Paper 3571-2019

Text Analytics at British Airways

Simon Cumming, British Airways PLC;

ABSTRACT

This paper descibes how British Airways, a leading British airline company, has been using SAS® Text Miner and SAS® Contextual Analysis to derive insight from textual data sources. We shall describe the 'Topic Discovery Tool', which is an application of SAS® Enterprise Guide developed at British Airways to provide a customised user interface to Text Miner. Some of the challenges of text analytics, as well as some tips and tricks will be discussed.

INTRODUCTION

This paper presents a user's view of experiences in applying text analytics to real-world business problems in British Airways. We will describe some of the business questions which are amenable to text analytics and how topic extraction and contextual rule approaches have been applied.

Text analytics methods can be split into statistical and semantic approaches. The statistical, or 'bag-of-words' approach uses a matrix describing which terms occur in which documents (a term / document matrix) and applies matrix factorisation to elicit structure and insights. This is the method used within the SAS® tools. A semantic approach requires a more complex representation which encodes the meanings and relationships of different terms in a more knowledge-based way.

In text analytics, the word 'document' can be used loosely; in some applications a 'document' is a whole corpus of text, for example a chapter of a book; in others it is a sentence or a specific comment. In this paper, a document refers to a textual comment, usually from a survey or log. It is usually written by a single author, within a specific context, but may refer to one or more issues.

Text analytics applications typically split into:

- *Descriptive text mining*, using clustering or topic extraction to understand the main issues dealt with in the text. With clustering, each 'document' maps on to one and only one cluster, whereas a document may relate to zero, one or more topics.
- *Categorization* against an existing category structure. One can either write specific rules to do this or train a data-driven propensity model if there is a labelled example training set. Some researchers and companies have also looked at 'active learning' approaches which combine elements of unsupervised and supervised learning.

The SAS® products referenced in this paper are:

- SAS® Text Miner (v14.2) , which is an add-on to SAS® Enterprise Miner, and enables the parsing and filtering of text and creation of clusters or topics, or use of a predictive model.
- SAS® Contextual analysis (v14.2) , which contains a syntax (LITI) for writing rules to create concepts and categories, with functionality to distinguish between different contexts (*e.g* for the rule to fire in the presence or absence of another nearby term or concept).

Sometimes it is beneficial to use two or more of the above approaches in tandem to gain a more complete insight or to work around some of the challenges of processing real-world natural language.

Other SAS® text mining products, including SAS® Sentiment Analysis Studio and SAS® Content Categorization are not dealt with in this paper. Text analytics products under SAS® Viya are not dealt with here.

BACKGROUND

Airline-related applications of text analytics.

Over the past fifteen years, a number of authors have published work on the use of text analytics in an airline or aviation context. These studies typically fall into the categories of air safety data analysis (for example Ananyan and Goodfellow, 2004) or gaining insight from customer feedback.

An example of use with customer feedback data is to be found in Tolety and Choudhary, 2016. The authors illustrated the use of SAS® Text Miner functionality on American Airlines reviews on Tripadvisor and other sites. They created a predictive model to assign comments to 'good' and 'bad' categories.

TEXT ANALYTICS AT BRITISH AIRWAYS

British Airways has access to rich sources of customer feedback data through a regular customer survey ('Customer Voice', provided by Maritz) conducted when a passenger has completed a flight. There is also a feedback survey for customers using the BA.com website. In addition we gather detailed feedback from cabin crew, for example on quantity and quality of food, and any issues on-board, and from engineers on the interior condition of the cabin. Verbatim comments are also available from airline colleagues across all departments via an internal social media platform (Yammer).

British Airways has, since 2015, applied text analytics, using SAS® Text Miner and SAS® Contextual Analysis, to a number of these sources. See Sankaran, S., & Taylor, G., 2015.

To facilitate pooling insight from different data sources, a generic categorization structure ('OneTree') was set up, based on an existing analysis code structure for complaint categorization. This was extended to include business areas such as Engineering and crew. A set of tables was created in the corporate data warehouse to hold input for, and output from, the text categorization process. Category rules have been set up in SAS® Contextual Analysis, and the generated code from these is scheduled and regularly run to perform the categorization.

This is an extensible and flexible structure and approach, and has shown promise, although there have been some issues around extending from a structure which has well-defined ownership in a specific part of the business, to include categories which 'belong' in different parts of the business. It is important here to get the governance structure right.

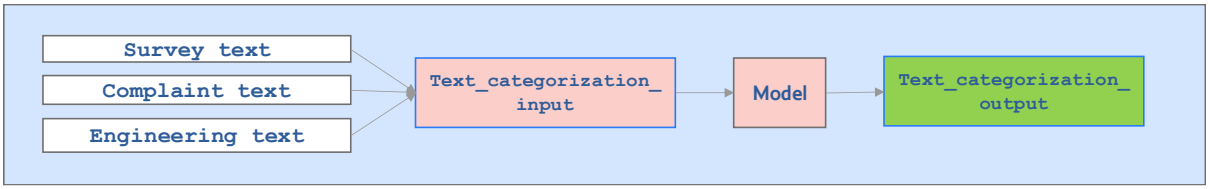


Figure 1. Text categorization table structure (schematic)

In parallel with this, BA uses SAS® Text Miner on a project basis, to create topics from text comments from the Customer Voice survey and cabin crew feedback. More recently this work has been extended to feedback on usability gathered from the airline's website. We have used User-defined Topics in Text Miner to perform Sentiment Analysis, using a variant of the AFINN_SENTIMENT table provided by SAS.

Text analytics business questions

Examples of some of the questions we have used text analytics to help tackle over the last two years are:

- Has the introduction of a new service routine in business class led to a significant increase in meal delivery times?
- Which customer touchpoints are particularly important to high value customers?
- What are the customer perceptions of the premium check-in experience at London Heathrow Terminal 3? [Most British Airways flights from London Heathrow depart from Terminal 5, with recently re-designed premium areas. Terminal 3 is an older building used for some services].
- What are the latest issues concerning usability of the BA.com website, and how are these evolving over time? Are there specific problems with particular browsers, operating systems, *etc.* ?

- What do customers need and value in terms of special and regional meals (dietary and religious requirements)?
- What are the perceived issues and specific needs of customers travelling with children?
- What are the main topics of conversation on the company social media channel (Yammer), and how is sentiment evolving? How are colleagues reacting to various company initiatives?
- Can we screen out from customer feedback any inappropriate comments on cabin crew [flight attendants]?
- Can we categorise and prioritise cabin defects for engineering attention?
- Can we better forecast the likely work required in a maintenance check?
- Can textual feedback lead to a better way of handling disruption?

The objectives and challenges of text analytics.

Analysis of natural language text will never be a perfect science, due to the huge variety of ways in which people can express themselves, and the intrinsic ambiguity in a language such as English.

Furthermore, one needs to be aware of the risk of certain different kinds of bias (even if it is subconscious) in working with text responses. Firstly, only a proportion of customers will respond to a survey or provide feedback. This sample is not necessarily random or unbiased, as those who are dissatsfied are much more likely to respond. Secondly there is a tendency for the analyst to be drawn to anecdotal or subjective evidence, especially where the sample sizes are small. Therefore it is important to relate the findings back to some base numbers, and where possible perform some statistical testing.

Because we will never achieve a 100% answer in text analytics, the 'Pareto' question arises of how much time to invest to produce a satisfactory result. Usually the analysis requires some customisation, e.g. of synonym lists, multi-word term lists, etc., and some iterative re-work.

