ADSURGO
ascend with analytics

Harness the Power of JMP®: Big Data and Social Media for Competitor Analytics
JMP Discovery Conference 2015

James Wisnowski, Adsurgo LLC
Flor Castillo, SABIC
Andrew Karl, Adsurgo LLC
Heath Rushing, Adsurgo LLC

Abstract:

Relative to the global population, there are now twice as many sensors (e.g., smartphones) in the Internet of Things and an equivalent number of social media accounts. The resultant explosive growth in data volume represents great opportunities for those who can find a signal in this vast collection of unstructured and uncertain noise. A recent project presented the need for JMP users to efficiently search and analyze large volumes of open source material on innovative technologies, associated maturity levels, current applications and primary competitors. Unique challenges included the need to extract the text from these diverse website architectures (absent the extraneous code) and from attached PDFs and PowerPoint files that are not always in a usable format. Once the corpus was developed, the text was imported into JMP to quickly identify critical themes of interest and associated documents. These steps were accomplished using an intuitive JMP interface to text mining routines in R. Text analytics using JMP data visualization, cluster analysis and decision tree capabilities enabled better characterization of the market over traditional methods. This talk will demonstrate a series of JMP scripts and platforms to effectively search scores of relevant websites and monitor social media outlets (such as Twitter), assemble a large collection of unstructured data, and answer focused research questions.

Introduction:

Though the term competitive intelligence (CI) has only been used in the last few decades to describe the acts of a corporation to acquire information about other organizations posing business threats, the practice has been around for centuries. Analogous to military intelligence, we can think of a business focusing on the external forces of competitors, products, and customers to accurately assess what strategy it needs to execute in order to thrive in challenging environments. We can think of competitive intelligence in terms of both the strategic environment governed by economic and market forces as well as the tactical intelligence required everyday (e.g matching a competitor's price). Counter-competitive intelligence is where business are also going to want to protect their own proprietary and other sensitive information. Tools and data sources need to keep pace with the dynamically evolving methods to effectively gain knowledge of business rivals and anticipate their future responses. This paper illustrates the value of JMP software from SAS Institute to both acquire and analyze data in the competitive market analytics space.

Competitive Intelligence Cycle:

The Society of Competitive Intelligence Professionals (SCIP) defines CI as *a systematic and **ethical** program for gathering, analyzing, and managing external information that can affect [the firm’s] plans, decisions, and operations. Specifically, [CI] is the **legal** collection and analysis of information regarding the capabilities, vulnerabilities, and intentions of business competitors, conducted by using information databases and other ‘**open** sources’ and through **ethical** inquiry.* The bold words highlight that this practice is not espionage or illegal, and in fact should be part of any business’s strategy.

The competitive intelligence cycle in Figure 1 is one of many constructs based on the military intelligence model to describe the process of finding exploitable market information on other businesses and applying analytics to discover actionable information. It starts with the definition phase figuring out exactly what dimensions you really want to investigate so you can develop a solid plan of attack. The second phases focuses on obtaining the correct data sources to shape the research. The third phase seeks to apply analytical methods to transform the raw data into insights ultimately producing a compelling story describing the landscape. The last phase involves feeding this information to the right teams who can act upon the intelligence.

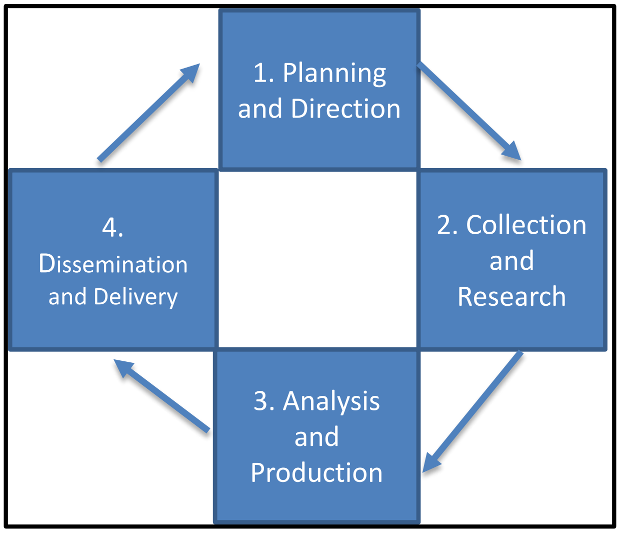

1. Planning and Direction
2. Collection and Research
3. Analysis and Production
4. Dissemination and Delivery

Figure 1. The Competitive Intelligence Cycle

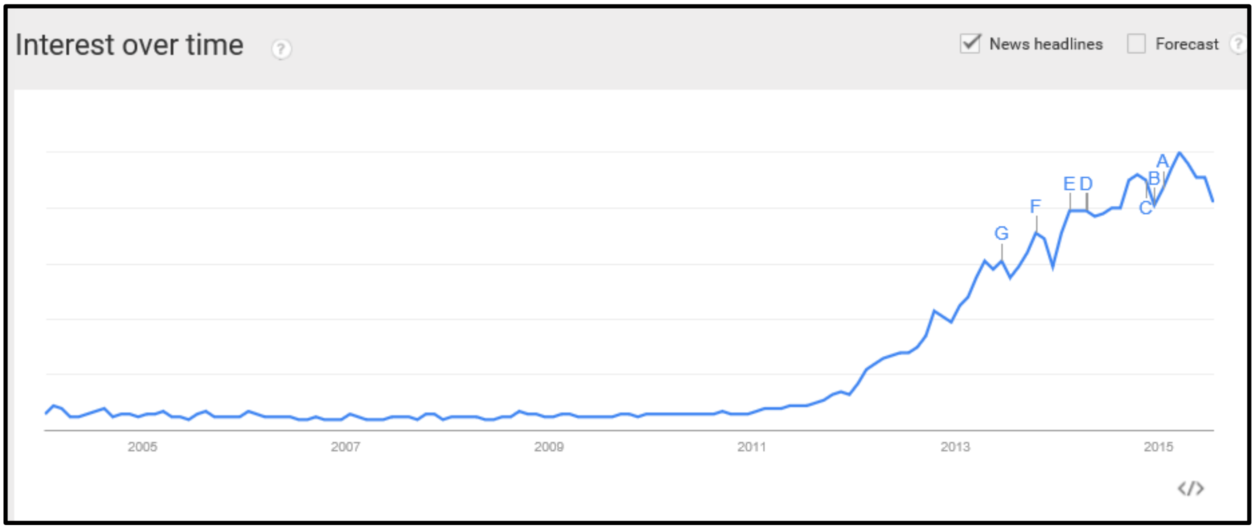
Data is what fuels the CI process. In the past, only CI professionals had access to meaningful data and sites. That has now changed with Big Data where everyone in the organization can bring something of value from their domain of expertise by focused “trawling.” Sources of CI data include annual reports, speeches, news releases (internal and external), financial analyses, trade shows, patent filings, company website, and social media to name a few. Often these are “open source”, though you may need special permissions to gain access. Numerous proprietary databases exist that may even have some analytics already available such as Dunn and Bradstreet’s Hoover, Salesforce, DataSift, Thomson Reuters, etc. Much of the raw data will come from internet searches followed by careful extraction of the pertinent data elements. CI collection efforts will necessarily have the challenge of unstructured text data from disparate sources and formats requiring significant effort to assemble a coherent data set for analytics.

