



University of **HUDDERSFIELD**

University of Huddersfield Repository

Kaklauskas, Arturas, Seniut, Mark, Amaratunga, Dilanthi, Lill, Irene, Safonov, Andrej, Vatin, Nikolai, Cerkauskas, Justas, Jackute, Ieva, Kuzminske, Agne and Peciure, Lina

Text Analytics for Android Project

Original Citation

Kaklauskas, Arturas, Seniut, Mark, Amaratunga, Dilanthi, Lill, Irene, Safonov, Andrej, Vatin, Nikolai, Cerkauskas, Justas, Jackute, Ieva, Kuzminske, Agne and Peciure, Lina (2014) Text Analytics for Android Project. *Procedia Economics and Finance*, 18. pp. 610-617. ISSN 2212-5671

This version is available at <http://eprints.hud.ac.uk/id/eprint/23018/>

The University Repository is a digital collection of the research output of the University, available on Open Access. Copyright and Moral Rights for the items on this site are retained by the individual author and/or other copyright owners. Users may access full items free of charge; copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational or not-for-profit purposes without prior permission or charge, provided:

- The authors, title and full bibliographic details is credited in any copy;
- A hyperlink and/or URL is included for the original metadata page; and
- The content is not changed in any way.

For more information, including our policy and submission procedure, please contact the Repository Team at: E.mailbox@hud.ac.uk.

<http://eprints.hud.ac.uk/>

4th International Conference on Building Resilience, Building Resilience 2014, 8-10 September
2014, Salford Quays, United kingdom

Text Analytics for Android Project

Arturas Kaklauskas^{a*}, Mark Seniut^a, Dilanthi Amaratunga^b, Irene Lill^c, Andrej Safonov^d,
Nikolai Vatin^e, Justas Cerkauskas^a, Ieva Jackute^a, Agne Kuzminske^a, Lina Peciure^a

^a*Vilnius Gediminas Technical University, Vilnius, LT-10223, Lithuania*

^b*Global Disaster Resilience Centre, University of Huddersfield, UK*

^c*Tallinn University of Technology, Tallinn, 19086, Estonia*

^d*Moscow State Industrial University, Moscow, Russia*

^e*Saint-Petersburg State Polytechnical University, Saint-Petersburg, Russia*

Abstract

Most advanced text analytics and text mining tasks include text classification, text clustering, building ontology, concept/entity extraction, summarization, deriving patterns within the structured data, production of granular taxonomies, sentiment and emotion analysis, document summarization, entity relation modelling, interpretation of the output. Already existing text analytics and text mining cannot develop text material alternatives (perform a multivariant design), perform multiple criteria analysis, automatically select the most effective variant according to different aspects (citation index of papers (Scopus, ScienceDirect, Google Scholar) and authors (Scopus, ScienceDirect, Google Scholar), Top 25 papers, impact factor of journals, supporting phrases, document name and contents, density of keywords), calculate utility degree and market value. However, the Text Analytics for Android Project can perform the aforementioned functions. To the best of the knowledge herein, these functions have not been previously implemented; thus this is the first attempt to do so. The Text Analytics for Android Project is briefly described in this article.

© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/3.0/>).

Selection and/or peer-reviewed under responsibility of the Centre for Disaster Resilience, School of the Built Environment, University of Salford.

*Corresponding author. *Email address:* arturas.kaklauskas@gmail.com

Keywords: text analytics, text mining, Text Analytics for Android Project, qualitative and quantitative analysis

1. Introduction

Text mining, sometimes alternately referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the divining of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities) (Machine Learning Market 2013).

Text analytics software can help by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analyzed with traditional data mining techniques. With an iterative approach, an organization can successfully use text analytics to gain insight into content-specific values such as sentiment, emotion, intensity and relevance. Because text analytics technology is still considered to be an emerging technology, however, results and depth of analysis can vary wildly from vendor to vendor (BusinessAnalytics 2013).

