

AN ALGORITHM FOR EXPLORATORY ANALYSIS
AND NORMALIZATION OF BIG DATA WITH PANDAS

Mariya Zhekova

Received on April 8, 2023
Presented by I. Ivanov, Member of BAS, on July 31, 2023

Abstract

The digitization of business processes and the extraction of answers to user requests for big data are modern problems that are of great interest to scientists and researchers. The data generated so far, located in various corpora, is much more than can be analyzed. Therefore, they are collected, identified, cleaned and normalized to be used most adequately. Segmentation, assumptions and hypotheses contribute to the degree of satisfaction with the returned result. The research proposed a general method for collecting, cleaning and normalizing data from various sources, structurally modelling it into appropriate models, then testing hypotheses and analyzing the obtained results to conclude large academic data that will benefit the business in making management decisions. This is possible with the means of computational linguistics and with the help of Python data manipulation libraries.

Key words: big data, data mining, big data analysis, educational analytics, natural language processing

1. Introduction. The digitization of business processes and the extraction of conclusions from big data are modern problems that engage the thinking and work of many scientists and researchers. The data generated so far, located in different corpora, related to the research problem is much more than can be used and analyzed. Therefore, when data is collected, organized and cleaned, it is done to use it most adequately, to make informed decisions and informed assumptions. The segmentation, assumptions and hypotheses that are made with the

DOI:10.7546/CRABS.2023.11.09

data contribute to the degree of satisfaction with the returned result. Cleaning, preparing and follow-up help make logical decisions about business, software, how you work, scientific work and decisions that affect or change people's behaviour and development.

The article offers a sequence of techniques and methods for ordering, cleaning, validating, analyzing, and modelling data through built-in Python libraries and experiments in the field of Higher Education.

The purpose of the document is to present a comprehensive approach to collecting, obtaining and validating, arranging and normalizing academic and educational data from various sources, their structural modelling to appropriate models, subsequent hypothesis testing, generalization and analysis of the obtained results to improve quality of the collected data, the effectiveness of the teaching and learning processes, as well as the success rate in the management of large volumes of educational data.

This is possible with the means of computational linguistics and with the help of the libraries for reading and working with data in Python – Pandas and Dask, as well as the library for connecting to databases – Pyodbc.

A 5-step methodology is proposed that can be followed in intelligent data analysis and management in information systems with different purposes and functionality, such as:

- combining data from different origins;
- cleaning, arranging and normalizing data;
- intelligent data analysis and management;
- making management decisions;
- concluding;
- retrieving answers to user questions.

2. Basic paradigms. Big data is a concept that describes large volumes of structured and unstructured data that grow exponentially over time and are complex to analyse with traditional techniques [1]. In addition to storage and access, they are an alternative to traditional solutions related to analysing, interpreting and extracting meaning and value from databases.

Gartner researchers in 2012 first defined big data as: “*High-volume, high-velocity, high-variety* information assets that require new forms of processing to enable improved decision-making, insight discovery, and optimization of processes. Subsequently, the characteristics of *value and veracity*, which comprehensively define big data, are included in the definition [2]. The latest research in the field adds qualities such as *variability, visualization, scope, venue, version control, viscosity, validity, robustness, visibility, vocabulary, ambiguity, complexity*, etc. [3].

The volume characteristic refers to large amounts of data of different types and structures from a variety of sources. Variety refers to the different types of data collected in the form of textual information, images, and sound files, using sensors, calculations, social networks, forums, files, etc. They can be structured,

semi-structured and unstructured. Velocity measures the speed of data collection, processing, analysis, transfer and distribution. Value is the characteristic that is associated with the process of extracting valuable information from large data sets [2]. For customers, the value translates into cost savings, better efficiency and improved functionality. According to CAPP et al. [4], high levels of veracity are related to the generation of profitability (value) by organizations.

The scope of the research is in a field that unites disciplines such as Big Data analytics, Business analytics, and Educational analytics, which deal with the study of Big Data and the identification of characteristics and regularities in them (Data mining). The focus is on the arrangement and analysis of data, as well as on the comparison of expectations and the obtained result in generalization. To achieve the set goal, all possible means were used – analytical approaches, statistical analysis, database management systems, software solutions, existing tools, libraries and tools.