The analysis and production phase typically organizes the multiple sources of data into any of a number of constructs. The well-known SWOT technique will look at competitors' strengths, weaknesses, opportunities, and threats. A PESTLE method evaluates the political, economic, social, technological, legal, and environmental dimensions for an industry and rivals. Harvard Professor Michael Porter's 5 Forces (barriers to entry, current rivals, substitute threats, supplier bargaining power, buyer bargaining power) has endured the test of time. Porter also has a Four Corners model (drivers/culture, current strategy, management assumptions, and capabilities) to predict competitors' future moves. Other analytical frameworks form matrices with rivals as the columns and key performance indicators such as distribution channels, technological edge, pricing, market share, customer focus, financial stability, workforce, facilities, partnerships and so forth as the row. The challenge with any of these frameworks is accurately populating the data elements as many are qualitative assessments dependent how well you can find credible data sources. We discuss a few methods in JMP that can help find and explore this critical data.

Competitive Intelligence Data from Website Analyses:

Avanash Kausik describes many tools available to analyze a web's traffic to get surprisingly insightful information on sources, demographics, bounce dynamics, and product sentiment. He breaks these down into the site-centric which look at competitor websites versus eco-system centric that looks at the industry and specific technologies. Examples of site-centric applications are Google Analytics, Compete, SimilarWeb, Marketing Grader, Majestic, and AdBeat (for on-line advertising metrics). These tools provide detailed information about each site (e.g sources, bounce rate, pages visited, keywords,…), though unfortunately it is only free for your own site. One example of an eco-system centric tool is Google Trends (*https://www.google.com/trends/)*. Figure 2 shows the presence of the term "Big Data" over time which, not surprisingly, has taken off in the past 3 years. On the plot are letters that choose a representative story from the web at that point in time that may provide insight as to why there is a peak or valley. Additionally, the top geographic regions and correlated search terms (e.g. big data Hadoop) are provided.

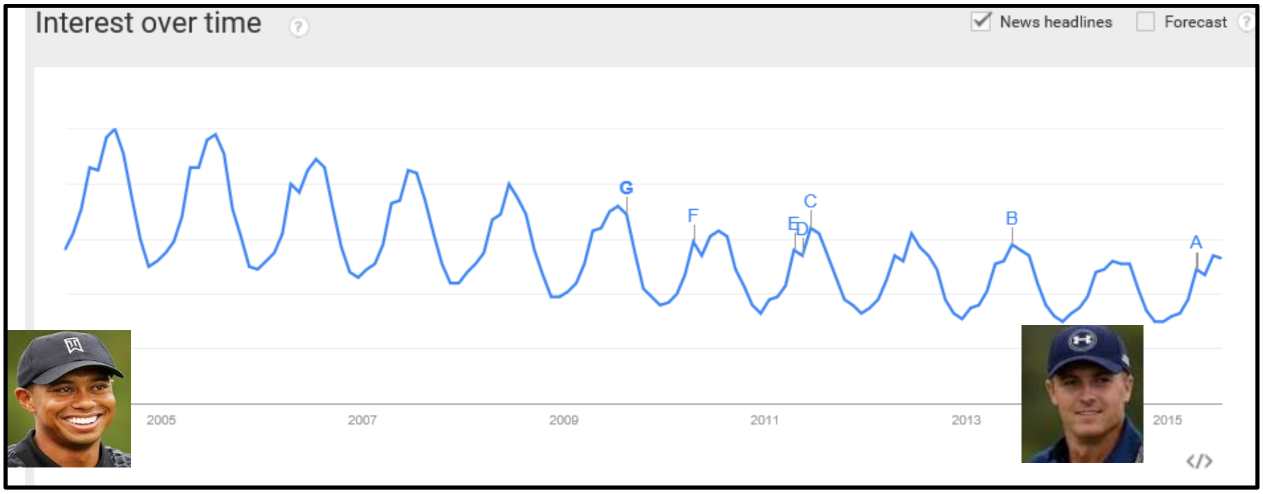
Google Trends can export the data which allows us to use the capabilities of JMP for our own analyses. As an example, Jordan Spieth is a 22 year old golfer who has had an outstanding year recently reaching the world's #1 ranking. It is not uncommon to hear in sports circles that golf as a sport is declining in interest and maybe Spieth can turn this trend around. The plot in Figure 3 certainly shows not only a downward trend, but also a very cyclical pattern that may be of use in our competitive analytics and marketing roles.



Interest over time

☑ News headlines □ Forecast

Figure 2. Google Trends Time Series Plot of "Big Data"

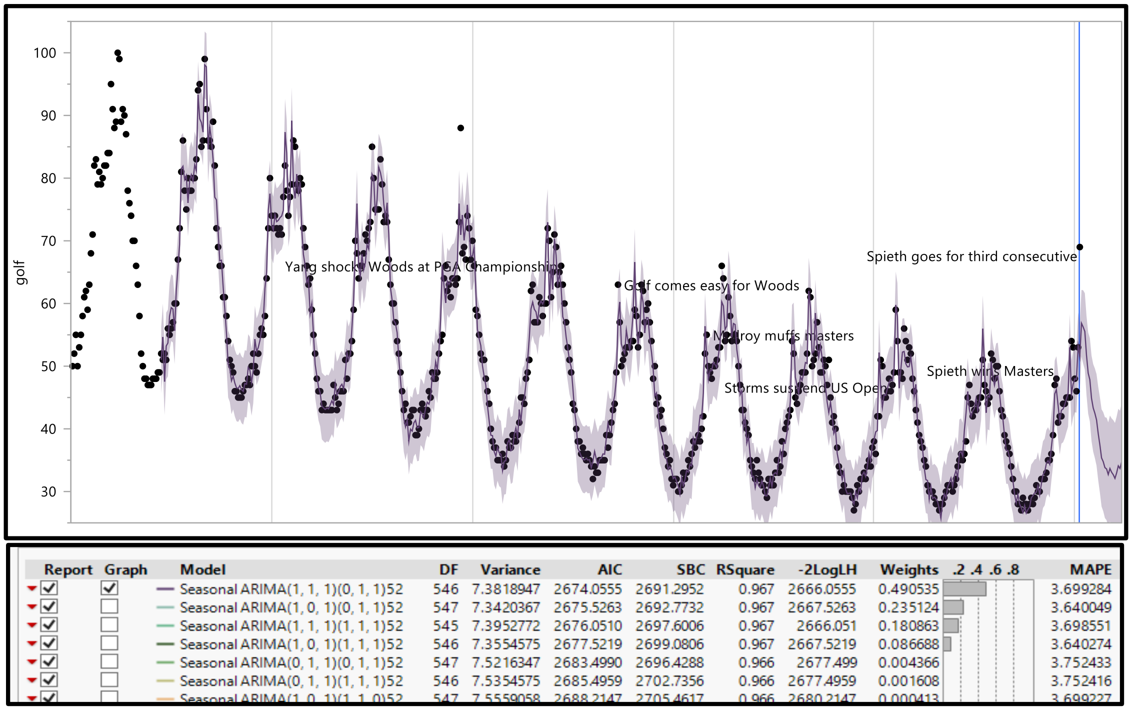


Interest over time

☑ News headlines □ Forecast

Figure 3. Google Trends Time Series Plot of "Golf"

From the data export, we can use JMP to create our own custom time series plot in Graph Builder that labels those events more clearly and we can forecast the growth in the future using seasonal ARIMA as shown in Figure 4. Corporations can use these types of advanced analytics to predict where the markets are going and timing of advertising campaigns.



