International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 7– No. 2, April 2014 – www.ijais.org

Tokenization and Filtering Process in RapidMiner

Tanu Verma
Student
CSE, ITM University

Renu
Student
CSE, ITM University

Deepti Gaur
Associate Professor
CSE, ITM University

ABSTRACT
Text mining is defined as a knowledge-intensive process in which a user interacts with a document collection. As in data mining[2,4,9], text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns. A key element of text mining is its focus on the document collection. A document collection can be any grouping of text-based documents. Most text mining solutions are aimed at discovering patterns across very large document collections. The number of documents can range from the many thousands to millions. In this paper, we will see how text mining is implemented in Rapidminer.

Keywords
Text mining, Tokenize, Filtering, Stop words, Stemming.

1.INTRODUCTION
Text mining [11, 12] is the analysis of data contained in natural language text. Text mining can help an organization derive potentially valuable business insights from text-based content such as word documents, electronic mail as well as postings on social media streams. Mining unstructured data with natural language processing (NLP), statistical modeling and machine learning techniques can be a challenge, because natural language text is often inconsistent. It suffers from ambiguities caused by inconsistent syntax and semantics.

2.TEXT MINING
In this paper Process Documents from Files operator is used. It generates word vectors from a text collection stored in multiple files. Parameters used in this operator are :-

- text directories:- In this list arbitrary directories can be specified and All the files that matches the given file ending will be loaded and assigned to the class value provided with the directory. file pattern: A pattern for the file to be read.
- extract text only:- If checked, structural information like xml or html tags will be ignored and discarded.
- use file extension as type:- If checked, the type of the files will be determined by their extensions. The unknown extensions will be considered as text files.
- content type:- The content type of the input texts.
- encoding:- The encoding used for reading or writing files.

The JISC and National Centre for Text Mining explain how "text mining involves the application of techniques from areas such as information retrieval, data mining, information extraction and natural language processing. All of these various stages of a text-mining process can be combined into a single workflow".

- Information retrieval (IR) systems match a user's query to documents in a database or collection. The first step in the text mining process is to find the body of documents that are relevant to the research question(s).
- Natural language processing (NLP) analyzes the text in structures based on human speech and allows the computer to perform a grammatical analysis of a sentence to "read" the text.
- Information extraction (IE) [3,5,6,7] involves structuring the data that the NLP system generates.
- Data mining (DM)[1,8,13] is the process of identifying patterns in large sets of data, to find that new knowledge.

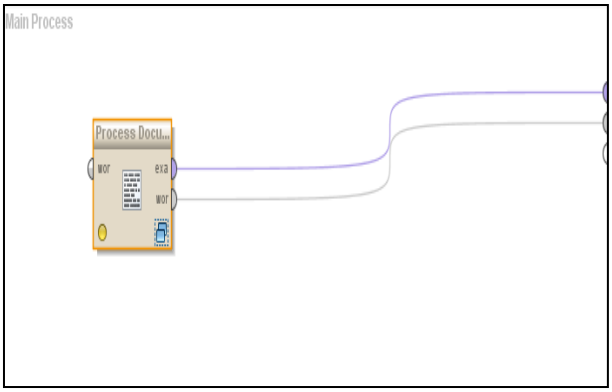

Fig. 1. Processing document from files in RapidMiner

Figure 1 shows the 'Process Documents From Files' in RapidMiner. In the parameter on the right hand side we have a field 'text directories' where we have to enter the text file which we want to tokenize and filter. The text file should be in a folder. Fig. 2 shows the insertion of text file which has to be tokenize. We have 2 column, the first one is class name(we can give any class name) and the second is directory(which we have to select from the specific location).

3. TOKENIZE
Tokenization is the process of breaking a stream of text up into phrases, words, symbols, or other meaningful elements called tokens. The goal of the tokenization is the exploration of the words in a sentence. Textual data is only a textual interpretation or block of characters at the beginning. In information retrieval require the words of the data set. So we require a parser which processes the tokenization of the documents. This may be trivial as the text is already stored in machine-readable formats. But Still there are some problems that has been left, for e.g., the removal of punctuation marks as well as other characters like brackets, hyphens, etc. The main use of tokenization is identification of meaningful keywords. Another problem are abbreviations and acronyms which need to be transformed into a standard form.

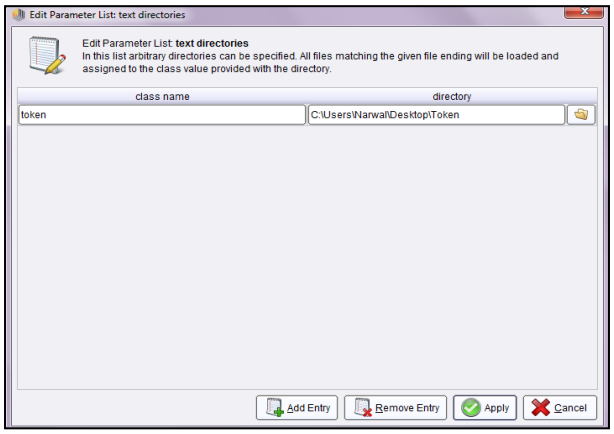

Fig. 2. Insertion of text file to process

- Tokenize :- This operator splits the text of a document into a sequence of tokens. There are several options to define the splitting points. The options are as follows:
- mode:-This selects the tokenization mode. Depending on the mode, split points are chosen differently. The Range is non letters, specify characters, regular expression and the default value is non letters
- characters:- The incoming document will be split into tokens on each of this characters. For example enter a '.' for splitting into sentences. The Range is string and the default value is '.:'
- expression:- This regular expression defines the splitting point. The Range is string.

