process known as *text mining*. Text mining is a useful tool for evaluating and quantifying responses to open-ended questions. The process helps to identify themes that cannot otherwise be determined from closed-ended questions. The purpose of identifying themes is to capture information based on what respondents consider important, not what researchers consider important. For large surveys (over 5,000 responses in this case), text mining by hand is impractical. R&P used text mining software to expedite the process of capturing common themes reported by employers about skills needed to be successful in jobs for which employees were newly hired. The process was then supplemented by reviewing *concepts* in individual records.

Methodology

R&P used PASW Text Analytics for Surveys 4 (*TAS*), text mining software from SPSS Inc. (http://www.spss.com/software/statistics/text-analytics-for-surveys/). Data originated from the New Hires Survey (see questionnaire in the Appendix) for fourth quarter 2009 and first quarter 2010 and were entered into a SQL server database using a form in Visual Basic.

Responses were then imported into TAS and concepts were extracted from the text. The five skills from questions six through ten of the questionnaire were entered into a *library*. The five skills are:

- Service orientation – actively looking for ways to help people.
- Critical thinking – using logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions or approaches to problems.
- Reading comprehension – understanding written sentences and paragraphs in work related documents.
- Technology design – generating or adapting equipment and technology to serve user needs.
- Operation and control – controlling operations of equipment or systems.

These skills were also entered as *categories*. In addition to the five skills R&P added 19 other skills listed in O*NET OnLine (http://online.onetcenter.org/find/descriptor/browse/Skills/) as categories for a total of 24 skills. Selected skills were combined with one another because of content overlap (see Table 2 for a complete list of skills, definitions, and skills that were combined with one another). R&P used O*NET to describe skills known in a nationally-known context. O*NET skills are used by employers, job seekers, career counselors, and researchers to help assess skills.

The first pass through the data R&P took the *extraction results* that were synonymous placements and grouped them into the 24 categories. For example, if an employer reported that

critical thinking was the most important skill for the job then the response was placed into the critical thinking category. The software did not always pick up on which categories to place concepts. R&P then reviewed the data to determine which categories to place records. “Common sense” is an example of a concept which TAS did not automatically place in the category of critical thinking. R&P then took this concept and placed it into the critical thinking category. In some instances concepts did not appear in the list of extraction results to be placed into categories. Records were reviewed individually and, where necessary, records were *forced into* categories. Alternatively some records were placed by TAS in categories where they did not belong. In these instances records were *forced out*. Although TAS expedites the process of categorizing concepts through *rules*, *libraries*, and other resources, the process of determining which categories concepts belong to is a subjective process (SPSS, Inc., 2010, p. 2). Of the 5,723 records, 5,331 9(93.2%) were placed in one or more categories. These are the records R&P used to develop frequency distributions and chi square statistics.

Results

Table 1 shows that 17.4% (926) of newly hired employees were in environmental jobs (those that involve activities and duties related to increasing energy efficiency, utilizing or developing renewable energy resources, or preserving and/or restoring the environment some or all of the time). Included in Table 1 are the five skills from the questionnaire (service orientation; critical thinking; reading comprehension; technology design; and operation and control); the table for the remaining skills are available online at http://doe.state.wy.us/LMI/ . Of the five skills, critical thinking had the highest frequency of importance for environmental jobs (247 of 926; 26.7%)

followed by service orientation (185 or 20.0%). Note that, although critical thinking was more important for environmental jobs relative to non-environmental jobs, critical thinking is still a skill important for all jobs.

Figure 1 illustrates a subset of Table 1. It shows the percent of jobs for which selected skills were reported as important by whether or not the jobs involve any environmental tasks (values are the row percents the middle column of Table 1). As an example, 19.9% (second value down in the column) of environmental jobs reported the skill of operation and control as important compared to 14.1% of jobs that do not involve environmental tasks (fifth value down in the column). Of the five skills, service orientation was the most frequently reported skill as important for jobs that do not involve environmental skills (28.9%). In contrast, critical thinking is the most important for jobs that involve environmental tasks at least some of the time (26.7%).

Figure 1 clearly shows the difference in the percent of jobs requiring the selected skills relative to the jobs involvement in environmental tasks. While visually demonstrating the differences is a good first step, mathematically proving that the differences are significant requires the application of a statistical technique called a Chi Square. Our assumption is that there is no difference between the skill requirements of a environmental job and a non-environmental job, in scientific terms this is referred to as the null hypothesis. The chi square tests our assumptions by comparing what we expect (i.e. no difference) to what we actually observe (the results of the survey questionnaire). The output of the chi square analysis is a probability or likelihood (always presented as p) that the differences observed were do to random chance, the lower the p

value the more significant the differences. All of the skills that were presented in Figure 1 were statistically significant; the results of each test are discussed below.