Text analytics is the process of deriving information from text sources. It is used for several purposes, such as: summarization (trying to find the key content across a larger body of information or a single document), sentiment analysis (what is the nature of commentary on an issue), explicative (what is driving that commentary), investigative (what are the particular cases of a specific issue) and classification (what subject or what key content pieces does the text talk about) (Gartner).

Text mining is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. Text mining is vast area as compared to information retrieval. Typical text mining tasks include document classification, document clustering, building ontology, sentiment analysis, document summarization, Information extraction etc. Where as information retrieval typically deals with crawling, parsing and indexing document, retrieving documents (Stackoverflow 2013).

Text mining concerns itself with discovering structure and patterns in unstructured data – usually text. There are many different approaches to this task, some focus on ancillary structures such as taxonomies and ontologies, some focus on semantics and natural language processing, while others use various algorithms to categorise and summarise. For example, KNIME Text Processing is a plug-in to the (free) KNIME data mining suite supports a six step text processing process which starts with the reading and parsing of text, followed by named entity recognition, filtering and manipulation, word counting and keyword extraction, bow and vector representation, and finally visualization (Butleranalytics 2013).

Presently, text mining is in a loosely organized set of competing technologies that function as analytical “city-states” with no clear dominance among them. To further complicate matters, different areas of text mining are in different stages of maturity. Some technology is easily accessible by practitioners today via commercial software, while other areas are only now emerging from academia into the practical realm (Miner et al. (2012).

Research shows that various researches have specialised in depth the different and very important areas of text analytics and mining (blogs and social networks (Shenghua and Li 2013; Marwick 2014; Mostafa 2013; Boulos et al. 2010), students' online interaction (He 2013), digital libraries (Nguyen 2014, Fagan 2014), process industry (Liew et al. 2014), medicine (Anholt et al. 2014), legal, business intelligence, security (Truyensand Van Eecke 2014), etc. A brief analysis of above research follows.

Social media have been adopted by many businesses. More and more companies are using social media tools such as Facebook and Twitter to provide various services and interact with customers. To increase competitive advantage and effectively assess the competitive environment of businesses, companies need to monitor and analyze not only the customer-generated content on their own social media sites, but also the textual information on their competitors' social media sites. The results reveal the value of social media competitive analysis and the power of text mining as an effective technique to extract business value from the vast amount of available social media data. Recommendations are also provided to help companies develop their social media competitive analysis strategy (Shenghua and Li 2013).

Marwick (2014) investigate the use of Twitter at a major conference of professional and academic anthropologists. Using R Marwick (2014) identify the demographics of the community, the structure of the community of Twitter-using anthropologists, and the topics that dominate the Twitter messages. A key finding is that the transformative effect of Twitter in academia is to easily enable the spontaneous formation of information-sharing communities bound by an interest in an event or topic (Marwick 2014).

Blogs and social networks have recently become a valuable resource for mining sentiments in fields as diverse as customer relationship management, public opinion tracking and text filtering. In fact knowledge obtained from social networks such as Twitter and Facebook has been shown to be extremely valuable to marketing research companies, public opinion organizations and other text mining entities. However, Web texts have been classified as noisy as they represent considerable problems both at the lexical and the syntactic levels. Mostafa (2013). Mostafa (2013) research results indicate a generally positive consumer sentiment towards several famous brands. By using both a qualitative and quantitative methodology to analyze brands' tweets, Mostafa (2013) study adds breadth and depth to the debate over attitudes towards cosmopolitan brands.

Boulos et al. (2010) explore Technosocial Predictive Analytics (TPA) and related methods for Web "data mining" where users' posts and queries are garnered from Social Web ("Web 2.0") tools such as blogs, micro-blogging and social networking sites to form coherent representations of real-time health events. Boulos et al. (2010) introduce to commonly used Social Web tools such as mashups and aggregators, and maps their exponential growth as an open architecture of participation for the masses and an emerging way to gain insight about people's collective health status of whole populations.

Students' online interaction data suggests that a combination of using data mining and text mining techniques for a large amount of online learning data can yield considerable insights and reveal valuable patterns in students' learning behaviors (He 2013).

