Book Review

Working with Text. Tools, Techniques and Approaches for Text Mining. Emme L. Tonkin & Gregory J.L. Tourte. Chandos Publisher, Cambridge (MA). 2016. 330 pp. (ISBN 978-1-84334-749-1).

This book published, in the Chandos Information Professional Series and written by 16 contributors, has as its objective to introduce methods, results, and case studies in text mining. And certainly, the first questions are: What is text mining? What can we expect from text-mining applications? What is specific to text mining or can it be said that text mining is simply data mining algorithms applied to text?

This book can be subdivided into three parts. The first two chapters form a general introduction to the text-mining domain and briefly describe the main natural language processing (NLP) tools required to achieve the targeted result. Chapters 3 and 4 present the general legal and ethical backgrounds and the corresponding problems that must be taken into consideration when collecting and distributing text collections or corpora. In the third part, composed of chapters 5 to 12, each chapter mainly corresponds to a case study focusing on a few NLP tools or describing a system used to solve a problem. In both cases, the main purpose is to illustrate the benefit of applying text-mining approaches in a given domain.

As described in the first and second chapters, various applications can be included under the relatively broad definition of text mining. From a historical perspective, we can include the detection of fake documents (e.g., as done by Lorenzo Vella in 1439 with the *Donation of Constantine*, incorrectly dated from the third century AD), authorship attribution, or generating concordances of a given work. Nowadays, the problem is more related to information management, with the need to design and implement tools able to filter, index, extract, or summarize available information. But as underlined by the authors, text mining corresponds to a more complex problem than data mining. Why? Various reasons support this claim. Text is unstructured (or semistructured when considering its logical structure reflected by XML tags). In the database field (structured data), each value must respect a specified domain (e.g., a value for the age must be positive and smaller than 120). Moreover, semantic elements in a text could be implicit, and all natural language is faced with different forms of ambiguities together with the problems generated by the synonymy and polysemy attached to many words or expressions.

In the third part, each chapter tends to follow the same structure. After a general introduction explaining the motivations and interests for a domain, some historical or past research is presented. I like such logical subdivisions. When following the development of a strategy to solve a question, the elements are introduced one by one, usually from the simplest to the more complex. In numerous chapters, the authors introduce examples to help the reader to understand the presented text-mining process. In the last sections of most chapters, the authors present some evaluation of the proposed strategies and compare them with one baseline or a few variants. The average performance values are usually given in a fair way, without manipulating parameter settings to achieve the highest values. Moreover, examples of common errors and drawbacks of the proposed tools give the reader a more precise picture of the limits of current technology. Each chapter ends with a rich reference list providing access to more recent works in the field.

Let's have a closer look at each chapter.

As indicated by the authors in the first chapter, text mining implies the processing of natural language text data, with the objective to extract, structure, and analyze their contents and to discover new and unknown patterns. With this objective, the authors want to provide a large overview of relevant technologies, areas of research and applications, together with their potential future directions. This definition clearly underlines the inherent interdisciplinary process behind an effective text-mining application requiring expertise in multiple domains (linguistics, information extraction, machine learning, man–machine interaction, etc.).

The second chapter briefly introduces the main aspects of a text-mining system showing that even in an interdisciplinary domain, applying text mining can be done cheaply and could be reasonably easy. The first problem is to define the raw material, namely, the text needed to generate a corpus. This harvesting is subject to legal and ethical concerns, as described in greater detail in chapters 3 and 4. The collected data must then by annotated in order to allow for a fair evaluation of the different text-mining strategies. As mentioned by the authors, the users and their expectations must be included in the first steps of a successful text-mining project.

Chapter 3 presents the differences between law and ethics and describes the possible limits when developing a data/text-mining application. Although it is difficult to present the legal framework for all countries, the authors (A. Charlesworth, E.L. Tonkin) have opted to present a few key legal issues (e.g., data as property, personally identifying information, limits when collecting, sharing, and using the data, etc.). Instead of developing a long legal and ethical argument, the authors provide a list of questions or considerations when beginning a text-mining application. This provides a useful checklist of legal and ethical issues that might arise during the project.

Chapter 4 ("Responsible Content Mining") focuses on the legal questions when crawling texts from the Internet, as, for example, when mining a set of scientific journals. Such a harvest can be done within a noncommercial research perspective, but as mentioned by the authors, the situation is not, in general, fully clear and varies from country to country. Therefore, the authors recommend obtaining the needed permissions. A few examples are given with some details to more concretely illustrate the underlying questions and problems. For example, the permissions can be guaranteed for a few persons inside a specific period of time. The question related to the publication of the results must also be taken into account. A proposed best practices guideline closes this presentation.