Stopword Elimination: - The most common words that unlikely to help text mining such as prepositions, articles, and pro-nouns can be considered as stopwords. Since every text document deals with these words which are not necessary for application of text mining. All these words are eliminated. We can choose any group of word for this purpose. It also reduces the text data and helps to improve the system performance. For e.g., "a","is", "you", "an".

Stemming: - Stemming also known as lemmatisation is a technique for the reduction of words into their stems, base or root. Many words in the English language can be reduced to their base form or stem e.g. like, liking, likely, unlike belong to like. Moreover, names can be transformed into root by removing the "s", for e.g., During the stemming process the variation "Stem's" in a sentence is reduced to "Stem" and this removal may lead to an incorrect stem or root. However, if these words are not used for human interaction then, these stems do not have to be a problem for the stemming process. But the stem is still useful, because all other inflections of the root are transformed into the same root.

4. FILTERING

Filtering helps you to provide the flexibility when you want to design your data sources and mining structure so that a single mining structure can be created based on the comprehensive data source view. For training and testing different models, filters can be created to use only a part of that data and no need to build a different structure for each subset of data. We can use filter by length, Content, English, dictionary and Region etc.

In this paper tokens are filtered by length. This operator filters tokens based on their length (i.e. the number of characters they contain). Parameters used in this operator are:

- min chars:- The minimal number of characters that a token must contain to be considered.
- max chars:- The maximal number of characters that a token must contain to be considered.

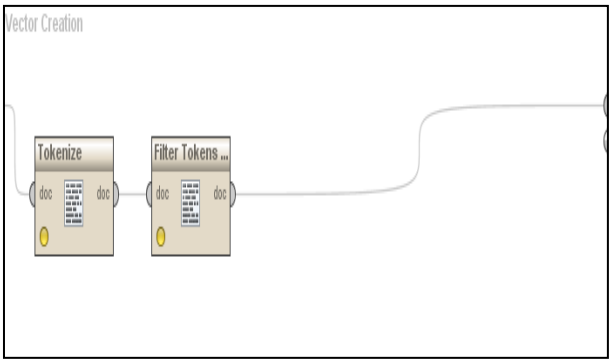

Fig. 3. Tokenization and Filteration in RapidMiner

Figure 3 shows the tokenization and filteration in RapidMiner. In this we use two operators "Tokenize" and "Filter Token by Length".

5. RESULT AND ANALYSIS

Now, we run it and get the output. Fig. 3 shows the output in Rapid Miner. The list of tokens i.e. words, phrases, symbols or other meaningful elements becomes input for Tokenization process such as parsing or text mining. The document occurrences and total occurrences of the tokens is given in the result summary. In this paper, we filter the tokens by length.

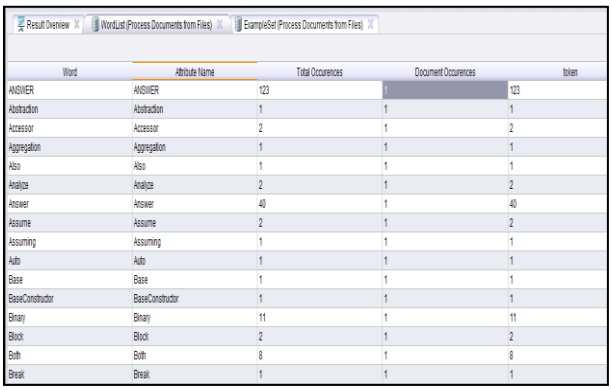

Fig. 4. Result of Text Mining

6. CONCLUSION AND FUTURE SCOPE

Text Mining is a growing applications field and an area of research, using techniques from well-established scientific fields such as data mining, natural language processing, case-based reasoning, statistics [10], machine learning[5, 8], information retrieval [3] and knowledge management. In this paper, we have presented an approach that uses an

automatically learned Information extraction system to extract a structured database.

7. REFERNCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in Proceedings of the 20th International Conference on Very Large Databases (VLDB-94), Chile, Sept. 1994.

[2] Margaret H. Dunham, Data Mining "Introduction and Advanced Topics".

[3] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval" ACM Press, New York, 1999.

[4] Agrawal , T. lmielinski and A. Swami " Database mining: A performance perspective", IEEE Transactions on knowledge and Data Eng. , vol. 5, no. 6.

[5] M. E. Califf, editor. Papers from the Sixteenth National Conference on Artificial Intelligence(AAAI-99) Workshop on Machine Learning for Information Extraction, Orlando, FL, 1999. AAAI Press.

[6] M. E. Califf and R. J. Mooney, " Relational learning of pattern-match rules for information extraction" in Proceedings of the 16th National Conference on Artificial Intelligence(AAAI-99), pages 328–334, Orlando, FL, July 1999.

[7] C. Cardie, "Empirical methods in information extraction", AI Magazine, 18(4):65–79, 1997.

[8] C. Cardie and R. J. Mooney, "Machine learning and natural language (Introduction to special issue on natural language learning)" Machine Learning, 34:5–9, 1999.

[9] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publisher, 722

[10] Yang Y M, "An evaluation of statistical approach to text categorization [R]" in Technical Report CMU - CS - 97-127. Computer Science Department, Carnegie Mellon University, 1997

[11] C. Choi and Y. Park "R&D proposal screening system based on text-mining approach", Int. J. Technol. Intell. Plan., vol. 2, no. 1, pp.61 -72 2006

[12] H. C. Yang and C. H. Lee "A text mining approach for automatic construction of hypertexts", Expert Syst. Appl., vol. 29, no. 4, pp.723 -734 2005

[13] Agrawal R, Imielinski T and Swami A, "Mining association rules between sets of items in large database[M]", Washington, DC: SIGMOD, 1993.207-216.