With the availability of open and structured data, digital libraries become an important source of data in recent data mining techniques. The inherent structure of digital libraries comes with information about date, authorship, involved institutions, geographic context, and large volumes of text (Nguyen 2014). As an illustration of digital libraries analysis, Nguyen (2014) focus on a space of scientific publications related to some science discipline and extract patterns and text-related information from the dataset. Nguyen (2014) discuss text preparation techniques in R, the use of the latent Dirichlet allocation to classify content, and the analysis of text features, like co-occurrence matrices. Simple similarity measures between authors and between papers is used to illustrate cluster cohesion within the dataset (Nguyen 2014).

As the demand for library assessment grows, academic libraries are becoming more interested in Web analytics. Data are automatically gathered and provide information about a wide variety of online interactions (Fagan 2014). Fagan (2014) discusses how commercial web metrics might be adapted for use in academic libraries. Major key performance indicators used in the commercial sector are reviewed in the academic library context (Fagan 2014).

Four main sectors of the process industry are studied by Liew et al. (2014): oil/petrochemicals, bulk/specialty chemicals, pharmaceuticals, and consumer products. Liew et al. (2014) study reveals that the top sustainability focuses of the four sectors are very similar: health and safety, human rights, reducing GHG, conserving energy/energy efficiency, and community investment.

The so-called text or data "mining" enables this huge amount of medicine information to be managed, extracting it from various sources using processing systems (filtration and curation), integrating it and permitting the generation of new knowledge (Piedra et al. 2014). Using commercially available text-mining software (WordStat™), Anholt et al. (2014) developed a categorization dictionary that could be used to automatically classify and extract enteric syndrome cases from the warehoused electronic medical records (Anholt et al. 2014).

Currently authors participated in the Android (Academic Network for Disaster Resilience to Optimise educational Development) project. Android project is being carried out with the financial assistance of the EU Life Long Learning programme, under the Erasmus networks action. ANDROID is concerned with what resilience is, what it means to society, and how society might achieve greater resilience in the face of increasing threats from natural and human induced hazards.

Already existing above text analytics and text mining cannot develop text material alternatives (perform a multivariant design), perform multiple criteria analysis, automatically select the most effective variant according to different aspects (popularity of a text (citation index of papers (Scopus, ScienceDirect, Google Scholar, etc.) and

authors (Scopus, ScienceDirect, Google Scholar, etc.), Top 25 papers, impact factor of journals, supporting phrases, document name and contents, density of keywords), calculate utility degree and market value. However, the Text Analytics for Android Project can perform the aforementioned functions. To the best of the knowledge herein, these functions have not been previously implemented; thus this is the first attempt to do so.

The structure of this paper is as follows. Following this introduction, Section 3 provides a brief review of the Text Analytics for Android Project. Finally the concluding remarks appear in Section 4.

2. Text Analytics Model and System for Android Project

The essence of this research involves the Text Analytics Model that is designated to select the most rational, integrated text material from a library of documents. It covers the inputting of bag of concepts space; selecting, processing and indexing information in accordance with the inputted bag of concepts space and User Model; formulating the results of the retrieval and finally showing them to the user. Further, after selecting, processing and indexing documents, it covers the selecting out of composite parts (chapters/sections/paragraphs) of the documents under analysis and, after that, performing the multi-criteria analysis of the composite parts. This is followed by the designing of alternative variants of the selected information and performing a multi-criteria analysis of the summarised integrated alternatives of the text by which the retrieval results are then formulated.

Once the selecting, processing and indexing of information has been completed, the selecting out of the composite parts of the documents and their multi-criteria analysis are performed. Further alternative variants are designed, these are analysed and the most rational alternative is selected. All this makes the text analytics system more flexible and more informative, since it selects out electronic information as much by area as by coverage.

The multi-criteria analysis of the most rational text materials from a library of documents under analysis covers the complex determination of criteria weights taking into account their quantitative and qualitative characteristics. It includes a multi-criteria evaluation of the text materials defining the utility and market value of the text materials.

