УДК 004.912 : 004.048

Bisikalo Oleg V.
Slobodian Roman V.

ANALYSIS OF THE BIG DATA FUNDAMENTAL AND THEIR APPLICATION FOR PROCESS OF DETECTION OF HIDDEN CONTENT IN TEXT INFORMATION

Vinnytsia national technical university

Abstract

All the countless processes in surrounding world generates a huge stream of information about their progress. Some of this information can be collected, stored and analyzed for solving some urgent issues as of all humanity, as of a separate person.

Data collection is one of the key components for all areas of research. Depending on discipline data collection method may varies, but their goal is almost always the same – capture pure information that after some processing allows to answer relevant questions.

One of the most urgent nowadays issues is the detection of hidden content in everyday communication between people. This issue can be solved or significantly simplified by using tools of Big Data's gathering, preservation and processing.

Keywords: Big Data, Text tone analysis, Natural language processing, Jupyter Notebook.

Анотація

Множина різноманітних процесів у навколишньому світі генерує величезний потік інформації про їхній прогрес. Частина цих відомостей може бути зібрана, збережена та проаналізована для вирішення певних актуальних питань для всього людства або окремої людини.

Збір даних є одним з ключових компонентів для всіх галузей дослідження. Залежно від дисципліни метод збору даних може змінюватися, але їх мета майже завжди однакова – отримання ключової інформації, що після деякої обробки дозволяє відповісти на важливі питання.

Однією з найбільш актуальних проблем сьогодні є виявлення прихованого змісту в повсякденному спілкуванні між людьми. Ця проблема може бути вирішена або значно спрощена за допомогою інструментів збору, збереження та обробки Великих Даних.

Ключові слова: Великі дані, Аналіз тональності тексту, Обробка природної мови, Jupyter Notebook.

Introduction

Data gathering – is the process of gathering and measuring information on targeted variables in an established system. This process allows to answer relevant questions and evaluate outcomes. Data collection is one of the key components for all areas of research. Depending on discipline data collection method may varies, but their goal is almost always the same – capture pure information that after some processing allows to answer relevant questions [1].

Regardless of the scope of research or the benefits of determining data (quantitative or qualitative), precise data collection is important for maintaining the integrity of the research. The selection of appropriate data collection tools (existing, modified or redefined) and a clear definition of the rules for their proper use reduces the likelihood of errors.

There are numerous amounts of techniques that being used for gathering data. At the same time, information that being gathered can come from different sources [2].

The ability to collect Big Data about progress of processes opens new perspectives for their analysis. Because of the analysis, conclusions can be obtained that will help to explore and better understand the processes, increase the efficiency of resource allocation, and so on. But for the processing of large volumes of information, in addition to the development of special analytical software, it is necessary to have the appropriate infrastructure for their preservation, processing and transmission [3].

To preserve large volumes of data, the infrastructure used for this purpose becomes of importance. Typically, its development and support require large amount of resources that the average company or group of researchers cannot afford. The problem lies in both, the lack of disk space and the architectural incompatibility of traditional storage for tasks of storing, accessing and processing large volumes of data [4].

The problem of data processing speed is that the data itself is almost useless if it is not processed quickly enough. For example, the results of data processing for real-time operation should not be processed longer than time, that required to take certain actions in a real situation.

The problem of data heterogeneity lies in the fact that data can be collected in different ways and from different sources. It means that it can have different degrees of importance in solving a problem, as well as require different storage formats.

The problem of data security is that data should not be lost or accessed without of permission to do that.

The need of large amounts of data processing is not new. While exploring relationships between the data warehouse and its content, i.e. data, clearly understands that in order to optimize the processing of a large set of data, data should be processed in parallel. Having broken one large set of data into several smaller ones and processing them in parallel, overall speed is increasing. For even greater optimization, large datasets can be stored immediately in parts on distributed file systems or in a distributed database. This will help to avoid or at least minimize operations at the stage of preparing information for processing. This will greatly save time required for the overall execution of the processing process.

Results of research

Most of the information that being collected from various sources using their specific gathering methods contains some informative data about processes progress. After analysis of this information obtained results can be used for making decisions about the state of the processes under research.

The statement mentioned above was taken as a basis for the natural language analysis with the purpose of search of hidden content in it.

Now, there are several methods that can be used to process and analyze text information that was gathered by recognition of natural language [5].

The most common are:

- Graphematic analysis;
- Morphological analysis;
- Keywords extraction;
- Thematic classification;
- Named entities extraction;
- Retrieving Relationships;
- Tone analysis.

