Low Cost Text Mining as a Strategy for Qualitative Researchers

Jeremy Rose and Christian Lennerholt
University of Skövde, Sweden
jeremy.rose@his.se
christian.lennerholt@his.se

Abstract: Advances in text mining together with the widespread adoption of the Internet have opened up new possibilities for qualitative researchers in the information systems and business and management fields. Easy access to large amounts of textual material through search engines, combined with automated techniques for analysis, promise to simplify the process of qualitative research. In practice this turns out not to be so easy. We outline a design research approach for building a five stage process for low tech, low cost text mining, which includes insights from the text mining literature and an experiment with trend analysis in business intelligence. We summarise the prototype process, and discuss the many difficulties that currently stand in the way of high quality research by this route. Despite the difficulties, the combination of low cost text mining with qualitative research is a promising methodological avenue, and we specify some future paths for this area of study.

Keywords: big data, business intelligence, qualitative research method, social media analysis, text mining, text analytics

1. Introduction

Of the two representation systems that characterize research in the social sciences - words and numbers - qualitative research is the form that relies primarily on the collection, analysis and interpretation of words. Collections of words – texts – have traditionally played an important role for the qualitative researcher. However the developed world has moved on: texts are now predominantly digital, the largest repository for them is the Internet, and algorithms embedded in software routinely perform the work of text analysis. In this article we examine the potential of automated analytics to contribute to the work of a qualitative researcher. Using design science as a simple methodological framework, we investigate the current state of text mining (as an introduction for the less technical researcher), look at the analytical process as understood by text miners and undertake our own experiment in trend analysis for business intelligence. The aim of the paper is to develop a prototype process for qualitative researchers who wish to experiment with text mining as a vehicle for making parts of their research process more cost effective. We articulate this aim with two research questions:

> RQ1 - how can a low cost, low-tech Internet text retrieval and analysis process be conducted to facilitate qualitative research?
> RQ2 - what are the advantages and disadvantages of such an approach when compared to traditional ways of collecting and analysing qualitative data?

It will be important that the researcher can still use their hard-earned sense-making skills, does not have to replace their existing skill set with an entirely new one, and that costs associated with collecting and analysing data are reduced, not simply replaced by the costs of programming, software and reconstructing texts. These advances will be important for qualitative researchers in the era of big data analytics as they compete for legitimacy and attention (not to mention funding) with quantitative researchers skilled in downloading and processing very large volumes of unstructured data. It will also be interesting for more technically oriented text miners as they seek to understand the complex overlay of assumptions and interpretations that surround automated analytics.

We use the simple design science steps proposed by Vaishnavi and Kuechler (2015) to describe the problem area in more detail, and then to suggest, develop and test a prototype low tech, low cost text mining process. The paper is organised as follows. Section 2 outlines the design science research approach and the following sections follow the shape of this approach. Section 3 describes the problem, in terms of potential advantages of text mining in making qualitative researchers' work more cost effective; section 4 contains an initial suggestion for the process; section 5 develops the suggestion by investigating the relevant literature, section 6 evaluates the process through a (rather unsuccessful) experiment in using it to examine business intelligence trends in blogs; section 7 presents the process together with the principle learning from the literature and the

Reference this paper as: Rose J and Lennerholt C, "Low Cost Text Mining as a Strategy for Qualitative Researchers" *The Electronic Journal of Business Research Methods Volume 15 Issue 1 2017*, (pp2-16) available online at www.ejbrm.com

experiments. The conclusion addresses the answers to the research questions and presents future avenues for methodological research.

2. A design science approach to process building

A process (in this case a process for low cost text mining as a part of qualitative research) can be understood as a designed artefact for which design science is a suitable research approach. Design science 'addresses research through the building and evaluation of artefacts designed to meet the identified … need' (Hevner et al. (2004)). An important strand of the literature is directly concerned with how to enact design research; a normative literature often expressed as principles or process models (Rose et al., 2010). Vaishnavi and Kuechler (2015) propose a simple process model involving five steps, between which iteration is encouraged.

1. Awareness of Problem - the identification of a problem that needs a (better) solution. The awareness of the problem results in an initial solution proposal.
2. Suggestion - a suggestion to how the initial solution proposal should be realized. The suggestion results in a tentative design.
3. Development - the realization of the tentative design, resulting in an artefact.
4. Evaluation - an evaluation of whether or not the initial problem has been solved by the developed artefact. The evaluation of the artefact results in some sort of feedback or performance measures for assessing it.
5. Result/conclusion - the conclusion should be based on the evaluation and should explain the quality of the solution – an assessment of how successful the design research process has been.

Table 1 describes the research approach used for designing a low cost text mining approach suitable for qualitative researchers.

Table 1: Design research approach to a low cost text mining process suitable for qualitative researchers

Design science step	Our research approach
Awareness of Problem	Generated from the authors' extensive experience of qualitative research, and awareness of rapid advances in the field of text mining
Suggestion	An initial process suggestion is developed from text miners' own accounts of their process in the literature, adapted for the low cost qualitative researcher context
Development	The process suggestion is developed through a small literature study, in which the research process of selected well-published and relevant text mining articles is analysed against the suggested process to discover strengths and pitfalls
Evaluation	Evaluation is through a case or discovery experiment (Alberts and Hayes, 2002, Alberts and Hayes, 2005). The method is suitable for early stages of the elaboration of a research problem, where little is known about conditions, constraints or variables, and where it is impossible to isolate particular variables for manipulation. Discovery experiments are designed to encourage creative and innovative problem solving. We conduct an experiment using the process to analyse trends for business intelligence from blog data.
Result/conclusion	The low cost text mining process is described in terms of the purpose, principle techniques and tools, outcomes and major problems for each step.

