JMP® ACADEMIC CASE STUDY

JMP041: Text Exploration of Patents

Word Cloud, Data Visualization, Term Selection

Produced by

Markus Schafheutle
office@schafheutle.co.at

Muralidhara A, JMP Global Academic Team
muralidhara.a@jmp.com

jmp® STATISTICAL DISCOVERY

Text Exploration of Patents

Word Cloud, Data Visualization, Term Selection

Key ideas:

This case study requires the use of unstructured data analysis to understand and analyze the text related to patents filed by different companies using JMP Text Explorer.

Background

A patent is a form of intellectual property that gives its owner the legal right to exclude others from exploiting the patented technology, including making, using or selling the patented invention. Organizations invest a lot of resources in inventing a new technology or a design, and this continuous effort is vital to its future success. Acquiring a patent is important for the company after it develops any innovative product or solution as it enhances and protects the value of the product or service.

Patent protection is granted for a limited period, generally 20 years from the filing date of the application. Patents are territorial rights, meaning the exclusive rights are only applicable in the country or region in which a patent has been filed and granted, in accordance with the law of that country or region. A patent is typically issued to the individual inventor and is granted by a national or regional patent office. In most countries, if an employee develops an invention as per employment contract, the invention (and the related patent rights) will belong to the enterprise.

Google Patents (https://patents.google.com/) is a search engine that indexes patents and patent applications spanning more than 350 years. It indexes more than 87 million patents and patent applications with full text from 17 patent offices from the US, Europe, China, Japan, Korea, Canada, UK, Russia and other countries. These documents include the entire collection of granted patents and published patent applications from each database, which belong in the public domain.

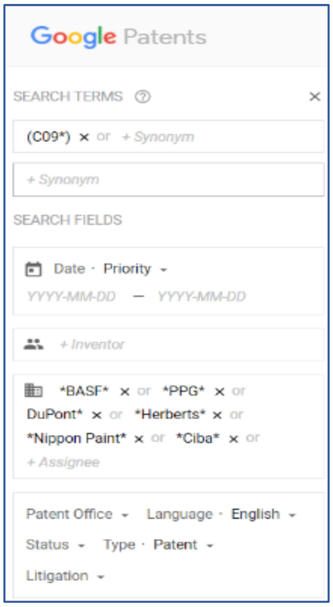
The International Patent Classification (IPC) is a hierarchical patent classification system used in over 100 countries to classify the content of patents in a uniform manner. The classification is updated regularly. The Cooperative Patent Classification (CPC) is an extension of the IPC and is divided into nine sections (A-H and Y), which in turn are subdivided into classes, subclasses, groups and subgroups. There are approximately 250,000 classification entries.

The Data

The patent application data was extracted from Google patents advanced search website (https://patents.google.com/advanced) for IPC code C09, which mostly includes dyes, paints, polishes, natural resins and adhesives and compositions as part of the patent. The organizations or assignees selected were a group of chemical companies, namely BASF, PPG, DuPont, Herberts, Nippon Paint and Ciba, as shown in Exhibit 1. The search was also restricted to English language and patent type for further analysis.

It is important to search for "Family" to prevent duplication of the same patent in different countries. Data was extracted and downloaded from Google Patents for these companies and their local subsidiaries and potential predecessors, resulting in roughly 1,330 hits. Please note that this information might change, as it is real-time information.

Exhibit 1 Google Patents Website



Google Patents

SEARCH TERMS ⓘ ✕

(C09*) ✕ or + Synonym

+ Synonym

SEARCH FIELDS

Date · Priority ▾
YYYY-MM-DD — YYYY-MM-DD

+ Inventor

BASF ✕ or *PPG* ✕ or DuPont* ✕ or *Herberts* ✕ or *Nippon Paint* ✕ or *Ciba* ✕ or
+ Assignee

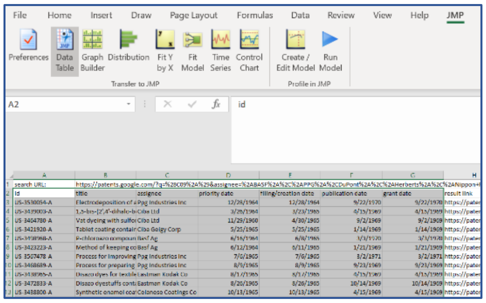
Patent Office ▾ Language · English ▾
Status ▾ Type · Patent ▾
Litigation ▾

By clicking the download button, the data is downloaded as a .csv file to a local desktop.

