State of Art of Data Mining and Learning Analytics Tools in Higher Education

https://doi.org/10.3991/ijet.v15i21.16435

Mohammed Salihoun
Ecole Marocaine des Sciences de l'Ingenieur, Casablanca, Morocco
salihoun.med@gmail.com

Abstract—In this decade, the use of learning management systems (LMS) does not cease to increase, becoming one of the most popular approaches adopted and widely used in the learning process. Learners' online activities generate a huge amount of unused data that is wasted because traditional learning analyses are not able to process them. In this regard, a large collection of applications/tools have emerged to conduct research in educational data mining (EDM) and / or learning analysis (LA). This study looks into the recent applications/tools of Big Data technologies in education and presents some of the most widely used, accessible, and powerful tools in this field of research. The majority of these tools are for researchers with the purpose of conducting research on educational data mining and learning analysis.

Keywords—Learning Analytics, LMS, Big Data, Educational Data Mining, Text Mining, Modeling

1 Introduction

In recent years, community learning environments have multiplied and have provided new directions for educational improvements to educational research. Recent learning methods like Flipped Classroom [1] largely depend on learners' online activities like discussion forums, online chats, instant messaging, etc… through a LMS (Moodle, Claroline, Google Classroom, etc.). Usually, huge amount of data is created and generated by learners' activities through a LMS. These data can be used in developing the learning environment helping the learners in learning and improving the overall learning experience. In this way, several frameworks and models [2] have been proposed for online learning management systems to improve the learning experience. But data created from learner's activities in educational institutions is so enormous [3] and traditional processing techniques cannot be used to process them. The limitations of traditional data processing applications have enabled communities of educational data mining (EDM) and learning analysis (LA) to become an alternative to basic approaches to working with educational data [4] [5].

2 Big Data in Learning

2.1 Big data

The term "Big Data" refers to any set of data [6] that is too large and moves too fast, becoming too difficult and too complex to be dealt by traditional data-processing application software. Indeed, the processing of data produced by learning environments has become a real challenge, which has made it necessary to use Big Data technologies and tools to handle them.

Generally, Big Data has come to be identified by several of fundamental characteristics. Key among them are [7]:

Table 1. Properties of Big Data

Volume	**Size of data plays a very crucial role in determining value out of data.**
Velocity	How fast the data is generated and processed to meet the demands, determines real potential in the data.
Variability	The inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively
Variety	Data in diverse format both structured and unstructured.

2.2 Big data in higher education

Big Data techniques and tools can be used in a variety of ways in learning analytics as listed in table 2 [8]:

Table 2. Different axes of applying Big Data in Learning Analytics

Performance Prediction	**By analyzing learner's interaction in a learning environment with other learners and tutors.**
Attrition Risk Detection	By analyzing the learner's behavior using measures implemented in the beginning of the learning process to detect the risk of students dropping out.
Data Visualization	By using visualization techniques to easily identify the relations and trends in the data just by looking on the visual reports.
Intelligent feedback	By providing through LMS intelligent and immediate feedback to students in response to their inputs in order to improve their interactions and performances.
Course Recommendation	By recommending best courses to students based on their interest and activities. That will ensure that students are not misguided in choosing fields.
Student skill estimation	By analysing interaction of the learner with the system or in the message boards or discussion forums.
Behavior Detection	By analyzing the learner's behavior in community-based activities or games which help in developing a student model.

3 General Summary of the Existing EDM / LA Tools

The main objective of this article is trying to discuss the most commonly used, most available, and most powerful tools accessible for the researcher interested in conducting Educational Data Mining and Learning Analytics research. This overview

will allow us to draw a guideline that could be used afterwards to perform an analysis of one of these tools or to explore a research question. Among the major challenges in the fields of EDM is the transformation of raw and incomplete data streams into significant variables. This transformation remains a complex process, since data usually comes in different forms that are not ready for analysis; the data need to be cleaned in order to remove cases and values that are actively incorrect, in this way data can be converted into a more significant format [9].

As a first step, we will start by listing some tools well suited for the manipulation, cleaning, and formatting of data as well as for feature engineering and data creation. We will also discuss the role of programming and querying languages in this task of data manipulation and formatting. In the second part, once data have been cleaned, transformed into significant format and structured appropriately for analysis, the problem for an EDM or LA researcher is how to analyse these data, what models can be constructed and what relationships can be mapped and explored from this manner. Several tools and packages are well suited for testing, analysis, and modelling data will be discussed later. In the last part of our discussion, after the analysis has been conducted and the model has been validated by researchers, the issue is to visualize information in the significant way. We will debate some tools and packages that allows data scientists the capacity to create informative graphs, diagrams models, networks, charts, and other manners of visualized information.

In the next section, we will discuss in details tools that are relevant to these types of specialized data, especially those frequently used in EDM due to their relevance and their popularity to researchers and practitioners.

3.1 Data manipulation and feature engineering

The process of data mining can begin, when datasets have been cleaned and prepared from their raw state. This is a recurring problem, and more complex when data miners have to work with log data or learning management system (LMS) data recorded in forms that are not directly amenable to analysis. Educational data is generally known by their messy, sometimes incomplete or some parts have to be merged; and usually in different and unusual formats. For example, if a tutor is interested to identify off-task learners [10] [11], a part of the information can be found in the logs file of the system as a raw time stamp. In this case, Baker [12] & Veeramachaneni et al. [13] recommend the feature engineering process to create new variables in order to conduct the desired analyses. In follows, we present tools that can be used for cleaning, organizing, and creating data. We will discuss for each tool, their advantages and their utility for restructuring large data sets and creating and managing new and more useful variables from existing variables.

EDM workbench: It is a tool with the aim to address the limitations of Excel and Google Sheets about the specific tasks such as the generation of complex sequential features and data labelling [14]. EDM Workbench allows the user to define the set of features by which the data should be grouped into subsets of learner-tutor transactions (referred to as ‘‘clips’’). Creating features in the EDM Workbench is based on XML and the extraction of several features used in existing literature and intelligent tutoring

systems (ITS). We can mention some features such as (the time the learner spent on the problem, the number, and the proportion of correct, wrong, or help actions for the current skill for the last n steps, for the skill, or for the learner, etc). In addition, it allows data labelling by creating text replays and printed sections of human behaviour [15]. These latter are coded by researchers or other domain experts in terms of categories of behaviour or other labels of interest. Finally, EDM Workbench supports sampling, reliability checking, synchronization and organization between features extracted and labels.