3. Problem awareness: costs of Qualitative research, advances in text mining

Qualitative research employs a variety of research approaches including action research, case study, ethnography, qualitative surveys, and grounded theory. Researchers collect data through a variety of techniques including interviews, observation and fieldwork. Written data sources can include documents, company reports, memos, letters, reports, email messages, faxes, and newspaper articles. These empirical materials can include visual material such as pictures and videos, but usually the emphasis is on collections of words and written descriptions - texts. Texts are analysed using a bewildering variety of strategies (analogous to the many statistical tests available to a quantitative researcher). Miles and Huberman (1994), for instance, explain more than 40 data analysis techniques. Most forms of textual empirical work display similar characteristics, including:

- empirical material that is not easily quantifiable
- the search for transferable meanings, such that the material can be summarised and generalised

- representations of (summarised) data within the materials, by means of notes, narratives, graphs, diagrams or matrices, amongst other techniques
- the analysis of patterns in the data which signify more than individual episodes in the material
- the development of concepts, categories, constructs, variables, or codes which are able to summarise and generalise the material
- the development of relationships, associations, conditions for the concepts, codes, variables, etc.
- formulation of theory derived from patterns, concepts and relationships

The qualitative researcher is judged by their ability to make sense of volumes of text that would be overwhelming to an untrained person. However qualitative data are not cheap to acquire and process. Most of the costs are incurred as researcher time, and they include

- making contact and interacting with research subjects and organisations
- collecting documents, making observations, conducting interviews, taking notes, collecting emails, examining web-sites
- putting data in textual forms (normally digital) for analysis, transcribing interviews and notes
- buying analysis software of various kinds
- undertaking the analyses, for example coding.

These costs are substantial; we estimate the costs of a moderate size interview survey conducted recently by one of the authors (Rose et al., 2016) as approximately 1000 hours, or 20 weeks of researcher time. Researchers have long turned to software to automate parts of these tasks; in our survey part of the analysis was facilitated by a content analysis software application.

Recent developments in text mining, and the widespread availability of large quantities of empirical material (textual and non-textual) on the Internet prompt consideration of whether these resources are well-spent. This may be an old-fashioned way for qualitative researchers to use their time, and it might be more effective to use the internet as a primary empirical resource and as automate much as possible of the analysis work with text mining techniques. This strategy eliminates most of the work in our estimate: materials are downloaded direct from browser to computer in a digital form, large volumes of text and other materials are easily collectable, no investment in making contacts, interviewing or transcribing, analysis automated or partially automated with algorithms and software. The potential advantages of an internet/text mining research strategy for qualitative researchers are considerable.

4. Process suggestion: the internet as a Text source; text mining and analytics

Text mining and analytics covers that part of digital investigation which deals with primarily unstructured or semi-structured web content, as opposed to structured content in databases, or web structure and usage mining (Kosala and Blockeel, 2000). Mining develops the analogy of extracting something valued (information, intelligence) from a larger mass of less valued material (e.g. the world wide web), whereas analytics in this context is associated with data processing partially or wholly automated with software. Taken together, they might encompass statistical natural language processing, information extraction, question-answering systems, opinion mining, sentiment/affect analysis, web stylometric analysis, multilingual analysis, text visualization (Chen et al., 2012), subjectivity analysis, market sentiment analysis, topic modelling (Pang and Lee, 2008), dialogue act classification (Kaiser and Bodendorf, 2012), text summarization and a variety of other techniques. According to He et al. (2013), text mining is largely data driven and 'its main purpose is to automatically identify hidden patterns or trends in the data and then create interpretation or models that explain interesting patterns and trends in the textual data.' Wikipedia defines text analytics as 'a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation.' Many advances in these techniques are driven by research in computational linguistics, natural language processing and machine learning. Text mining and analytics are sometimes associated with big data (Chen et al., 2012, Xiang et al., 2015), as automation becomes imperative with larger volumes of data. He et al. (2013) for example observe that 'social media data are usually large, noisy and unstructured' and that 'it would be tedious and time-consuming if we had to manually code a large amount of social media data.'

4.1 Text miners' process

The existing text mining literature is primarily concerned with developing the algorithmic techniques for analysis, rather than conducting business and information systems research. However several text-mining researchers describe their own process. Kosala and Blockeel (2000) propose the following steps:

- resource finding (retrieving relevant documents)
- information selection and pre-processing (removing noise and irrelevance from the text and presenting it in a form that can be processed by a computer algorithm)
- generalization (discovering patterns)
- analysis (validation and/or interpretation of the patterns)

The first two steps are represented as data capture by Vitolo et al. (2015), to be followed by two steps concerned with identifying patterns: processing (running software algorithms on the data) and visualisation (software techniques for making patterns in data more obvious to the researcher). The final outcome of the process is specific for Xiang et al. (2015): actionable intelligence – insights that can lead to decisions or improvements in the outside world. Their process reads: text pre-processing, processing/analysis, actionable intelligence. A final contribution is from web content analysis in a more familiar qualitative context (Romano et al., 2003):

- elicitation (finding and downloading relevant data)
- reduction (analysis activities: selection, coding, clustering the data)
- visualization (as described above)

Though differing in detail, these accounts show some obvious common features, which are adapted for the process design suggestion in the next section.

4.2 Process suggestion for qualitative research using text mining

The process suggestion below includes four steps that are common to all the text miners' processes: finding the data, retrieving it, analysing it and attributing meaning to the analysis. The first step (finding the problem to address) is implied but never formally stated and the resulting five-step model is adapted to the qualitative researcher's task area.

- *Problematisation* - formulating a research problem (objective or question) that can be solved (achieved or answered) using some combination of text mining and qualitative data analysis
- *Choosing data sources* - defining a set of relevant internet text sources with the potential to solve the research problem
- *Retrieving text* - downloading, cleaning (removal of irrelevant text, or repair of damaged or difficult to analyse text), organisation and storage of the set of relevant text sources, such that they can be analysed
- *Analysing text* - enacting some (combination of) automated and/or manual analyses on the text body judged likely to solve the research problem
- *Interpreting results* - solving the research problem by attributing meaning to the results of the text analyses.

5. Development: five process steps in the text mining literature

We investigate these process steps in some articles published in quality research outlets, in order to sensitize ourselves to issues experienced by fellow researchers. We selected articles with extended empirical examples and a focus on methodological issues. The issues are clarified at the end of each sub-section.

5.1 Problematisation

In this section we look at how text mining researchers characterize their research problems. Xiang et al. (2015) investigate hotel guest satisfaction, through the medium of travel advice sites facilitating customer reviews and ratings. Moon et al. (2014) also focus on customer reviews, determining the impact of product reviews on sales. Attitudes (opinions, sentiment) registered in social media text are a common topic: three pizza chains are compared though their Twitter and Facebook sites by He et al. (2013), and Kaiser and Bodendorf (2012)

study attitudes towards iPhones in online forums. Apple's support communities for iPods were mined by Abrahams et al. (2014), who provide a framework for product defect detection. The derivation of demographic profiling for twitter users is mapped by Ikeda et al. (2013). Several of our examples use article databases as their sources, for detecting terrorism related texts in news articles (Choi et al., 2014), and for automating conventional scientific literature studies (Moro et al., 2015, Delen and Crossland, 2008). In many articles the empirical examples are more demonstrators for a new analytical technique (or combination of techniques) than serious investigations of external phenomena, and they often display a data-driven (qualitative research: grounded) approach, working from the available data to derive results, rather than a hypothetico-deductive approach – working from theory to derive hypotheses and testing them. This leads to many very open and exploratory research questions, for example: 'What patterns can be found from their Facebook sites respectively? What patterns can be found from their Twitter sites respectively?' (He et al., 2013). It also meant that problematisation was often poorly related to substantive theory in the problem area.