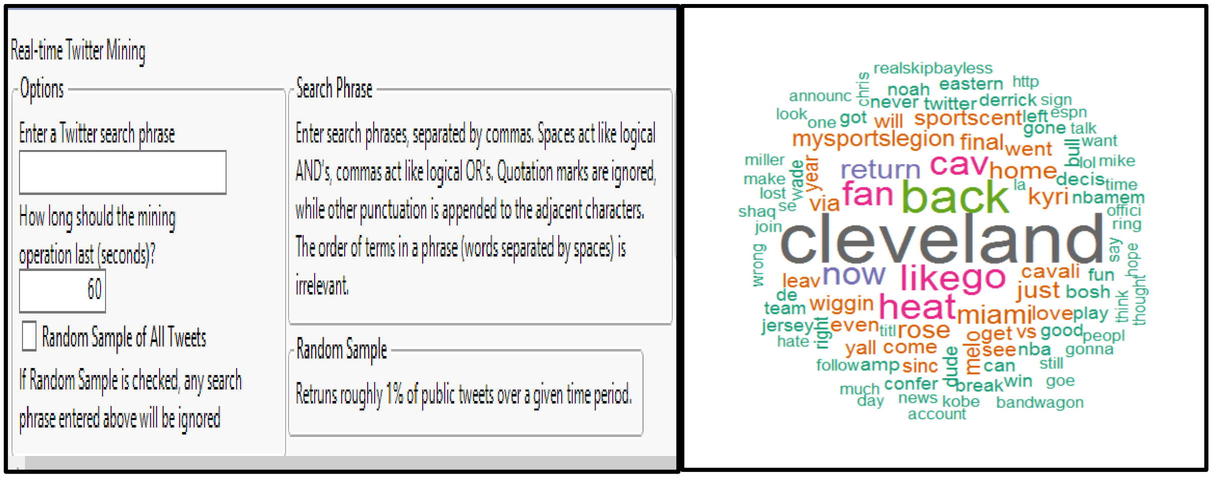
Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	Weights	.2 .4 .6 .8	MAPE
☑	☑	Seasonal ARIMA(1, 1, 1)(0, 1, 1)52	546	7.3818947	2674.0555	2691.2952	0.967	2666.0555	0.490535		3.699284
☑	☐	Seasonal ARIMA(1, 0, 1)(0, 1, 1)52	547	7.3420367	2675.5263	2692.7732	0.967	2667.5263	0.235124		3.640049
☑	☐	Seasonal ARIMA(1, 1, 1)(1, 1, 1)52	545	7.3952772	2676.0510	2697.6006	0.967	2666.051	0.180863		3.698551
☑	☐	Seasonal ARIMA(1, 0, 1)(1, 1, 1)52	546	7.3554575	2677.5219	2699.0806	0.967	2667.5219	0.086688		3.640274
☑	☐	Seasonal ARIMA(0, 1, 1)(0, 1, 1)52	547	7.5216347	2683.4990	2696.4288	0.966	2677.499	0.004366		3.752433
☑	☐	Seasonal ARIMA(0, 1, 1)(1, 1, 1)52	546	7.5354575	2685.4959	2702.7356	0.966	2677.4959	0.001608		3.752416
☑	☐	Seasonal ARIMA(1, 0, 1)(1, 1, 0)52	547	7.5559058	2688.2147	2705.4617	0.966	2680.2147	0.000413		3.699227

Figure 4. ARIMA Group Time Series Plot and Forecasts for “Golf”

Competitive Intelligence Data from Social Media:

The last decade has seen an explosion in a new kind of intelligence data that can be assimilated from social media sites. Communications of all types are being recorded now more than ever and we have the opportunity to find CI signals amongst all this noisy, unstructured data. Blogs and bulletin boards are outstanding hunting grounds for gathering industry and company-specific information. You can follow companies, employees, technology groups, job opportunities, former employees on LinkedIn and other professional sites. Posts to sites like Facebook, Twitter, and Instagram also offer opportunities to understand trends in markets and competitors. The challenge is efficiently collecting this unstructured data characterized by a new informal English-like language with many ways to express the same word—think of how you would text someone the word *tonight*.

We wrote a JMP Scripting Language (JSL) script that calls the statistical programming language R to download all (or a random sample) of the Twitter feeds for a specified duration relative to a key word. The script collects data from the present time to a pre-specified period in the future, although Twitter archives to collect retrospective data do exist for a cost. The data returned includes the actual Tweet, location, time, and other demographic fields. Many companies collect real time Tweets to see public reaction to major product releases—whether their own or the competition. As an example, we ran the script for 5 minutes after LeBron James announced he was returning to Cleveland, as all of our Twitter data on products and companies is proprietary. The resulting wordcloud in Figure 5 gives insight into the most frequently occurring terms (largest font) after removing common words (the, and, or, …) and offensive language. Additionally, we conducted sentiment analysis on the Tweets by crosstabulating a list of approximately 1,500 words generally associated with positive sentiment, a list of approximately 2,000 negative words with the all of the words collected across 4,000 Tweets in that 5 minute period. A single Tweet may have multiple positive and negative words or it may have none of either. The grand tally was 3,533 positive words and 1,626 negative words. The wordcloud and the sentiment analysis are conducted using a JMP Add-In for text mining that also calls R for text analytics routines.



Real-time Twitter Mining

Options

Enter a Twitter search phrase

How long should the mining operation last (seconds)?

60

□ Random Sample of All Tweets

If Random Sample is checked, any search phrase entered above will be ignored

Search Phrase

Enter search phrases, separated by commas. Spaces act like logical AND's, commas act like logical OR's. Quotation marks are ignored, while other punctuation is appended to the adjacent characters. The order of terms in a phrase (words separated by spaces) is irrelevant.