Python and Jupyter notebook: Several programming languages allows the manipulation of data and engineering of features. One of these programming languages is Python which is considered the most suited for these purposes, especially in engineering context dependent or temporal features compared to Excel and Sheets. Jupyter Notebook [16] is a web-based interactive computational environment having a useful feature that allows creating and sharing document including data cleaning and transformation, numerical simulation, data visualization, statistical modeling, machine learning, etc. Nonetheless, visual inspection of data and features created in Excel or Google Sheets is easier than in Jupyter. Among the difficulties encountered by data scientists is the high time generated to identify missing data, duplicates or unusual values such as JSON files (JavaScript Object Notation) produced by several online learning platforms and Massively Open Online Course (MOOC). Such files unusual data formats can be handled with Python. Even if, Python is more powerful in accommodating larger data and those involving nested loops than the spreadsheet tools covered above, but it faces data size limitations and becomes slower during processing data.

Structured Query Language (SQL): SQL is a language used in programming and designed to manage data in some (but not all) databases. It is mostly useful for extracting exactly the desired data, sometimes integrating (joining) across multiple database tables. SQL, Hadoop [17] or Spark [18] are database languages that allows significantly fast processing for basic tasks such as (selecting a specific subset of learners or obtaining data from a specific date range) than any of the tools aforementioned. In addition, SQL can work effectively in combination with Excel & Python in sorting and filtering task.

3.2 Algorithmic analysis

This step consists to analyse and model datasets and validate the resulting models when features have been engineered, then outcome variables have been labelled, and finally data have been sampled and appropriately structured. A big collection of modelling frameworks and algorithms will be detailed in the following section. All these tools are used to model and predict processes and relationships in pedagogical data.

RapidMiner: RapidMiner is a data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics with an extensive set of classification and regression algorithms as well as algorithms for clustering, association rule mining, and other applications [19]. Among the various data mining tools that exist, RapidMiner's

graphical programming language is the more powerful of them, since it allows, for example to conduct cross validation at multiple levels (such as learner-level/course-level cross validation) using the BatchCrossValidation operator, which becomes an advantage over the other graphical languages in data mining packages. In order to help the user to evaluate the goodness of a model, a large range of metrics is available by RapidMiner in this sense Models here are generated from mathematical models based on RapidMiner code or XML files [20]. Integration of a programming application program interface (API) into RapidMiner's graphical programming language allows the possibility to achieve more and different tasks. Moreover, it integrates all of the algorithms available in WEKA, detailed below. RapidMiner is a free open source software and is available for free for academic use, also it's a commercial product, where licenses are available through the publisher Rapid-I. A wide set of tutorials is available in the official website in learning how to use the graphical programming language.

WEKA: WEKA is a free and an open source software including a set of algorithms related to machine learning. It offers tools for data mining tasks such as regression, classification, association rules mining, clustering, and visualization [21]. Data mining algorithms can be invoked by users through a graphical user interface (GUI), a command line, or by invoking algorithms from a Java API. GUI does not give users access to all advanced functions than the command line interface and APIs. The integration of PMML (Predictive Modeling Markup Language) files support into the Weka scoring plugin and a new PMML classifier scoring plugin for the Weka KnowledgeFlow have been completed. From Weka 3.6.0, PMML models can be run from the Classify panel in Weka's Explorer user interface and from the command line. Learning to use WEKA is supported by a book by Witten, Frank, Pal and Hall [22], now in its fourth edition. The WEKA website also hosts an active mailing list, tutorials, wikis, and bug reports.

SPSS: SPSS is a widely used program for statistical analysis in social science including a large packages of regression frameworks, statistical tests, factor analyses and correlations. SPSS Modeler was created by IBM in order to build predictive models and conduct other new analytic tasks [23] [24]. From the outset, one of its main goals was to eliminate the unnecessary complexity of data transformations and to make complex predictive models very easy to use. In addition, a functionality had been added for using the target class in feature selection, which is not available in many other packages. Even if, SPSS is considered such as a complete statistical analysis tool, but it faces modelling limitations compared to others tools in this section. SPSS remains less flexible in terms of customization and also not documented as well. SPSS is available commercially at the official website of IBM.

KNIME: KNIME is a free and open-source data analysis and reporting platform generally similar to WEKA and RapidMiner [25] [26]. It integrates many components for data mining and incorporates all of WEKA's algorithms. In addition, KNIME offers several algorithms in different areas such as Social network analysis (SNA) and sentiment analysis. One of the biggest advantages of KNIME is its capacity to incorporate data from multiple sources (e.g., a database of learners, a word document of text responses, and a csv file of engineered features, etc.). Finally, several extensions

of KNIME allows the interfacing with programming language such as (Python, R, Sql & Java).

Orange: Orange is an open-source data visualization, machine learning and data mining toolkit [27]. It contains fewer algorithms compared to the other tools mentioned before, but offers many commonly used algorithms, such as random forests, kNN and naive Bayes. Orange remains much easier for understanding the interface by using color-coded widgets to make simple the difference between data input and cleaning, visualization, regression, and clustering. Also, Orange offers the possibility to customize visualization modules for the presentation of model results in the best way. Compared to the other tools cited in this section, Orange is limited in the scale of data and may be better suited as a tool for smaller research projects.

KEEL: KEEL (Knowledge Extraction based on Evolutionary Learning) is an open source Java software tool that can be used by EDM researchers for a large number of different knowledge data discovery tasks [28]. KEEL provides a simple GUI based on data flow to design experiments with different datasets and computational intelligence algorithms (paying special attention to evolutionary algorithms) in order to assess the behaviour of the algorithms [29]. It contains a wide variety of classical knowledge extraction algorithms, pre-processing techniques (training set selection, feature selection, discretization, imputation methods for missing values, among others), computational intelligence-based learning algorithms, hybrid models, statistical methodologies for contrasting experiments and so forth. It allows to perform a complete analysis of new computational intelligence proposals in comparison to existing ones. KEEL has relatively less support for new users than most other data mining packages, though there are help features and a user manual. KEEL is open source and free for use under a GNU license. Moreover, KEEL has been designed with a two-fold goal: research and educational.