In summary a large variety of themes and issues can be addressed with text mining, but the exploratory data-driven character of the research style may make it difficult to engage with the central theoretical issues of a research field.

5.2 Choosing data sources

Textual data is available from a staggering variety of sources, both internal to organisations (social network feeds, emails, blogs, online forums, survey responses, corporate documents, news, and call center logs) and external. In research on developing and testing new algorithmic techniques in the text analytics field the choice of data source may be relatively insignificant, however this is not the case in research in other fields where text mining is used as the research method. The choice of data sources becomes important because most of the analytical techniques used are statistical and/or algorithmic in nature, implying that the composition of the sample text sources is likely to affect the analysis results (Krippendorff, 2004). The data sources should ideally be *representative* and *relevant*, where an unspoken assumption in many studies is that the chosen sources represent what happens on the net, and that what happens on the net also represents the physical world.

In respect to *representativeness,* Krippendorff (2004) identifies a variety of sampling techniques as appropriate to text; however none of the text mining articles that we investigated addressed this issue. They made an unspoken assumption that patterns they found in their Internet data selection would naturally hold true for larger populations.

If the sampling problem is less critical in the case of big data analyses, where the volume of data sometimes ensures better representativeness than other research techniques, the *relevance* problem is more acute. Human analysts both organize the collection of data so that it is relevant, (with for example a questionnaire) and intuitively sort and discard irrelevant text during analysis. Automated techniques do not do this except where they are programmed to do it, so the presence of significant quantities of noise in the data will also affect the analysis results. A commonly used way of ensuring some degree of relevance was to choose a themed network site, blog site or discussion forum such as Apple's support communities (Abrahams et al., 2014), or iPhone blogs (Kaiser and Bodendorf, 2012). Sometimes a specialised search engine was available, such as Google scholar for scientific articles (Delen and Crossland, 2008). Basic sampling and relevance reasoning was sometimes explicit: 'we chose Xanga (www.xanga.com) as our source of blog data … Xanga is the second most popular blog hosting site after the Google-owned Blogger (www.blogger.com). It is also ranked 17th in traffic (visit popularity) among all Web sites in English' (Chau and Xu, 2012). However the dominating logic for choice of sources in the articles we investigated is convenience, of which a large part resides in the retrieval and download mechanisms available.

The qualitative researcher will need to address issues of representativeness and relevance.

5.3 Retrieving text

In retrieving text two problems predominate, those of *structure* and *noise*. Where some researchers are prepared to manually download PDF files, most found an API or web crawler an acceptable solution and some went to considerable lengths to code rule sets and procedures in commercial web mining tools (O'Shea and Levene, 2011). There is also a legal and ethical context for what may be scraped, and its probable that

copyright is widely abused. Although commonly described as unstructured, textual data on the web often contains many forms of *structure*. In addition to the naturally occurring grammatical, lexical and genre structures of text (scientific articles have titles, authors, abstracts and key words), most web text is annotated (with html), it contains other significant relationships such as URLs and links, and some is thematically structured (with xml, for example). Expedia reviews (Xiang et al., 2015) have rating systems, various kinds of tagging and labelling, including date stamps and author tags, and the free text review typically contains less text than the tags. The structure can either be retained and exploited (by transposing it to a database for example), or partially or wholly removed.

Structure that is superfluous to the analysis task (for instance the headers of email messages) is *noise*, as are many other textual features that may get downloaded – annotations, spacing, links, repetitions, trademarks, commercial interests promoting products and services, and advertisements. Automation is required for these tasks, otherwise the costs of retrieving, cleaning, and storing data are considerable – for instance in programming or training a web crawler, removing superfluous annotation characters, checking for download errors, introducing analysis tags (such as a unique identifier for individual blogs), concatenating files, OCR processing and other mundane tasks. Structure and noise are related to relevance and representativeness; unwanted structure and text compromise statistical analysis. Removing them can involve many manual, even subjective decisions, as in this product defect study: 'the full set of all discussion threads, as of June 2010 was crawled and extracted ... in total, 1,560,379 ... Threads containing less than 50 words or less than two postings were excluded ... we … filtered the dataset to identify threads … related to defects ... determined the top 200 component description keywords ... ascertained the number of hits for this shortlist … sorted the threads from highest use … to lowest ... finally, we re-balanced the dataset by sampling an equal number of threads from each brand' (Abrahams et al., 2014).

The notion that text in large quantities is easily available from the Internet is therefore naïve; researchers must deal with structure and noise.

5.4 Analysing text

Most of research into text mining goes into the development of algorithms for processing text. Though academic fields often develop a consensus about how to use research methods (for instance which statistical tests are appropriate for variance analysis of questionnaire data), text mining is a field in rapid development, and no such consensus yet exists. All algorithmic techniques rely on assumptions that should condition the interpretation of their results, most have accuracy limitations, and some rely on intensive computing resources unavailable to most researchers. A simple analysis technique is word frequency counting, which relies on two limiting assumptions: that word frequency has some significance in a corpus of texts, and that words have no relationship to each other – the text is treated as a bag of words (unigrams). Bigrams or other n-grams can also be used. Accuracy is often measured by precision (the fraction of retrieved instances that are relevant), and recall (the fraction of relevant instances that are retrieved) or their weighted average: F score. The most common evaluation technique is no compare the performance of the algorithm with results obtained from human evaluators. In the case of this very simple algorithm (word frequencies), recall and precision can be expected to be close to 100%. Since many common words have no special significance (the, it) a stop-word list can be used, and since words often have several variations (like, likes), stemming reduces variants to the basic stem. Simple calculations (word frequency) can be used in metrics such as tf-idf (term frequency-inverse document frequency) - intended to reflect how important a word (for instance a search term) is to a corpus of texts - and incorporated in many search engines.