3. Algorithm for preliminary analysis and processing of big data.

The data in a collection is not always normalized and cleaned of errors. In many cases, they are incomplete, duplicated or not relevant. Sometimes they are inaccurate, incorrect or wrongly filled (applied or imported). Other times, they are of different structures and origins, and come from different sources, such as people, devices, programs, etc. In other words, the data may be inconsistent, incoherent and irrelevant and need to be merged, reformatted and organized to find patterns and better analyze the newly obtained data sets.

Sometimes data analysis is done in real-time, other times the data is pre-fetched and processed over some time. Data is not static, it is transformed as a result of the activities and operations performed on it. They are updated and changed over time at the pace of work processes in a given educational institution and are constantly changing.

The techniques for extracting the information (data mining) from big data can be: (i) sampling, evaluation and testing of statistical hypotheses; (ii) search algorithms, modelling techniques and theories of artificial intelligence, pattern recognition and machine learning; (iii) optimization, evolutionary computation, signal processing and visualization [5]. But before we talk about data processing and extraction, it is important to do some preliminary actions to clean and normalize the initial collections that we have or that are the object of analysis.

The Algorithm of Big Data Pre-Preparation and Analysis includes the following stages/steps:

Stage 1: Collection of data from various sources, incl. from existing databases, means and tools for natural language processing (NLP);

Stage 2: Alignment and validation of existing data using the Pandas library;

Stage 3: Data modelling – regression, classification; formation of linguistic models with metadata for the objects of the considered domain area (DO);

Stage 4: Application of knowledge in NLP – joint application of known

techniques for structuring data, for converting unstructured into structured data, algorithms for working with text, images, geographic data and networks;

Stage 5: Visualization and exploratory data analysis – introduction and testing of initial hypotheses, experiments, analysis of obtained results and further action strategies.

As a result of the applied actions, the obtained data is normalized and convenient for visualization, access, processing and management, which is a step towards automation and digitization of work processes, as well as achieving better efficiency, productivity, sustainability and data coherence.

3.1. Stage 1 – Gathering data from various sources. Data analysis is based on the analysis methods based on data collection: text analytics, audio and video analytics, social media analytics, predictive analytics, etc. [2]. The places/sources can be different – text files, excel files, web services, relational SQL databases, NoSQL databases, etc.

Before it can be validated and translated into a usable form, metadata must be acquired. Before processing, data must be collected (received) and converted into an understandable format (data consolidation principles). Text analyses are applied when searching and retrieving information from the database of a given information system. Any kind of news, blogs, forums, social media, platforms and their computer processing based on linguistic, statistical or machine learning methods are included in the concept of text analysis. Further reasoning in the study is based on this data collection method.

Data can be structured or unstructured. For those that are tabular, i.e. have a pattern, we know what the columns are called and what type the data is in them. They also have lines that say what information those columns contain for each item. If they are unstructured (e.g. images, plain text, audio, web pages) and have no pattern, the data must be manipulated, we must make it so (feature extraction). But in most cases, the data follows some pattern, is classified, and can be read immediately. The libraries for reading and manipulating data in Python are Pandas and Dask.

Reading data from a local file – In Pandas, a file can be read using the `read_*`() functions, which differ in the type of their arguments. They make it easy to work with different formats, e.g. if instead of True and False, the table contains “Yes” and “No”; or the actual table starts at line 30 of the file, or there are blank lines that should be omitted, or there are no column names in the file. In general, `pd.read_table()` is the read function. All others (`read_csv()`, `read_fwf()`, etc.) just apply some settings.

Reading data from a CSV file. CSV files are normal text files that can be read and processed using the Pandas library. It has compatibility with a large number of data sources. Main reading function is `read_csv`, but besides it there are other functions like `read_table()`, `read_fwf()`, `fixwithformat`, `read_sql()`, `read_parquet()`, `read_json()`, `read_excel()` etc. `Read_csv()` has many param-

ters. And all of them can help to read different types of data. You can set the header of the table data, the type of separator (delimiter) – whether it is a comma, a semicolon and/or a space, which columns to extract (usecols), which column is an index, etc. Pandas as a library has two indexes, a column index and a row index. Therefore, the columns represent an index for Pandas.

Reading data from Excel. When an Excel file consists of a single Sheet, it does not need to be specified. When we have data, it is good to have metadata about it to understand what information is hidden in the question. The read function in this case is `read_excel()`.