Text Analytics Model permits selecting the maximally rational information in the coverage that the user desires. The designing of alternative variants provides the user with an opportunity to supplement and/or correct the already inputted bag of concepts space, modify the weights and then repeat the search. In other words, the user by using User Model is provided an opportunity to intervene in the occurring retrieval and to redirect it; thus the retrieval takes into account the user-selected priorities and the existing situation.

The designing of alternative variants from the selected text materials contained in a library of documents covers the following stages: a) development of a table of codes of text materials from a library of documents, b) rejection of inefficient versions, c) computer-aided development of summarised, integrated text alternatives based on the codes compiled during Stage a), d) development of summarised, integrated text alternatives and the conceptual and quantitative information describing them and e) development of a summary decision-making table of all the obtained summarised, integrated text alternatives and relevant conceptual and quantitative information overall.

At the beginning of a search, a user is able to submit the following kinds of search requirements: The user indicates the goal or goals for the search – research, practical or cognitive. The user notes the possibilities of interest to him/her while conducting the search: research literature (books, academic articles and the like), practical literature or popular literature; the user requests or selects bag of concepts space (see Figure 1); the user establishes various limitations (volume of the material under search by pages, desired time for reading a lecture by minutes and the like).

To limit the amount of search results showing the pages that include the concepts in question (or to restrict the search by the duration of reading), tick the option Advanced search options below the button Search. Additional fields appear: Approximately ... pages and Approximately for: ... minutes. You will also see round buttons to choose search either by the number of pages (default) or by the duration of reading.

ANDROID | **DISASTER RESILIENCE NETWORK**

Search Content Systems

Enter concept:

☐ Only PPT

Add concept **Quick search**

Strategies :: 1
 Earthquake :: 5
 Risk :: 1
 Management :: 1
 Types of earthquake :: 1,9
 Seismic zones :: 2,8

Disaster
 Preparedness
 Earthquakes by continent
 Earthquakes by country
 Bridge disasters caused by earthquakes
 Deaths in earthquakes
 Earthquake engineering
 Earthquake templates
 Earthquake stubs

☒ Add with synonyms

Delete **Delete ALL** **Add selected concept from the list**

New concept: Weight for concepts adding:

Weight:

Add a concept **Search** **Restore concepts**

Fig. 1. User window of the Text Analytics for the Android Project

The Agent subsystem accumulates information about a user and stores his/her individual data. This information can be explicit (year of birth or university graduation) or implicit. The main skills of a user are implicit. They consist of informal and unregistered knowledge, practical experiences and skills. Such data are very important because they describe a user's experience. Information about a user's existing education, needs and the like accumulate in the Agent subsystem.

As a user's historical search information is being analyzed, his/her initial search requirements can be refined (or made more specific). In this case, the user's behavior is under analysis; for example, which documents the user does or does not select for review, how often a document is viewed and how much time is spent looking at it along with use of the drag function are all under observation. This may partially be called the analysis of user conducted searches, the agent function. The Agent subsystem accumulates statistical information about the previous searches conducted by a user in a matrix form: bag of concepts space of a search; results of a search; how many times a user modified the initial search before suitable results were gained; the most popular resources and Internet website addresses employed by the user; how many times did a user read the selected material and how much time was spent doing so.

This way the automatic search is actually personalized by applying the historical information gathered by the Agent subsystem: bag of concepts space under search is refined (or made more specific); information about the user's education, work experience and search needs are considered; the user's most frequently employed resources, Internet website addresses and authors are considered; the user's opinion regarding the significance of the documents gained by the results of a search are considered.