Text analytics provide varying ways of accomplishing tasks familiar from statistics, machine learning and natural language processing, many of which also overlap with data mining techniques. Machine learning may be supervised (dependent on a training set of data pre-coded by a human), semi-supervised or un-supervised. Common text analysis tasks include classification, clustering, entity recognition, relationship extraction, text summarization, question answering and sentiment analysis. *Classification* algorithms assign documents to predefined categories representing their content (or other parameters such as style, author, language, spam). Common methods include naïve Bayes, support vector machines, k-nearest neighbour, decision trees, expectation maximization, latent semantic indexing. *Clustering* analysis has the advantage of uncovering unanticipated trends, correlations, or patterns from data (He et al., 2013) using techniques such as decision tree construction, rule induction, clustering, logic programming, and statistical algorithms. A related task is topic modelling (Blei, 2012). *Information extraction* algorithms structure textual data, often through *entity*

recognition (ER) – classifying text stumps into categories, and *relation extraction* (RE) – specifying pre-defined relationships (Gandomi and Haider, 2015). *Text summarization* techniques aim to convey key information in single or multiple documents. Extractive summarization involves determining the most salient text units through their location and frequency and concatenating them. Abstractive techniques rely on Natural Language Processing (NLP) to parse the text and derive semantic information (Gandomi and Haider, 2015). *Question answering* (QA) systems (such as Apple's Siri and IBM's Watson) provide answers to questions posed in natural language, making them reliant on NLP techniques, at least for parsing the questions. Modern systems tend to use NLP techniques to provide semantically derived queries, trawl the web to extract candidate answers, and rank them to formulate answers. *Sentiment analysis* (opinion mining) techniques analyze text for opinions about phenomena such as products and services. Texts can be classified as negative or positive, or into more complex rating systems. Fundamental techniques are classification, regression and ranking (Pang and Lee, 2008).

Tasks normally have many algorithmic variations for their solutions - some are minor variations, others have completely different strategies. Different algorithms (obviously) produce different results. For example, Lu et al. (2011) compare four related topic modelling strategies: latent Dirichlet allocation (LDA), local LDA, multi-grain LDA, and segmented (using a Poisson Dirichlet process). Considerable mathematical sophistication is required to distinguish the different assumptions built into these variations. In addition many algorithmic techniques require input parameters (the number of neighbours to be considered, which weightings, or how many topics to be generated) seed words, dictionaries or pre-coded learning materials, requiring a degree of experimentation from the text miner. These make it difficult to independently assess the accuracy of the results. Some algorithms, such as LDA, are known to be sensitive to initial conditions. Assessing accuracy is in any case problematic; where a result coded by humans is available for comparison, accuracy expressed as precision and recall should reach 70% to be worth publishing, and 80% is considered excellent. Somewhere between a fifth and a third of results are in these cases erroneous. However in many areas agreement between humans (inter coder reliability) is known to be poor, making the measurement standard unreliable, and in many cases human solutions are not available.

One response to accuracy problems is to combine techniques, for example Abrahams et al. (2014) combine the analysis of lexical, stylistic, social, sentiment, product and semantic features with distinctive terms. Automated text analysis techniques may be combined with other algorithmic strategies, with more conventional analysis of structured data, or with traditional qualitative analysis (e.g. coding). Here there is often a trade-off: accuracy is improved at the cost of transparency in the process, and the overlaps and combination weightings become difficult to understand, or questionable.

The difficulties facing the naïve researcher in the analysis step are therefore considerable, and can be summarised in these questions:

- which analysis task is appropriate for the research problem?
- which algorithm(s) should be chosen for this task?
- which starting parameters should be employed (where these are necessary)?
- how reliable are the results?

5.5 Interpreting results

The ability to process a large volume of textual data is only meaningful if the results can contribute to some business or research activity. *Visualization* and *sense making* are key components of interpretation. *Visualization* involves 'organized, compressed assemblies ... that permit conclusion drawing and action´ (Romano et al., 2003) such as extended text, matrices, graphs, and charts. Visualizations organize data into accessible compact forms in which analysts can identify patterns. Contemporary visualisations include tag clouds and visual text analysis systems (Dou et al., 2013). Cui et al. (2011) develop a visualisation system for evolving topics as they develop over time in social networking systems. Despite advanced visualisation techniques, *sense-making* problems associated with the results of text analysis are considerable. For instance, a (typical) topic in the Delen and Crossland (2008) analysis of scientific literature is represented by the keywords: error, discipline, mis, major, methodology, field, value, time, future, set. A considerable interpretive leap, on the basis of extensive domain knowledge is required to understand what such a theme might be about. Many studies lack a domain theory background, which also limits sense making. Since the techniques

are still evolving, studies tend to focus on justifying the usefulness of the text-mining techniques employed, and the interpretation of the text mining results becomes a secondary consideration. Xiang et al. (2015) explain the root of many of these difficulties: whereas conventional methods usually start from a set of predefined hypotheses (derived from existing scientific knowledge, which also set the frame for interpretation), big data analytics reveal patterns that the researcher must later try to interpret or explain, sometimes by fitting them to a theoretical context.

Any assumption that text mining will necessarily provide self-explanatory results is also naïve: the text miner, like any other researcher, must strive to attribute meaning to their results.

6. Evaluation of the text mining process: business intelligence trend analysis experiment

We applied the low cost, low-tech text mining process to a relevant example. The low cost strategy means that we should not learn new programming skills, buy expensive software, use techniques with high learning curves or invest more than 15-20 hours of analysis time on a particular text mining technique without productive sense-making results. In addition we should use only the computing resources available to us on our standard specification laptops. We allotted roughly one half man-day per week over the course of an academic year. We chose to conduct trend analysis for business intelligence derived from blogs. Business intelligence 'incorporates the collection, management and, reporting of decision-oriented data as well as the analytical technologies and computing approaches that are performed on that data' (Davenport and Harris, 2007). However the experiment was rather frustrating.

6.1 Problematisation

Trend scouting aims at discovering topics and opinions which are of current interest and which may evolve in the future. The web is an important platform for trend scouting since trends in consumer behaviour often arise in online communities before spreading through society (Kaiser and Bodendorf, 2012). In our text mining experiment we asked the research question: what current trends in Business Intelligence can be revealed by mining the Internet? We chose an open question, for which the answer could reasonably be data driven, and had no formal hypothesis or theoretical background for the study, though we had enough background in the subject to be able to interpret the results.