By open the .csv file using Excel, it will appear in the format shown below, with the first row representing the URL details of the search followed by the column names in the second row.

The JMP Add-In for Excel provides new capabilities to JMP and Excel users on Windows, as shown in Exhibit 2. Use the JMP Add-In for Excel to transfer a worksheet from Excel to the following data table directly.

Exhibit 2 JMP Add-In Tab for Excel



A2 | id

	A	B	C	D	E	F	G	H
1	search URL:	https://patents.google.com/?q=%28C09%2A%29&assignee=%2ABASF%2A%2C%2APPG%2A%2CDuPont%2A%2C%2AHerberts%2A%2C%2ANippon+						
2	id	title	assignee	priority date	filing/creation date	publication date	grant date	result link
3	US-3530054-A	Electrodeposition of a	Ppg Industries Inc	12/28/1964	12/28/1964	9/22/1970	9/22/1970	https://pater
4	US-3439003-A	1,5-bis-[2',4'-dihalo-b	Ciba Ltd	3/26/1964	3/23/1965	4/15/1969	4/15/1969	https://pater
5	US-3464780-A	Vat dyeing with sulfo	Ciba Ltd	11/29/1960	4/30/1965	9/2/1969	9/2/1969	https://pater
6	US-3421920-A	Tablet coating contain	Ciba Geigy Corp	5/25/1965	5/25/1965	1/14/1969	1/14/1969	https://pater
7	US-3498968-A	P-chloroazo compoun	Basf Ag	6/16/1964	6/8/1965	3/3/1970	3/3/1970	https://pater
8	US-3423223-A	Method of keeping co	Basf Ag	6/12/1964	6/11/1965	1/21/1969	1/21/1969	https://pater
9	US-3567478-A	Process for improving	Ppg Industries Inc	7/6/1965	7/6/1965	3/2/1971	3/2/1971	https://pater
10	US-3468689-A	Process for preparing	Ppg Industries Inc	8/9/1965	8/9/1965	9/23/1969	9/23/1969	https://pater
11	US-3438965-A	Disazo dyes for textile	Eastman Kodak Co	8/17/1965	8/17/1965	4/15/1969	4/15/1969	https://pater
12	US-3472833-A	Disazo dyestuffs cont	Eastman Kodak Co	8/26/1965	8/26/1965	10/14/1969	10/14/1969	https://pater
13	US-3438800-A	Synthetic enamel coa	Celanese Coatings Co	10/13/1965	10/13/1965	4/15/1969	4/15/1969	https://pater

Select the cells to transfer into JMP, including any cells that you want to use as column names, from A2 to G1315. The first row of the selection is part of the column name. Select the destination (e.g., data table). If only one cell (or no cell) is selected, the entire Excel worksheet is transferred to JMP. If you are using a Mac, use the Excel wizard to import the data to JMP.

The other two variables – result link and representative figure link – are internet links for the line items in the data; they are not part of the analysis so they can be ignored or deleted. Inventor/Author column can also be ignored. A new JMP data table will be created from the Excel data for further analysis. Now we have imported the patent data as a JMP data set.

Data cleansing

Close observation of each variable is required to check for data quality and consistency. Some of the records in the assignee column are not English so they need to be recoded. JMP has robust recoding and data cleaning capabilities.