Spark MLLib: Apache Spark is a framework for wide-scale processing of data across multiple computer processors, in a distributed fashion [30]. Spark can connect with several programming languages, including Java, Python, and SQL, through an API, allowing these languages to be used for distributed processing [31]. Even if MLLib's functionality is still somewhat limited, and it is essentially a programmatic tool (reducing its usability to nonprogrammers), its distributed nature makes it an efficient and rapid choice.

R Analytical tool to learn easily: Rattle is a popular and free open source GUI-based data mining tool using Gnome graphical interface from Togaware [32]. Rattle supports unsupervised and supervised data mining and machine learning models. It allows the dataset to be partitioned into training, validation and testing. Also, data in Rattle can be summarized visually.

3.3 Visualizations

After the data extraction and analysis phases, the visualization phase comes to support both analysts and practitioners in deriving meaning from data [33] [34] [35] [36]. We will introduce in this section, some general tools that may have relevant implications of data analysis, especially tools and methods that allows visual analytics. These

tools enable building interactive visual interfaces in order to gain knowledge and insight from data as well as communicating important implications for learning to learners and tutors.

Tableau: Many products for interactive data analysis and visualization are offered by Tableau [37]. These products have been widely applied in learning environments to analyse learner's data, provide actionable information, improve pedagogical reports and tutoring practices. Using Tableau requires no programming knowledge to analyse enormous amounts of data from various sources [38]. This advantage makes it easy to have a range of visualizations for a larger community. The visualizations are displayed to users on a are dynamic real-time way and based on rich and interactive dashboards. Another advantage for this tool is its ability to import data from different standardized formats for data storing (e.g., data warehouses, databases, log files, etc…). One of the limitations of Tableau is its incapacity to support relational data mining or predictive analytics, furthermore, it's a commercial tool, which does not allow extensions and integration of other software platforms.

D3.js: D3.js is a JavaScript library for manipulating data driven by documents, producing dynamic, it helps researchers to build interactive data visualizations in modern web browsers [39]. Even though, D3.js is free and open source, does not require installation, supports code reuse, and has the ability to build wide range of kinds of data visualization, but it's always a big challenge to adopt it for educational research purposes. D3.js is facing by the problem of compatibility with some browsers (e.g., Internet Explorer) as well as some performance limitations for larger datasets [40]. In addition, to guarantee privacy and data security, the data pre-processing is required in order to hide data from users of visualizations. In parallel, several data visualization tools exist to provide different ways for presenting data visually and building interactive dashboards. We can cite some tools such as JavaScript InfoVis Tool kit, Raw, jpGraph, Chart.js and Google Visualization API. We didn’t detail these tools in this section because they have been less frequently used by EDM and LA researchers compared to D3.js.

3.4 Specialized EDM & LA applications

Until now, we have listed and discussed general-purpose tools/applications for EDM modeling and analysis. Sometimes, analysis goals and types of data require more specialized algorithms that are not offered by these general-purpose tools. In this case, we will cover some of the most popular tools widely used by practitioners and researchers that accomplish these aims.

SNA: Social network analysis (SNA) is a sociological approach, based on the study of network theory applied to social networks. It builds and designs social relations with nodes and links. Nodes are usually the social actors or institutions, and the links are the relationships and the interactions between these nodes. SNA is commonly used with the aim to analyze collaborative social networks (e.g., learner interaction within MOOCs, in social media, or in online courses within LMS).

Gephi: It is popular and interactive software writing in Java for graph and network analysis and visualization in real-time [41]. It provides easy and broad access to net-

work data and allows for specialization, filtration, navigation, manipulation and clustering. Gephi offers also a Java API to manipulate social network graphs, calculate several measures (e.g., average path, density, and degree of centrality), and to execute algorithms (e.g., graph clustering and giant connected component extraction) usually used in SNA. Gephi is available on several operating systems, and commonly applied in LA research. Although it is used under GPL license.

EgoNet: It is a free SNA program that helps the user to create survey, to collect and analyses all the egocentric network data (all social network data of a website on the Internet), and offer data matrixes and general network measures in order to be used for other analysis by other programs [42]. Furthermore, since members provide information about network structure from their perspective (hence "ego" in the name), a collection of analysis tools are offered by EgoNet to better understand the overall network structure, with the possibility to interrogate a member of the network with further questions.

NodeXL: It is a free package for Microsoft Excel that facilitates the exploration of network graphs from a large variety of input data formats [43]. The basic version of NodeXL provides a set of tools for filtering and visualizing the data, NodeXL Pro propose additional features that extend NodeXL Basic, allowing an easy access to the various social media network data streams (e.g., Twitter, YouTube, and Flickr), offering advanced network metrics, and powerful report generation.

Pajek: It is a free desktop tool for complex analysis of a large variety of huge networks including the analysis of networks of social interactions [44]. It is widely used in LA research and academia for SNA. Pajek allows the network partitioning, information flow analysis and community detection. Recent version of Pajek is designed called Pajek-XXL that is able to work with enormously huge networks networks (with millions of nodes and more). Pajek is available on Windows, Mac OS & Unix.

Sonia: It is a social network an open-source platform with an image animator specialized for longitudinal analysis of networks [45]. It has the ability to retrieve the information about the time relationships occurred or at least the order in which relationships developed by members. This allows a better visualization of network changes over time. The final result is an animated image of structural changes over time, which can be exported into QuickTime video format. Sonia project is based on Java programming language, developed by Stanford University and can be used in all major operating systems.

NetworkX: It is a Python package that allows to create, manipulate, and analyze complex network processes, structures, and dynamics [46]. NetworkX is widely used in academic research and offers a wide range of advanced functionalities to manipulate networked data, such as graph clustering, graph reduction, community detection, network triads analysis, link prediction (finding missing links, e.g., missing Facebook connection among two friends), and others.