Table 1 shows the results of chi square tests to determine if the relationship between each of the skills and employment in environmental jobs was significant. Critical thinking was found to be significant at the .0001 level, as was operation and control and service orientation. This means that the odds of the result being caused by chance are .01%. Technology design was significant at the .001 level while reading comprehension was significant at the .05 level. Other relationships found to be significant included instructing, negotiation, coordination, and complex problem solving at the .05 level, troubleshooting at the .01 level, and science at the .0001 level (see Table online at http://doe.state.wy.us/LMI) . Cell chi square values indicate how much a particular cell contributes to the total chi square. The higher the cell chi square, the more it adds to the overall chi square value. For example, for operation and control, the cell chi square was highest for "job involves environmental tasks at least some of the time" with a value of 14.081. This tells us that, of the four cells of the cross tabulation, "operation and control" crossed with "job involves environmental tasks at least some of the time" had the most influence on the chi square value of 20.0669, and thus the result that operation and control is significant at the .0001 level. The overall chi square results indicate that several skills, such as critical thinking, are important for all jobs, not just environmental jobs. However, in the case of environmental jobs, relative to jobs generally, critical thinking is more important.

Although employers were asked to report which one skill was most important, many reported that two or more skills were important. Table 3 shows the number of co-occurrences of skills

reported as important. The table tells us that multiple skills are important for employees to have, not just one. Critical thinking had a high frequency of occurring with other skills. Of the first five highest co-occurrences, critical thinking appears four times with operation and control, service orientation, reading comprehension, and technology design.

Figure 2 on page ? (called a category web graph) describes relationships between reading comprehension, critical thinking, technology design, operation and control, and other skills. These are the four skills reported as most important for employees working in environmental jobs. The thicker the line between two skills the more responses two skills had in common. As the figure illustrates, there is a strong association between reading comprehension, critical thinking, technology design, and operation and control. In addition these skills are strongly linked to service orientation. These are the five skills asked about in the closed-ended questions on the questionnaire. The concept map demonstrates that, for any given job, many skills are important

Discussion

Although we asked employers to report the one skill that was most important, as Figure 2 shows, multiple skills may be equally important for a new hire to have. This is where text mining is helpful in providing information that could not be obtained from the closed-ended questions. Information such as that learned from the survey may help to inform prospective employees about the skill sets that they need to successful in the labor market.

It is possible that asking employers about five skills predisposed respondents to answer the open-ended question using one or more of the five skills. The three most frequently occurring responses to the open-ended question were service orientation (1,496), critical thinking (1,021), and operation and control (821). However, the fact that skills other than those listed on the survey (instructing, negotiation, coordination, complex problem solving, troubleshooting, and science) were significant suggests that other skills are important for newly-hired employees to have as well.

Conclusion

R&P used text mining software to help determine what skills newly hired employees need to be successful in the labor market, in particular for environmental jobs. The skills on the questionnaire were found to be significantly related to whether or not a job was environmental, as were other skills on which employers reported. Many skills are related to one another and frequently described as equally important by employers. This information can help employees identify the skill sets they need to be successful in Wyoming's labor market. In addition, this information helps to inform educators, policy makers, and training providers about the skills necessary to get a job. The usefulness of the approach on the questionnaire will be expanded to investigate skills needs described in the Wyoming at Work database (https://www.wyomingatwork.com/). The Wyoming at Work database is an online resource overseen by the Wyoming Department of Workforce Services for employers and job seekers. Employers can register job openings and job seekers can post resumes for review by employers. R&P will review and analyze employer job postings for skills needed by new hires using the

categories already developed in TAS. As with the questionnaire, R&P is seeking to determine if there are any differences between environmental job postings and other jobs. We anticipate that using the previously constructed categories, rules, and libraries in TAS will expedite our analysis.

In addition to the Wyoming at Work database, there are other questions for future research. One is exploring whether or not listing skills in advance of the open-ended question about the most important skills predisposes respondents to answering in a certain way. Another question to be addressed is whether knowledge and abilities should be included in addition to or in lieu of skills.

References

Knapp, L. (Forthcoming). Trends Article – Methodology. *Wyoming Labor Force Trends*.

SPSS, Inc. (2010). *IBM SPSS Text Analytics for Surveys 4.0 User's Guide*.

WYSAC. (2010). *Cognitive Interviews for the Wyoming Department of Employment: Testing a Job Skills Questionnaire*, by T. Furgeson & M. Dorssom. (WYSAC Technical Report No. SRC-1014). Laramie, WY: Wyoming Survey & Analysis Center, University of Wyoming.