Select Cols>Recode. Notice that New Column is selected by default to place the recoded values in a new column. Two new columns, namely Prior Country and Assignee Category, are created using ID and Assignee columns respectively. The first two letters of ID will give the Prior Country. Assignee Category has been created by carefully analyzing the Assignee column and recoding it for the Special and Non-English letters. Click on the + mark next to those new columns to see the formula. To learn about recoding, please visit Recoding in JMP.

The details of each variable are defined in Exhibit 3.

Exhibit 3 Variables and Descriptions

Variable Name	Description
ID	Patent IDs start with the abbreviation of the country.
Prior Country	Short name of the country where the patent is filed.
Title	Title of the patent.
Assignee	Name of the company to which the patent has been assigned.
Assignee Category	The derived column from Assignee column.
Priority Date	Earliest filing date from within a family of patent applications. When the matter involves just one patent application, the priority date would be the filing date of the single application,
Filing/Creation	Date the patent office acknowledges as the application date for a patent on an invention.
Publication Date	Date on which a patent application is first published. It is the date on which the document is made available to the public, thereby making it part of the state of the art. European patent applications are generally published 18 months after the date of filing or earliest priority date.
Grant Date	Date on which the official body grants the patent.

Even if the search was for just five assignees, there is much more information in the data set, perhaps as a result of joint ventures, mergers, etc. of the companies. If the individual history of the companies can be determined, the assignees could be recoded to the final ones.

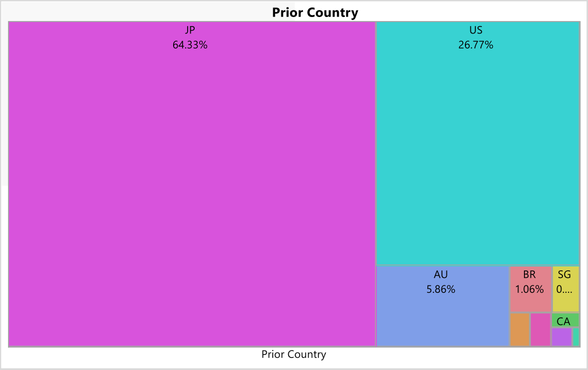
The Task

- To explore the data visually for basic insights.
- To understand the countries of focus for filing the patents by the assignees.
- To analyze the title of the patents collectively and find dominant words associated with the titles.
- To identify the unique topics by company/assignee to understand the principal differences between assignees of interest.

Distribution

Let’s explore the data using Graph Builder. Since the patent IDs start with the abbreviation of the country, we can use Graph Builder to learn which countries are of the greatest interest to the assignees.

Exhibit 4 Tree Map

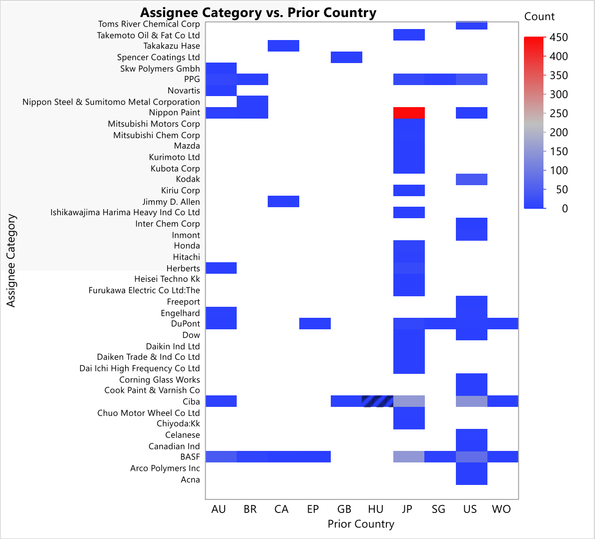


To create, Graph>Graph Builder>Drag Prior Country to X Axis. Select Tree Map, which shows the response summarized by many categories. To squarify the squares by sizes, right-click on the tree map then Layout>Squarify.