The social networks adapting pedagogical practice: It is a software tool that performs real-time social network analysis and visualization of discussion forum activity within a Learning Management System (e.g., Desire2Learn, Blackboard, and Moodle) [47]. Data formed through learner's posting and replying interactions from HTML pages of LMS discussions can be exported for further analysis or visualized within

SNAPP. Visualization and analysis can be realized with the use of several various graph layout algorithms. In addition, SNAPP has the ability to explore the evolution of learner social networks, to identify structural holes, to analyze the highly active/inactive users and to compare analysis of several discussion forums.

R packages: Network, sna, igraph, statnet and ergm: The network package [48] has the ability to build and modify network objects, extract network metrics, and visualize network graphs. There are several packages for SNA [49] in the R programming language, that are used with the network package to get some functionalities commonly needed for SNA, such as network regression, graph generation networks, calculation of node and network metrics, and others. Another package written in the C programming language with additional language bindings for the R and Python programming languages called igraph [50] is usually used for SNA. It can be used for builidng and modifying social networks from a large variety of input formats (e.g. Gephi, Pajek, GraphML), visualizing graphs, calculating of node and network properties. In addition, it supports different network analysis such as graph clustering, block modeling, community detection, and others. Another package for SNA is statnet package [51] which is a collection of packages for network analysis with the latest improvement in the statistical modeling of networks. Statnet provides a set tools for the visualization, representation, simulation and analysis of various forms of network data. This extensive functionality is supplied by a central Markov chain Monte Carlo algorithm (MCMC). Finally, the ergm package [52] can provide the same functionalities of stanet and can also be used for statistical modeling of social networks using Exponential Random Graph Models (ERGMs).

Cytoscape: It is an open source tool developed on the Java platform allowing the visualization of complex networks and the integration of these latter with any type of attribute data [53]. It can be applied in various types of problem domains (e.g., bioinformatics, semantic web and sna). A basic set of features is offered by the Cytoscape core distribution allowing data analysis, visualization, and integration, which is then extended by using several user-supplied modules (formerly called plugins). Cytoscape can be applied within several operating systems.

Text mining: Text mining is a speedily expanding field of text mining, which consists of deriving high quality information from text. Various applications and APIs are available for the tagging, processing, and identification of textual data. Text analysis tools can process text parts of speech, sentence structure, and semantic word meaning. Also, some tools have the capacity to detect representational relationships between different sentences and words. We will cite below a set of popular tools allowing the treatment and the textual analysis.

Linguistic Inquiry & Word: It is a graphical and easy-to-use computerized text analysis that calculates the degree to which various categories of words are used in a text [54]. LIWC organizes the words into dozens of linguistic and psychological groups that tap social, cognitive, and affective processes. LIWC tool has been broadly used and validated in several and various empirical studies.

Coh-Metrix: It is a system for calculating cohesion and coherence metrics using indices of linguistic and discursive representations of a text [55] [56]. Coh-Metrix calculates the coherence of texts on more than 100 measures of text divided into 11

categories. With his multiple tags, Coh-Metrix allows assessing deep text cohesion, such as measures of referential cohesion or narrativity. Coh-Metrix remains better for the analysis of text features and relationships in the data.

ConceptNet: It is a free semantic network, based on a multilingual knowledge with the aim to help computers understand the meanings of words used by people [57]. The main objective of ConceptNet is to develop an enormously large graph of "commonsense" knowledge (e.g., "keyboard is a computer hardware"), which can be then used for understanding and processing natural text. With a wide knowledge base, ConceptNet has the ability to categorize textual documents according to corpora topics, to analyze sentiments (e.g., detecting emotions in the text), and to summarize text among other uses.

AlchemyAPI: It is an IBM-owned tool based on machine learning (more precisely, deep learning) to do natural language processing (in particular, semantic text analysis, including sentiment analysis) [58]. It affords It allows content processing in the form of standard text documents or web resources (i.e., accessible through URL) and supports response formats such as JSON, XML, and RDF. AlchemyAPI remains a commercial platform, and gets paid per API call, but it provides free access for up to 1,000 calls per day. According to the studies of AlchemyAPI's performance [59] [60] [61], it turned out that the best results are obtained when the tool is used for the semantic analysis of long articles, such as research articles or blogs.

TAGME: It is a powerful tool that have the ability to identify in unstructured text a significant short-phrases (called "spots") and to link them to a pertinent Wikipedia page in a fast and effective way [62]. That is, TAGME assigns (if possible) a Wikipedia concept to each of the term sequences in the analyzed text. According to the studies of TAGME's performance and compared to other solutions [63], it turned out that the best results are obtained on short text segments and a comparable precision/recall results on longer text. TAGME offers an API to be integrated with other applications.

Apache stanbol: It is an open source tool and reusable collection of components allowing semantic content management with the goal to bring semantic technologies into existing CMS and for text mining [64]. Apache Stanbol remains an easy tool to set up and run it on small set of instances allowing the possibility to incorporate a domain-specific ontology in the annotation process. This is extremely useful when working with locally defined concepts specific to a given educational context. In addition, Apache Stanbol supports and integration with different CMS, and the text annotation in various languages.

Natural language processing (NLP) tool kits (Apache OpenNLP, Python analysis and Stanford CoreNLP): These represent an important part of the text mining toolset and they are naturally used in the preprocessing stage of the analysis, by (i) splitting paragraphs into individual sentences, or words; (ii) extracting syntactic dependences between words; (iii) assigning categories to each word; (iv) reducing derived words to their root word; (e) extracting named-entity, (i.e., names of people, places, institutions, monetary amounts, dates); and (f) resolving coreference (resolution of pronouns to their target nouns). Among the most popular NLP tools, we can cite Apache OpenNLP tool kit [65], a Java-based NLP tool kit that supports most of the common NLP tasks listed above. Also, Python NLTK [66] which is an NLP library for Python

programming language with very same abilities. Another useful toolkit is Stanford CoreNLP [67] which aside from providing a Java API, also provides a stand-alone command line interface and a set of "wrappers" for other programming languages (e.g., Python, R, C#, Ruby, JavaScript and Scala).