Table 1: Wyoming Environmental Jobs by Importance of Selected Skills[a], 2010

	Operation and Control		
Environmental Job	Not Reported as Important	Reported as Important	Total
Job involves environmental tasks at least some of the time	742	184	926
Row Pct	80.1%	19.9%	
Cell Chi-Square	2.5007	14.081	
Job does not involve any environmental tasks	3,785	620	4,405
Row Pct	85.9%	14.1%	
Cell Chi-Square	0.5257	2.9599	
Total	4,527	804	5,331
Percent	84.9%	15.1%	100.0%

Statistic	DF	Value	Prob
Chi-Square	1	20.0669	<.0001

	Technology Design		
Environmental Job	Not Reported as Important	Reported as Important	Total
Job involves environmental tasks at least some of the time	893	33	926
Row Pct	96.4%	3.6%	
Cell Chi-Square	0.1872	8.4921	
Job does not involve any environmental tasks	4,323	82	4,405
Row Pct	98.1%	1.9%	
Cell Chi-Square	0.0394	1.7852	
Total	5,216	115	5,331
Percent	97.8%	2.2%	100.0%

Statistic	DF	Value	Prob
Chi-Square	1	10.5038	0.0012

	Reading Comprehension		
Environmental Job	Not Reported as Important	Reported as Important	Total
Job involves environmental tasks at least some of the time	880	46	926
Row Pct	95.0%	5.0%	
Cell Chi-Square	0.1423	3.6495	
Job does not involve any environmental tasks	4,251	154	4,405
Row Pct	96.5%	3.5%	
Cell Chi-Square	0.0299	0.7672	
Total	5,131	200	5,331
Percent	96.3%	3.8%	100.0%

Statistic	DF	Value	Prob
Chi-Square	1	4.5888	0.0322

Table 1 continued

	Service Orientation		
Environmental Job	Not Reported as Important	Reported as Important	Total
Job involves environmental tasks at least some of the time	741	185	926
Row Pct	80.0%	20.0%	
Cell Chi-Square	6.9995	18.558	
Job does not involve any environmental tasks	3,130	1,275	4,405
Row Pct	71.1%	28.9%	
Cell Chi-Square	1.4714	3.9012	
Total	3,871	1,460	5,331
Percent	72.6%	27.4%	100.0%

Statistic	DF	Value	Prob
Chi-Square	1	30.9304	<.0001

	Critical Thinking		
Environmental Job	Not Reported as Important	Reported as Important	Total
Job involves environmental tasks at least some of the time	679	247	926
Row Pct	73.3%	26.7%	
Cell Chi-Square	7.5329	33.113	
Job does not involve any environmental tasks	3,664	741	4,405
Row Pct	83.2%	16.8%	
Cell Chi-Square	1.5835	6.9608	
Total	4,343	988	5,331
Percent	81.5%	18.5%	100.0%

Statistic	DF	Value	Prob
Chi-Square	1	49.1897	<.0001

[a]Skills are from O*NET. See http://online.onetcenter.org/find/descriptor/browse/Skills/ for more information.

Table 2: O*NET Skills Used in Wyoming New Hires Survey

Skill	Definition
Active learning	Understanding the implications of new information for both current and future problem-solving and decision-making.
Active listening	Giving full attention to what other people are saying, taking time to understand the points being made, asking questions as appropriate, and not interrupting at inappropriate times.
Complex problem solving	Identifying complex problems and reviewing related information to develop and evaluate options and implement solutions.
Coordination	Adjusting actions in relation to others' actions.
Critical thinking	Using logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions or approaches to problems.
Equipment maintenance	Performing routine maintenance on equipment and determining when and what kind of maintenance is needed; included with operation and control.
Equipment selection	Determining the kind of tools and equipment needed to do a job; included with operation and control.
Installation	Installing equipment, machines, wiring, or programs to meet specifications.
Instructing	Teaching others how to do something.
Judgment and decision making	Considering the relative costs and benefits of potential actions to choose the most appropriate one.
Learning strategies	Selecting and using training/instructional methods and procedures appropriate for the situation when learning or teaching new things; included with critical thinking.
Management of financial resources	Determining how money will be spent to get the work done, and accounting for these expenditures; included with management of material resources.
Management of material resources	Obtaining and seeing to the appropriate use of equipment, facilities, and materials needed to do certain work.
Management of personnel resources	Motivating, developing, and directing people as they work, identifying the best people for the job; included with management of material resources.
Mathematics	Using mathematics to solve problems.
Monitoring	Monitoring/assessing performance of yourself, other individuals, or organizations to make improvements or take corrective action; included with critical thinking.
Negotiation	Bringing others together and trying to reconcile differences.
Operation and control	Controlling operations of equipment or systems.
Operation monitoring	Watching gauges, dials, or other indicators to make sure a machine is working properly; included with operation and control.
Operations analysis	Analyzing needs and product requirements to create a design; included with critical thinking
Persuasion	Persuading others to change their minds or behavior.
Programming	Writing computer programs for various purposes; included with critical thinking.
Quality control analysis	Conducting tests and inspections of products, services, or processes to evaluate quality or performance; included with critical thinking.
Reading comprehension	Understanding written sentences and paragraphs in work-related documents.
Repairing	Repairing machines or systems using the needed tools.
Science	Using scientific rules and methods to solve problems.
Service orientation	Actively looking for ways to help people.
Social perceptiveness	Being aware of others' reactions and understanding why they react as they do.
Speaking	Talking to others to convey information effectively.

