Demo: Using RapidMiner for Text Mining

Yordan Shterev

NMU, Veliko Turnovo, Bulgaria
jshterev@abv.bg

Abstract. In this demo the basic text mining technologies by using RapidMining have been reviewed. RapidMining basic characteristics and operators of text mining have been described. Text mining example by using Navie Bayes algorithm and process modeling have been revealed.

Keywords: Text Mining, RapidMiner, Text Processing, Tokenization, Naive Bayes

1 Introduction

Data and information are mainly in text format and very small part is in figures. There are a lot of books, documents, web pages, e-mails, blogs, news, summaries, papers etc. in digital format. In addition the quantity of information both digital and hard copies increases exponentially. Quick access to them is needed. It is estimated that approximately 80-85% of the information in data bases is natural languages [1,4]. However, we are not able to perceive and process such a big quantity of information. As a result, we are exposed to a lot of information which increases in the course of time. That is why text analysis is topical nowadays [2,3,5]. That is why text analysis follows data analysis. On the other hand, text analysis helps the development of web mining (information analysis in web space), social networks analysis.

RapidMiner is able to process and analyze data, analyze text and web as well. It is number one amongst non-commercial software for data processing in recent years.

Text mining and its essence, the tasks for text analysis, some related algorithms and the characteristics of RapidMiner for text analysis have been discussed in this paper.

2 Essence of Text Mining

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).

Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting it from different written resources. A key element is the linking together of the extracted information to form new facts or new hypotheses to be explored further by more conventional means of experimentation [1,2]. In

text mining, the goal is to discover heretofore unknown information, something that no one yet knows and so could not have yet written down.

Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. The difference between regular data mining and text mining is that in text mining the patterns are extracted from natural language text rather than from structured databases of facts. Databases are designed for programs to process automatically; text is written for people to read.

The fundamental limitations of text mining are: first, that it will not be able to write programs that fully interpret text for a very long time, and second, that the information one needs is often not recorded in textual form.

The term token is used here. It is a string of characters, categorized according to the rules as a symbol (e.g., identifier, number and comma). The process of forming tokens from an input stream of characters is called tokenization.

In computer science, lexical analysis is the process of converting a sequence of characters into a sequence of tokens. A program or function which performs lexical analysis is called a lexical analyzer.

In computing, stopwords are words which are filtered out prior to, or after, processing of natural language data (text). It is controlled by human input. There is not one definite list of stop words which all tools use, if even used [2,5].

3 RapidMiner Possibility for Text Mining

RapidMiner is a software packet with open code for data mining, web mining, text mining.

This main group contains operators to load and process non-structured textual data and transform such data into structured forms for further analysis. RapidMiner version 5.2.006 has the next groups of text mining operators [6,7]:

1. Tokenization.
2. Extraction - Operators filter tokens from documents and to filter document collections.
3. Stemming - The mapping of distinct morphological variations of a word to a common form by reducing variants of such a word to the same word stem.
4. Transformation - Operators transform documents and tokens in documents.
5. Utility - 1 operator about convert a word list into a data set.
6. Other operators about reading, creating, writing and processing documents.

4 Conclusion

One of the leading products for data analysis, including text analysis with open code RapidMiner has been presented. Some basic technologies for text analysis by RapidMiner, as well as the algorithms and operators have been presented. Some operators for text processing and text analysis have been discussed.

It is necessary to apply text analysis on a variety of documents with different purpose and structure, and in more natural languages as well. What is more, text analysis should be used for text information processing in our libraries, taking into consideration the fact that they are being computerized nowadays.

References

1. Nong Ye, The Handbook of Data Mining, Arizona State University, Lawrence Erlbaum As-sociates, publishers Mahwah, New Jersey London, 2003.
2. Brook Wu, Handbook of Research on Text and Web Mining Technologies, Information Sci-ence Reference - New York, 2009.
3. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, An Introduction to Information Retrieval, Online edition 2009 Cambridge UP.
4. I. Rish, An Empirical Study of the Naive Bayes classifier, Computer Science, November 2, 2001.
5. Ian H.Witten, Katherine J. Don, Michael Dewsnip, Valentin Tablan, Text-mining in a digital library, Published online: 2003 – Springer-Verlag 2003.
6. RapidMiner 4.4, User Guide Operator Reference Developer Tutorial, Rapid-I GmbH, Stockumer Str. 475, 44227 Dortmund, Germany, http://www.rapidminer.com/, Copyright 2001-2009 by Rapid-I, March 14, 2009.
7. http://rapid-i.com/