From the Exhibit 4, it can be concluded that the assignees have a majority of the patents belonging to Japan (64%) followed by US (26%).

One can also create a heatmap to understand the distribution of patents between country and assignee category.

Exhibit 5 Heat Map

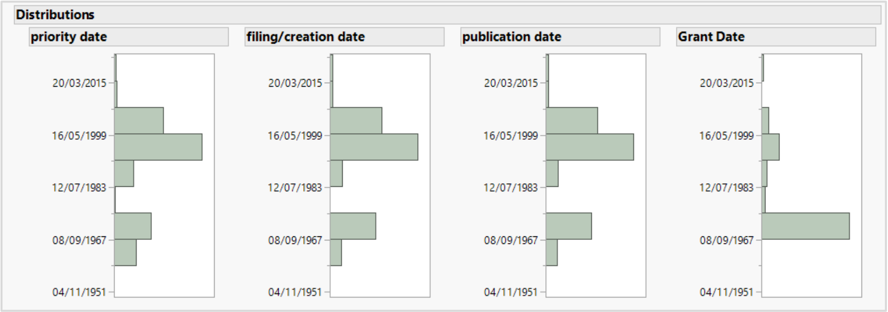


To create, Graph>Graph Builder>X = Prior country, Y = Assignee Category. Select Heat Map, which shows the counts of patents for the specified X and Y categories.

It can be observed from Exhibit 5 that Nippon has the highest number patents and that they belong to Japan. One can also observe that BASF has filed its patents in more countries than other companies.

Next, let's explore the distribution of patents based on the dates.

Exhibit 6 Histogram



Analyze>Distribution>Y Columns = Priority Date, Filing Date, Publication Date and Grant Date. Click OK to get a distribution of patents over a period.

From Exhibit 6, it can be observed that after the first small peak around 1968, a second prolific range occurred between 1984-2004. Since then, a small number of individual applications followed. From a granting perspective, 1968 had the greatest number of patents.

Text analysis terminologies and phases

Text Explorer uses a bag of words approach. Except for phrases, the order of words is ignored, and the analysis is based on the count of words and phrases. Text analysis is often an iterative process, which revolves around curating and analyzing the list of terms. Text analysis uses some unique terminology, which is summarized below.

Exhibit 7 Terminology Related to Text Analysis

Term or Token	Smallest piece of text, like a word in a sentence. The basic unit of analysis for text mining is a term. However, terms can be defined in many ways, including through the use of regular expressions.
Phrase	A short collection of terms that is a sequence of tokens that appear more than once. Each phrase will be considered as to whether it should be a term. JMP has options to manage phrases that are specified as terms in and of themselves.
Document	Collection of words. The unstructured text in each row of the text column corresponds to a document.
Corpus	Collection of documents.
Stop Words	Excluded words from the analysis. The platform has a default list of stop words, but you can also add specific words as stop words.

The text is processed in three stages: tokenizing, phrasing and terming.

The process of breaking the text into terms is called tokenization. It involves converting text to lowercase, applying a tokenizing method (either Basic Words or Regex) to group characters into tokens, and then recoding tokens based on specified recode definitions. Note that recoding occurs before stemming.

The process of combining words with identical beginnings (stems) by removing the endings that differ is called stemming. For example, "dyed," "dyeing," and "dyes" would all be treated as the term: "dye." When a phrase is stemmed, each word in the phrase is stemmed as it would be stemmed as a stand-alone term.

The phrasing stage involves collecting phrases that occur in the corpus (collection of documents) and then specifying individual phrases that are to be treated as terms. Phrases cannot start or end with a stop word, but they can contain a stop word.