Table 2 continued

Systems analysis	Determining how a system should work and how changes in conditions, operations, and the environment will affect outcomes; included with critical thinking.
Systems evaluation	Identifying measures or indicators of system performance and the actions needed to improve or correct performance, relative to the goals of the system.
Technology design	Generating or adapting equipment and technology to serve user needs.
Time management	Managing one's own time and the time of others.
Troubleshooting	Determining causes of operating errors and deciding what to do about it.
Writing	Communicating effectively in writing as appropriate for the needs of the audience.

Adapted from O*NET OnLine, http://www.onetonline.org/find/descriptor/browse/Skills/

Table 3: Number of Co-Occurrences of Skills Reported as Important for the Wyoming New Hires Survey

Category 1 (Total Responses)	Category 2 (Total Responses)	Number
Critical Thinking(1021)	Operation & Control(821)	146
Service Orientation(1496)	Critical Thinking(1021)	119
Reading Comprehension(202)	Critical Thinking(1021)	81
Operation & Control(821)	Service Orientation(1496)	70
Critical Thinking(1021)	Technology Design(116)	69
Reading Comprehension(202)	Service Orientation(1496)	67
Reading Comprehension(202)	Operation & Control(821)	66
Technology Design(116)	Operation & Control(821)	66
Reading Comprehension(202)	Technology Design(116)	65
Technology Design(116)	Service Orientation(1496)	65
Critical Thinking(1021)	Judgement & Decision Making(95)	33
Critical Thinking(1021)	Time Management(436)	25
Active Listening(507)	Critical Thinking(1021)	22
Critical Thinking(1021)	Complex Problem Solving(63)	7
Time Management(436)	Judgement & Decision Making(95)	6
Critical Thinking(1021)	Social Perceptiveness(75)	5
Judgement & Decision Making(95)	Operation & Control(821)	5
Service Orientation(1496)	Judgement & Decision Making(95)	5
Service Orientation(1496)	Active Listening(507)	4
Math(83)	Critical Thinking(1021)	3
Operation & Control(821)	Time Management(436)	3
Service Orientation(1496)	Time Management(436)	3
Critical Thinking(1021)	Personnel Management(15)	2
Critical Thinking(1021)	Coordination(61)	2
Repairing(59)	Critical Thinking(1021)	2
Active Learning(19)	Service Orientation(1496)	1
Active Learning(19)	Critical Thinking(1021)	1
Active Listening(507)	Judgement & Decision Making(95)	1
Active Listening(507)	Social Perceptiveness(75)	1
Active Listening(507)	Operation & Control(821)	1
Critical Thinking(1021)	Installation(24)	1
Critical Thinking(1021)	Persuasion(9)	1
Installation(24)	Time Management(436)	1
Installation(24)	Social Perceptiveness(75)	1
Instructing(48)	Critical Thinking(1021)	1
Judgement & Decision Making(95)	Social Perceptiveness(75)	1
Judgement & Decision Making(95)	Persuasion(9)	1
Math(83)	Service Orientation(1496)	1
Reading Comprehension(202)	Active Listening(507)	1
Repairing(59)	Service Orientation(1496)	1
Repairing(59)	Operation & Control(821)	1
Service Orientation(1496)	Complex Problem Solving(63)	1
Service Orientation(1496)	Social Perceptiveness(75)	1
Speaking(18)	Critical Thinking(1021)	1
Systems Evaluation(73)	Critical Thinking(1021)	1
Systems Evaluation(73)	Time Management(436)	1
Time Management(436)	Social Perceptiveness(75)	1

DRAFT

Figure 1: Percent of Wyoming Jobs With Selected Skills Reported as Important by Whether or Not Job Involves Environmental Tasks, Fourth Quarter 2009-First Quarter 2010