Consider them and analyze the possibility of using them for search of hidden content in information.

The results of the analysis of the text by one or another method can be used to construct tools for working with the Big Data model, which will allow to determine with certainty accuracy the fact of the presence of hidden information in the analyzed text.

Unfortunately, due to the disparity and partial or complete incompatibility between the tools that implement one or another of the above methods of processing natural text, must be made a choice of an integrated solution that will find the most complete answer to the problem.

On the role of this solution, due to its architectural features and opportunities for use, might be used IBM Cloud Platform.

The process of analysis of the text is complex and take place in several stages. To obtain a comprehensive result, recognized from natural language text must be processed using natural language processing and tone analysis tools.

Natural language processing process consist of next steps:

- Data gathering and cleaning;
- Definition of key entities in data;
- Data analysis;
- Analysis response saving or transferring.

Tone analysis is slightly different from natural language processing process. It is a process of analysis of authors' attitude to a certain thing, process or event, expressed in the text.

By performing natural language processing and text tone analysis and combining their results together will show about what author is talking about when saying some words, and does these words mean exactly that, what we used them to mean.

Conclusions

Both, natural language processing and tone analysis processes are complicated, hard and expensive to implement from scratch. That's why as a tool for text analysis some parts of IBM Cloud are used.

IBM Cloud is a provider of services for solving various types of data analysis. It allows to process necessary natural language and tone analysis with Bid Data in near to real time period, corresponding to SCV paradigm.

The «Watson developer cloud», one of the parts of IBM Cloud, provides interfaces for low-level integration with IBM Cloud using APIs.

For the purpose of this work we consider using the Natural Language Understanding and Tone Analysis tools.

Both, Natural Language Understanding and Text Tone Analysis tools can be accessed using API and provided for Python developers «Watson developer cloud» module.

The Natural Language Understanding Service performs the processing of a natural language for the analysis of semantic signs of any text. It can analyze data presented in various forms, such as plain text, HTML, or data that can be accessed through a common URL. The Natural Language Understanding tool returns the results for the requested functions that were specified in the API call [6].

Using the Natural Language Understanding tool, we can analyze the semantic features of the text recognized as the natural language input, including:

- Categories;
- Concept;
- Emotions;
- Subjects;
- Keywords;
- Metadata;
- Relations;
- Semantic roles;
- Mood.

The second service, Tone Analyzer, uses linguistic analysis tools to recognize and identify the emotional color of the tact and language tones in it. This service can analyze the tone at several levels: a document, and sentences [7]. The evaluation of the document's tone is a summary estimate of the tone of the sentences of the text.

This service can be used to understand how written messages are perceived. It can also be used to study the tone of the object of observational communications. With similarity to Natural Language Understanding service, Tone Analyzer can work with JSON, plain text, or HTML input that contains your written content to the service. The service accepts up to 128 KB of text, which is about 1000 sentences.

After processing, service returns JSON results that report the tone of provided in a call input.

Once both services finished text analyze, the results obtained from them are being merged and send to the Azure DevOps Cloud.

References

1. Data collection [Electronic resource]. – Access mode: https://bit.ly/2ytyCl5 – Title from the screen.
2. Data Collection Techniques [Electronic resource]. Access mode: https://bit.ly/2mNeAPI – Title from the screen.
3. Где хранить Большие Данные [Electronic resource]. – Access mode: https://bit.ly/2EdUG9x – Title from the screen.
4. Большие данные и их хранение [Electronic resource]. – Access mode: https://bit.ly/2A2fQUa – Title from the screen.
5. Обработка текста [Electronic resource]. – Access mode: https://bit.ly/2OSGpDo – Title from the screen.
6. Natural Language Understanding API Reference [Electronic resource]. – Access mode: https://ibm.co/2Ol7lwh – Title from the screen.
7. Tone Analyzer API Reference [Electronic resource]. – Access mode: https://ibm.co/2RJzFWU – Title from the screen.

Slobodian Roman V. – student of the group 3ACIT-17m, control computer system department, Vinnytsia national technical university.

Supervisor – Bisikalo Oleg V. – Dr. of Sci., professor, Dean of the Faculty of Computer Systems and Automatics, Vinnytsia national technical university.

Слободян Роман Віталійович – студент групи 3АКІТ-17м, кафедра комп’ютерних систем управління, Вінницький національний технічний університет.

Науковий керівник– Бісікало Олег Володимирович – доктор технічних наук, професор, декан факультету комп’ютерних системта автоматики, Вінницький національний технічний університет.