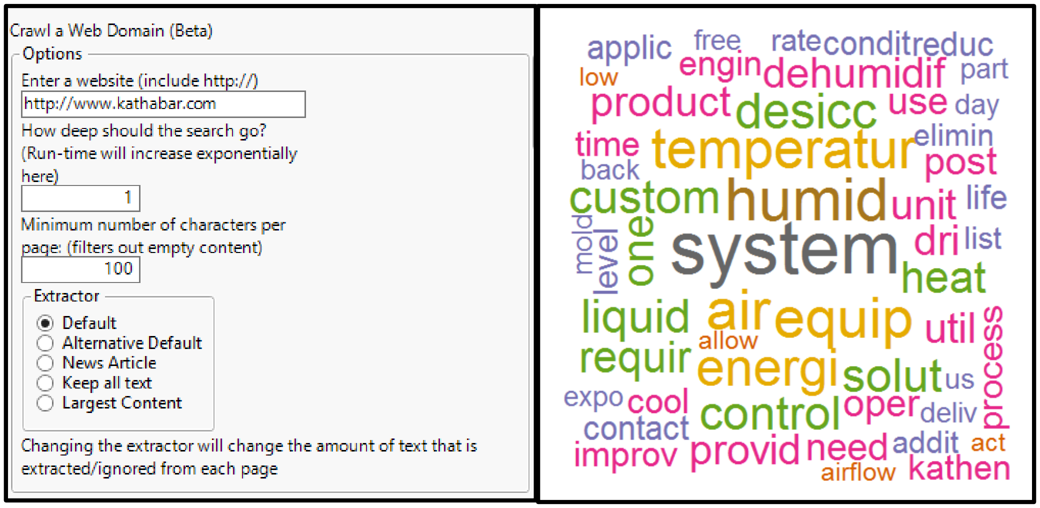
Random Sample

Retruns roughly 1% of public tweets over a given time period.

Figure 5. Twitter JSL Script Interface and Wordcloud on “LeBron” in Summer 2014

Competitive Intelligence Data from Scraping Competitor Websites:
The architecture of a company’s website alone is a useful piece of strategic intelligence as they carefully choose what to reveal in their online presence both in terms of current capabilities and customers as well as what they are working on for future improvements. Other exceptional data opportunities include meetings/conferences, leadership, links to whitepapers, links to news stories, best practices, technology discussions, job postings, and key words in search engine optimization. One method to get this data is to manually copy and paste the text of interest. Alternatively, an automated web scraping utility has the advantages of being much quicker and comprehensive, though too much or extraneous information may be returned. We developed JSL code to scrape the text data from any website while keeping track of metadata like the location of the text on the website. The Spider script once again calls R and only requires you to put in the website of interest. You can choose the type of scraping you want to conduct since websites differ in construction. Scraping options include selecting all the text, a news story or the default setting. You can also select how deep you want to crawl where level 0 is the webpage you specify, 1 is that page plus all linked pages, and so forth.

As an example, a popular green energy technology is liquid desiccant air conditioning which achieves a cooling effect by removing water molecules rather than the conventional chilled air or refrigerant systems that remove humidity by lowering temperatures below the dew point. A quick Google search on commercial liquid desiccant air conditioning shows two leading companies, Advantix Systems and Kathabar. We scraped both sites with the JSL script to create a JMP data table with each row representing a webpage. Figure 6 displays the Spider interface along with the resultant word cloud from the text mining Add-In.



Crawl a Web Domain (Beta)

Options

Enter a website (include http://)

http://www.kathabar.com

How deep should the search go? (Run-time will increase exponentially here)

1

Minimum number of characters per page: (filters out empty content)

100

Extractor

● Default
○ Alternative Default
○ News Article
○ Keep all text
○ Largest Content

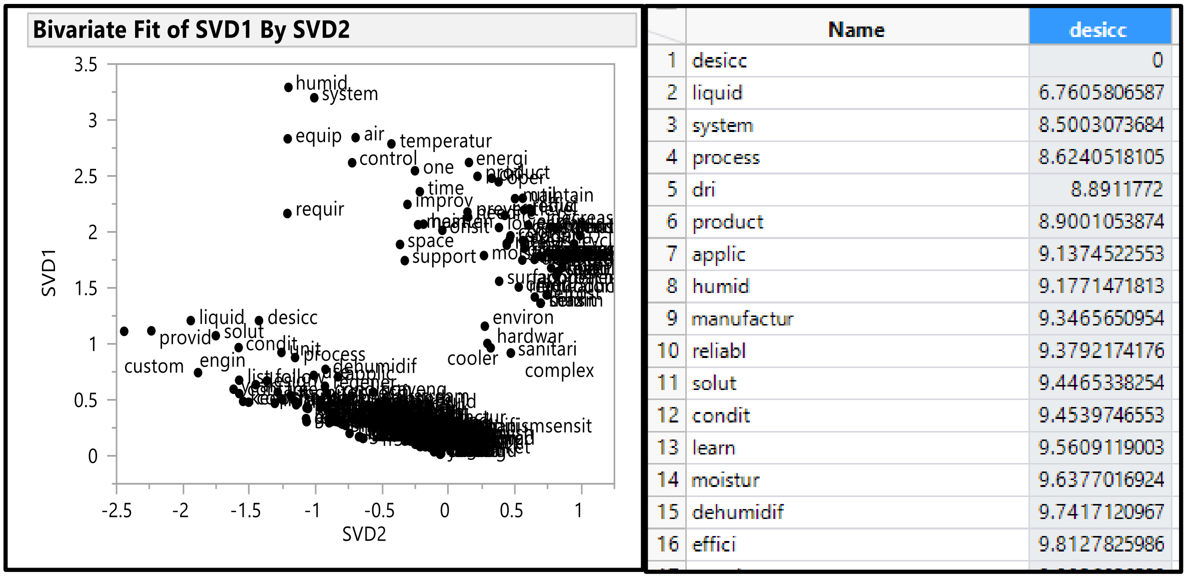
Changing the extractor will change the amount of text that is extracted/ignored from each page

Figure 6. JSL Spider Interface and Resulting Wordcloud from www.kathabar.com

The wordcloud once again represents the frequency of the words. Though this raw word count is useful, it does not provide much intelligence as to what words occur together and what themes may be present. To translate unstructured text into a flat file for statistical analysis beyond word counts, it is common to form a Document Term Matrix (DTM) that has each document (webpage for our example) as a row and every unique word that is used across the entire corpus as a separate column. The matrix entries are the frequency each term occurs in the particular row though it may be better to just use a binary representation indicating if the word appears or not. A weighted scheme based on the number of occurrences the word has across the rows is another commonly applied transformation. That is, a word that appears in every row may not have much discrimination power to be helpful analytically. The problem with the DTM is that it is usually very large and very sparse; therefore, standard statistical methods may fail to detect useful relationships.

One solution to the large and sparse DTM is to project it into a much smaller dimension space much like principal components analysis (PCA). A better solution is taking the singular value decomposition (SVD) of the DTM in a lower dimension than the number of columns or words (reduce it to say 100 to 300). This SVD can be thought of as DTM M=UDV'. The eigenvectors in the U matrix represent the documents, the eigenvectors in the V matrix represent the words, and the singular values on the diagonals of the D matrix are from the eigenvalues that explain the proportion of variability explained for each SVD column.