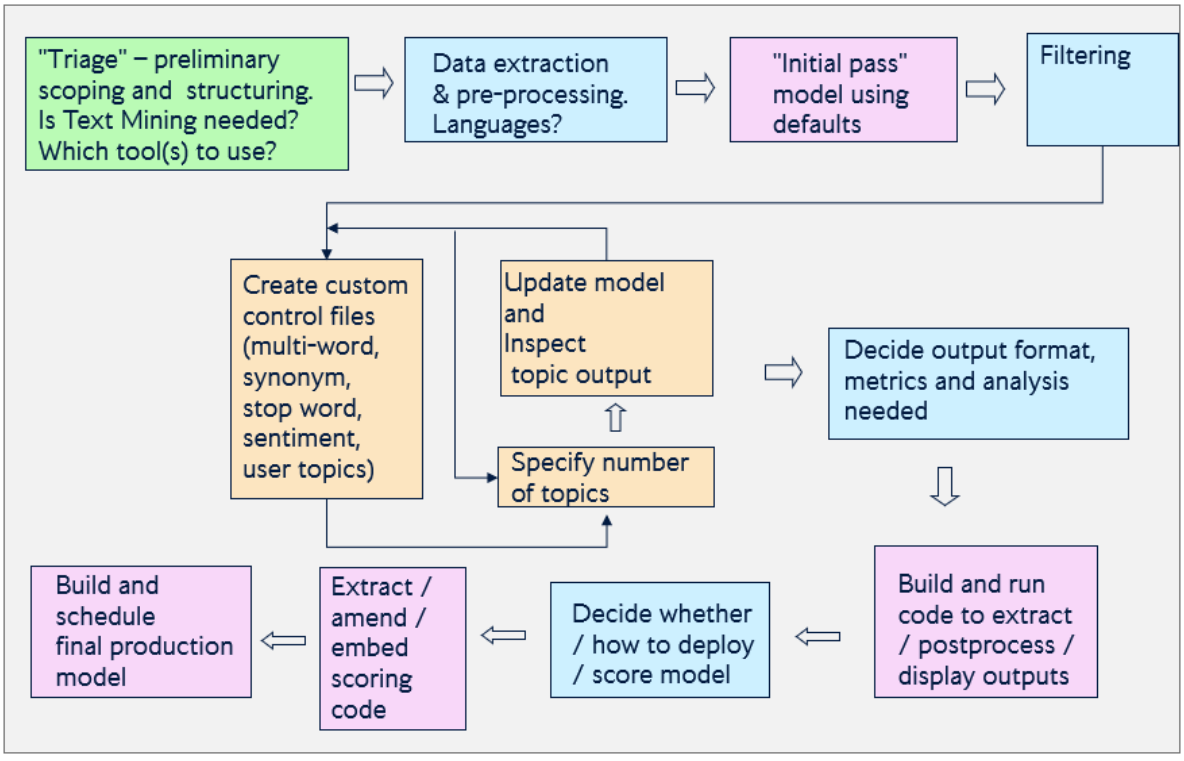


Figure 2. Iterative process for topic extraction

Topics over time.

A frequent request is to pick out the most prevalent topics and monitor over time whether new topics are emerging or the number of mentions of existing topics is increasing or declining. The complication here is that we need to be able, simultaneously, to measure the movement of topics and detect new ones. A number of approaches are possible, but none is wholly satisfactory.

1. Running a separate data-driven topic analysis each month (or time period) will bring up new topics which may be similar to previous months' but not exactly the same. It is possible using the datasets behind Text Miner to compare topics over time, by simply spotting how many of the defining terms are the same.
2. If we are just looking at historic data, we can train the model on a number of previous months and just monitor how the incidence of topics varies over time.
3. Alternatively if we use user-defined topics, we can be definite about which issues we are measuring but that will not capture new problems. So the best approach perhaps is a combination of the two.
4. Wright (2016) presents a further, intriguing idea for combining text topics and time series clustering, and gives examples of SAS® code nodes added to the Text Miner flow to perform time series similarity measurement and dimension reduction.

SAS® TEXT MINER

Text miner (within Enterprise Miner) enables a flow to be set up to perform the main components of a text analysis.

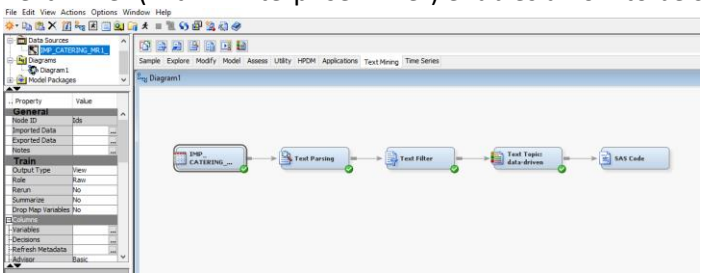


Figure 3. Typical SAS® Text Miner flow for a text topic extraction project

The usual stages are…

1. Data source / input data node: specifies which variable in which table to analyse.
2. Text parsing node (Figure 4): options to select which parts of speech to ignore and which to process; location of multi-word term list, synonym list and stop word list.
3. Text filter node (Figure 5, Figure 6: options for weightings, dictionaries and filtering. The Filter Viewer enables interactive selection of terms.
4. Text topic node: to specify the number of topics and provide a table to define any user-defined topics. There is also the option of creating clusters, where each 'document' maps onto one and only one cluster, but here, for practical purposes we have found the topics node more useful, because each document (survey response in this instance) can map onto zero, one or more topics; and one can mix user-defined and data-driven topics.

