# Big Data in text processing: an old problem with a new approach

**Handor Ten***

*International Information Technologies University, Manas Str./Zhandosov Str., 34 A /8A 050040 Almaty, Kazakhstan*

**Corresponding author's e-mail: handor.ten@gmail.com*

**Abstract**

Data is everywhere around us. We use existing data and produce new data every day, every hour and every minute. In addition, most of this data is unstructured. It includes data from the Internet, data from tons of smart devices that around us, data from social nets and messengers that we are using every day in order to communicate with our friends and coworkers. Significant part of this volume is textual information and quite often, it is necessary to somehow store and process it. However, we cannot use usual data management systems and approaches in order to do so and one (but not the only one) of the reasons is that the amount of this data is very, extremely huge. Therefore, we need slightly new modern approaches and techniques to solve this problem. This paper covers the existing solutions in text mining sphere using Big Data technologies.

*Keywords:* Big Data, Map Reduce, Hadoop, Text Mining, Distributed Processing

## 1 Introduction

Nowadays we face with data every day in our life. We study, work, communicate and relax somehow using different types of data. Moreover, today each of us can not only consume this data but also produce new data. The development of the Internet provide a great opportunity to try ourselves as content creators. We can write posts in blogs, upload gigabytes of video material on YouTube, exchange emails and messages through email services, messengers and social nets and, of course, we can write comments approximately on every information resource. The amount of textual information is extremely huge and it is increasing every second. Traditional technologies cannot process this amount of data because it is Big Data. Therefore, we need to use approaches and technologies that are appropriate for Big Data.

## 2 Basic Information

This part of the paper describes shortly main terms and definitions that are connected with Big Data. What is Big Data? How to identify that some data is "Big"? What is 4 V-s that are the main characteristics of Big Data? What is Hadoop and what is it destination? And some more terms that are necessary to know in order to understand this article.

## 3 Existing solutions

Of course, many algorithms and approaches exist that are traditionally used in order to process textual information. Moreover, nowadays these algorithms and approaches successfully combined together with modern Big Data technologies and solutions in order to do old job but on a large scale. This paper investigates such kind of solutions. For example, usage of Knuth Morris Pratt algorithms with the help of Hadoop Distributed File System [1], basic principles of Hadoop and major resources that it uses [2], concrete solutions that are implemented by big companies such as HETA [3] etc.

This paper covers basic principles of these solutions, analyze their advantages, disadvantage, and how the experience from existing solutions can help in implementing slightly new approaches.

## 4 Conclusion

This paper collects information about existing solutions that can help in further work in the sphere of text mining using Big Data technologies and approaches.

## References

[1] Ramya A, Sivasankar E 2014 Distributed pattern matching and document analysis in big data using Hadoop MapReduce model *International Conference on Parallel, Distributed and Grid Computing*

[2] Pandey K, Gadwal A, Lakkadwala P 2016 Hadoop multi node cluster resource analysis *Symposium on Colossal Data Analysis and Networking (CDAN)*

[3] Nicolas V, Da Silva A, Picard M 2014 HETA: Hadoop environment for text analysis *International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)*