6.2 Choosing data sources

We chose, for convenience, major themed blogs: Smart Data Collective, TechnoSocial Blog, Integrate The Clouds, Key2 Consulting, Data Doghouse, Business Intelligence: Process, People and Products, BI Scorecard, David Menninger, Meta S. Brown, Peter James Thomas and msSQLgirl. There are several blogs about business intelligence and we chose these because they are relatively large, with many current and recently updated posts. The blogs were written in the period 2010-15, with the largest concentration from 2103-4. We have no economically feasible means of investigating how representative this text sample is, and ignored this issue. We also ignored copyright and other potential legal issues. Choosing a themed blog ensures some degree of relevance: however there were many issues with noise deemed likely to interfere with statistical analysis, from surplus html characters, to blog headers and identifiers, all the way through to commercial interests repeatedly using the blog as a marketing channel.

6.3 Retrieving text

The blog was scraped using the free crawler tool Import.io. The tool was trained to recognise the structure of the different sites and retrieve (mainly) the blog entries with some header information. Many of the sites crawled chose to display only the first 5-6 lines of the blog on their top-level page (with an option to click through to the rest), and this is all we collected. The results were downloaded as an Excel binary .xsl. file and, after initial cleaning, transferred to a text file. The downloads had a certain structure (Input, Result, Number, Widget, Data Origin, Result Row, Source Page URL, data=blog text), which was of no particular analytical value to us (see Appendix 1). A variety of unpredictable noise was removed manually at this stage, particularly advertising, page formatting info, URL's, links, some company names and authors. There are also many download and OCR errors – too numerous to correct manually. The resulting corpus file contains approximately 12600 blog entries (example in appendix), 4736 pages, more than three million words.

6.4 Analysing text

We performed a simple frequency analysis of single words by importing the text into a familiar free cloud tool (Hypertext) after removing stop (common) words. There were many problems with noise, and with clearly irrelevant words (e.g. company names, blogger names), which we removed iteratively until we had a result that contained no obviously irrelevant words. We had to make an instinctive guess about how many words should be in the picture. We duplicated the same analysis using RapidMiner's text analysis package; the addition of stemming produced similar, but not identical results. We tried to generate n(2)grams to catch terms such as 'big data' but our free starter package ran out of memory. We tried a variety of free cloud-based tools (Voyant, Textalyser, Google spreadsheet text analysis add on), but these could not handle a large volume of text, or required payment. We also found a simple program (VOSviewer) that generated word clusters, which we tried to use for topic analysis. Here the algorithmic process was black boxed. The package offered a variety of settings and alternatives including (for every cluster map) two counting methods, minimum term occurrences, and number of terms to be selected. The cluster generation algorithm turned out to be sensitive to these initial conditions, allowing the generation of substantially different visualisations from the same data set (see next section). CasualConc offered many further analysis possibilities: word concordance, collocation, clustering, facilitating a nearly limitless investigation; we used the starting term 'business intelligence.' Concordance provided more than 6000 contexts for the term in the text corpus. Collocation provided lists of adjacent words or phrases (such as the bigrams: business analytics, cloud computing, information management, data warehousing, business performance, industry analysis, predictive analysis, big data). Five word clusters (frequency > 100) including business intelligence were

- performance tags business analytics
- big data business analytics
- information management it performance
- cloud computing data integration
- business technology chief information

6.5 Interpreting results

We used some simple word cloud software to visualise the most frequent terms (**Figure 1**) in our word frequency analysis.

Figure 1: Word cloud of most frequent single word terms in the BI blog dataset

The word cloud provides some pointers to possible trends in business intelligence but does not, in itself, make a viable analysis. The word *data* appears more than 78000 times in the text making some kind of sampling process necessary for deeper (qualitative) inspection of the text. We experimented with three – looking at random instances in the text, systematic sampling (taking every hundredth appearance) or looking for clusters of the word in particular blogs. Frequent references to *data* were neither surprising or (for our purpose) interesting, so we used our intuition to look at promising word combinations: *data integration* appeared over 9000 times, and *big data* nearly 7000 times, making them candidates for trends. *Cloud* appeared nearly 50,000 times, and although about 7000 of these were in the repeating header of a cloud themed blog site (Integrate

the Clouds) this seemed an excellent trend candidate. Qualitative investigation showed that *integration* was nearly always in the context of *data integration, system integration* or *cloud integration* (though this last may have been a statistical distortion caused by the themed blog site mentioned above). *Analytics* is a term that causes no great surprise, however about a third of these references were to *predictive analytics* – another candidate trend. *Social* appeared more than 5000 times, *media* more than 3000 – we used our intuition to arrive at the term *social media* – another trend candidate. However, with more than 10^{83} available combinations of the 50 terms in the picture, intuition seemed a poor way of making a comprehensive analysis, even when supported by the clustering analysis in the next paragraph. The appearance of the term *collective* was intriguing, but only the product of poor data cleaning. It appeared three times in the header of every blog at one of the larger sites Smart Data Collective, accounting for nearly all of its appearances. This also ruled out the term *smart*. Other meaningless statistical blips were *loading leave* and *blog*.