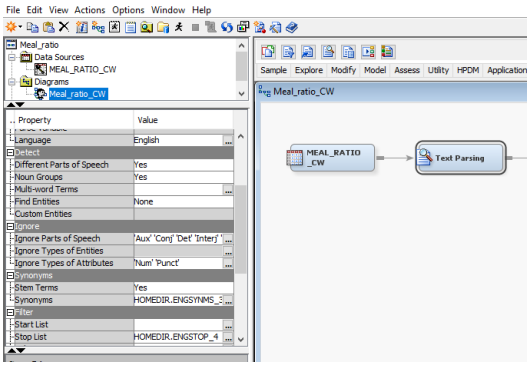


Figure 4. Text parsing node

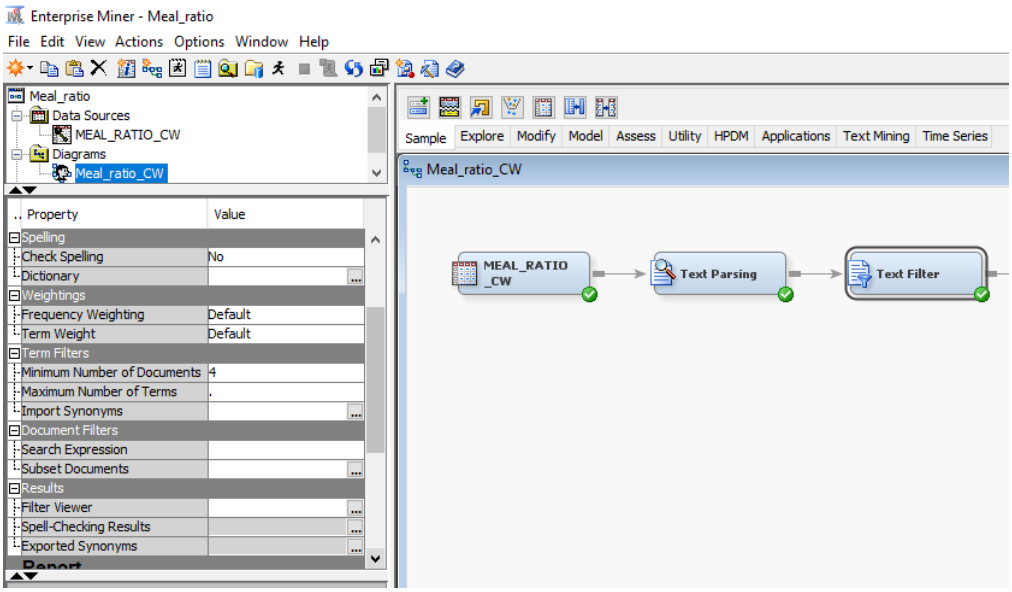


Figure 5. text filter node

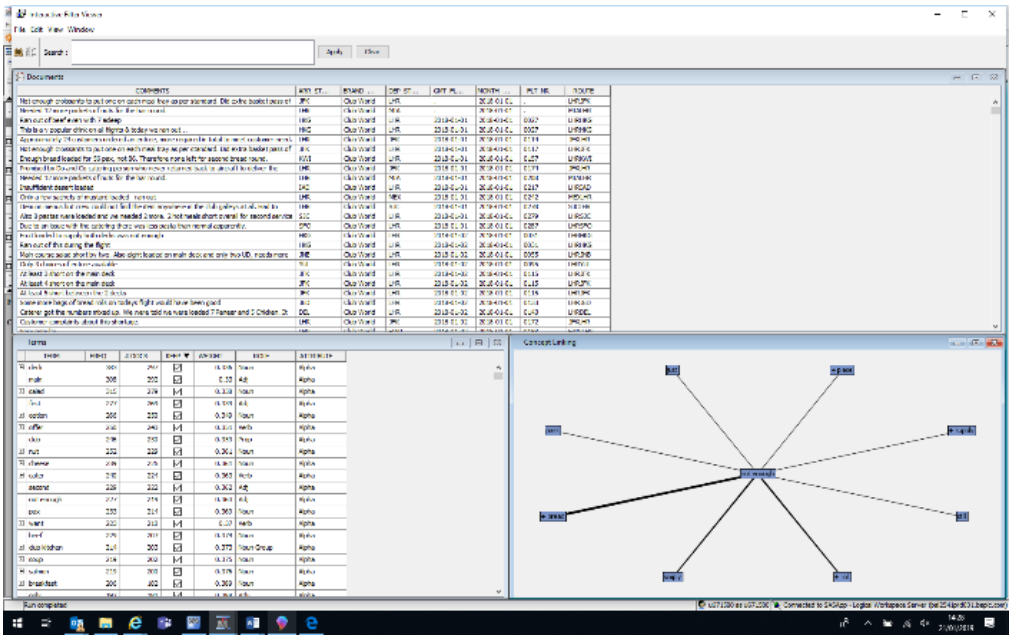


Figure 6. Output from text filter node including 'concept linking'

Back-end tables in Text Miner

Text Miner generates SAS® tables (datasets) which can be useful to feed into user-defined displays and analyses. Examples are: The Texttopic_topics table (see Figure 7) gives, for each topic, the top 5 terms and the number of documents for which that topic is relevant.

Topic ID	Topic	# Docs
1	+bread,+roll,+soup,+bag,+bread roll	465
2	club,+kitchen,+club kitchen,+trolley,+load	411
3	+deck,main,upper,main deck,+upper deck	301
4	+choice,first,+customer,+refuse,choice	509
5	enough,not,+load,+pasta,loaded	589
6	+starter,+salmon,+salmon starter,+smoke,+ratio	431
7	popular,popular choice,+choice,+option,always	459
8	+scone,+tea,afternoon,+afternoon tea,+sandwich	237
9	+cheese,+biscuit,+goat,+cracker,+chutney	283

Figure 7. Example of topics dataset from catering reports analysis (some columns have been hidden)

Texttopic_train* (see Figure 8): gives a copy of the input dataset, together with weightings for each document for each topic, and indicator variables to describe which topics are featuring for each document. The labels on the topic variables give the top five terms occurring in each topic.

*There is a naming convention such that the output of the *n*th topic node in a workspace is denoted by texttopic<*n*>_train, e.g. texttopic2_train is the output of the second text topic node.

	_1_0_+pasta,increase,+reduce,+ratio,+chicken	_1_0_not enough,+load,+cabin,+pasta,upper	_1_0_+soup,+flask,+load,+offer,+bread	_1_0_choice,+meet,first,+customer,+demand	Document	Comments
25	0	0	0	0	25	Very popular
26	0	0	0	0	26	Take the second choice off or load a few extra to cover adhoc veg requirement. The pea and bean salad was not appetising and tasted awful
27	0	0	0	0	27	No nuts left from previous sector. Or offloaded by catering in error.
28	0	0	0	0	28	Most pax wanted full meal.good night express not popular so very tight for catering.
29	1	0	0	0	29	More pasta pls
30	0	1	0	0	30	No dry stores loaded for Club cabin so no specialty teas etc.

Figure 8. Example of part of a topic_train dataset. Note that the columns for the topics are given labels by Text Miner, using the top 5 terms in the topic.

THE TOPIC DISCOVERY TOOL

The **'Topic Discovery Tool'** (TDT) has been created at British Airways (with assistance from Aquila Insight) in SAS® Enterprise Guide using a standard exported flow from SAS® Text Miner, similar to the one displayed in Figure 3).

The TDT consists of a sequence of SAS® macros embodied in an Enterprise Guide flow. These are primarily set up to read and analyse text from the regular 'Customer Voice' survey. The TDT can also be used in an *ad-hoc* way on other sources of textual data.

It extracts data from the data warehouse for a specified period of time, and for specified variables containing the text responses for particular touchpoints in the customer survey, and creates a standard input set. It then runs generated code from Text Miner to extract the topics, and performs various standard post-processing tasks to give an enhanced set of outputs. It uses the texttopic_topics and texttopic_train output datasets described above (see Figure 7Figure 7, Figure 8), to provide a standardized and user-friendly way to analyse text sources rapidly. A schematic of the structure of the macros is given in Figure 9, and examples of the outputs in Figure 10 and Figure 11.

The TDT has proven very beneficial in terms of identifying 'pain points' *i.e.* parts of the customer journey which are generating issues; also for identifying areas which are performing well and where customers are particularly pleased. In particular, it is possible to analyse in detail specific issues which are not categorized in existing systems and for which data is otherwise difficult to obtain, or to examine the effect of particular changes to product or service.

Insights from this analysis can be used directly to inform specific decisions or business cases, *e.g.* approaching particular catering suppliers about issues with food and drink products on board or fine-tuning the quality and quantity of refreshments and the speed and pace of service.