LightSIDE: It is an open source platform based on the WEKA toolkit to support text-mining. It allows creating a set of features usually used for educational text, especially, creating variables for individual words, punctuation, line length, bigrams (similair and adjacent words), and word stemming [68]. LightSIDE offers also a simplified interface for error analysis that can help researchers to iteratively improve their text mining solution.

Process and sequence mining: In addition to more classical methods to educational data analysis, such as course persistence or predicting learning outcome, researchers also have the objective to track sequences of learner activities to understand learning approaches and processes [69] [70]. In this section, we will present two tools for this type of application which are commonly used to support EDM and LA research. They are generally used to perform analysis, but they also allow for some level of data preprocessing.

ProM: It is an independent framework developed with Java that supports a wide range of process mining techniques [71]. It supports running process mining in a distributed setting or through batch processing, and providing a clear specification of expected inputs and outputs for each of the supported implementations. Furthermore, new plugins can be added at run-time, to be integrated simply into the analysis process. Another advantage for ProM is the integration with existing information systems does not require programming task. The most current version is ProM 6.9.

TraMineR: It is a R-package that allows mining, description and visualization of states or events sequences data [72]. Some of the primary features of TraMineR for the analysis and visualization of state sequence data include (i) handling of longitudinal data and conversion between several sequence formats, (ii) plotting sequences (frequency plot, density plot, and more), (iii) sequence transversal characteristics by age point (transversal state distribution, transversal entropy), (vi) individual longitudinal characteristics of sequences (length, time in each state, longitudinal entropy, complexity and more).

PSLC Datashop: It is a free multifunctional web application which offers a secure place to store & access research data and it supports various kinds of research [73]. DataShop has functionalities allowing a focus on learner-tutor interaction data with a learning curves & error reports provide summary and low/high level views of learner performance (e.g. hint use, latent knowledge, response times, and other variables of interest). Additionally, it offers performance profiler aggregates across various levels of granularity (problem, dataset levels, knowledge components, etc.).

4 Conclusion

EDM and LA is speedily changing area, and several tools are emerging constantly. We have reviewed in this article, about 40 tools commonly used for datamining and

analytics in the area of education. The main objective of this state of art is to give help to researchers interested in learning about these emerging methods in terms of theoretical and practical application and use.

According to this state of art, it has been found that no one tool is ideally suited to conducting the entire process of analyzing most data sets from start to finish. For this, researchers and practitioners using EDM and LA have the obligation to use different tools that are suited to different tasks. For instance, data generated by a popular MOOC can easily reach more than 50 system transactions. A researcher may use (SQL) to select only data of a particular semester, then use (Excel) to refine this dataset in order to calculate total learner time in the system. After he will use (RapidMiner) to fit a predictive model and (NodeXL) to analyze the relationship between forum posts and replies. Then, he will use (CohMetrix) to get an overall textual quality of posts and replies by that learner. At the end, this researcher may use (Gephi) to visualize the most interesting clusters of learner found within the social network data.

We have presented in this paper, a collection of tools that researchers in the fields of EDM and LA currently use, and are represented in aggregate across the different groups of scientists working in this field. As we mentioned above, each tool represents different approaches to different problems, with their own particular strengths and weaknesses. The combination of these tools can be a useful discovery and a best way to perform complex analyses.

5 Acknowledgement

The authors would like to thank SMARTiLAB / EMSI the High School of engineering group.

6 References

[1] Tucker, B. (2012). The flipped classroom. Education next, 12(1), 82-83.

[2] Raghunathan, S., & Kaur, A. (2011). Assessment of online interaction pattern using the Q-4R framework. In the International Lifelong Learning Conference.

[3] Ray, S. (2013). Big data in education. Gravity, the Great Lakes Magazine, 8-10.

[4] Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In Learning analytics (pp. 61-75). Springer, New York, NY. https://doi.org/10.1007/978-1-4614-3305-7_4

[5] Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert systems with applications, 33(1), 135-146. https://doi.org/10.1016/j.eswa.2006.04.005

[6] Wikipedia, "Big data --- Wikipedia, The Free =Encyclopedia", https://en.wikipedia.org/wiki/Big_data. Accessed 2020.

[7] Daniel, B. (2015). Big Data and analytics in higher education: Opportunities and challenges. British journal of educational technology, 46(5), 904-920. https://doi.org/10.1111/bjet.12230

[8] Sin, K., & Muthu, L. (2015). Application of Big Data in Education Data Mining and Learning Analytics--A Literature Review. ICTACT Journal on Soft Computing, 5(4). https://doi.org/10.21917/ijsc.2015.0145
[9] Huda, M., Maseleno, A., Mat Teh, K., Don, A., Basiron, B., Jasmi, K., Mustari, M., Nasir, B., & Ahmad, R. (2018). Understanding Modern Learning Environment (MLE) in Big Data Era. International Journal of Emerging Technologies In Learning (IJET), 13(05), pp. 71-85. https://doi.org/10.3991/ijet.v13i05.8042
[10] Baker, R. S., Corbett, A. T., & Wagner, A. Z. (2006). Human classification of low-fidelity replays of student actions. In Proceedings of the educational data mining workshop at the 8th international conference on intelligent tutoring systems (Vol. 2002, pp. 29-36).
[11] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., . & Ivanov, P. (2016, May). Jupyter Notebooks-a publishing format for reproducible computational workflows. In ELPUB (pp. 87-90). [15']
[12] Hadoop. Hadoop Homepage. http://hadoop.apache.org/.
[13] Spark. Spark Homepage. http://spark-project.org/.
[14] Rodrigo, M., Mercedes, T., d Baker, R. S., McLaren, B. M., Jayme, A., & Dy, T. T. (2012). Development of a Workbench to Address the Educational Data Mining Bottleneck. International Educational Data Mining Society.
[15] Baker, R. S., Corbett, A. T., & Wagner, A. Z. (2006). Human classification of low-fidelity replays of student actions. In Proceedings of the educational data mining workshop at the 8th international conference on intelligent tutoring systems (Vol. 2002, pp. 29-36).
[16] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., . & Ivanov, P. (2016, May). Jupyter Notebooks-a publishing format for reproducible computational workflows. In ELPUB (pp. 87-90). [15']
[17] Hadoop. Hadoop Homepage. http://hadoop.apache.org/.
[18] Spark. Spark Homepage. http://spark-project.org/.
[19] Dwivedi, S., Kasliwal, P., & Soni, S. (2016, March). Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime). In 2016 Symposium on Colossal Data Analysis and Networking (CDAN) (pp. 1-8). IEEE. https://doi.org/10.1109/cdan.2016.7570894
[20] Naik, A., & Samant, L. (2016). Correlation review of classification algorithm using data mining tool: WEKA, Rapidminer, Tanagra, Orange and Knime. Procedia Computer Science, 85, 662-668. https://doi.org/10.1016/j.procs.2016.05.251
[21] Singhal, S., & Jena, M. (2013). A study on WEKA tool for data preprocessing, classification and clustering. International Journal of Innovative technology and exploring engineering (IJItee), 2(6), 250-253.
[22] Weka. DATA MINING, Practical Machine Learning Tools and Techniques. https://www.cs.waikato.ac.nz/ml/weka/book.html
[23] McCormick, K., Abbott, D., Brown, M. S., Khabaza, T., & Mutchler, S. R. (2013). IBM SPSS modeler cookbook. Packt Publishing.
[24] Wendler, T., & Gröttrup, S. (2016). Data mining with SPSS modeler: theory, exercises and solutions. Springer. https://doi.org/10.1007/978-3-319-28709-6_1
[25] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., & Wiswedel, B. (2009). KNIME-the Konstanz information miner: version 2.0 and beyond. AcM SIGKDD explorations Newsletter, 11(1), 26-31. https://doi.org/10.1145/1656274.1656280
[26] Kataria, L. (2013). Implementation of Knime-Data Mining Tool. International Journal of Advanced Research in Computer Science and Software Engineering, 3(11).
[27] Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., & Štajdohar, M. (2013). Orange: data mining toolbox in Python. The Journal of Machine Learning Research, 14(1), 2349-2353.