Looking at a bivariate plot of the first two eigenvectors from the V matrix is helpful to pick up major themes in the text data as well as find words that group together. Figure 7 shows the bivariate plot of the first 2 SVDs for the Kathabar example. Here we see the words grouping around *providing a custom liquid desiccant solution* and another group suggesting *low energy equipment to control air temperature and humidity.*



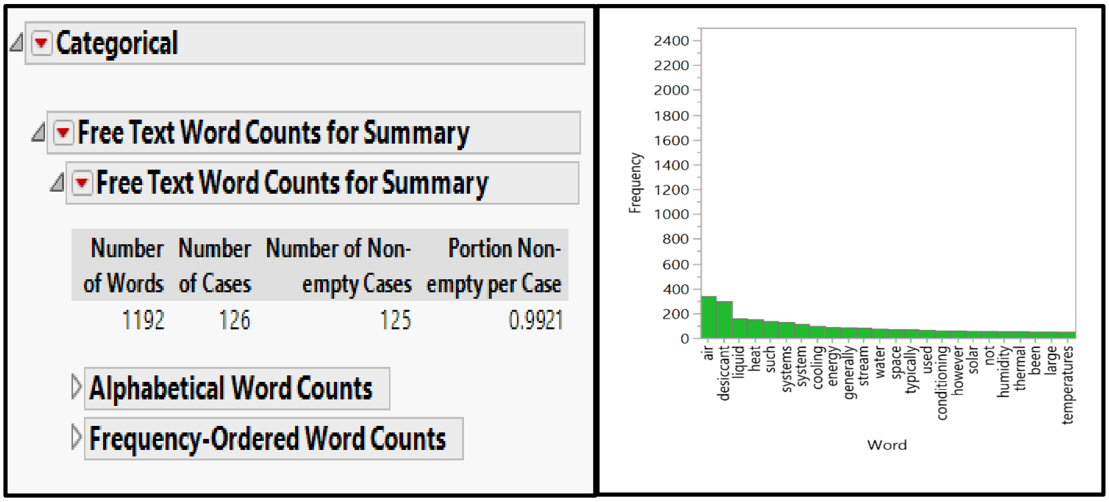
	Name	desicc
1	desicc	0
2	liquid	6.7605806587
3	system	8.5003073684
4	process	8.6240518105
5	dri	8.8911772
6	product	8.9001053874
7	applic	9.1374522553
8	humid	9.1771471813
9	manufactur	9.3465650954
10	reliabl	9.3792174176
11	solut	9.4465338254
12	condit	9.4539746553
13	learn	9.5609119003
14	moistur	9.6377016924
15	dehumidif	9.7417120967
16	effici	9.8127825986

Figure 7. First to SVD Eigenvectors from Kathabar V Matrix and Multivariate Distances for Desiccant.

We can also use multivariate distances from the eigenvectors of the V matrix from the reduced rank SVD. These distances show words that group together which allows you to choose any word in the corpus to find closely associated terms. In right panel of Figure 7, we can see the words closest to *desiccant* are not surprisingly: *liquid, system, process, dry, product, application, reliable,* and *manufacture.*

Competitive Intelligence Data from Patents:

Patent filings offer a good barometer of a corporation's technological profile. Fortunately, there are many great patent search engines available that give you free access to all of the abstracts, drawings, and text fields to collect data on the competitors protected property and understand the state of technology. PATENTSCOPE from the World Intellectual Property Organization (https://patentscope.wipo.int/search/en/result.jsf) was used to extract the abstracts for liquid desiccant air conditioning patents. Each paragraph of an abstract formed a separate row in the JMP data table. Using the Consumer Research platform (**Analyze>Consumer Research>Categorical> Response Roles=Multiple, Column of text data assigned to Free Text**), we can see the word counts and quickly make a Pareto diagram of the top terms to appear in the patent abstracts.



Categorical

Free Text Word Counts for Summary

Free Text Word Counts for Summary

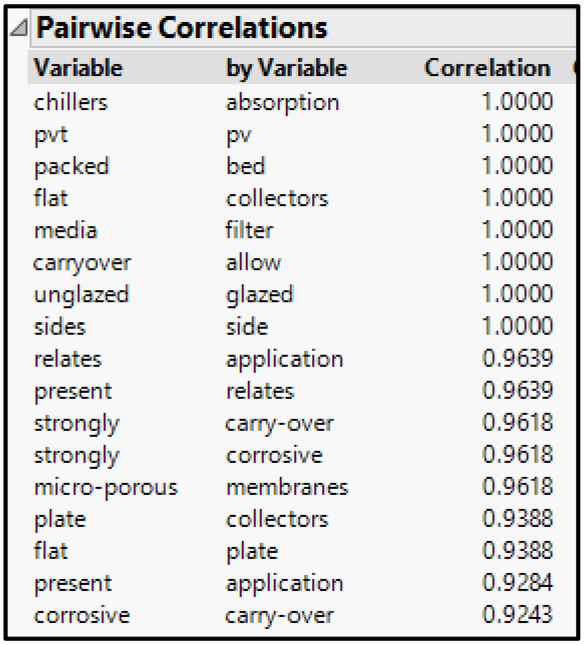
Number of Words	Number of Cases	Number of Non-empty Cases	Portion Non-empty per Case
1192	126	125	0.9921

Alphabetical Word Counts

Frequency-Ordered Word Counts

Figure 8. JMP Free Text Interface and Word Frequency Count for Patent Data.

We can also use JMP to form indicator columns for the most frequent 200 (for example) words to see if certain terms correlate using the **Analyze>Multivariate Methods>Multivariate>Pairwise Correlations**. Figure 9 highlights some of the word pairs that are highly correlated like *media filter, strongly corrosive carry-over, microporous membranes* and so forth.



Pairwise Correlations

Variable	by Variable	Correlation
chillers	absorption	1.0000
pvt	pv	1.0000
packed	bed	1.0000
flat	collectors	1.0000
media	filter	1.0000
carryover	allow	1.0000
unglazed	glazed	1.0000
sides	side	1.0000
relates	application	0.9639
present	relates	0.9639
strongly	carry-over	0.9618
strongly	corrosive	0.9618
micro-porous	membranes	0.9618
plate	collectors	0.9388
flat	plate	0.9388
present	application	0.9284
corrosive	carry-over	0.9243

Figure 9. Pairwise Correlations from Multivariate Platform of Patent Data

Topic extraction using the JMP text mining add-in provides additional information. The plot of the first two SVDs in Figure 10 gives the general idea of what these patents are trying to do: a technology that requires the use of liquid desiccant (stop word) to dehumidify the air resulting in cooling effect that is most applicable in humid climates; the risk is most liquid desiccants have is the use spray over a filter media that increases risk of carry-over into air space; these patents use a membrane to eliminate this risk.