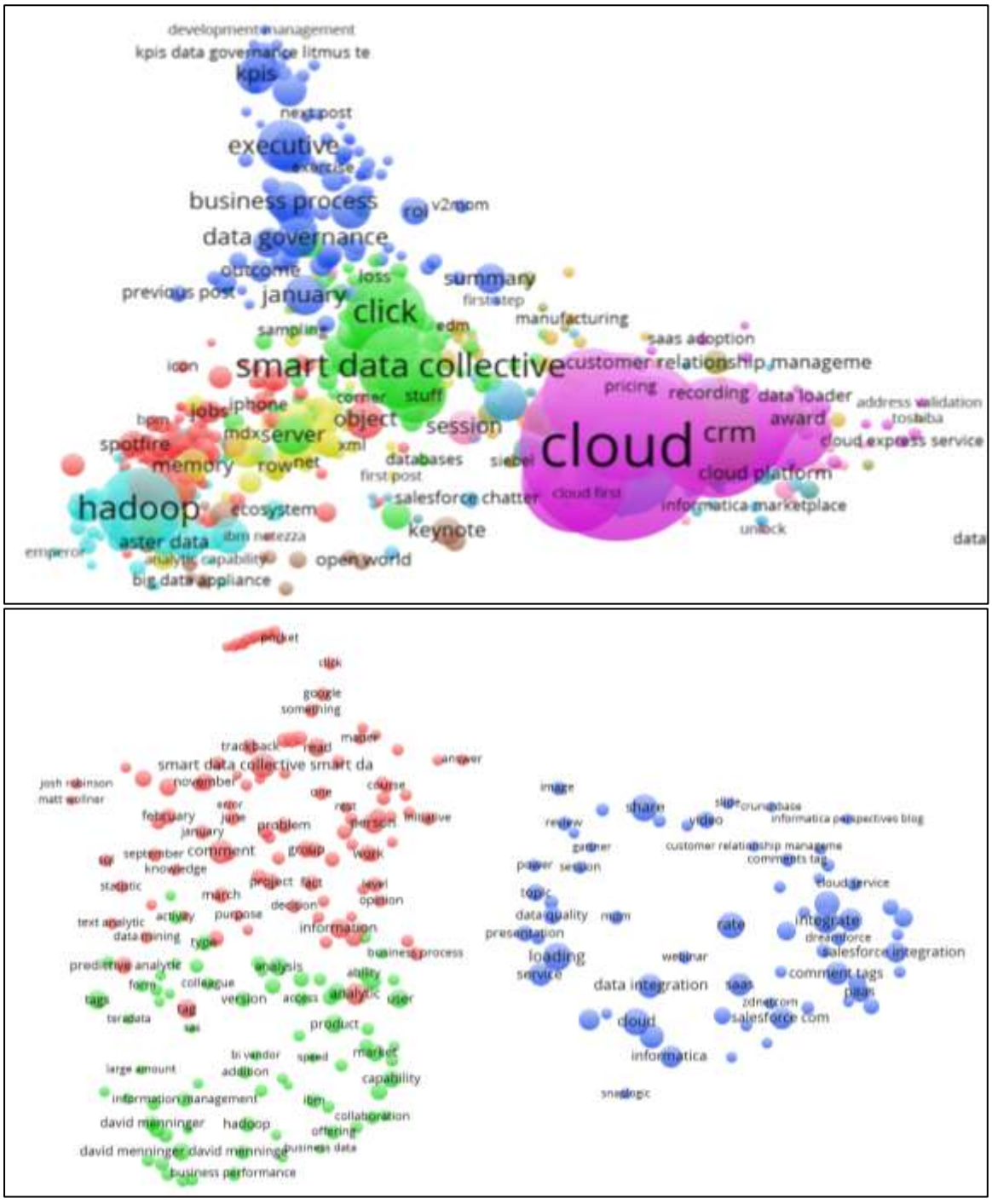

Figure 2: Two cluster maps of the BI blog data set from VOSviewer showing sensitivity to initial conditions (different parameter settings)

Our VOSviewer results produced impressive coloured cluster charts, and the unknown algorithm did produce bigrams – 2 (or more) word phrases. Some of these were helpful (*data governance, customer relationship management, business process*), some of them still the result of poor data cleaning (*smart data collective, next post*), many rather irrelevant to our purpose (*open world* - the name of Oracle's annual trade fair) and some quite mysterious (*KPIs Data Governance Litmus Test*). We hoped that statistical clustering would provide some insight into BI trends but these maps turned out to be exceptionally difficult to interpret. The two considerably different cluster maps in figure 2 are both produced from the same text corpus by changing some of the anonymous algorithm's starting parameters: *minimum term occurrences*, and *number of terms to be selected*. We had no way of determining the optimal starting parameters from the very large number of possible combinations. In figure 2 (top diagram), it may be possible to discern a purple cluster that might be something to do with the cloud, a blue cluster referring to the business, management and governance side, possibly a turquoise cluster relating to big data technologies – however some groupings are fairly baffling. What should one make of cluster 11: *benioff, big data appliance, cloud offering, configuration, exadata, exalytic, hardware, keynote, memory capability, noSQL, open world, oracle database, public cloud, ram, social network, Wednesday?* Moreover the analysis program is a black box, so that the statistical process is not available for inspection, and it is sensitive to initial conditions – in this case its settings. It may be possible to produce a meaningful output by many hours of experimentation, and cleaning or manipulating the data – but this is not a low cost strategy. More importantly, it's not methodologically secure in research terms.

We arrived at some candidates for business intelligence trends including: data integration, systems integration, cloud integration, social media, big data, business management and governance. However, since we had no secure method for arriving at a convincing result in a reasonable amount of time we abandoned our investigation. We compared our data analyses with a commercially produced list of BI trends for 2015 from Tableau Business Intelligence (http://get.tableau.com/campaign/business-intelligence.html) - no research method reported. We could observe some commonality (*governance, social intelligence, analytics, data integration, cloud analytics*) and some differences (*journalism, mobility, self-service*).

7. Results: a generic process for low cost text mining

In this section we note the strengths and weaknesses of the process we developed, as shown in our evaluation experiment in trend analysis for business intelligence. Most of the issues we identified in our literature review in section 5 also became problematic in the experiment.

7.1 Problematisation

Though we had no problem finding an open-ended investigation topic for our relatively trivial example, it doesn't fit well with the more conventional hypothetico-deductive style of research prevalent in our research fields. Here investigations are more normally based on theoretical premises or literature studies, and the expectation is that research questions are focused, clearly defined and theory based. There may or may not be an explicit hypothesis, but examples of data-driven statistically generated theory are still rare. Its not immediately clear how the text mining techniques we studied can be adapted to hypothesis testing, or to focused research questions, and we found few examples of this in the literature.

7.2 Choosing data sources

The strength of the text mining approach is the easy availability of very large quantities of digital text through the Internet: however this can be also a weakness in statistical approaches. We chose some thematically targeted blogs to help improve relevance; nevertheless our blog set included large amounts of irrelevant data. Nor is it feasible to establish the total population of relevant blogs, so the sampling process is insecure, and representativeness unknown. Validity often rests on the argument that there is a lot of data. Qualitative researchers are used to smaller quantities of data collected in some well-targeted way.

7.3 Retrieving text

When compared to interviewing subjects and transcribing the results, retrieving text from the Internet is relatively quick and easy, though not without costs. Whereas the interviewer has the opportunity to create their own structure in the resultant texts through the design of questions, it's difficult to know what to do with the many pre-existing structures of the web. However the predominant problem is noise. An experienced interviewer will gently guide their research subject back on track if they wander; bloggers are free to discuss what they choose and site designers add many other noise elements. Since noise undermines statistical

analysis this is not an insignificant problem: efficiency gains in downloading text are reversed if the removal of noise cannot be cheaply automated.