[28] Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., & Herrera, F. (2011). Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic & Soft Computing, 17. https://doi.org/10.1109/nwesp.2011.6088224

[29] Alcalá-Fdez, J., Sanchez, L., Garcia, S., del Jesus, M. J., Ventura, S., Garrell, J. M., . & Fernández, J. C. (2009). KEEL: a software tool to assess evolutionary algorithms for data mining problems. Soft Computing, 13(3), 307-318. https://doi.org/10.1007/s00500-008-0323-y

[30] Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., . & Xin, D. (2016). Mllib: Machine learning in apache spark. The Journal of Machine Learning Research, 17(1), 1235-1241.

[31] Assefi, M., Behravesh, E., Liu, G., & Tafti, A. P. (2017, December). Big data machine learning using apache spark MLlib. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 3492-3498). IEEE. https://doi.org/10.1109/bigdata.2017.8258338

[32] Williams, G. J. (2009). Rattle: a data mining GUI for R. The R Journal, 1(2), 45-55.

[33] Duval, E. (2011, February). Attention please! Learning analytics for visualization and recommendation. In Proceedings of the 1st international conference on learning analytics and knowledge(pp. 9-17). https://doi.org/10.1145/2090116.2090118

[34] Recker, M., Krumm, A., Feng, M., Grover, S., & Koedinger, K. (2016). Educational data mining and learning analytics. The Center for Innovative Research in CyberLearning.

[35] Tervakari, A. M., Silius, K., Koro, J., Paukkeri, J., & Pirttilä, O. (2014, April). Usefulness of information visualizations based on educational data. In 2014 IEEE global engineering education conference (EDUCON) (pp. 142-151). IEEE. https://doi.org/10.1109/educon.2014.6826081

[36] Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. American Behavioral Scientist, 57(10), 1500-1509. https://doi.org/10.1177/0002764213479363

[37] Jones, B. (2014). Communicating Data with Tableau: Designing, Developing, and Delivering Data Visualizations. " O'Reilly Media, Inc.".

[38] Deardorff, A. (2016). Tableau (version. 9.1). Journal of the Medical Library Association, 104(2), 182-183.

[39] Jain, A. (2014). Data visualization with the D3. JS Javascript library. Journal of Computing Sciences in Colleges, 30(2), 139-141.

[40] Bao, F., & Chen, J. (2014, May). Visual framework for big data in d3. js. In 2014 Ieee Workshop on Electronics, Computer and Applications (pp. 47-50). IEEE. https://doi.org/10.1109/iweca.2014.6845553

[41] Hernández-García, Á., González-González, I., Jiménez-Zarco, A. I., & Chaparro-Peláez, J. (2016). Visualizations of online course interactions for social network learning analytics. International Journal of Emerging Technologies in Learning (iJET), 11(07), 6-15. https://doi.org/10.3991/ijet.v11i07.5889

[42] Lauren, B., & Pigg, S. (2016). Networking in a field of introverts: The egonets, networking practices, and networking technologies of technical communication entrepreneurs. IEEE Transactions on Professional Communication, 59(4), 342-362. https://doi.org/10.1109/tpc.2016.2614744

[43] Bozkurt, A. (2017). Book Review: Analyzing Social Media Networks with NodeXL-Insights from a Connected World. Contemporary Educational Technology, 8(2), 191. https://doi.org/10.30935/cedtech/6195