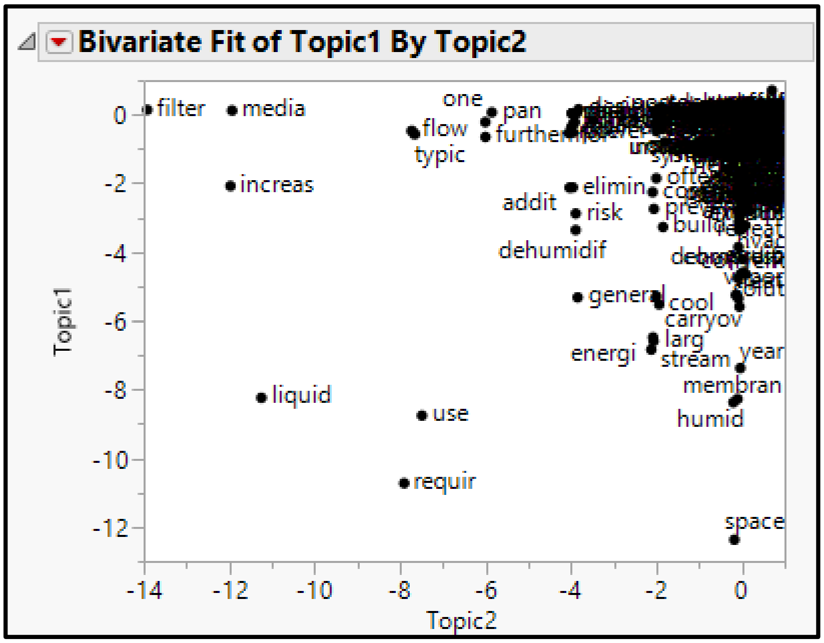


Figure 10. First Two SVD Vectors for Patents

Market Intelligence from Technical Journal Articles: Many technological advances are detailed in academic literature which suggests these would be very helpful to paint the technological landscape. To better understand the state of liquid desiccant technology and the readiness for commercial applications, we downloaded 45 recently refereed technical articles from respected journal publications in the clean energy field from the last three years. These portable document files (pdf) were difficult to translate into text then to a JMP data table because the text characters are really more images and the large amount of mathematical equations do not translate into text well. We experimented with numerous methods (both open source and proprietary) to read these pdf files cleanly and concluded using AdobeAcrobat Standard or Pro to save the pdf as a MS Word Rich Text File (rtf) using the Optical Character Recognition (OCR) function was a good intermediate step. Next, we developed a MS Word macro to open up a file of Word documents and save them as text files to be used in our macro that reads a folder of text files into a JMP table to form the corpus. We also had to update our JSL script for the text mining add-in to have the option to remove the non-alphanumeric characters. The references were removed as they unnecessarily interfered with the text analytics.

The word frequencies in Figure 11 from the JMP text mining add-in tell a similar story as the patents: this green energy technology is a system to cool air reducing energy consumption by regenerating the liquid desiccant (LiCl for example) to dehumidify hence cool the air where the membrane separates the toxic desiccant from the air stream.

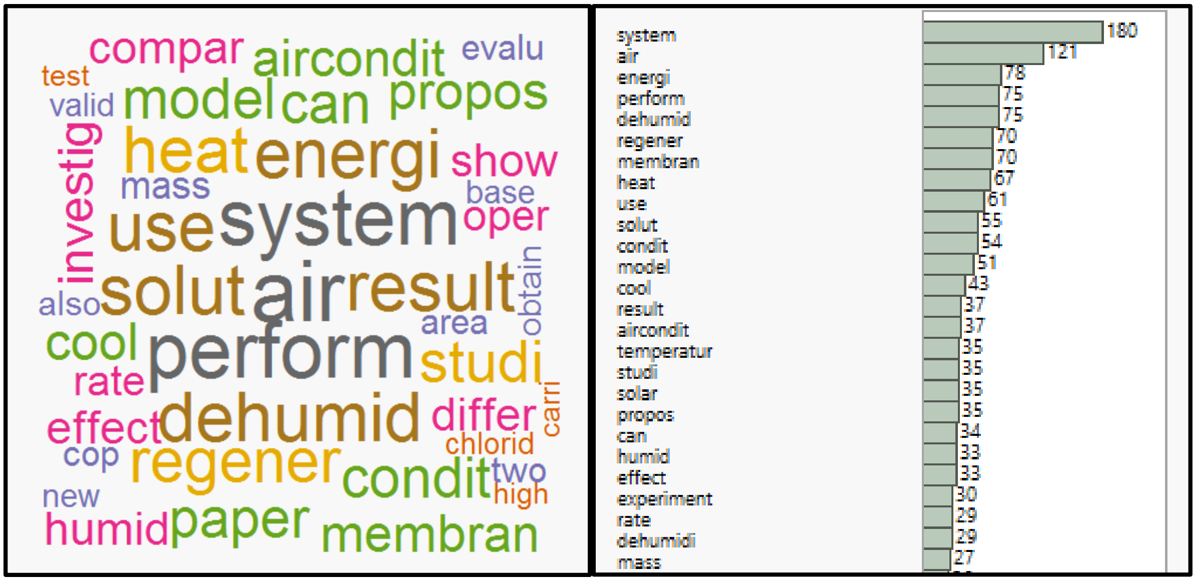
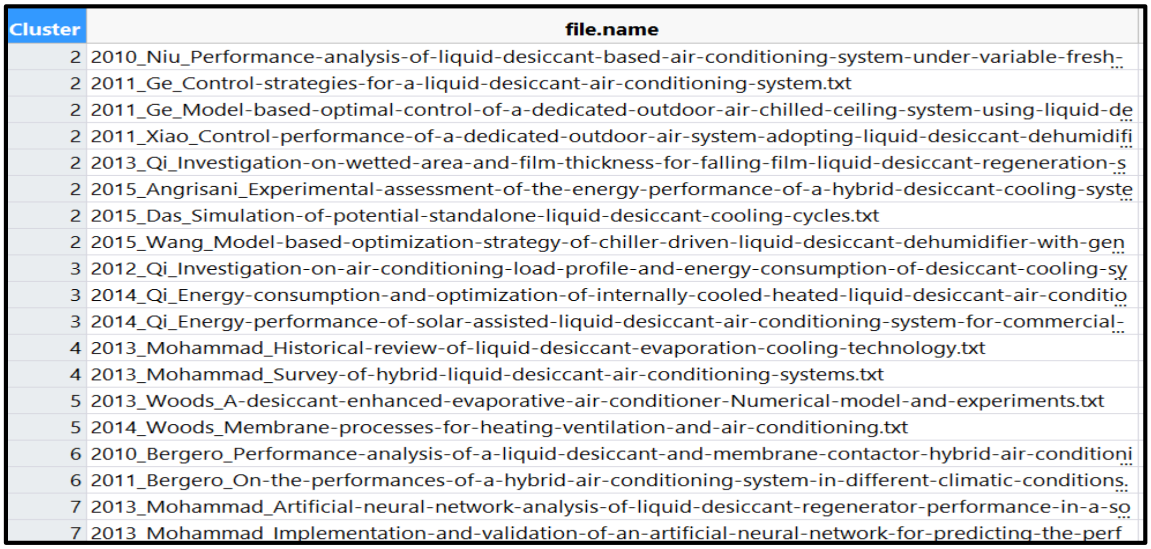


Figure 11. Journal Article Word Frequencies for Liquid Desiccant Membrane.

