Seven Questions to ask a Text Analytics Vendor

Perhaps you have noticed that there really are no successful text analysis systems which are in general use on people's desktops. It is fair to ask why this is so.

It isn't that people don't have to absorb larger amounts of text. One might guess that the *basic approach* taken by the makers and vendors is not appropriate to what people want to get from text data.

So let's examine what text offers most people, then compare this with what text analytics software tries to do.

Text tells the story.

> Text tells us the story. A good story lays out the ideas and characters with their attributes. We read the text to set the scene - to explain the situation which we have dropped in on. It is like the first episode of a TV series. After that, we read on to see how the characters and ideas interact. These are changing relationships. A survey or report is no different. We need to see what issues, products or services are front of mind for the authors or responders, what attributes they assign to these issues and products, and how they see the relationships. We then move on to start answering questions and fixing problems. This is how we apply the knowledge gained.

In concrete terms, the process goes like this:

1. **We discover the concept space of the situation from the text.**
2. **We discover the explanations, or insights, from the text.**
3. **We can then act on these insights to alter the system.**

In the space of Text Analytics, Step 1 is important and neglected. You cannot understand the situation without understanding the background ideas.
Step 3 is almost totally ignored. Text data can tell you the story so you can fix the system. What else would you really want to do with it?

You cannot understand an IT text book using the concepts from political science. You would struggle to paint a sea scape with a palette suitable for a children's cartoon. Unfortunately, this problem is insidious and leads to

mistakes which we fail to notice. Why? Because if we naively analyse some data with a set of ideas which we know well and expect will apply to the data, we may never see that we are missing a quite different perspective.

Many text analysis systems will not automatically extract a *clear* set of the concepts and actors which characterise the text. Text analysis systems which come with predefined sets of categories, dictionaries, and entity lists are a menace. You cannot risk interpreting your data filtered through an understanding created by someone who is not familiar with your data and your situation, *even if the answer looks simple and neat*. This leads to our first question for a vendor:

Question 1: Does the analysis system's set of categories, entities, and concepts reflect a real understanding of my data and my situation?

Some systems which use predefined categories are manually tuned by the vendor during pre-sales. The vendor's consultants will sift through your data and construct extensive lists of terms, pattern matchers, and possibly rules. The analysis will then look ok at that time, but things change. New issues will arise in your business, and the terms and entities will change over time. How much effort did the vendor put in to customising the system? This leads to Question 2:

Question 2: How much time and effort did the vendor invest in tuning the category dictionaries, rules, and entity lists before go-live? When your data inevitably changes, can you afford to repeat this process to maintain the fidelity of your analysis?

If the analysis system does not use predefined categories, it may use document or word clustering. Many such systems do not produce clear or validated concepts. Remember that for easy and regular use, the discovered patterns of meaning need to be stable and clear. Don't be fooled by people who say that this sort of system works because it looks attractive and even compelling. There are ways to check whether discovered term clusters are real measures of meaning, or whether they are wasting your time. This is called cross-validation Here are some questions for vendors who offer term or document clustering, or other concept map solutions:

Question 3: If the product uses document clustering: how does the system scale with vast numbers of documents?; if a document contains several different ideas, can it be placed in two topics at once?; if I cut up the same documents into different chunks, would the pattern of clusters be similar?

Question 4: Do the discovered patterns portray the meaning of the documents? If two distinct documents have the same meaning, but are written in different languages, styles, and formats, do the two maps reveal similar patterns?

Quantitative, categorical, and numeric data mining is really good for establishing metrics and testing to see if these pre-defined metrics change. It is also really good for predicting whether a pre-selected situation is matched, such as customer churn probability.

Text analysis on the other hand excels at *telling you* what is happening. Because text is human communication - that is what it is for. So why waste this extremely valuable and rich source of intelligence to get another 3% in a black-box predictive model?

Think of it this way. If your metrics show your sales are rising, everyone feels great. If your metrics show you your results are falling off a cliff, how do you work out how to fix the system? This is the *feedback* you need for controlling a system. Your text data will tell you how to turn things around faster and more accurately than almost any other source of management information.

Unfortunately, this is where most text analysis systems fail, or don't even bother. Here are some questions:

Question 5: Does the system suggest chains of meaning which are well supported by the data, and which I can understand and explain to a manager? In other words, is it an explanatory model?

Question 6: Can I test hypotheses (educated guesses) in the concept space?

Question 7: How does a simple list of terms tell me much about the reasons for what is happening, without having to do a whole lot of guessing or having to read large amounts of text after all?

The Final List of Questions

Question 1: Does the analysis system's set of categories, entities, and concepts reflect a real understanding of my data and my situation?

Question 2: How much time and effort did the vendor invest in tuning the category dictionaries, rules, and entity lists before go-live? When your data inevitably changes, can you afford to feasibly repeat this process to maintain the fidelity of your analysis?

Question 3: If the product uses document clustering: how does the system scale with vast numbers of documents?; if a document contains several different ideas, can it be placed in two topics at once?; if I cut up the same documents into different chunks, would the pattern of clusters be similar?

Question 4: Do the discovered patterns portray the meaning of the documents? If two distinct documents have the same meaning, but are written in different languages, styles, and formats, do the two maps reveal similar patterns?

Question 5: Does the system suggest chains of meaning which are well supported by the data, and which I can understand and explain to a manager? In other words, is it an explanatory model?

Question 6: Can I test hypotheses (educated guesses) in the concept space?

Question 7: How does a simple list of terms tell me much about the reasons for what is happening, without having to do a whole lot of guessing or having to read large amounts of text after all?

I hope this helps. People are still doing a lot of writing and talking trying to tell you things. I think we need to listen more carefully, understand what they are saying, and then act thoughtfully.