Reading data from Web Services. Web services operate on the HTTP protocol and provide data in several formats. The most commonly used are JSON and XML. The read function is `pd.read_json()`, but using the `orient = "index"` parameter, the dataset can be better ordered.

Reading data from SQL files. Relational databases store their data in tables, i.e. the data has a model, and contains rows and columns. Columns are named and their data type is known. The rows tell what information these columns contain for each item. To read an SQL file, it is necessary to install and import the Pyodbc database connection library, and then connect to the database.

3.2. Stage 2 – Prepare, arrange, clean and validate the data. The algorithm for performing Experiment analysis includes the steps of identifying the sample, identifying the population from which it came, assembly methods, making sure that each group will be represented, cleaning up errors and/or omissions, determining a control group, with which the input and output data of the experiment are controlled. Cleaning, organizing, and preparing are steps that bring data into a convenient and workable form.

In [6], BREUR concludes that data with a large number of records leads to quantitative and qualitative statistical analysis, inferences and predictive results, while data with higher complexity (larger number of fields) leads to a high error rate and false discoveries.

Identification of the sample. The question to be answered when receiving and processing big data is not how certain data is taken, but why exactly it is taken, and rather why the rest is removed. They may have very high entropy, may carry a lot of information, and may have wrong values or missing values. If there are wrong ones, they should be replaced either with nothing or with some average or other appropriate value.

Tidying. Pandas rely on the data being ordered in a meaningful way and the columns each containing one variable of the same data type. The rules when working on creating, using and manipulating ordered (structured, tabular) data sets are related to their main characteristics: columns (attributes, characteristics, variables), rows (each row is one observation) and values (one observation per each characteristic). All other data is unordered.

In the ordering stage, there are still no models, no inputs, no outputs, only

rows and columns. Accordingly, the intersection of one row and one column (of one observation and one attribute) is a cell value. In the best case, it is good to have one value in one cell, the so-called scalar quantity. But that is not always the case. This means that a column in which there are several tags (collection) contains messy data.

Identifying variables. Initially, the variables that find a place in the table are defined. A table in which one variable is the row argument, the second variable is the column argument, and the third with the values in it is called a pivot table. This kind of table is very convenient for viewing data, but not too inconvenient for working with data.

An allocation of values stored in a variable. A very important characteristic of sets is that each element is unique. Duplicates in rows are not allowed, which is part of the ordering. Observation must be unique and not repeated. If there are variables that are different in meaning, the values of which are combined in one cell, then they must be separated and classified. The process of arranging data in a standard and meaningful way is known as normalization (third normal form).

Cleaning. The data cleaning stage includes activities such as validation, transformation, error cleaning, and feature identification. It is possible to reformat, merge or split data into more than one column (for clarity). Checks are performed for duplicate row or column names, missing (not available NA) or broken data. Data can be combined, grouped and/or sorted. Validation is checking that the data meets some pre-set or generally accepted requirements. For example, an email address can be validated when we send a return email to the user and ask for confirmation that it exists.

3.3. Stage 3 – Data modelling and classification. Visualization. You cannot have a good model without good data processing for it. Modelling is a process that requires an ordered and clean set of data. This stage includes actions such as merging, linking and comparing the data, to better visualize or prepare for post-processing. Linguistic and conceptual models for processes and phenomena, as well as hypotheses for the synthesis of requests to the database, including experiments, comparison and control of the way of work, are characteristic of the field of Higher Education.

The linguistic model, part of the process of extracting information from big data, represents an interrelated set of concepts that have a place in the database and their synonymous representations in natural language that form user queries. It is an essential element in the natural language process and in extracting information from information systems with large databases. The modelling of a domain area and the creation of a linguistic model is discussed in detail in previous articles by the authors.

3.4. Stage 4 – Knowledge application and database query synthesis. In stage 4, familiar techniques such as structuring unstructured (non-tabular) data, algorithms for working with text, images (base information), geographic

data and networks (graphs, data structures that explain related components) are applied. This cannot be achieved without concepts of quality code, knowledge of structuring data, testing, debugging, and organizing studies so that they are repeatable (reproducible research) and argumentative (evidence-based research).