The following factors determine a rational text:

- Citation index of papers (Scopus, ScienceDirect, Google Scholar) (Figure 2)
- Citation of authors (Scopus, ScienceDirect, Google Scholar, etc.) (Figure 2)

- Top 25 papers
- Impact factor of journals (Figure 2)
- Popularity of a text (citation index, number of readers, time spent reading)
- Reputation of the documents
- Supporting phrases
- Document name and contents
- Density of keywords

The following factors determine a rational text:	Paragraph 1	Paragraph 2	Paragraph 3
Citation of papers:			
Citation of papers (Web of Science)	11	5	5
Top 25 papers	-	-	-
Impact factor of journals	1,603	0,983	0,983
Density of keywords (% of a text):			
energy	0,611932687404382	0,42164441321153	0,53003533568905
buildings	1,87891440501043	1,24826629680999	0,202474690663668
Citation of authors:			
Author 1			
Web of Science			
Sum of the Times Cited	24	96	96
Sum of Times Cited without self-citations	10	86	86
Citing Articles	16	80	80
Citing Articles without self-citations	10	74	74
Average Citations per Item	2,18	7,38	7,38
H-index	3	5	5
Google			
Citations	4760	792	792
H-index	-	11	11
i10-index	-	13	13

Fig. 2. User window of the Text Analytics for Android Project for the analysis of the citation index of papers (Scopus, ScienceDirect, Google Scholar), citation of authors and impact factor of journals.

The system was developed as Web application using Microsoft Visual Studio 2010 (.Net Framework 4), C# as a main programming language and MS SQL Server 2012 as database platform. An example of the fragment of the rational text analytics result is presented in Figure 3.

☒ Pages ☐ Time
 Approximately pages
 Approximately for : minutes

Second level keywords

[Earthquakes by continent](#) [Earthquakes by country](#) [Bridge disasters caused by earthquakes](#) [Deaths in earthquakes](#) [Earthquake engineering](#) [Seismic zones](#) [Types of earthquake](#) [Earthquake templates](#) [Earthquake stubs](#) [Earthquake prediction](#) [Earthquake sensitive](#) [Global Earthquake Model](#)

SEARCH RESULT DOCUMENT

1 Great East Japan Earthquake and tsunami United Nations Environment Programme. Great East Japan Earthquake and tsunami. [2013 02 20]. Available on the Internet: .

[Great East Japan Earthquake and tsunami](#)

[Great East Japan Earthquake and tsunami](#)

2 On 11 March 2011, a 9.0 magnitude earthquake off the north-eastern coast of Japan – the strongest ever

Fig. 3. Fragment of an User window of the Text Analytics for the Android Project results

3. Conclusions

Research shows that various researches have specialised in depth the different and very important areas of text analytics and mining (blogs and social networks (Shenghua and Li 2013; Marwick 2014; Mostafa 2013; Boulos et al. 2010), students' online interaction (He 2013), digital libraries (Nguyen 2014, Fagan 2014), process industry (Liew et al. 2014), medicine (Anholt et al. 2014), legal, business intelligence, security (Truyensand Van Eecke 2014), etc). Currently authors participated in the Android (Academic Network for Disaster Resilience to Optimise educational Development) project. In order to increase the efficiency and quality of the delivery of training, teaching and research activities: a Text Analytics for the Android Project has been developed. The Text Analytics for Android Project covers the following methods developed by authors (Kaklauskas 1999): Method for a complex determination of the weights of the criteria taking into account their quantitative and qualitative characteristics; Multiple Criteria Method for a Complex, Proportional Evaluation of Text Materials; Method for a multi-criteria, multivariant design of summarised, integrated text variants; Method for a determination of the utility degree and market value of the text materials. The developed System is also practically used in two distance MSc study programmes of Vilnius Gediminas Technical University (Real Estate Management; Construction Economics and Business).

In the future we are intended to integrate Text Analytics for the Android Project with biometrics technologies. We will perform a comparable research, aiming to base the choice of learning materials on a students' learning productivity and the level of interestingness.