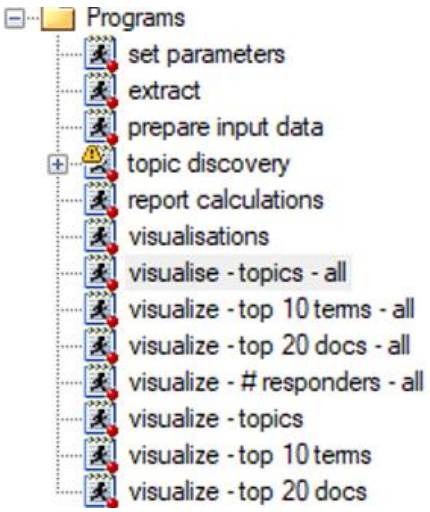
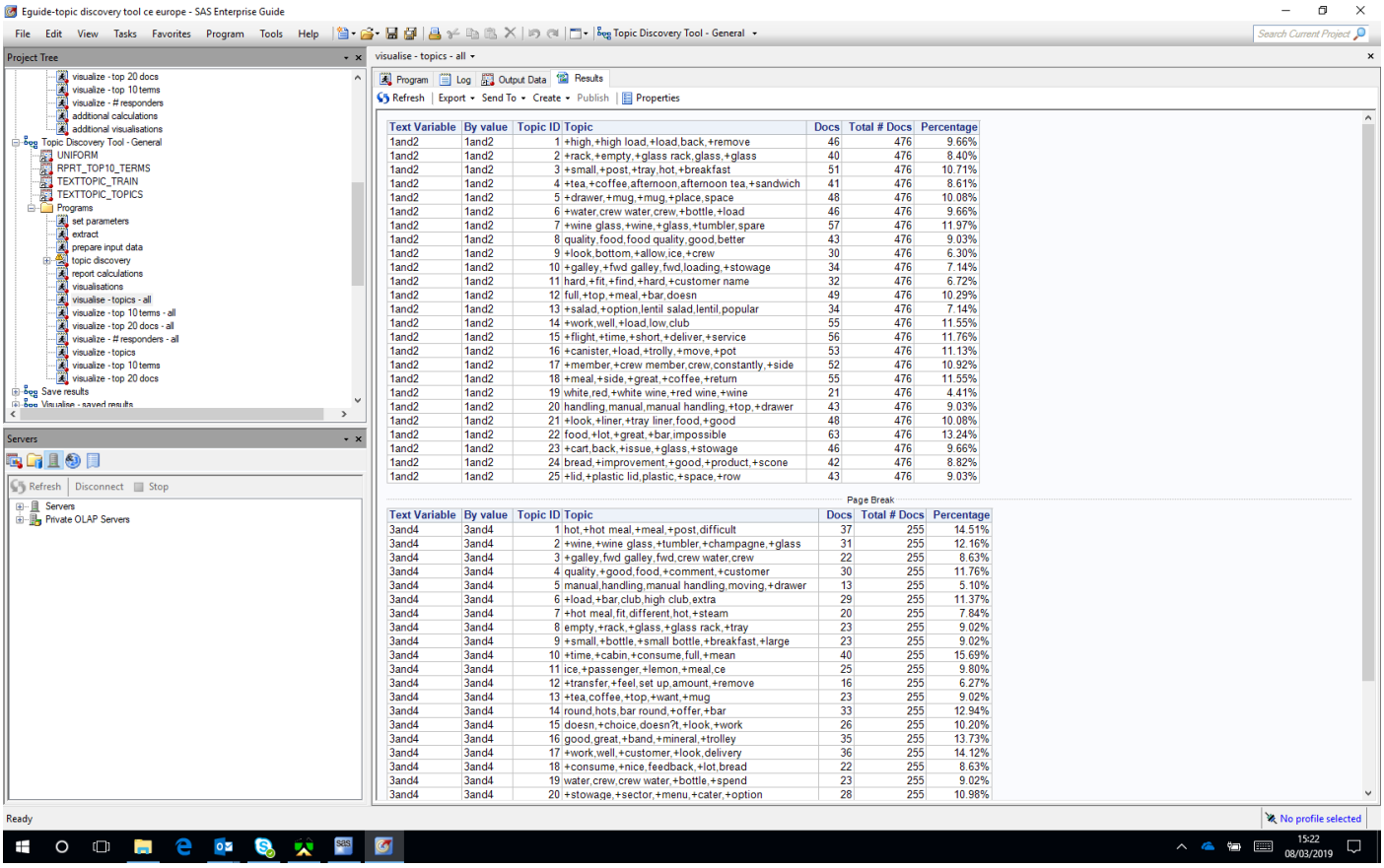


Figure 9. Sequence of macros in Topic Discovery Tool



Text Variable	By value	Topic ID	Topic	Docs	Total # Docs	Percentage
1and2	1and2	1	+high,+high load,+load,back,+remove	46	476	9.66%
1and2	1and2	2	+rack,+empty,+glass rack,glass,+glass	40	476	8.40%
1and2	1and2	3	+small,+post,+tray,hot,+breakfast	51	476	10.71%
1and2	1and2	4	+tea,+coffee,afternoon,afternoon tea,+sandwich	41	476	8.61%
1and2	1and2	5	+drawer,+mug,+mug,+place,space	48	476	10.08%
1and2	1and2	6	+water,crew water,crew,+bottle,+load	46	476	9.66%
1and2	1and2	7	+wine glass,+wine,+glass,+tumbler,spare	57	476	11.97%
1and2	1and2	8	quality,food,food quality,good,better	43	476	9.03%
1and2	1and2	9	+look,bottom,+allow,ice,+crew	30	476	6.30%
1and2	1and2	10	+galley,+fwd galley,fwd,loading,+stowage	34	476	7.14%
1and2	1and2	11	hard,+fit,+find,+hard,+customer name	32	476	6.72%
1and2	1and2	12	full,+top,+meal,+bar,doesn	49	476	10.29%
1and2	1and2	13	+salad,+option,lentil salad,lentil,popular	34	476	7.14%
1and2	1and2	14	+work,well,+load,low,club	55	476	11.55%
1and2	1and2	15	+flight,+time,+short,+deliver,+service	56	476	11.76%
1and2	1and2	16	+canister,+load,+trolly,+move,+pot	53	476	11.13%
1and2	1and2	17	+member,+crew member,crew,constantly,+side	52	476	10.92%
1and2	1and2	18	+meal,+side,+great,+coffee,+return	55	476	11.55%
1and2	1and2	19	white,red,+white wine,+red wine,+wine	21	476	4.41%
1and2	1and2	20	handling,manual,manual handling,+top,+drawer	43	476	9.03%
1and2	1and2	21	+look,+liner,+tray liner,food,+good	48	476	10.08%
1and2	1and2	22	food,+lot,+great,+bar,impossible	63	476	13.24%
1and2	1and2	23	+cart,back,+issue,+glass,+stowage	46	476	9.66%
1and2	1and2	24	bread,+improvement,+good,+product,+scone	42	476	8.82%
1and2	1and2	25	+lid,+plastic lid,plastic,+space,+row	43	476	9.03%

Page Break

Text Variable	By value	Topic ID	Topic	Docs	Total # Docs	Percentage
3and4	3and4	1	hot,+hot meal,+meal,+post,difficult	37	255	14.51%
3and4	3and4	2	+wine,+wine glass,+tumbler,+champagne,+glass	31	255	12.16%
3and4	3and4	3	+galley,fwd galley,fwd,crew water,crew	22	255	8.63%
3and4	3and4	4	quality,+good,food,+comment,+customer	30	255	11.76%
3and4	3and4	5	manual,handling,manual handling,moving,+drawer	13	255	5.10%
3and4	3and4	6	+load,+bar,club,high club,extra	29	255	11.37%
3and4	3and4	7	+hot meal,fit,different,hot,+steam	20	255	7.84%
3and4	3and4	8	empty,+rack,+glass,+glass rack,+tray	23	255	9.02%
3and4	3and4	9	+small,+bottle,+small bottle,+breakfast,+large	23	255	9.02%
3and4	3and4	10	+time,+cabin,+consume,full,+mean	40	255	15.69%
3and4	3and4	11	ice,+passenger,+lemon,+meal,ce	25	255	9.80%
3and4	3and4	12	+transfer,+feel,set up,amount,+remove	16	255	6.27%
3and4	3and4	13	+tea,coffee,+top,+want,+mug	23	255	9.02%
3and4	3and4	14	round,hots,bar round,+offer,+bar	33	255	12.94%
3and4	3and4	15	doesn,+choice,doesn?t,+look,+work	26	255	10.20%
3and4	3and4	16	good,great,+band,+mineral,+trolley	35	255	13.73%
3and4	3and4	17	+work,well,+customer,+look,delivery	36	255	14.12%
3and4	3and4	18	+consume,+nice,feedback,+lot,bread	22	255	8.63%
3and4	3and4	19	water,crew,crew water,+bottle,+spend	23	255	9.02%
3and4	3and4	20	+stowage,+sector,+menu,+cater,+option	28	255	10.98%