After performing the SVD, we can cluster like journal articles by using the values of the eigenvectors from the U matrix. The documents cluster very much along author lines as well as general research areas within liquid desiccant (Figure 12). For topic extraction, we use the V matrix from the SVD. Some selected topics (Figure 13) across these journals are: 1) comparative experiments validating liquid desiccant as an air conditioning solution and increase in efficiency from regeneration method that saves energy; 2) experiment to predict mass transfer rates and efficiency ratios while allowing different inlet parameter values; 3) alternative method to remove vapor using hybrid electric compressor and liquid desiccant.



Cluster	file.name
2	2010_Niu_Performance-analysis-of-liquid-desiccant-based-air-conditioning-system-under-variable-fresh-...
2	2011_Ge_Control-strategies-for-a-liquid-desiccant-air-conditioning-system.txt
2	2011_Ge_Model-based-optimal-control-of-a-dedicated-outdoor-air-chilled-ceiling-system-using-liquid-de...
2	2011_Xiao_Control-performance-of-a-dedicated-outdoor-air-system-adopting-liquid-desiccant-dehumidifi...
2	2013_Qi_Investigation-on-wetted-area-and-film-thickness-for-falling-film-liquid-desiccant-regeneration-s...
2	2015_Angrisani_Experimental-assessment-of-the-energy-performance-of-a-hybrid-desiccant-cooling-syste...
2	2015_Das_Simulation-of-potential-standalone-liquid-desiccant-cooling-cycles.txt
2	2015_Wang_Model-based-optimization-strategy-of-chiller-driven-liquid-desiccant-dehumidifier-with-gen...
3	2012_Qi_Investigation-on-air-conditioning-load-profile-and-energy-consumption-of-desiccant-cooling-sy...
3	2014_Qi_Energy-consumption-and-optimization-of-internally-cooled-heated-liquid-desiccant-air-conditio...
3	2014_Qi_Energy-performance-of-solar-assisted-liquid-desiccant-air-conditioning-system-for-commercial-...
4	2013_Mohammad_Historical-review-of-liquid-desiccant-evaporation-cooling-technology.txt
4	2013_Mohammad_Survey-of-hybrid-liquid-desiccant-air-conditioning-systems.txt
5	2013_Woods_A-desiccant-enhanced-evaporative-air-conditioner-Numerical-model-and-experiments.txt
5	2014_Woods_Membrane-processes-for-heating-ventilation-and-air-conditioning.txt
6	2010_Bergero_Performance-analysis-of-a-liquid-desiccant-and-membrane-contactor-hybrid-air-conditioni...
6	2011_Bergero_On-the-performances-of-a-hybrid-air-conditioning-system-in-different-climatic-conditions...
7	2013_Mohammad_Artificial-neural-network-analysis-of-liquid-desiccant-regenerator-performance-in-a-so...
7	2013_Mohammad_Implementation-and-validation-of-an-artificial-neural-network-for-predicting-the-perf

Figure 12. Cluster Analysis Results from Liquid Desiccant Journal Articles

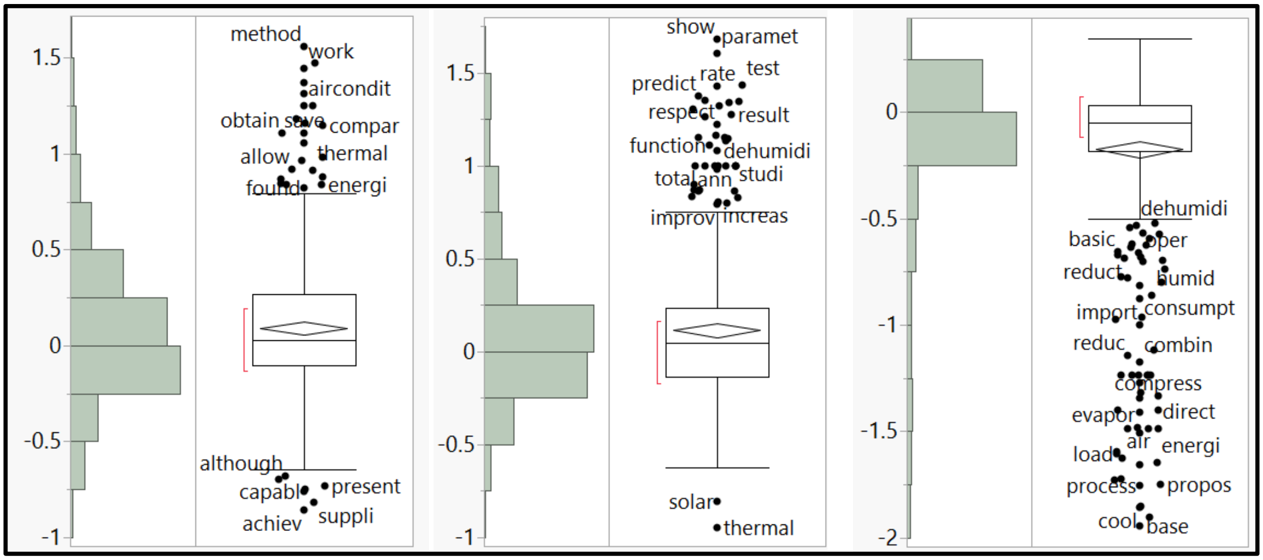


Figure 13. Topics from the Abstracts of Liquid Desiccant Air Conditioning Membrane

The plots of the first two eigenvectors from the V matrix (Figure 14) shows some general themes from these research efforts based on the abstracts: energy regeneration, improve dehumidification, simulation, mass transfer, experiment prediction, model, temperature and membrane, and thermal process with water vapor.