References

- Abul Seoud R.A., Mabrouk M.S., 2013. TMT-HCC system: a tool for text mining the biomedical literature for hepatocellular carcinoma (HCC) biomarkers identification. *Computer Methods and Programs in Biomedicine* 112 (3), 640-648.
- Anholt, R.M., Berezowski, J., Jamal, I., Ribble, C., Stephen, C., 2014. Mining free-text medical records for companion animal enteric syndrome surveillance. *Preventive Veterinary Medicine* 113 (4), 417-422.
- BusinessAnalytics, 2013. Text analytics. Available in the Internet [2014-04-30]: <<http://searchbusinessanalytics.techtarget.com/>>.
- Butleranalytics, 2013. 5 Free Text mining Tools. Available in the Internet [2014-04-30]: <<http://butleranalytics.com/5-free-text-mining-tools/>>.
- Fagan, J. C., 2014. The Suitability of Web Analytics Key Performance Indicators in the Academic Library Environment. *The Journal of Academic Librarianship* 40 (1), 25-34.

- Gartner. Text Analytics. Available in the Internet [2014-04-30]: <<http://www.gartner.com/it-glossary/text-analytics>>.
- 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.
- Hettne, K. M., Kleinjans, J., Stierum, R. H., Boorsma, A., Kors, J. A., 2014. Chapter 6.3 - Bioinformatics Methods for Interpreting Toxicogenomics Data: The Role of Text-Mining, in *Toxicogenomics-Based Cellular Models*, Academic Press, New York, pp. 362.
- Kaklauskas, A., 1999. Multiple criteria decision support of building life cycle: research report presented for habilitation: technological sciences, civil engineering (02T) / Vilnius Gediminas Technical University. Vilnius: Technika, pp. 119 p.
- Kaklauskas, A., Zavadskas, E., Babenskas, E., Seniut, M., Vlasenko, A., Plakys, V., 2007. Intelligent library and tutoring system for Brita in the PuBs project. 4th International conference on cooperative design, visualization, and engineering, September 16–20, 2007, Shanghai, China. Springer-Verlag Berlin, p. 157–166.
- Kamel Boulos, M. N., Sanfilippo, A. P., Corley, C. D., Wheeler, S., 2010. Social Web mining and exploitation for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance. *Computer Methods and Programs in Biomedicine* 100 (1), 16-23.
- Liew, W. T. Adhitya, A., Srinivasan, R., 2014. Sustainability trends in the process industries: A text mining-based analysis. *Computers in Industry* 65 (3), 393-400.
- Machine Learning Market, 2013. CSML Projects. Available in the Internet [2014-04-30]: <<http://www.csml.ucl.ac.uk/projects/>>.
- Marwick, B., 2014. Chapter 3 - Discovery of Emergent Issues and Controversies in Anthropology Using Text Mining, Topic Modeling, and Social Network Analysis of Microblog Content, in *Data Mining Applications with R* (by Zhao Y., Cen Y.), Academic Press, New York, pp. 514.
- Miner, G., Elder, J., Hill, T., Nisbet, R., Delen, D., Fast, A., 2012. Chapter 2 - The Seven Practice Areas of Text Analytics, in *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, Academic Press, New York, pp. 1000.
- Mostafa, M. M., 2013. More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications* 40 (10), 4241-4251.
- Nguyen, E., 2014. Chapter 4 - Text mining and Network Analysis of Digital Libraries, in *Data Mining Applications with R* (by Zhao Y., Cen Y.), Academic Press, New York, pp. 514.
- Piedra, D., Ferrer, A., Gea, J., 2014. Text Mining and Medicine: Usefulness in Respiratory Diseases. *Archivos de Bronconeumología (English Edition)* 50 (3), 113-119.
- Stackoverflow, 2013. Text mining and information retrieval. Available in the Internet [2014-04-30]: <<http://stackoverflow.com/>>.
- Truys, M., Van Eecke, P., 2014. Legal aspects of text mining. *Computer Law & Security Review* 30 (2), 153-170.
- Witten, I. H., Bainbridge, D., Nichols, D. M., 2009. Chapter 9 - Visions: Future, past, and present, in *How to Build a Digital Library* (Second Edition), Academic Press, New York, pp. 565.
- Wu, H., 2013. Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior* 29 (1), 90-102.