[44] De Nooy, W., Mrvar, A., & Batagelj, V. (2018). Exploratory social network analysis with Pajek: Revised and expanded edition for updated software (Vol. 46). Cambridge University Press. https://doi.org/10.1017/9781108565691
[45] Moody, J., McFarland, D., & Bender-deMoll, S. (2005). Dynamic network visualization. American journal of sociology, 110(4), 1206-1241. https://doi.org/10.1086/421509
[46] Hagberg, A., Swart, P., & S Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Lab.(LANL), Los Alamos, NM (United States). https://doi.org/10.2172/425288
[47] Bakharia, A., & Dawson, S. (2011, February). SNAPP: a bird's-eye view of temporal participant interaction. In Proceedings of the 1st international conference on learning analytics and knowledge (pp. 168-173). https://doi.org/10.1145/2090116.2090144
[48] Butts, C. T. (2008). network: a Package for Managing Relational Data in R. Journal of statistical software, 24(2), 1-36.
[49] Butts, C. T., & Butts, M. C. T. (2019). Package 'sna'.
[50] Csardi, M. G. (2013). Package 'igraph'. Last accessed, 3(09), 2013. Handcock,
[51] M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data.Journal of statistical software, 24(1), 1548–7660. https://doi.org/10.18637/jss.v024.i01
[52] Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. Journal of statistical software, 24(3), nihpa54860. https://doi.org/10.18637/jss.v024.i03
[53] Kofia, V., Isserlin, R., Buchan, A. M., & Bader, G. D. (2015). Social Network: a Cytoscape app for visualizing co-authorship networks. F1000Research, 4. https://doi.org/10.12688/f1000research.6804.2
[54] Hayati, H., Chanaa, A., Khalidi Idrissi, M., & Bennani, S. (2019). Doc2Vec &Naïve Bayes: Learners' Cognitive Presence Assessment through Asynchronous Online Discussion TQ Transcripts. International Journal of Emerging Technologies in Learning (IJET), 14(08), pp. 70-81. https://doi.org/10.3991/ijet.v14i08.9964
[55] Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. Educational researcher, 40(5), 223-234. https://doi.org/10.3102/0013189x11413260
[56] McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Metrix. Cambridge University Press. https://doi.org/10.1017/cbo9780511894664.006
[57] Liu, H., & Singh, P. (2004). ConceptNet—a practical commonsense reasoning tool-kit. BT technology journal, 22(4), 211-226. https://doi.org/10.1023/b:bttj.0000047600.45421.6d
[58] Turian, J. (2013). Using AlchemyAPI for enterprise-grade text analysis. AlchemyAPI: Denver, CO, USA.
[59] Cornolti, M., Ferragina, P., & Ciaramita, M. (2013, May). A framework for benchmarking entity-annotation systems. In Proceedings of the 22nd international conference on World Wide Web (pp. 249-260). https://doi.org/10.1145/2488388.2488411
[60] Jean-Louis, L., Zouaq, A., Gagnon, M., & Ensan, F. (2014, December). An assessment of online semantic annotators for the keyword extraction task. In Pacific Rim international conference on artificial intelligence (pp. 548-560). Springer, Cham. https://doi.org/10.1007/978-3-319-13560-1_44
[61] Jovanovic, J., Bagheri, E., Cuzzola, J., Gasevic, D., Jeremic, Z., & Bashash, R. (2014). Automated semantic tagging of textual content. IT Professional, 16(6), 38-46. https://doi.org/10.1109/mitp.2014.85

[62] Ferragina, P., & Scaiella, U. (2011). Fast and accurate annotation of short texts with wikipedia pages. IEEE software, 29(1), 70-75. https://doi.org/10.1109/ms.2011.122
[63] Ferragina, P., & Scaiella, U. (2010, October). Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 1625-1628). https://doi.org/10.1145/1871437.1871689
[64] Gangemi, A. (2013, May). A comparison of knowledge extraction tools for the semantic web. In Extended semantic web conference (pp. 351-366). Springer, Berlin, Heidelberg.
[65] Morton, T., Kottmann, J., Baldridge, J., & Bierner, G. (2005). Opennlp: A java-based nlp toolkit. In Proc. EACL.
[66] Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. arXiv preprint cs/0205028.
[67] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations (pp. 55-60). https://doi.org/10.3115/v1/p14-5010
[68] Mayfield, E., & Rosé, C. P. (2013). LightSIDE: Open source machine learning for text. In Handbook of Automated Essay Evaluation (pp. 146-157). Routledge.
[69] Beheshitha, S. S., Gašević, D., & Hatala, M. (2015, March). A process mining approach to linking the study of aptitude and event facets of self-regulated learning. In Proceedings of the fifth international conference on learning analytics and knowledge (pp. 265-269). https://doi.org/10.1145/2723576.2723628
[70] Bogarín, A., Romero, C., Cerezo, R., & Sánchez-Santillán, M. (2014, March). Clustering for improving educational process mining. In Proceedings of the fourth international conference on learning analytics and knowledge (pp. 11-15). https://doi.org/10.1145/2567574.2567604
[71] Verbeek, H. M. W., Buijs, J. C. A. M., Van Dongen, B. F., & van der Aalst, W. M. (2010). Prom 6: The process mining toolkit. Proc. of BPM Demonstration Track, 615, 34-39.
[72] Gabadinho, A., Ritschard, G., Mueller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. Journal of Statistical Software, 40(4), 1-37. https://doi.org/10.18637/jss.v040.i04
[73] Stamper, J., Koedinger, K., d Baker, R. S., Skogsholm, A., Leber, B., Rankin, J., & Demi, S. (2010, June). PSLC DataShop: A data analysis service for the learning science community. In International Conference on Intelligent Tutoring Systems (pp. 455-455). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-13437-1_112

7 Author

Mohammed Salihoun obtained the doctorate degree in computer science in 2018 at EMI, (Ecole Mohammadia des Ingénieurs, Mohammadia School of Engineers) of the Mohammed V University (UM5) of Rabat, Morocco. He has been teaching computer sciences since 2012. His areas of interests are: Elearning, Big Data.

Article submitted 2020-06-20. Resubmitted 2020-07-26. Final acceptance 2020-07-29. Final version published as submitted by the authors.