The terming stage creates the term list from the tokens and phrases that are produced from the previous stages. For each token, the terming stage performs the following operations:

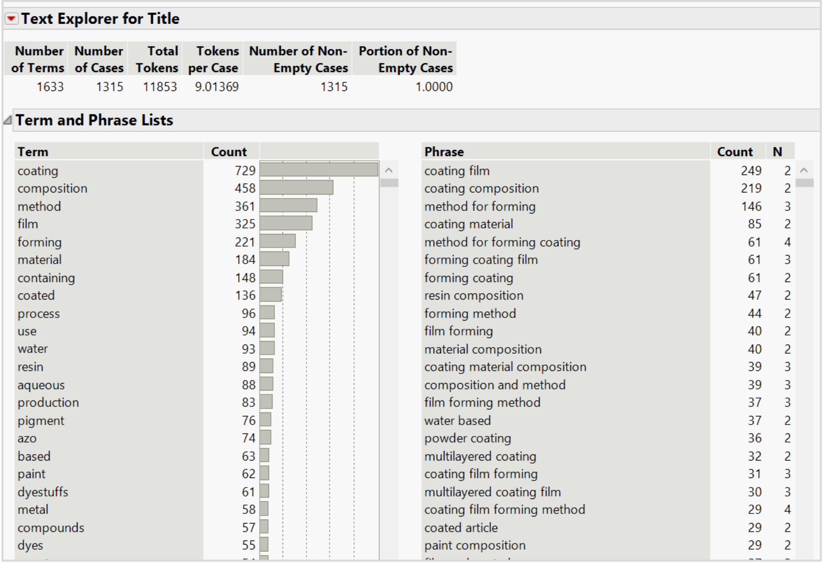
Although stop words are not eligible to be terms, they can be used in phrases. There is also a need for recoding terms, which is useful for combining synonyms into one common term.

JMP has standard built-in stop words, however using the iterative procedure and a close observation of the words would help identify additional stop words.

Text analysis of the title

The procedure to identify stop words and recoding the synonyms is an iterative process encompassing terms lists, word clouds, latent class analysis and SVD.

Exhibit 8 Terms and Phrases



Text Explorer for Title

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-Empty Cases	Portion of Non-Empty Cases
1633	1315	11853	9.01369	1315	1.0000

Term and Phrase Lists

Term	Count
coating	729
composition	458
method	361
film	325
forming	221
material	184
containing	148
coated	136
process	96
use	94
water	93
resin	89
aqueous	88
production	83
pigment	76
azo	74
based	63
paint	62
dyestuffs	61
metal	58
compounds	57
dyes	55

Phrase	Count	N
coating film	249	2
coating composition	219	2
method for forming	146	3
coating material	85	2
method for forming coating	61	4
forming coating film	61	3
forming coating	61	2
resin composition	47	2
forming method	44	2
film forming	40	2
material composition	40	2
coating material composition	39	3
composition and method	39	3
film forming method	37	3
water based	37	2
powder coating	36	2
multilayered coating	32	2
coating film forming	31	3
multilayered coating film	30	3
coating film forming method	29	4
coated article	29	2
paint composition	29	2

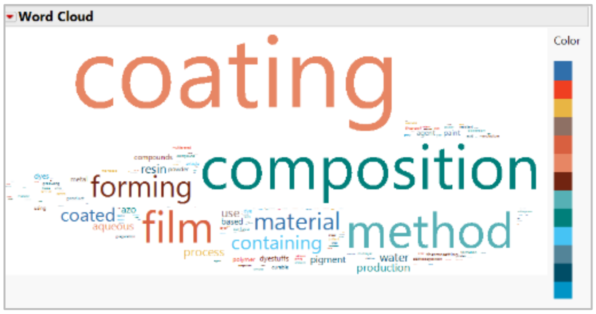
To create, Analyze>Text Explorer>Text Columns =Title, Stemming = No stemming, Tokenizing = Regex. Select Regex and click OK.

Each document is broken into initial units of text called tokens. Text Explorer's report window contains the Summary Counts report and the Term and Phrase Lists report. Its report for title contains the summary statistics, while the Term and Phrase Lists report contains tables of terms and phrases found in the text after tokenization has occurred.