After the fourth chapter, the reader can select the chapter s/he would like to read first. No strict sequence is given. Each chapter is self-contained and presents an application using text-mining tools in a given context with an emphasis on a few main objectives (favoring to focus on a specific method or a set of NLP tools).

Chapter 5 presents a complex system used to allow a more user-friendly and efficient access to biological and medical full-text services (Europe PubMed Central). The biomedical domain represents an active field of research for many NLP applications, and thus it is not a surprise to find a case study in this field. For example, after giving a gene name (e.g., MDM2) to the search engine, the user can be overloaded by too many answers. Therefore, the main idea behind *EvidenceFinder*, a search aid, is to suggest a more precise context for the submitted query by generating a set of questions and then letting the user select the most appropriate one (e.g., the user's information needs could be interaction, regulation, inhibition by the corresponding gene, etc.). After selecting one of the proposed questions, the system produces an output in the form of a number of sentences extracted from research articles. In those sentences, *EvidenceFinder* is able to emphasize the verb related to the question, or an expression corresponding to the answer of the selected question. Such a system is built around a set of text-mining tools. First a sentence splitter decomposes the input document into a sequence of sentences. Then a syntactic parser defines the sentence structure in conjunction with a named entity recognizer (NER), used to detect the correct text span. Of course, such NERs are built in conjunction with a list of names in the target domain (e.g., protein, gene, drug, and disease names). The structure Subject-Verb-Object (SVO) is then stored. One of the main difficulties when developing such a system is the generation of a set of pertinent questions without closely similar formulations.

Chapter 6 exposes problems related to social media (e.g., Tweets, Facebook, LinkedIn, blogs) where the Internet language (Crystal, 2011) with its variations, multilingual content, and unusual or incorrect spelling impacts the effectiveness of various text-mining tools (e.g., an NER can achieve an accuracy of 85–90% on longer and traditional texts, but only 30–50% on microblogs such as tweets). Similar effectiveness variations can be observed when considering the performance of a POS tagger (from 97% [newswires] to 70–75% [with tweets]). In this perspective, this chapter describes the GATE open-source text-mining tools, focusing mainly on the information extraction (IE) aspect in which the main subtasks are NER, co-reference resolution, and relation extraction. Therefore, a typical IE system corresponds to the following sequence of tools: language identification, tokenization, sentence splitting, spelling correction, and word normalization, POS tagging, and NER. Examples are given to illustrate the effects of the different tools.

The problems described in chapter 7 are the development and evaluation of an NER applied to around 32,000 company names that can be found in news archives. In this application, one major constraint consists in obtaining a high precision and recall (the proposed system achieves a precision of 0.84 and a recall of 0.72). The system architecture is mainly based on existing software components (e.g., UIMA from IBM, Mallet, OpenNLP, etc.), with additional programs to combine them. One of the difficulties is to recognize variations of the same name (e.g., Microsoft Corp, Microsoft, MS, Microsoft Corporation, or even Micro$oft). In some cases, the correct full name could be rather problematic to define precisely (e.g., does "United" mean US, the flight company or another one?). Various heuristics must be defined to improve the effectiveness of the system. Given a string as input, the system works in a two-stage approach. First, a list of possible candidates is generated and ranked. Then the best one, if good enough, is selected (binary classification task). This second phase is not always easy and the context can be helpful. For example, having extracted the name "Walmart," the string "Apple" could still be a company name but certainly with a lower probability.

Chapter 8 (written by M. Zampieri) presents the problem of language identification (given a [short] text excerpt, can you identify the language?). This chapter explains with examples the techniques used to detect the written language. A first set of features corresponds to functional words, words used very frequently in a language (e.g., articles, prepositions, conjunctions, pronouns, and auxiliary verb forms). A second set of features consists of *n*-gram of characters (e.g., "ing," "th," "wha").

DOI: 10.1002/asi

An interesting aspect of this chapter is the accuracy rate analysis. This classification task can achieve an accuracy rate of 95–99%. As explained by the author, such percentages can be obtained when discriminating between two (or more) distinct languages (such as Spanish and English). However, the performance level decreases when faced with similar languages (e.g., Croatian, Slovene, and Serbian), or when discriminating between national varieties (such as Canadian, US, British, or Australian English). Electronic-mediated communication (texting, tweets, chats) or text obtained after applying an optical character recognition (OCR) process represent other cases in which language identification tasks can be more challenging. Therefore, the accuracy rate reported from text-mining applications must be taken with caution (Hand, 2006); changing the context or the language can certainly have a significant affect on the overall performance.

