Evaluation for Tools and Techniques for Text-Extraction

Naveen Kumar N. S.
Amrita Vishwa Vidyapeetham
Bangalore 560035, Karnataka, India
naveenkumar.nattanmy@gmail.com

Poorvajaa M. S.
Amrita Vishwa Vidyapeetham
Bangalore 560035, Karnataka, India
poorvajaa.poori2@gmail.com

Madhulika Varanasi
Amrita Vishwa Vidyapeetham
Bangalore 560035, Karnataka, India
varanasimadhulika123@gmail.com

M. Mahathi Bhargavi
M.L.R. Institute of Technology
Hyderabad 500043, Telangana, India
mahathi_bhargavi@yahoo.com

Rohan Boorugu
Lovely Professional University
Phagwara 144411, Punjab, India
rboorugu16@gmail.com

Dr. Shekar Babu PhD
Professor & Founding Head
Amrita Vishwa Vidyapeetham
Bangalore 560035, Karnataka, India
sb@amrita.edu

Abstract

Sustainability and Corporate Social Responsibility (CSR) has become important not only to corporations but to society and government as well. As a result, there has been a dramatic increase on how these corporations are informing and disclosing their social and environmental activities in which they are involved through their company annual reports. These annual reports are not only useful for companies but also to government, NGO's, society and many other stakeholders. In the past many of the annual reports were studied manually. But now with internet, there is a great opportunity to develop automation through software tools and techniques. Data Mining techniques and related techniques and softwares can help not only researchers but also corporations to extract text from multiple data sources like word processing documents as well as image documents like an image file or a PDF file. In this paper, the researchers first and designed the manual process and also performed the manual process of text extraction from annual reports. Having understood and performed the manual process the researchers analyzed the various tools and techniques available on net. They explored and analyze all the tools, scripts and software available on the net. The researchers analyzed through various parameters. The researchers developed a benchmarking model across all the tools and techniques after analyzing each of these software critically. These softwares were specifically analyzed for performance and feasibility and various parameters were evaluated within performance and feasibility.

Keywords
Text-Mining, Text Extraction, Software Tool

1.Introduction

With the explosion of Net 2.0 the size of data has grown exponentially. It is continuing to grow at an exponential pace. All the organizations, institutions and businesses are storing and archiving data in electronic format. Most of the data that is passing across the internet is in unstructured data format. (Sagayam 2012) [1]. Hence, to extract any textual based information from unstructured data formats it is required to use different computing methods and techniques. It is also not easy to find useful patterns through within vast amount of data (Padya et al., 2012) [2]. One of the traditional methods are data mining tools, however they are incapable of extracting textual data and unstructured data.

One of the other methods are Text Mining which is used for extracting texts in whole or selected that has useful information (Feldman 2007) [3]. Also not all the textual information can be stored in databases. The data is not stored in a structured manner like in traditional databases, but it is present across company reports and available in company's websites. Techniques used on database based structured data cannot be used on text. Since unstructured data is complex, we need to use more efficient and effective techniques to extract the required information. Hence, text mining tools are needed not only to extract but also to analyze these unstructured data.

Text mining is an interdisciplinary field based on information of text. Text mining can be used in most fields like digital media, social media platforms, healthcare, and cybersecurity. Text mining also includes text in natural natural language, that can be available in various structural formats. We can see that (Sateli et al.2012) [4] describes how text mining can be used for Natural Language Processing (NLP). In the software testing text mining to explore bugs (Malhotra et al.2013) [5]. For social media analytics to predict sentiments text mining is used (Jurek et al.2015) [6].

The researchers in this paper, try to explore the varied techniques used in text-mining, tools and algorithms. They also explore the various efficient and effective techniques to extract text. Moreover the authors would also like to explore the pros and cons of these tools and techniques.

The paper is organized in different sections. Different types of Tools are described in section 1. In section 2, description of the different types of tools and techniques are discussed. Section 3, concludes and outcomes.

2.Different Types of Tools:

1. Pytesseract

Tesseract within Python is primarily used as a tool to recognize optical character based information. The tool can from the images can recognize the text.

Pytesseract takes an image and outputs the text which is embedded in the image and some application as the interface. This allows us to expose the functionality in a more familiar way.

2. Nuance OmniPage Ultimate

OmniPage Ultimate is an innovative solution that converts paper, PDF files, and forms into documents we can automatically send, edit, or archive.

With OmniPage we can get great accuracy, support for basically any scanner, the best tools to customize our process, and automatic document routing.

3. Knime

KNIME is a data analytics based platform which is an open source software. The tool enables the user to understand data by designing workflows and reusable components.

We can blend data from any source, shape it, leverage machine learning and AI and discover insights.

4. Orange

The program provides a platform for experiment selection, recommendation systems, and predictive modeling and is used in biomedicine, bioinformatics, genomic research, and teaching.

It is used as a platform for testing AI based algorithms and to implement new techniques in the area of life sciences and biotechnology. In education, it was used for teaching machine learning and data mining methods to students of biology, biomedicine, and informatics.

5. Rattle

Rattle uses the graphical user interface (GUI) in the R Statistical Software and provides data mining functions. Rattle is generally used as to teach and learn the R software Language. Hence, Rattle is also used for statistical and model building.