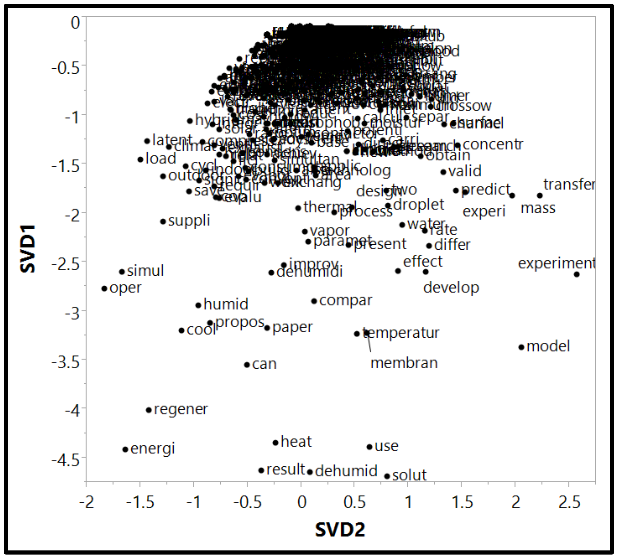
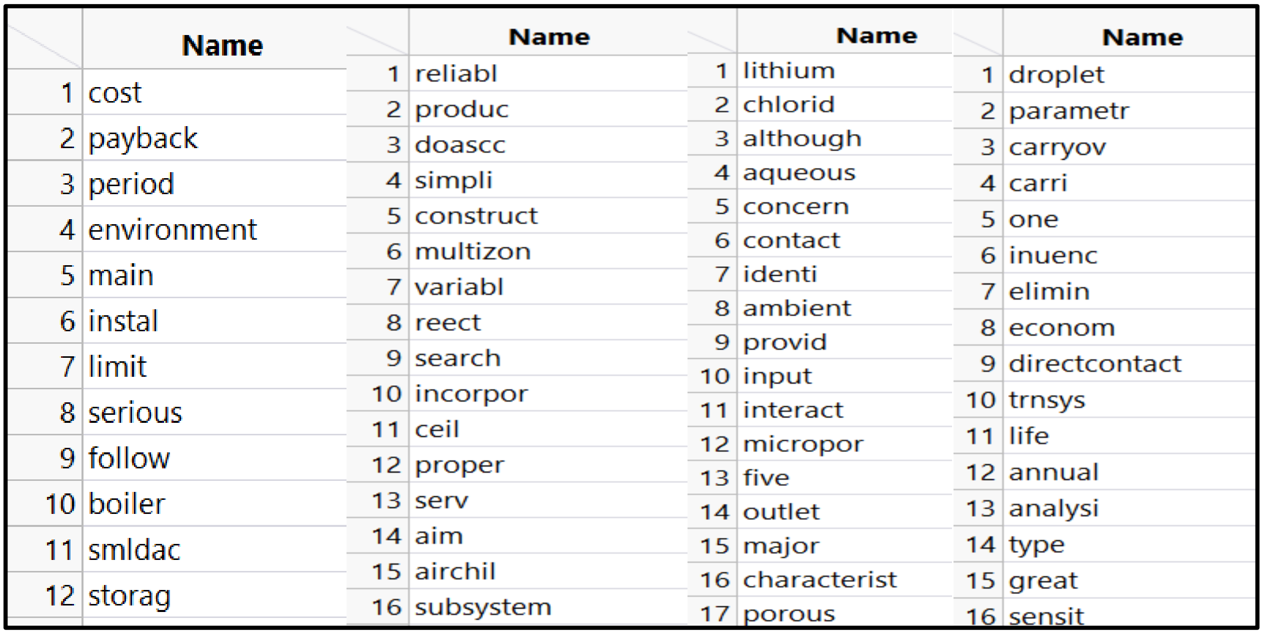


Figure 14. Bivariate Plot of First Two Eigenvectors of Liquid Desiccant Air Conditioning Journal Abstracts

Another useful technique in topic extraction is to look at the multivariate distances from the V matrix for particular words. These multivariate distances are computed by JMP during the cluster analysis and are obtained by saving the distance matrix. You can select a specific word and find out which words are most closely related to it. A few words of interest shown in Figure 15 provide the following information from the abstracts of these journal articles:

- Cost—concern is payback period, main installation, boiler, and storage big drivers
- Reliability—producing multizone and ceiling units with airchilling subsystem
- Lithium-dessicant is lithium chloride as aqueous solution; major concern is contact with ambient environment (toxic), microporous membrane is solution
- Droplets—coming in direct contact is harmful, need to eliminate to make economically feasible



	Name
1	cost
2	payback
3	period
4	environment
5	main
6	instal
7	limit
8	serious
9	follow
10	boiler
11	smldac
12	storag

	Name
1	reliabl
2	produc
3	doascc
4	simpli
5	construct
6	multizon
7	variabl
8	reflect
9	search
10	incorpor
11	ceil
12	proper
13	serv
14	aim
15	airchil
16	subsystem

	Name
1	lithium
2	chlorid
3	although
4	aqueous
5	concern
6	contact
7	identi
8	ambient
9	provid
10	input
11	interact
12	micropor
13	five
14	outlet
15	major
16	characterist
17	porous

	Name
1	droplet
2	parametr
3	carryov
4	carri
5	one
6	influenc
7	elimin
8	econom
9	directcontact
10	trnsys
11	life
12	annual
13	analysi
14	type
15	great
16	sensit

Figure 15. Multivariate Distances from the V Matrix for Liquid Desiccant Journal Abstracts.

Summary:
There are many frameworks for Competitive Intelligence analysis such as SWOT, Porter's 5 Forces and 4 Corners, that depend on current and dependable data input. CI data sources have skyrocketed and become democratized so employees across the enterprise can find exploitable information on their rivals. The problem is the data is not necessarily easily retrievable, is mostly noisy and unstructured. Analytic tools from JMP and JSL scripts are helpful to collect and make sense out of data from sources such as websites, social media, patents, and technical journal articles. Like the military analolgy, the goal is to transition from unstructured open source information to "classified" intelligence reports that impact the strategic direction of the organization.

References:

Baglama, James and Reichel, Lothar. *Augmented Implicitly Restarted Lanczos Bidiagonalization Methods.* 2005, SIAM Journal of Scientific Computing, Vol. 27, pp. 19-42.

Barbera, Pablo. streamR: Access to Twitter Streaming API via R. [Online] 2014. http://CRAN.R-project.org/package=streamR.

Czepiel, John and Kerin, Roger. *Competitor Analysis*. [Online] 2011. http://pages.stern.nyu.edu/~jczepiel/Publications/CompetitorAnalysis.pdf.

Evangelopoulos, Nicholas, Zhang, Xiaoni and Prybutok, Victor R. *Latent Semantic Analysis: five methodological recommendations.* 2012, European Journal of Information Systems, Vol. 21, pp. 70-86.

Fellows, Ian. wordcloud: Word Clouds. *CRAN.* [Online] 2014. http://cran.r-project.org/web/packages/wordcloud/index.html.

Kaushik, Avanish. *Crushing it With Competitive Intelligence Analysis: Best Metrics, Reports*. [Online] 2015. http://www.kaushik.net/avinash/competitive-intelligence-analysis-tools-metrics-reports-techniques/.

Lang, Duncan Temple. RCurl: General network (HTTP/FTP/...) client interface for R. *CRAN.* [Online] 2015. http://cran.r-project.org/web/packages/RCurl/index.html.

Martin. *Strategy frameworks: competitor analysis and competitive intelligence*. [Online] 2015. http://www.entrepreneurial-insights.com/competitor-analysis-competitive-intelligence/.

Metayer, Estelle. *Top analysis techniques your competitive intelligence or strategic planning team should master.* [Online] 2015 http://competia.com/50-competitive-intelligence-analysis-techniques.

Scrapinghub. Scrapy. *A Fast and Powerful Web Crawling Framework.* [Online] 2015. http://scrapy.org/.

World Intellectual Property Organization Patentscope. [Online] 2015. https://patentscope.wipo.int/search/en/result.jsf