Figure 10. Topic output in Topic Discovery Tool (Enterprise Guide version)

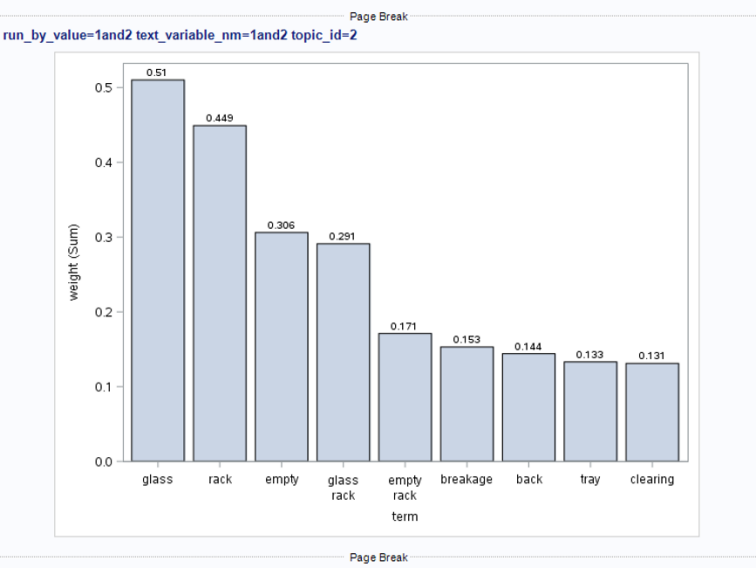


Figure 11. Example chart output from Topic Discovery Tool

Macros for routing outputs to Powerpoint, Excel and Word.

Aside from using the Topic Discovery Tool, it is useful to surface the various output tables from Text Miner into reports which can either be sent to Microsoft Office products, viz Excel, Word or Powerpoint, or routed to a data warehouse (Teradata) table for input into Tableau for a more interactive report.

I have found it useful to write SAS® macros which access the output tables from Text Miner or Contextual Analysis and set out the following.

1. Topic details, including percentages of overall number of documents
2. Top terms in each topic
3. A time series showing the evolution of topic instances.
4. Overlap of topics between different runs (e.g. different time periods) expressed as the number of main terms they have in common.

SAS® procedures PROC TEMPLATE and PROC ODSTABLE can be used for formatting the output (specifying custom font size, colour, typeface etc.).

Examples of projects using Text Miner or the Topic Discovery Tool.

1. High value customers

The purpose here was to determine which customer touchpoints were giving rise to particular issues for high-value customers. Here a high value customer is defined as having a high percentile total revenue for the last three years. Text analysis was used to compare directly the sentiment scores for high value customers versus those for customers in general. An index measure was derived, and statistical testing used to determine significance. Food and drink, aircraft cleanliness, online check-in and in-flight entertainment were found to be the elements where the sentiment of high-value customers differed particularly from customers in general. An example output is given in Figure 12.

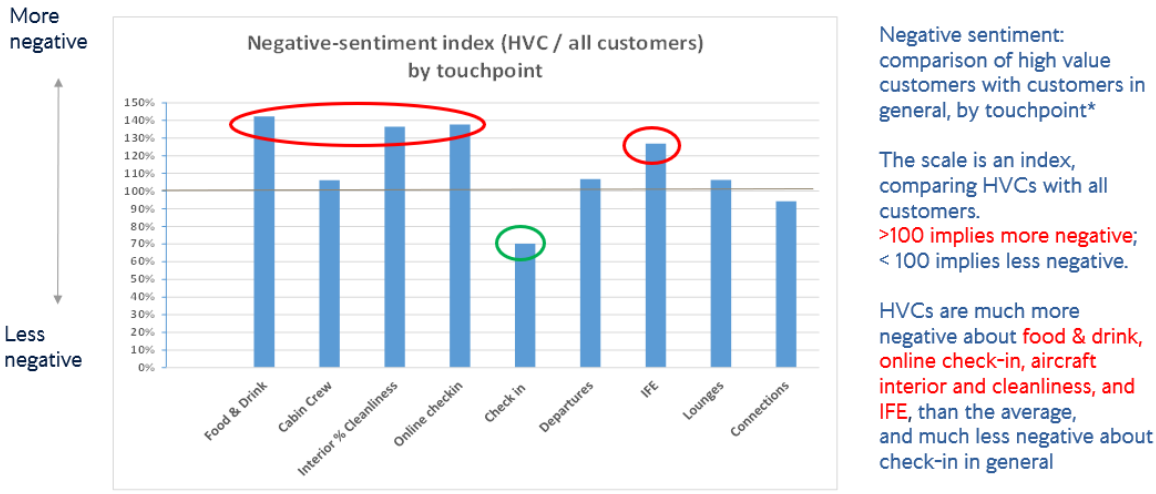


Figure 12. Index of negative sentiment for different touchpoints for high value customers versus customers in general

2. Website feedback

A recently-implemented button on the British Airways website (BA.com) asks the customer to give feedback on the usability of the site. Topics were built from the text responses using Text Miner, and analysed by operating system, browser type and device type to pinpoint any specific issues requiring attention, and to show which topics are arising or declining over time. Example results are given in Figure 13.

Topic no	OS	GNU/Linux	MacOSX	Win10	Win7	MacOSX	Win10	Win8.1	Win10	Win10
	Browser_type	Chrome	Chrome	Opera	IE	Safari	Chrome	IE	Firefox	IE
	Device	Desktop	Desktop	Desktop	Desktop	Desktop	Desktop	Desktop	Desktop	Tablet
	count	***26***	***502***	***24***	***365***	***1375***	***1535***	***43***	***192***	***81***
1	return,return flight,flight,outbound,outbound flight	0	0.03	0.04	0.08	0.05	0.06	0.05	0.06	0.05
2	email,address,email address,password,register	0.08	0.06	0.17	0.06	0.04	0.07	0.02	0.07	0.05
3	payment,card,authorisation,screen,credit	0.27	0.12	0.13	0.05	0.12	0.09	0.14	0.11	0.04
4	price,hotel,car,want,pay	0.12	0.08	0	0.08	0.07	0.09	0.07	0.09	0.11
5	page,load,home page,home,search	0	0.08	0.08	0.07	0.14	0.09	0.12	0.07	0.11
6	account,avios,password,log,number	0.15	0.1	0.13	0.12	0.09	0.12	0.09	0.13	0.14
7	booking,complete,error,incomplete,unable	0.15	0.1	0	0.09	0.11	0.09	0.14	0.09	0.16
8	easy,navigate,find,information,clear	0	0.03	0.08	0.04	0.03	0.04	0	0.06	0.06
9	book,flight,seat,time,flight	0.04	0.12	0.04	0.13	0.18	0.13	0.16	0.08	0.14
10	profile,update,detail,update,date	0.04	0.12	0	0.04	0.09	0.06	0	0.06	0.04
11	date,select,outbound,time,option	0.15	0.09	0.08	0.12	0.1	0.13	0.05	0.09	0.06
12	add,passenger,baggage,want,option	0.15	0.13	0.04	0.07	0.08	0.11	0.09	0.09	0.07
13	site,web,web site,want,avios	0.04	0.06	0.13	0.09	0.08	0.07	0.05	0.07	0.09
14	avios,point,avios point,pay,find	0.04	0.06	0.08	0.07	0.05	0.07	0.07	0.06	0.04
15	pay,seat,ticket,time,choose	0.04	0.07	0.04	0.04	0.06	0.07	0.12	0.06	0.05

Figure 13. Negative and positive sentiment and incidence of specific topics, by operating system, browser type and device type. Data for Jan / Feb 2019.