6. Keel

KEEL contains a wide variety of tools and techniques within which is useful for preprocessing, extraction, computational based learning algorithms, statistical models for experiments and hybrid model building.

It allows to perform a complete analysis of new computational intelligence proposals in comparison to existing ones. KEEL is used for research areas as well as for educational purposes.

7. TopOCR

The aim is to make it easy to acquire documents and texts from business cards, newspapers, books and magazines without having to carry around a bulky notebook computer and scanner.

It can handle images with mixed text and graphics and it can even tolerate skew and uneven lighting.

8. Mallet

Mallet is widely used if data within the documents are needed to be categorized or classified. Within Mallet there are tools which are used for extraction of text.

Table 1. Tools Description

TOOLS	DESCRIPTION

PYTESSERACT	● Tesseract within Python is primarily used as a tool to recognize optical character based information. The tool can from the images can recognize the text. ● Python-tesseract is similar to Google's Tesseract OCR. It is also useful as a stand-alone invocation script to tesseract. It can support a variety of image types like Pillow and Leptonica imaging libraries, tiff, bmp, png, and others. ● Pytesseract takes an image and outputs the text which is embedded in the image and some application as the interface. This allows us to expose the functionality in a more familiar way. ● Additionally, if used as a script, Python-tesseract will print the recognized text instead of writing it to a file.
NUANCE OMNIPAGE ULTIMATE	● Seamlessly share documents: Omni Page ○ Codirect ™ can automatically send converted files to a predefined destination, multiple destinations or on-demand to anyone, anywhere. ○ It can also watch a shared folder and automatically send converted files into appropriate workflow. ● Language fluency: OmniPage Ultimate recognizes more than 120 languages, so you can process, edit and store documents from anywhere in the world. ● OmniPage Ultimate recognizes languages based on the Latin, Greek and Cyrillic alphabets as well as Chinese, Japanese and Korean languages.
KNIME	● KNIME is a data analytics based platform which is an open source software. The tool enables the user to understand data by designing workflows and reusable components. ● We can blend data from any source, shape it, leverage machine learning and AI and discover insights.
ORANGE	● The program provides a platform for experiment selection, recommendation systems, and predictive modeling and is used in biomedicine, bioinformatics, genomic research, and teaching. ● Orange is for visualization of data at a component level. ● It is used as a platform for testing AI based algorithms and to implement new techniques in the area of life sciences and biotechnology. In education, it was used for teaching machine learning and data mining methods to students of biology, biomedicine, and informatics. ● Visual programming is implemented through an interface in which workflows are created by linking predefined or user-designed widgets, while advanced users can use Orange as a Python library for data manipulation and widget alteration.
RATTLE	● Rattle GUI is a free and open source software (GNU GPL v2) package providing a GUI for data mining using the R statistical programming language. ● Rattle uses the graphical user interface (GUI) in the R Statistical Software and provides data mining functions. Rattle is generally used as to teach and learn the R software Language. ● Hence, Rattle is also used for statistical and model building.

KEEL	● KEEL contains a wide variety of tools and techniques within which is useful for preprocessing, extraction, computational based learning algorithms, statistical models for experiments and hybrid model building ● It allows to perform a complete analysis of new computational intelligence proposals in comparison to existing ones. KEEL is used for research areas as well as for educational purposes.
TOPOCR	● TopOCR brings together a powerful collection of ● the latest Neural Net OCR and image processing ● technology for scanning books, magazines and ● newspapers with document cameras. ● TopOCR combines sophisticated real-time image ● processing with three specialized OCR ● Engines together with an easy to use Image Editor ● and Word Processor/Spell Checker. It also ● provides a single-click Real-Time Document Camera Image Preview and Capture Dialog that ● makes document positioning very helpful.
MALLET	● Mallet is a statistical based natural language processing it is also with very similar to Java. Mallet can be utilized for categorizing text within the documents, group or cluster texts, extraction of information. ● Mallet is widely used if data within the documents are needed to be categorized or classified. ● Within Mallet there are tools which are used for extraction of text

Conclusion

The authors explored the various techniques and tools and through the comparative and detailed analysis they evaluated these tools and techniques. The author in this paper evaluate their features towards text mining.

References

R.Sagayam, A survey of text mining: Retrieval, extraction and indexing techniques, International Journal of Computational Engineering Research, vol. 2, no. 5, 2012.

N. Padhya, D. Mishra, R. Panigrahi et al., "The survey of data mining applications and feature scope," arXiv preprint arXiv:1211.5723, 2012.

Feldman, Ronen, and James Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data.* Cambridge University Press, 2007.

Sateli, B., Angius, E., Rajivelu, S. S., & Witte, R. (2012). Can text mining assistants help to improve requirement specification. *Mining Unstructured Data(MUD 2012), Canada.*

Malhotra, Ruchika, et al. "Severity Assessment of Software Defect Reports using Text Classification." *Intrenational Journal of Computer Applications* 83.11 (2013).

Jurek, Anna, Maurice D. Mulvenna, and Yaxin Bi. "Improved lexicon-based text sentiment analysis for social media analytics." *Security Informatics* 4.1 (2015): 1-13.