The process of NLP and extracting information from big data begins with freely asking a natural language user question. One should extract as much as possible about the context, the explicit and implicit information hidden in the question, the type of outputs, etc. Research is done, all metadata about the language objects from the question is extracted and a conclusion is made about the type/model/pattern of the given text question. Part of the scientific process is creating a hypothesis (null or alternative). This is not the question Why? or How? This is a conjecture that can be answered with a specific DB query pattern and question that can be scientifically verified or rejected depending specifically on the parameters in the input text.

3.5. Stage 5 – Exploratory analysis – hypotheses, experiments, analysis of results and follow-up. Predictive analytics takes into account current, available or existing old data. It is based on statistical methods and finds patterns in the data as well as relationships between them.

A central place in the proposed method for NLP of Big Data is hypothesis testing. These are experiments that generate data. Apart from them, the methods (mathematical and logical), the ideas, the ways of collecting and storing the data (data warehousing), the verification and arrangement of the data, the quality of the data (data engineering), how they are analyzed (analytics), visualized and presented, are important in order to decide how to use them together to get responses from the system.

After the analysis and formulation of a hypothesis, a database query can be generated until a result is obtained from the system. The obtained result is again analyzed if there is a way to even compare it with another type of search with the same parameters. And so it is checked whether the hypothesis is true or not, and only then after an experiment has passed, can the final results be analyzed or the models updated or improved if necessary. Then the results either confirm the hypothesis (True) or do not confirm it (False). But in both cases, the results should be explained to take measures to eliminate the problems.

4. Conclusion. In conclusion, we can say that the process of cleaning, ordering, validating, modelling and extracting useful information from big educational data follows the scientific method/approach and is completely logically based on the business logic and processes that take place in the considered domain area for taking decisions and their justification.

The process is very highly iterative. Each step in data preparation can lead to another. It can reverse the process, to change the strategy to something different, or forward, depending on the situation. As the data analysis progresses, it is understood what problem there is in the data (if any). Modelling brings infor-

mation needed for the cleaning stage. Because of some models, the data needs to be changed and presented differently. This also has an impact on the hypotheses that are made, and this can lead to filling in with new data (if necessary). Therefore, the proposed process/algorithm model is not so much a sequence of stages as a connectedness of individual stages. Preprocessing is essential for analyzing multivariate data sets before actually extracting information from them.

The proposed stages represent only the first step of research related to the intelligent analysis and processing of heterogeneous data to extract answers to freely asked user questions from Big Data. In subsequent papers, verified data models and techniques for automatically recognizing these models without knowing the data themselves will be proposed and developed.

REFERENCES

- [¹] AMANA M., P. LIU, M. ALARIQI (2022) Value Creation and Capture with Big Data in Smart Phones Companies, *Sustainability*, **14**(23), 15882, <https://doi.org/10.3390/su142315882>.
- [²] KOSELEVA N., G. ROPAITE (2017) Big Data in Building Energy Efficiency: Understanding of Big Data and Main Challenges, *Procedia Engineering*, **172**, 544–549, <https://doi.org/10.1016/j.proeng.2017.02.064>.
- [³] AROCKIA S., S. VARNEKHA, A. VENESHIA (2017) The 17 V's of Big Data, *International Research Journal of Engineering and Technology (IRJET)*, **4**(9), 329–333.
- [⁴] CAPPÀ F., R. ORIANI, E. PERUFFO, I. MCCARTHY (2021) Big Data for Creating and Capturing Value in the Digitalized Environment: Unpacking the Effects of Volume, Variety, and Veracity on Firm Performance, *J. Prod. Innov. Manag.*, **38**(1), 49–67, <https://doi.org/10.1111/jpim.12545>.
- [⁵] DONEVA R., S. GAFTANDZHIEVA, G. PASHEV (2022) Creating methods and models for determining typical schemas of structured and unstructured data, *Journal of Informatics and Innovative Technologies (JIIT)*, **5**(4), 13–28, (in Bulgarian), ISSN 2683-0930, Plovdiv, 2022, <https://journal.iiit.bg/wp-content/uploads/2023/04/JIIT-br-5-2022-min.pdf>.
- [⁶] BREUR T. (2016) Statistical Power Analysis and the contemporary “crisis” in social sciences, *Journal of Marketing Analytics*, Palgrave Macmillan, **4**(2), 61–65, doi: 10.1057/s41270-016-0001-3.

*Department of Computer Systems and Technologies
University of Food Technology
26 Maritsa Blvd
4002 Plovdiv, Bulgaria
e-mail: m_jekova@uft-plovdiv.bg*