CONTEXTUAL ANALYSIS

Apart from the use of SAS® Contextual Analysis to populate the 'OneTree' categorization structure, we have also used it in various projects requiring more in-depth or specific identification of concepts and categories.

Examples of use of Contextual Analysis

1. Duration of food and drink service

As part of the measurement of the effectiveness of some changes to the food and drink service in longhaul business class (Club World) in 2017-18, we used the verbatim feedback from the Customer Voice survey, to identify occasions where customers were mentioning that the meal service had taken a long time. Contextual analysis rules were used to capture the fact that customers could refer to the duration of service in a number of ways, e.g. 'dinner service was slow', 'I had to wait [x] minutes for the food' , etc. Some examples of concept definitions using the 'CLASSIFIER' syntax are given in Figure 14.

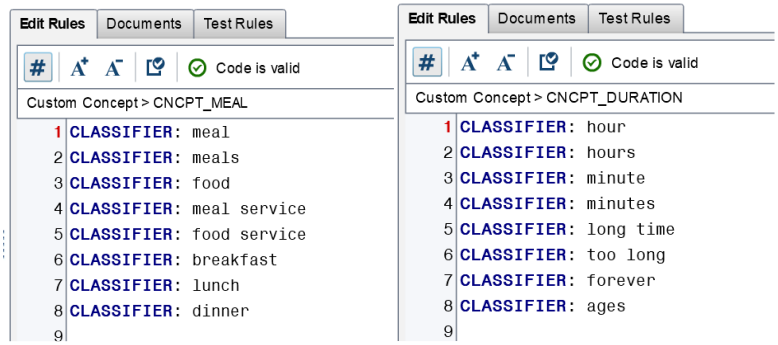


Figure 14. Example of 'meal' and 'duration' concepts

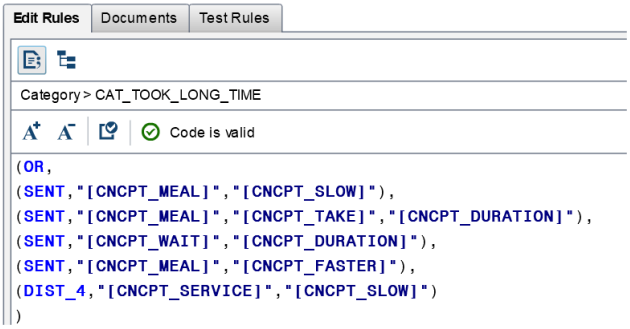


Figure 15. Example of categorization rule in Contextual Analysis

The concepts are combined using logical operators in a Category rule (see Figure 15). The SENT operator picks out where two terms or concepts occur in the same sentence. The DIST_4 operator picks out items within 4 words of each other. The "[…]" syntax allows us to refer to a Concept we have already defined (see Figure 14).

2. Use of context: food and drink quantity

Within SAS® Contextual Analysis, we can use contextual rules to pick out, from reports from crew for example, which food and drink items are in greater or lesser demand, and when an excess was loaded or where supply was insufficient. Figure 16 shows in schematic form the general approach.

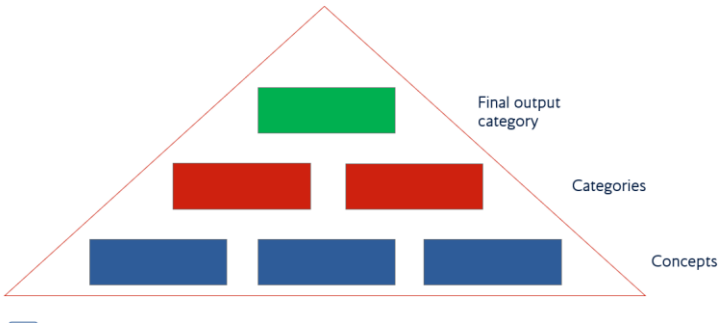


Figure 16. Structure for contextual analysis

Specifically, the approach here is:

1. Define the base concepts (using CLASSIFIER: rules to group terms together into lists), e.g. a concept for the items we are interested in, and a separate concept for different ways of expressing that there is not enough. (See Figure 17).

Custom Concept > CncptInsufficient

```
1  CLASSIFIER: not enough
2  CLASSIFIER: insufficient
3  CLASSIFIER: short
4  CLASSIFIER: didn't have enough
5  CLASSIFIER: only
6  CLASSIFIER: more
7  CLASSIFIER: few
8  CLASSIFIER: no
9  CLASSIFIER: run out
10 CLASSIFIER: ran out
11 CLASSIFIER: running out
```

Custom Concept > CncptItems

```
1  CLASSIFIER: nuts
2  CLASSIFIER: nut
3  CLASSIFIER: pretzels
4  CLASSIFIER: bread
5  CLASSIFIER: croissant
6  CLASSIFIER: croissants
7  CLASSIFIER: salad
8  CLASSIFIER: salad dressing
9  CLASSIFIER: pasta
10 CLASSIFIER: beef
```

Figure 17. Basic concept examples for contextual analysis

2. Set up *contextual* rules such as the one in Figure 18 which identifies the *items* of which not enough has been loaded, i.e it is picking out terms in the item list only when they occur with a term from the 'insufficient' list. Note the use of the C_CONCEPT keyword and the _c{..} syntax to say that we wish to flag the items. See Figure 19 for an example.

Custom Concept > CntxtInsufficient

```
1  C_CONCEPT: CncptInsufficient _c{CncptItems}
```

Figure 18. Example of C_CONCEPT rule

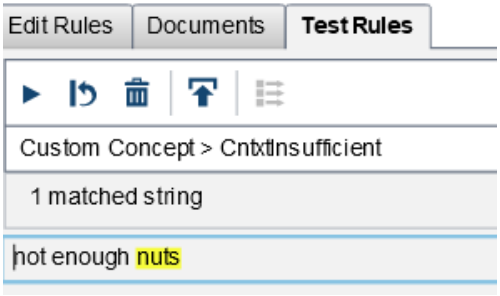


Figure 19. Testing the C_CONCEPT rule

Figure 19 shows an example of testing this rule using a typed-in string. Here the word 'nuts' is highlighted, as it follows the phrase 'not enough'.