7.4 Analysing text

In principle the application of text analysis is made insecure by the lack of an agreed, methodologically valid approach for the field. Instead there are a multitude of competing algorithmic techniques for multiple analytic purposes, many of them requiring significant researcher input. These remain largely black box techniques for everyone except algorithm developers and statisticians. In the social sciences there are statistical analysis techniques for survey data that are generally agreed to be methodologically valid, but no such consensus exists yet for text analysis. In practice, only rather simple analysis techniques are available at low cost. Where a variety of analysis techniques were available, the analysis possibilities quickly snowballed out of control. We seldom knew which analysis task was appropriate, which algorithm pre-programmed software was running, how we should set starting parameters, or how reliable the results were. Poor execution of the earlier retrieval stage often led to the junk in, junk out condition. Often we didn't have enough computing power available without extra cost. A further problem is that there is no standard method for combining statistical text analysis with more traditional qualitative analysis.

7.5 Interpreting results

Interpreting text mining results is much more complex than indicated in the literature. Many of our results required leaps in sense making that even hardened qualitative researchers were unwilling to make. More experience in problematisation, choosing sources, and retrieving and analysing text would no doubt make interpretation easier (for example by removing noise from the data). Iterative passes with the data where the analysis is refined to improve sense-making (for instance by tweaking starting variables for algorithms) may help. Here the complementary skills of the qualitative analyst may be important. However the hermeneutic styles of qualitative analysis (for example establishing a relationship between the whole text and its parts) were seriously challenged by our dataset. It was too large, and too diverse to develop any coherent overall impression.

Table 1 shows the generic process for low cost text mining for qualitative researchers, comprising the steps: problematisation, choice of data sets, retrieval of text, text analysis, interpretation of results. For each step, conclusion about its purpose, available tools and techniques and the principle issues that researchers should expect to encounter are specified.

Table 1: Generic process for low cost text mining for qualitative researchers

process step	*purpose*	*tools and techniques*	*outputs*	*principle issues*
problematisation	focus the area of investigation	guess + stepwise refinement, literature survey, existing market research	research question or topic	finding suitable open questions that can lead to publishable research
choice of data source(s)	establish a body of text to be mined	Google search algorithm or other techniques for prioritising web sources, sampling, specialised databases	text corpus	establishing relevance and representativeness
retrieval of text	retrieve, clean and store the text body	crawling, scraping, application programming interface, database calls, many simple free cloud and desktop tools available	clean digital text corpus ready for analysis	distinguishing meaningful and relevant text and structure from noise
text analysis	model the text	word count, topic modelling, sentiment analysis, (countless other techniques), some free tools available	various (often statistically derived) models of the text, text outputs suitable for qualitative analysis	choosing techniques and algorithms, statistical validity, junk in junk out
interpretation of results	learn in order to contribute to knowledge or guide decisions,	visualisation techniques, further qualitative analysis, relation of model to context or theory	complementary or guiding insights for qualitative analysis	visualisation, sense making

8. Conclusions

Quantitative text analysis has been around for a long time, but has mainly been the preserve of highly trained researchers using specialist techniques, and custom software. The Internet has made large quantities of textual data readily available, and the big data revolution has created the expectation that researchers should capitalise on new ways of creating knowledge. We set out to create a text mining process for qualitative researchers using a simple design science method. We developed a five-stage process and tested it using a small-scale experiment. The experiment demonstrated that the process was usable, but that most of the issues that were identified through literature search became significant problems in use.
We asked the questions

- how can a low cost, low tech Internet text retrieval and analysis process be conducted to facilitate qualitative research?
- what are the advantages and disadvantages of such an approach when compared to traditional ways of collecting and analysing qualitative data?

In response to our first research question, we note that the literature we studied concentrated on the text analysis techniques, but often assumed that the other stages were trivial – not our experience. In particular the problematisation and interpretation stages became difficult in the context of research, where open questions are seldom encouraged and interpretations must be carefully justified. In understanding the advantages and disadvantages of using such a process (RQ2) we acquired the learning that our labour intensive style of (qualitative) research was limited to pitifully small quantities of well-targeted data, and was extremely difficult to scale up. It was relatively easy to acquire a large volume of text from the Internet, however there was a clear trade off between quantity and quality. Our traditional practice provided the data we needed for analysis, but retrieving text from the Internet presented many additional problems. Whereas we understand the norms for qualitative analysis, we had to make many relatively insecure choices with the automated text mining tools, and lacked confidence that those choices would stand up to the scrutiny of experience reviewers. There are no established standards for conducting this kind of research that simplify the many methodological choices, and relatively few examples to follow. In addition there is little existing methodological help for combining quantitative and qualitative text analysis, apart from the general precepts of multi-method research.

Despite the difficulties, we expect the marriage of text mining with qualitative research to be a productive one in the future, and made inevitable by current trends. Two conditions for progress are the wider availability of better low cost software (this is expected to improve rapidly), and the emergences of some generally accepted methodological standards and procedures for text mining as research. Some avenues for future methodological research are:

- using text mining to suggest patterns in a large text corpus combined with selective or sampled qualitative investigation of the patterns (the research style of our example)
- using information extraction techniques to understand ontological structures of text
- using natural language analysis techniques to understand semantic structures
- using automatic text translation to enable cross cultural qualitative studies
- automatically summarizing a large text corpus to provide a manageable amount of text for qualitative analysis
- combining sentiment analysis with qualitative evaluation
- developing automatic question answering as a vehicle for hypothesis testing

The methodological challenges involved are considerable, but both text mining and qualitative research can be richer as a result.

References

Abrahams, A. S., Fan, W., Alan Wang, G., Zhang, Z. J. & Jiao, J. (2014) An integrated text analytic framework for product defect discovery. *Production and Operations Management*.

Alberts, D. S. & Hayes, R. E. (2002) Code of Best Practice for Experimentation. In: *DoD Command and Control Research Program, Office of the Assistant Secretary of Defense*. Washington, D.C.

Alberts, D. S. & Hayes, R. E. (2005) Campaigns of Experimentation: Pathways to Innovation and Transformation. In: *Command and Control Research Program, Office of the Assistant Secretary of Defense* Washinton. D.C.

Blei, D. M. (2012) Probabilistic topic models. *Communications of the ACM,* **55**(4), 77-84.