At a glance, you can see that there are 1,633 unique terms in 1,315 documents. In all, there are 11,853 tokenized terms. By default, the Terms List is sorted in descending count order; terms with the same frequency count are sorted alphabetically. The most common term is "coating," occurring 729 times.

A word cloud is a visual representation of text data. Each term is shown with different font size and color. The bigger font size means greater frequencies.

Exhibit 9 Word Cloud



Under the red triangle, choose Display Options>Show Word Cloud. Again, under the Word Cloud red triangle, choose Centered Layout for layout option and Arbitrary Colors for the coloring option.

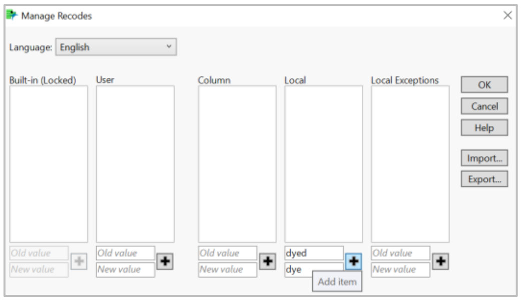
As specified earlier, text analysis is often an iterative process that revolves around curating and analyzing the list of terms. It involves close observation of the list of terms and phrases to identify stop words and the words for recoding.

Recoding

After a close observation of the terms and phrases, the following recodes are established. In Exhibit 10, the old values are recoded to new values.

Exhibit 10 List of Old Values and New Values for Recoding

Old Value	New Value
Dyed	dye
Dyeing	dye
Dyes	dye
Dyestuff	dye
Dyestuffs	dye
Multilayerd	multilayer
Multilayered	multilayer
Treated	treatment
Treating	treatment
Water	aqueous



Language: English

Local: dyed → dye

To create, click the red triangle next to Text Explorer for Title. Select Term Options>Manage Recodes.

Manage stop words

The Manage Stop Words window contains multiple lists of stop words that represent the different scopes (or locations) of specified stop words. Below each list is a text edit box and an add button. These controls enable you to add custom stop words to each scope. You can move stop words from one scope to another by dragging them. You can copy and paste items from one list to another list. Two buttons at the bottom of the window move the selected items from one scope to the next, either left or right. The X button removes the selected items from their current scope. You can edit existing items in a list by double-clicking on an item and changing the text.

A close observation of the initial result reveals the stop words that need to be included for next level of analysis.

Some form of the word “coating” occurs frequently, compared to other terms, but these terms do not provide much information to differentiate among documents. All of the terms that stem to “coating” are added to the stop word list. Similarly, the stop words listed in Exhibit 11 are determined and then added to the list of stop words.

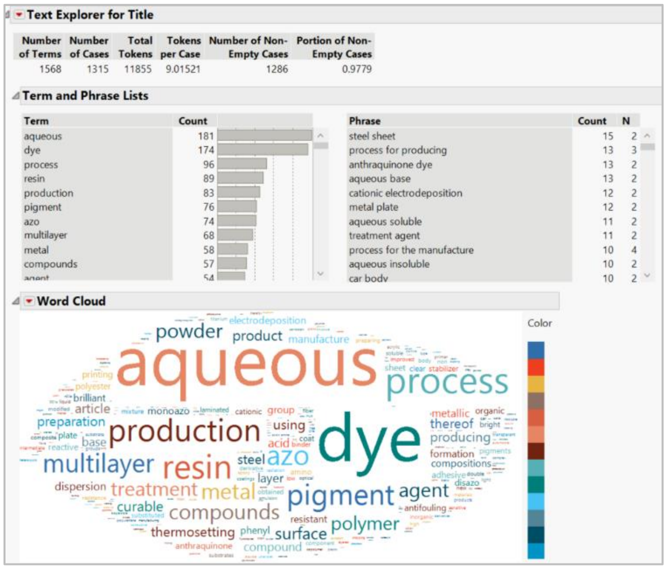
Exhibit 11 List of Stop Words

-9	-1,3,5	-2,5	1	1,5	2,6	4	5	based	forming
-1	-2	-3	1,2	10	3	4,4	6	coated	material
-1,1	-2,2	-4	1,3	12	3,4	4,5	6,14	coating	method
-1,1,1,4,4,4	-2,3	-4,6	1,4	13	3,4,	4,5,6,7	60	composition	paint
-1,2		-5	-6	2	62	4,8	7	containing	use
-1,3				2,2		40		film	