3. We can then define a category rule (Figure 20) which pulls through the contextual rule to the next level.

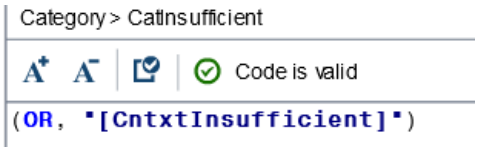


Figure 20. Category rule

4. If necessary we can define an overall 'output' category rule (Figure 21), which 'ORs' together other category rules. This is useful, for example, to measure how many of the comments we are successfully categorizing. Figure 22 shows the results output to a SAS® dataset. Note that to make meaningful sense of the output in business terms it would then be joined back to an input dataset with details of the flight date, flight number and route, etc., and some summary statistics derived.

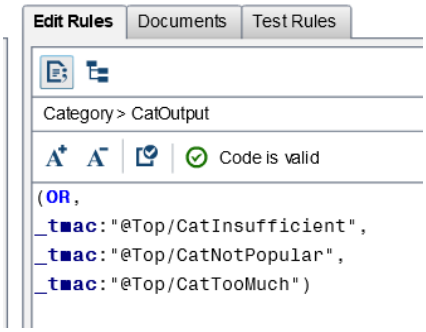


Figure 21. Referring to other category rules

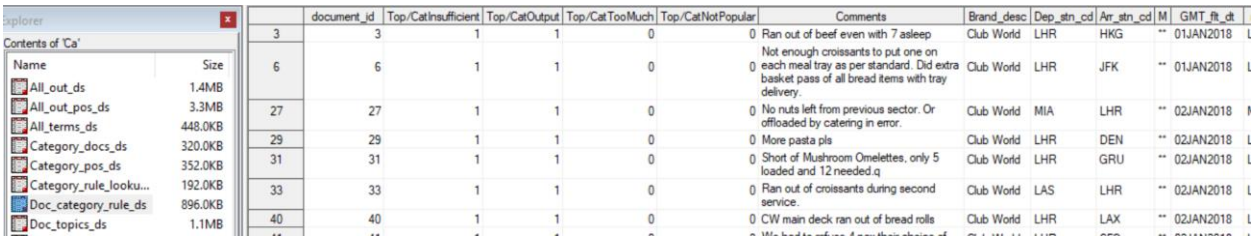


Figure 22. Example of Contextual Analysis output

TIPS AND TRICKS?

Text analytics is as much an art as a science and requires lateral thinking. Every practitioner has their own ways of approaching different issues.

- Lateral thinking

Think about whether we can use any information about the circumstances under which the data was captured, *e.g.* if it is from a problem-log we may be able to make reasonable assumptions about the content being mainly to do with faults, rather than having to laboriously enumerate all the ways someone could express this.

- Keeping it simple

In general, experience has shown that it is best to keep text analytics rules as simple as possible, and often live with an imperfect outcome especially if the requirement can be limited to comparing different subsets of data, or measurement over time.

- Use of context

Sometimes we do need to make full use of the contextual functionality in CA, for example to distinguish between 'seat 24A is broken' and 'the reading light by seat 24A is broken' and so on.

- It depends on the data!

The success or otherwise of textual processing depends on the quality and nature of the input data: short and focussed comments give rise to the best results; lengthy text documents such as complaints can be difficult to analyse cleanly.

Suitable use of the stop word list may be the key here, together with perhaps using regular expressions to identify any formulaic entities such as email addresses, flight numbers *etc.*

- Sentiment analysis

SAS has a specific tool for performing sentiment analysis [not licensed as part of British Airways' SAS® estate], but sentiment scores are also built into Contextual Analysis and sentiment analysis can be performed in Text Miner. This is done by setting up user-defined topics in the Text Topic node using a sentiment file provided (SAMPSIO.AFINN_SENTIMENT or a customised version of it). One has a choice of whether to use a sliding scale (from -1 being very negative through to +1 indicating very positive terms) or whether to use a binary (+/- 1) representation. Bartlett and Albright (2008) describe a number of further techniques for dealing with sentiment classification.

Many other commercially available software products and agency services also provide sentiment analysis in one form or another. There is no 'right answer' or universally accepted standard scale for sentiment. In addition to the well-known issues around sarcasm etc., there are some domains where 'negative' words imply a positive outcome. (Think for example of someone raising money for a charity which is helping people who are sick or disadvantaged in various ways).

It is preferable to stay within a relative or comparative framework rather than pinning too much importance on the absolute value.

<u>Deployment of text analytics code</u>

For the 'OneTree' categorization, which is regularly run on engineeng, survey and complaint data, we have extracted SAS® code from Contextual Analysis, together with the `.mco` and `.li` files which contain the concept and category rules, and this code is run weekly, the output being sent to Teradata tables in a corporate data warehouse.

OPPORTUNITIES FOR FUTURE TEXT ANALYTICS WORK AT BRITISH AIRWAYS

1. Feedback from crew. A newly-commissioned system, CITA (Crew Insight to Action) is capturing verbatim text from cabin crew on issues on board, such as quality and quantity of food and drink, etc. While the system has some semi-automatic categorization functionality, it is likely that text analytics will play a key part in improving the way in which this information s turned into immediately actionable insights.
2. Yammer (colleague social medium). Data derived from the Yammer APIs will shortly be available for analysis, to give understanding of current topics and the employee reaction to various business decisions.
3. Ways of extending from 'sentiment analysis' to provide a measure of 'importance' will also be key in obtaining the most useful and succinct insights.
4. There is an opportunity to extend the power of the text analytics using some of the ideas discussed in Albright, Punuru and Surratt (2013) regarding using rules set up in Contextual Analysis as custom entities in Text Miner.
5. There are also opportunities to explore enhanced methods for sentiment analysis (see Bartlett & Albright, 2008).
6. Twitter is another data source that could be used to understand customer sentiment, particularly during disruption. See Albright, Foley and Devarajan (2010).

DISCUSSION: SAS® AND OPEN-SOURCE TOOLS

A variety of text analytics functionality is readily available using open-source libraries in the python or 'R' languages. While these have a number of attractive features and may provide for more flexibility and customizability, the SAS® software suite allows us to work seamlessly within the same environment that we use for the majority of data analysis used to support business decisions.

However, as time progresses, use of open-source approaches for text analysis is increasing, and provides the ability to be more in control of the algorithms and flows involved. We are already starting to explore these techniques.

Future: Text analytics in SAS® Viya

The new product set in the Viya architecture eliminates the dichotomy between Text Miner and Contextual Analysis, and enables better interworking of SAS® and open-source code.

CONCLUSION

Text analytics provides a valuable method for assessing customer sentiment, to gain high-level insight into key issues and topics, and how these are evolving; also to probe into detail to provide specific actionable points, particularly where systematic scores or numerical data are not being captured otherwise. The SAS® tools Text Miner and Contextual Analysis provide a framework for deriving numerical scores from text input; these can then be processed using the usual SAS® methods. It is important to keep in mind how the text, and text-derived information, relates to overall numbers, and be aware of potential biases. It is useful to be able to surface the outputs and insights derived from text analytics in a number of different ways, both to a user interface and as a regular feed into another system or data warehouse.

APPENDIX: SAS® CODE AND MACROS

We have found it useful to write some utility macros to enable us to make use of the output datasets from Text miner and Contextual Analysis in a systematic way. Some examples follow.