Chau, M. & Xu, J. (2012) Business intelligence in blogs: Understanding consumer interactions and communities. *MIS Quarterly,* **36**(4), 1189-1216.

Chen, H., Chiang, R. H. & Storey, V. C. (2012) Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly,* **36**(4), 1165-1188.

Choi, D., Ko, B., Kim, H. & Kim, P. (2014) Text analysis for detecting terrorism-related articles on the web. *Journal of Network and Computer Applications,* **38**, 16-21.

Cui, W., Liu, S., Tan, L., Shi, C., Song, Y., Gao, Z. J., Qu, H. & Tong, X. (2011) Textflow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics,* **17**(12), 2412-2421.

Davenport, T. H. & Harris, J. G. (2007) *Competing on analytics: The new science of winning.* Harvard Business Press, Brighton, MA.

Delen, D. & Crossland, M. D. (2008) Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications,* **34**(3), 1707-1720.

Dou, W., Yu, L., Wang, X., Ma, Z. & Ribarsky, W. (2013) Hierarchical topics: Visually exploring large text collections using topic hierarchies. *IEEE Transactions on Visualization and Computer Graphics,* **19**(12), 2002-2011.

Gandomi, A. & Haider, M. (2015) Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management,* **35**(2), 137-144.

He, W., Zha, S. & Li, L. (2013) Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management,* **33**(3), 464-472.

Hevner, A. R., March, S. T., Park, J. & Ram, S. (2004) Design science in Information Systems research. *MIS Quarterly,* **28**(1), 75-105.

Ikeda, K., Hattori, G., Ono, C., Asoh, H. & Higashino, T. (2013) Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems,* **51**, 35-47.

Kaiser, C. & Bodendorf, F. (2012) Mining consumer dialog in online forums. *Internet Research,* **22**(3), 275-297.

Kosala, R. & Blockeel, H. (2000) Web mining research: A survey. *ACM Sigkdd Explorations Newsletter,* **2**(1), 1-15.

Krippendorff, K. (2004) *Content Analysis* Sage, Thousand Oaks.

Lu, B., Ott, M., Cardie, C. & Tsou, B. K. (2011) Multi-aspect sentiment analysis with topic models. In: *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, 81-88. IEEE.

Miles, M. B. & Huberman, A. M. (1994) *Qualitative data analysis (2nd edition).* Sage, Thousand Oaks.

Moon, S., Park, Y. & Seog Kim, Y. (2014) The impact of text product reviews on sales. *European Journal of Marketing,* **48**(11/12), 2176-2197.

Moro, S., Cortez, P. & Rita, P. (2015) Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications,* **42**(3), 1314-1324.

O'Shea, M. & Levene, M. (2011) Mining and visualising information from RSS feeds: a case study. *International Journal of Web Information Systems,* **7**(2), 105-129.

Pang, B. & Lee, L. (2008) Opinion mining and sentiment analysis. *Foundations and trends in information retrieval,* **2**(1-2), 1-135.

Romano, N. C., Donovan, C., Chen, H. & Nunamaker, J. F. (2003) A methodology for analyzing web-based qualitative data. *Journal of Management Information Systems,* **19**(4), 213-246.

Rose, J., Jones, M. & Furneaux, B. (2016) An Integrated Model of Innovation Drivers for Smaller Software Firms. *Information & Management,* **53**(3), 307-323.

Rose, J., Markfoged, K. & Andersen, J. L. (2010) Can design science be used for design? *Proceedings of the Fifth Mediterranean Conference on Information Systems*, Tel Aviv, AIS Association.

Vaishnavi, V. K. & Kuechler, W. (2015) *Design science research methods and patterns: innovating information and communication technology.* CRC Press.

Vitolo, C., Elkhatib, Y., Reusser, D., Macleod, C. J. & Buytaert, W. (2015) Web technologies for environmental big data. *Environmental Modelling & Software,* **63**, 185-198.

Xiang, Z., Schwartz, Z., Gerdes, J. H. & Uysal, M. (2015) What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management,* **44**, 120-130.

Appendix 1. sample blog download:

15 Techno Social Blog Techno Social Blog 15 http://rajasekar.net/page/3/ Predictive Analytics-A future Insight of Data Analysis February 9, 2008 Rajasekar Predictive analytics encompasses a variety of techniques from statistics and data mining that analyze current and historical data to make predictions about future events.Such predictions rarely take the form of absolute statements, and are more likely to be expressed as values that correspond to the odds of a particular event or behavior taking place in the future. In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions. It is been used with the applications involving CRM,Cross-Selling,Direct marketing ,Collection analytics not only that even helps to detect Fraud detection in credit card Apps. The statistical techniques used in Predictive Analytics are as follows Regression Techniques Linear Regression Model Discrete choice models Logistic regression Time series models Apart from these Statistical Techniques there are some Machine learning techniques are used such as Neural Networks and k-nearest neighbours The tool used to help with the execution of predictive analytics are SAS, S-Plus, SPSS and Stata and For machine learning/data mining type of applications, KnowledgeSEEKER, KnowledgeSTUDIO, Enterprise Miner, GeneXproTools, Clementine, KXEN Analytic Framework, InforSense are some of the popularly used options. WEKA is a freely available open-source collection of machine learning methods for pattern classification, regression, clustering, and some types of meta-learning, which can be used for predictive analytics. Recently Business Objects has announced a partnership with SPSS , a worldwide provider of predictive analytics software, announced the companies have entered into an original equipment manufacturer agreement in which Business Objects will offer its customers the ability to use SPSS predictive analytics data mining technology as part of the market-leading Business Objects TM XI platform. Users of Business Objects XI with predictive analytics data mining technology will be able to leverage business predictions to make more informed decisions that can help generate revenue, control expenses, and mitigate risk. Today SAS, the leader in business intelligence, has significantly enhanced its award-wining SAS Enterprise Miner, SAS Text Miner, and SAS Forecast Server software, bringing predictive analytics to their highest level yet. The newest release of SAS Enterprise Miner improves productivity through added interactive advanced visualization and new analytics. Fifteen new analytical tools improve the resulting predictive models, which can mean significant savings for customers with proactive marketing departments such as in retail or banking.With innovative new modeling algorithms, including gradient boosting, partial least squares and support vector machines, SAS Enterprise Miner users can build more stable and more accurate models and thus make better decisions faster and with more confidence. Source:Wikipedia and SAS,Business Objects Read More