To create, click the red triangle next to Text Explorer for Title. Select Term options>Manage Stop Words. Insert all the stop words one by one as part of local stop word list.

Exhibit 12 shows the revised term list, phrase list and word cloud after incorporating the recodes and stop words.

Exhibit 12 Revised Term List, Phrase List and Word Cloud



Text Explorer for Title

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-Empty Cases	Portion of Non-Empty Cases
1568	1315	11855	9.01521	1286	0.9779

Term and Phrase Lists

Term	Count
aqueous	181
dye	174
process	96
resin	89
production	83
pigment	76
azo	74
multilayer	68
metal	58
compounds	57
agent	54

Phrase	Count	N
steel sheet	15	2
process for producing	13	3
anthraquinone dye	13	2
aqueous base	13	2
cationic electrodeposition	12	2
metal plate	12	2
aqueous soluble	11	2
treatment agent	11	2
process for the manufacture	10	4
aqueous insoluble	10	2
car body	10	2

Word Cloud

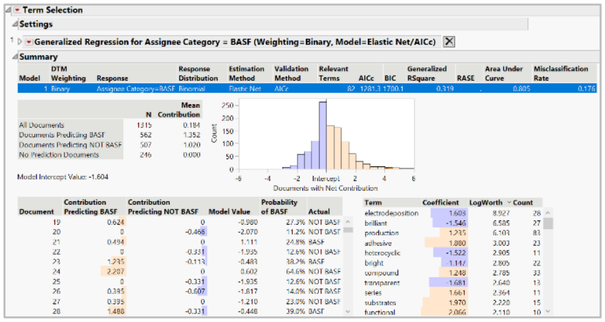
The revised result shows that there are 1,568 unique terms in 1,315 documents. In all, there are 11,853 tokenized terms. The most common term is “aqueous,” occurring 181 times.

Term selection

With term selection, you can understand the principal differences between assignees of interest. The term selection feature in JMP helps identify the preferred terms or the terms closely associated with the assignee, so you can see which term explains the which responses. It’s interesting to identify the key terms that are associated with an outcome of interest for assignees. It can also be applied as part of sentiment analysis.

The output will have many components. An output based on the generalized regression for the assignee category BASF is shown in Exhibit 12.

Exhibit 12 Term Selection



Term Selection

Settings

1 Generalized Regression for Assignee Category = BASF (Weighting=Binary, Model=Elastic Net/AICc)

Summary

Model	DTM Weighting	Response	Response Distribution	Estimation Method	Validation Method	Relevant Terms	AICc	BIC	Generalized RSquare	RASE	Area Under Curve	Misclassification Rate
1	Binary	Assignee Category=BASF	Binomial	Elastic Net	AICc	82	1281.3	1700.1	0.319	.	0.805	0.176

	N	Mean Contribution
All Documents	1315	0.184
Documents Predicting BASF	562	1.352
Documents Predicting NOT BASF	507	1.020
No Prediction Documents	246	0.000

Model Intercept Value: -1.604

Document	Contribution Predicting BASF	Contribution Predicting NOT BASF	Model Value	Probability of BASF	Actual
19	0.624	0	-0.980	27.3%	NOT BASF
20	0	-0.466	-2.070	11.2%	NOT BASF
21	0.494	0	1.111	24.8%	BASF
22	0	-0.331	-1.935	12.6%	NOT BASF
23	1.235	-0.113	-0.483	38.2%	BASF
24	2.207	0	0.602	64.6%	NOT BASF
25	0	-0.331	-1.935	12.6%	NOT BASF
26	0.395	-0.607	-1.817	14.0%	NOT BASF
27	0.395	0	-1.210	23.0%	NOT BASF
28	1.488	-0.331	-0.448	39.0%	BASF