1. Show top *n* documents for each topic.

```
%macro top_records /* for one topic at a time */
(
      emws=, /* workspace nr */
      topic_node=, /* number of text topic node within the workspace */

      topic_nr=, /* number of this topic */
      tag=, /* first part of dataset name for output */
      textvar=, /* name of text variable */
      n=, /* number of documents to retrieve */
      n_chars= /* retrieve first n_chars characters of each document */
);

proc sort data=emws&emws..texttopic&topic_node._train
(keep=&textvar texttopic&topic_node._raw&topic_nr.)
out= homedir.&tag._&topic_nr. ;
/* sort in order of weighting */
by descending texttopic&topic_node._raw&topic_nr. ;
run;
data &syslast;
/* pick out top n documents and first n_chars characters in those documents */
/*length topic_desc $25; */ /* include if we want the label as a column */
set &syslast(obs=&n);
/*topic_desc = &label; */ /* include if we want the label as a column */
&textvar = substr(&textvar,1,&n_chars);

run;
%mend;
%macro toprecs_all_topics /* just defines a do loop to call top_records n_topics times */
      (emws=,
      topic_node=,
      tag=,
      n_topics=,
      textvar=,
      n=,
      n_chars=);
            %do topic = 1 %to &n_topics.;
            %top_records
                  (
                        emws=&emws,
                        topic_node=&topic_node,
                        topic_nr=&topic,
                        tag=&tag,
                        textvar=&textvar,
                        n=&n,
                        n_chars=&n_chars
                  );
            %end;
%mend;
%toprecs_all_topics
      (
            emws=1, /* workspace */
            topic_node=6, /* number of text topic node */
            tag=my_output_, /* first part of dataset name for output */
            n_topics=17, /* number of topics */
            textvar=comment, /* input text variable */
```

```
            n=10, /* shows top n documents for topic in order of weight */
            n_chars=500 /* shows first n_chars characters of each document */
      );
```

2. Export results to Powerpoint

```
%macro to_ppt
/* formats and outputs a SAS®  dataset to Microsoft Powerpoint */
(
      lib=, /* library name */
      ds=,  /* datset name */
      file=, /* powerpoint file */
      colour=,
      font=, /* font size */
      header=,
      h_colour=, /* header colour */
      h_font= /* header font size */
);

      proc contents data=&lib..&ds.  out=contents(keep=name type) noprint; run;

      proc sql noprint;
            select strip(put(count(name),2.)) into :nvars from contents;
            select distinct name into :var1 - :var&nvars from contents;
            select distinct name into :vars separated by ' ' from contents;
      quit;

      proc template; /* to make sure the column headers come out with
      the right font size and colour and justification */
           define style styles.PowerPointNew;
             parent=styles.PowerPointLight;
             style header from header /
      /*         backgroundcolor = …*/
                 color = DarkBlue
                   just=l
      /*         fontfamily = "SAS Monospace"  */
                 fontsize = /*10pt*/ &font.
      /*         fontweight = bold*/
                 ;
           end;
      run;
            ods powerpoint file="&file" style=styles.PowerPointNew;
            /* note use of style= to access the template created above */

      proc odstable data=&lib..&ds. ;

         define header myheader1;
            text "&header.";
            style={color=&h_colour. fontsize=&h_font. just=l};
         end;

      column &vars.;
      %local i;
      %do i = 1 %to &nvars.;
         define &&var&i;
                  header = myheader1;
                  style={color=&colour. fontsize=&font. just=l};
         end;
      %end;
      run;
      quit;
```

```
      ods powerpoint close;
%mend;
```

ACKNOWLEDGMENTS

The author would like to acknowledge the valuable contributions of:
Dr Leen van Langenhoven, formerly of Aquila Insight (for the Topic Discovery Tool work); Pandeep Khatker, Stefan Jackson, Andrew Sankey and Dr Tom Ravalde from British Airways; Srikanth Sankaran, formerly of British Airways; and Matt Stainer and Kayne Putman from SAS (UK).

REFERENCES

Albright, R., Cox, J., Ning, J. 2016. "Getting More from the Singular Value Decomposition (SVD): Enhance Your Models with Document, Sentence, and Term Representations", *Proceedings of the SAS Global Forum 2016 Conference*. Available at http://support.SAS.com/resources/papers/proceedings16/SAS6241-2016.pdf

Albright, R., Foley, R., Devarajan, R. 2010. Listening to the Twitter Conversation, *Proceedings of the SAS Global Forum 2010 Conference*. SAS Institute, Inc.
Available at http://support.sas.com/resources/papers/proceedings10/355-2010.pdf

Albright, R., Punuru, J., Surratt, L. 2013. "Relate, Retain, and Remodel: Creating and Using Context-Sensitive Linguistic Features in Text Mining Models", *Proceedings of the SAS Global Forum 2013 Conference*. SAS Institute, Inc.
Available at https://support.sas.com/resources/papers/proceedings13/100-2013.pdf

Ananyan, S. and Goodfellow, M. 2004. "New capabilities of PolyAnalyst text and data mining applied to STEADES data at IATA" (Megaputer / IATA / Global Aviation Information Network)
Available at https://flightsafety.org/wp-content/uploads/2016/09/IATA_data_mining_report.pdf

Bartlett, J., Albright, R. 2008. "Coming to a Theater Near You! Sentiment Classification Techniques Using SAS® Text Miner" *Proceedings of the SAS Global Forum 2008 Conference*. SAS Institute, Inc.
Available at https://support.sas.com/resources/papers/proceedings/pdfs/sgf2008/152-2008.pdf

Bee Yee Liau, Pei Pei Tan. 2014. "Gaining customer knowledge in low cost airlines through text mining", *Industrial Management & Data Systems*, Vol. 114 Issue: 9, pp.1344-1359,
Available at https://doi.org/10.1108/IMDS-07-2014-0225

Blei, D M., Ng, A Y., Jordan, M I. 2003. In Lafferty, J. (ed.) "Latent Dirichlet Allocation". *Journal of Machine Learning Research.* **3** (4–5): pp. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

Chakraborty, G., Pagolu, M K. 2014. "Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining", *Proceedings of the SAS Global Forum 2014 Conference*. SAS Institute, Inc.
Available at http://support.sas.com/resources/papers/proceedings14/1288-2014.pdf

Landauer, T., *et al*. 1998, "Learning Human-like Knowledge by Singular Value Decomposition: A Progress Report". In Jordan, M.I., Kearns, M.J., & Solla, S.A. (Eds.), *Advances in Neural Information Processing Systems 10*, Cambridge: MIT Press, 1998, pp. 45–51.

Lange, K., Sethi, S. 2011. "What Are People Saying About Your Company, Your Products, or Your Brand?", *Proceedings of the SAS Global Forum 2011 Conference*. SAS Institute, Inc.

Peleadeau, N. and Stovall, C (2005) "Application of statistical content analysis text mining to airline safety reports" (Provalis Research Corporation / Global Aviation Information Network)

Available at https://flightsafety.org/wp-content/uploads/2016/09/Provalis_text_mining_report.pdf

Sankaran, S., & Taylor, G. 2015. "Using Text mining and Natural Language Processing to Automate the Classification of Passenger Complaints" SAS Analytics conference 2015, Rome

Stainer, M. 2018 "An approach to using SAS® Text Miner" (personal communication)

Tolety, R. and Choudhary, S. 2016. "Text Analysis of American Airlines Customer Reviews". 24th Annual Southeast SAS Users Group (SESUG) Conference 2016.

Wright, R. 2017. "Temporal Text Mining: A Thematic Exploration of Don Quixote", *Proceedings of the SAS Global Forum 2011 Conference*. SAS Institute, Inc.
Available at http://support.sas.com/resources/papers/proceedings17/SAS0523-2017.pdf

CONTACT INFORMATION

The author may be contacted at:

Dr Simon Cumming,
British Airways PLC.,
PO box 365, Harmondsworth,
Uxbridge,
Middlesex, England
UB7 0GB