8 Appendices

Table 3. Comparison of Data Mining Tools

Tools	Latest Version, Released Year	License	Operation System	Language	Website	Type	Availability
EDM Work-bench	1.0, July 6, 2013	-	Cros Platform	Java	http://penoy.admu.edu.ph/~alls/downloads-2	Data manipula-tion, Feature Engineering	-
Jupyter notebook	6.0.3, June 6, 2020	BSD license	Cross Platform	Python	https://jupyter.org/	Machine learn-ing, Data trans-formation, Data visualization	Open Source
RapidMi ner	9.7, June 2, 2020	APGL	Cross Platform	Language independent	www.rapidminer.com	Data Science, Machine Learn-ing, Predictive Analytics	Open Source
WEKA	3.8.3, Septem-ber 4, 2018	GNU	Windows, OS X, Unix	Java	www.cs.waikato.ac.nz/~ml/weka	Machine Learn-ing	Open Source
SPSS	26, April 9, 2019	Com-mercial	Windows, OS X, Unix	Java	https://www.ibm.com/fr-fr/products/spss-statistics	Statistical analysis, Data mining, Text analytics, Data collection	Shareware
Knime	4.0, June 27, 2019	GNU	Windows, OS X, Unix	Java	www.knime.com	Data Mining, Deep Learning, Data Analysis, Text Mining	Open Source
Orange	3.24.1, January 17, 2020	GPL v3	Cross Platform	Python, Cython, C++, C	www.orange.biolab.si	Machine Learn-ing, Data Mining, Data Visualization, Data Analysis	Open Source
Keel	3.0, April 9, 2018	GPL / GNU	Windows, OS X, Unix	Java	https://sci2s.ugr.es/keel/download.php	Data transfor-mation, Data analysis, Data visualization	Open Source
Spark MLLib	2.4.5, February 8, 2020	Apache v2	Windows, OS X, Unix	Java, Scala, Python, R	https://spark.apache.org/mllib/	Machine Learn-ing, Data Processing	Open Source
Rattle	5.1.0, Septem-ber 5, 2017	GNU	Windows, OS X, Unix	R	https://rattle.togaware.com/	Data mining, Statistical analysis	Open Source
Tableau	2020.2.1, May 28, 2020	Com-mercial	Windows, OS X, Unix	-	https://www.tableau.com/	Data analysis, Data visualiza-tion	Shareware
D3.js	5.16.0, April 20, 2020	BSD	Windows, OS X, Unix	JavaScript	https://d3js.org/	Data Pro-cessing, Data visualization	Open Source

Gephi	0.9.2, September 24, 2017	GNU v3	Windows, OS X, Unix	Java	https://gephi.org/	Network analysis, Data visualization	Open Source
EgoNet	2.0.1, May 28, 2017	GPL	Windows, OS X, Unix	Java	https://sourceforge.net/projects/egonet/	Data collection, Network analysis,	Open Source
NodeXL	1.0.1.238, April 8, 2013	Microsoft Public License	Windows	C#	https://nodexl.com/	Network analysis, Data visualization	Shared Source
Pajek	5.08, September 2nd, 2019	Free	Windows, OS X, Unix	-	http://mrvar.fdv.uni-lj.si/pajek/	Network analysis, Data analysis, Data visualization	Closed Source
Sonia	1.1, June 7, 2015	GPL v2	Windows, OS X, Unix	Java	https://sourceforge.net/p/sonia/wiki/Main_Page/	Network analysis, Data visualization	Open Source
NetworkX	2.4, October, 2019	BSD	Cross Platform	Python	https://networkx.github.io	Network analysis, Data visualization	Open Source
SNAPP	2.1, February 26, 2014	Free	Cross Platform	Java	http://www.snappvis.org/	Network analysis, Data analysis, Data visualization	Open Source
R Package (Network)	1.16.0, December 1st, 2019	GPL v2	Windows, OS X, Unix	R	https://cran.r-project.org/web/packages/network/index.html	Network analysis, graph visualization	Open Source
R Package (sna)	2.5, December 10, 2019	GPL v2	Windows, OS X, Unix	R	https://cran.r-project.org/web/packages/sna/index.html	Network analysis, Network visualization	Open Source
R Package (igraph)	1.2.5, March 19, 2020	GPL v2	Windows, OS X, Unix	R, Python, C/C++	https://igraph.org/	Network analysis, Network visualization	Open Source
R Package (statnet)	2019.6, June 14, 2019	GPL v3	Windows, OS X, Unix	R	http://statnet.org/	Network analysis, Network visualization	Open Source
R Package (ergm)	3.10.4, June 10, 2019	GPL v3	Windows, OS X, Unix	R	https://cran.r-project.org/web/packages/ergm/index.html	Network analysis, Network visualization	Open Source
Cytoscape	3.8.0, April 15, 2020	LGPL	Cross Platform	Java	www.cytoscape.org	Data analysis, Data integration, Data visualization	Open Source
LIWC	1.6, January 9, 2015	Commercial	Windows, OS X	-	http://liwc.wpengine.com/	Natural Language Processing Analysis Content Analysis	Shared Source
CohMetrix	3.0, August	-	Cross Platform	Java	http://cohmetrix.com/	Text analysis	-

	16, 2017						
ConceptNet	5.8, May 20, 2020	CCA 4.0	Unix	Python	http://conceptnet.io/	Text analysis	Open Source
AlchemyAPI	September 28, 2016	Commercial	-	Java, Python, JavaScipt, Ruby, PHP	www.alchemyapi.com	Machine learning, Deep learning, Natural language processing, Text analysis	-
Tagme	0.1.3, April 6, 2017	Apache v2	Cross Platform	Java	https://sobigdata.d4science.org/web/tagme/demo	Text analysis	Open Source
Apache Stanbol	1.0.0, October 23, 2016	Apache v2	Cross Platform	Java	https://stanbol.apache.org/	Text analysis	Open Source
Apache OpenNLP	1.9.2, December 11, 2019	Apache v2	Cross Platform	Java	https://opennlp.apache.org/	Machine learning	Open Source
Python NLTK	3.5, April, 2020	Apache v2	Windows, OS X, Unix	Python	http://www.nltk.org/	Text analysis	Open Source
Stanford CoreNLP	4.0.0, April 19, 2020	GNU	Windows, OS X, Unix	Java	https://stanfordnlp.github.io/CoreNLP/index.html	Text analysis	Open Source
LightSide	2.3.2, August 15, 2015	GPL / GNU	Windows, OS X, Unix	Java	http://ankara.lti.cs.cmu.edu/side/	Machine learning, Text analysis	Open Source
ProM	6.9, June 3, 2019	GNU v2	Cross Platform	Java	http://www.promtools.org/doku.php	Process Mining	Open Source
Traminer	2.2.0, April 22, 2020	GPL v2	Windows, OS X	R	http://traminer.unige.ch/	Sequence Mining	Open Source
PSLC Datashop	10.5.4, February 12, 2020	Proprietary	Cross Platform	-	https://pslcdatashop.web.cmu.edu/	Process and sequence mining	Open Data Repository