Term	Coefficient	LogWorth	Count
electrodeposition	1.603	8.927	28
brilliant	-1.546	6.585	27
production	1.235	6.103	83
adhesive	1.880	3.003	23
heterocyclic	-1.522	2.905	11
bright	-1.147	2.805	22
compound	1.248	2.785	33
transparent	-1.681	2.640	13
series	1.661	2.364	11
substrates	1.970	2.220	15
functional	2.066	2.110	10

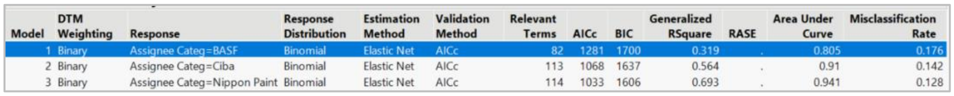
To create, Text Explorer>Title>Term Selection. Choose Assignee Category as Response Column. Under Target Level, choose assignee one by one, in this case BASF, to explore the terms closely associated with that organization. Click Run.

Sorting the term coefficients by decreasing logworth, the significant terms that are pro and con for the assignee can be found. The words associated highly with the assignee are highlighted: rose color for pro words and blue for con.

Close observation reveals the following pro terms for BASF: production, adhesive, compound and substrates. Thus, it can be concluded that for BASF, production of coatings, adhesives and electroconductive substrates are predominant focus areas.

Similarly, the term selection exercise can be replicated for CIBA and Nippon. For CIBA, the focus areas are water, insoluble pigments and fiber dying; for Nippon, the focus areas are coatings and coating performances.

Exhibit 13 Model Comparison



Model	DTM Weighting	Response	Response Distribution	Estimation Method	Validation Method	Relevant Terms	AICc	BIC	Generalized RSquare	RASE	Area Under Curve	Misclassification Rate
1	Binary	Assignee Categ=BASF	Binomial	Elastic Net	AICc	82	1281	1700	0.319	.	0.805	0.176
2	Binary	Assignee Categ=Ciba	Binomial	Elastic Net	AICc	113	1068	1637	0.564	.	0.91	0.142
3	Binary	Assignee Categ=Nippon Paint	Binomial	Elastic Net	AICc	114	1033	1606	0.693	.	0.941	0.128

One of the outputs of term selection is the list of model details. Considering the AUC values, the fit appears to be best for Nippon Paints (0.941), followed by CIBA (0.91) and BASF (0.805).

Summary

Statistical Insights

The case described the basic analysis of unstructured data using JMP Text Explorer by taking example of Patent data from google patents. Text mining is an iterative process which involved visualizing and organizing words, creating summaries, and performing word and phrase extraction.

Implications

The growth of the qualitative data or unstructured text data in digital form is growing exponentially. Applying text analysis using JMP is an easier, quick, and effective way to extract information to make meaningful insights and observations. Doing basic analysis like breaking into terms and phrases, visualizing through word cloud, and understanding the associations through Term selection helps further to apply multivariate techniques to the unstructured text data.

JMP® Features and Hints

This case used the Graph builder and Distribution platform to visualize the data. Text Exploration platform was leveraged to create word cloud and perform term selection.

Exercise

1. Extract the patent information for CPC code G10D for the following assignees:
   Yamaha Corporation
   Hoshino Gakki Co. Ltd.
   Roland Corporation
   Drum Workshop Inc.
   Randall L. May
2. Visualize the data using Graph Builder. Identify the common phrases and terms before and after incorporating stop words (if required).
3. Create a word cloud as described in the case study.
4. Identify the unique topics by company/assignee to understand the principal differences between assignees of interest.