With chapter 9, the authors present an application of text mining with historical documents dealing with trading information inside the British Empire (mainly between Canada and the UK). Here the main objective is to extract data and tables from digitalized documents written or printed from 1800 to 1940. The main challenges are the poor text quality of the OCR and the correct interpretation of tabular data. Clearly, working with old fonts, poor print and paper quality, together with the 19th century language, explains the poor OCR performance. The OCR output can also be corrupted by headers, page numbers, or even handwritten annotations in page margins. For tables, the row and column alignment is not always perfect, and the presence of various glyphs (to indicate the same entry such as "or") may render the correct identification of the quantities, dates, and proper nouns more problematic. To achieve better results, different postprocessing steps are required to correct the data and to generate the XML representation. As for other applications, the underlying computation used additional sources such as lexicon, gazetteers, and lists of locations and commodity names to achieve a better NER performance. In a final step, the XML documents are used to enrich a relational database (called Trading Consequences) providing access to the relationships between the commodities, locations (origin, destination, or transit), and dates.

The description of an automatic indexing system (HIVE) is the main topic of chapter 10. Based on a set of controlled vocabularies and an input text, this system generates a document representation (a list of terms reflecting its content). Those terms could be extracted from the document itself or from the provided controlled vocabularies. An interesting aspect is the emphasis on extracting not only isolated words but longer expressions or phrases. Based on a set of training documents (texts and their manually assigned descriptors), the system is able to learn the associations between the given subject-headings and the text. The final term selection is a function of multiple criteria such as the well-known *tf idf* (Manning, Raghavan, & Schütze, 2008), the term position inside the document, its presence in the manually assigned descriptors, or in the provided thesauri, as well as their presence in Wikipedia pages. A version in Spanish is available. The overall effectiveness of the system depends, however, on various parameters for which an optimal value must be fixed (e.g., based on a set of calibration tests).

Chapter 11 is an interesting illustration of the numerous problems related to the synonym issue present in all NLPs. The objective of the described system (PIMMS) is to mine documents on climate science to extract targeted information for improving related controlled vocabularies. For each sentence appearing in the input document, the system is able to generate the corresponding parse tree. However, the underlying domain needs to be taken into account. Working with chemical expressions, the parser must be adapted to this domain (e.g., the NER module must have access to chemical entity dictionaries). As an example, the authors present the difficulties encountered when extracting all temporal phrases (e.g., a decade ago). For example, the word "year" can appear in the following forms: y, yr, ya, a, cal. Some periods have their own name (e.g., Holocene) but they can appear in more complex expressions such as "last glacial maximum to the late Holocene." The recognition system is not perfect because such temporal phrases can be incorrectly classified as a location (e.g., Greenland interstadial). Based on their evaluations, the agreement between the manual and automatic phrase recognition is around 89%.

Chapters 12 and 13 end this book by providing an answer to the question "Is a linguist still useful in an NLP project?" (chapter 12). This question is pertinent when we remember the quote of Fred Jenilek (pioneer in automatic speech recognition technology) "Anytime a linguist leaves the group, the recognition rate goes up." And certainly, one of the main concerns in NLP domain is the small number of tools/corpora/research done in languages other than English (and very few other ones, such as Chinese or Spanish). To this, we can add the problem related to texts written in old English (or French, Italian, etc.). A second problem is the wrong impression that NLP tools can be built in a vacuum, without closely analyzing their errors. Selling or using a system that generates too many errors gives to the final user the impression of a "stupid tool" or a "stupid answer." User's confidence in the system is essential but can disappear quickly when encountering silly answers. Chapter 13 presents a short interview with one of the authors (E.L. Tonkin) with R. Garreta, cofounder of Tryolabs.com, a start-up developing NLP applications combining machine learning and software engineering. The text-mining market is characterized by high expectations. The targeted projects are however risky because customers want both a high accuracy rate and a fast response time.

A final remark. This book contains useful appendices. Appendix A provides lists of resources for text mining, software, libraries, frameworks, a relatively complete list of packages (with a short description of the most

DOI: 10.1002/asi

important ones), web-mining solutions, data-mining packages, web interfaces for text mining, and distribution and scaling solutions. Appendix B focuses on databases and vocabularies to allow the reader access to text collections or benchmarks to evaluate different solutions (e.g., the well-known Reuters data set or e-mails from the Enron corpus). Appendix C exposes the different systems providing effective visualization of corpora with maps, graphs, and other graphical tools.

In conclusion, this is an interesting book providing an overview of various text-mining applications, without entering into the technical details. Each chapter can be read separately after consulting the first two chapters. The examples are mainly given with English data and the reported experiments are presented with their advantages and drawbacks.

References

Crystal, D. (2011). Internet linguistics. London: Routledge.

Hand, D. (2006). Classifier technology and the illusion of progress. Statistical Science, 21, 1–14.

Manning, C.D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge, UK: Cambridge University Press.

Jacques Savoy
University of Neuchatel
Computer Science
ruen Emile Argand 11
2000 Neuchatel
E-mail: Jacques.Savoy@unine.ch

Published online 29 August 2017 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/asi.23899