International Journal on Data Science and Technology
2019; 5(1): 1-7
http://www.sciencepublishinggroup.com/j/ijdst
doi: 10.11648/j.ijdst.20190501.11
ISSN: 2472-2200 (Print); ISSN: 2472-2235 (Online)

Digital Language Mining Platform for Nigerian Languages (DLMP)

Emejulu Augustine Obiajulu[1], Okpala Izunna Udebuana[1, *], Nwakanma Ifeanyi Cosmas[2]

[1]Department of Communication and Translation Studies, National Institute for Nigerian Languages, Aba, Nigeria

[2]Department of Information Management Technology, Federal University of Technology, Owerri, Nigeria

Email address:

omejulus@yahoo.com (E. A. Obiajulu), izunna.okpala@yahoo.com (O. I. Udebuana), ifeanyi.nwakanma@futo.edu.ng (N. I. Cosmas)

[*]Corresponding author

To cite this article:

Emejulu Augustine Obiajulu, Okpala Izunna Udebuana, Nwakanma Ifeanyi Cosmas. Digital Language Mining Platform for Nigerian Languages (DLMP). *International Journal on Data Science and Technology.* Vol. 5, No. 1, 2019, pp. 1-7. doi: 10.11648/j.ijdst.20190501.11

Received: March 1, 2019; **Accepted**: April 8, 2019; **Published**: May 15, 2019

Abstract: Effective communication occurs when the receiver and sender both understand and synchronize the flow of information across board. The utility of language extends beyond human to human interaction and includes also, the use of syntactically formed programming languages to interact with digital systems. Nigeria has an estimate of over 450 languages, which makes it cumbersome to harmonize and put all into a single large repository for data mining. The goal of this paper is to firmly establish the importance of Information Technology in galvanizing Nigerian Languages and Mining scientific data thereof. The purpose of applying Information and Communication Technology (ICT) is to codify the process of extracting various underlying meanings in a language, processing the various idioms, proverbs and quaint statements in such language with the view of bringing out the creativity behind them. The authors explore the developmental stages and techniques of applying an artificial Intelligence system that scans through a given indigenous linguistic system to bring out the hidden facts therein. It is recommended that stakeholders in the 'digital humanities' adopt such mining platforms which helps in achieving greater insight into the diverse cultures and languages, in turn, promoting easy learning experience for indigenous languages.

Keywords: Artificial Intelligence, Language Mining, Nigeria, Communication

1. Introduction

Languages are central to exchange of information amongst individuals, groups and the global population. According to a credible Nigerian newspaper [1], Nigeria has over 400 languages of which 3 are major. Out of all the languages, perhaps twenty are severely endangered and as many as two hundred are threatened.

Linguists of all persuasions seem to agree that a language should be viewed as a system; a set of elements, each of which has a capacity of contributing to the workings of the whole [2]. Language forms a large part of the people's culture. It is through various languages that people express their folk tales, myths, proverbs and history [3]. Furthermore, effective communication occurs when the receiver and sender both understand and synchronize the flow of information across board. The utility of language extends beyond human to human interaction and includes the use of syntactically formed programming languages to interact with digital systems. A useful definition of communication should include the feeling that language has been widely studied and acclaimed as the most valuable human institution and is indispensable in all spheres of life [4]. In the same paper, Danladi cited Crystal, as saying; communication is conceptualized as having perhaps a ''magical,'' "mystical'' and "unique'' role in capturing the breadth of human thoughts and endeavour. Inadvertently, it means that for a country to function properly, it needs the cooperation and understanding of people from different ethnic groups.

Nelson Mandela remarked "If you talk to a man in a language he understands, it goes to his head, but If you talk to him in his language, it goes to his heart". He believed language is important, the more reason he learnt Afrikaans during the time he was imprisoned in Robben Island thereby unknowingly benefiting from The Foreign Language Effect.

Nigeria boasts of her many and diverse cultures and

languages, but steps need to be taken in order to preserve this precious cultural heritage of various ethnic groups, otherwise Nigeria risks losing them forever. Nigeria has Moribund/threatened languages i.e. languages not being used, nor transmitted to the younger generation e.g. *Basa-Kontagora, Bete, Baissa Fali, Defaka, Kiong, Kudu-Camo, Labir, Lere, Lufu, Polci/Luri, Njerep, Odut, Putai, Shau, Somyev, Vono, Ziriya*. Apart from Moribund languages, Nigeria also have Retreating languages, i.e. languages that appear to be dying from a particular area, but still flourishing in another area or inter-country boundaries. Example of retreating languages are *Abanyom, Abon, Abua, Abureni, Achipa, Adim, Aduge, Adun, Afade, Afo, Afrike, Gbo, Agbo, Ajawa, Akaju-Ndem, Akweya-Yachi, Alago* etc.

2. Methodology

The methodology employed in this paper is the descriptive research methodology. The authors explore the developmental stages and techniques of applying an artificial Intelligence system that scans through a given indigenous linguistic system to bring out the hidden facts therein. The observational and case study methods made the analysis easier.

3. Theoretical Framework

Language is very powerful and was used as a tool of control during the colonial days. It forms a large part of the culture of people. it is through language that people express their folk tales, myths, proverbs and history [3]. Language also covers a more potent and characteristic of human behaviour. Speech communication employs a host of expressive means ranging from linguistic to paralinguistic and extra linguistic features.

Scientific Components of Languages

i. Lexis, structure, stylistics, orthography and phonology: Phonetics and phonology are two related areas of linguistics that not only deals with sounds, but they deal with different aspects of sounds [5]. Language units bear a stylistic marker already before they are actually used, and so they tend to occur only in some types of texts e.g., terms, some foreign plural nouns, vulgarisms, participial constructions; these bearers of stylistic information which may come from any linguistic plane are also called stylemes/stylemy. For example, Igbo is a tonal language with two distinctive tones, high and low. In some cases a third (down stepped high tone) is recognized. The language's tone system was given by John Goldsmith as an example of auto segmental phenomena that go beyond the linear model of phonology laid out in The Sound Pattern of English. Igbo words may differ only in tone.

Examples are:

(a) Ákwá "cry"
(b) àkwà "bed"
(c) àkwá "egg"
(d) ákwà "cloth"

ii. Words, pronunciation, and the methods of combining words: The Power of Language Influences thought and action. The words we use to describe things to ourselves and others affects how we think and act. This powerful influence happens in all kinds of situations and most certainly with language related to teaching and learning where words and combination of words are used constructively.

iii. Cultural Universals: Language is heavily influenced by culture. As cultures come up with new ideas, they develop language components to express those ideas. The reverse is also true: the limits of a language can define what is expressible in a culture i.e. the limits of a language can prevent certain concepts from being part of a culture. Language and culture are inseparable, in that, language is intrinsic to the expression of culture. As a means of communicating values, beliefs and customs, it has an important social function and fosters feelings of group identity and solidarity. It is the means by which culture and its traditions and shared values may be conveyed and preserved.

iv. Language correlates intelligence: Referencing Nelson Mandela's quote "If you talk to a man in a language he understands, it goes to his head, but If you talk to him in his language, it goes to his heart." This purely shows a strong relationship between intelligence and languages. This can be viewed philosophically to be investigated by a systematic analysis of the necessary and sufficient conditions for the occurrence of various thoughts in human mind.

v. Instrument in communication, commerce, politics and education: In political and social policy, language functions as a vehicle of interaction and an instrument of communication with the use of communications, although it has always possessed an added cultural importance as a tool of the dominant ideology. That is to learn a language is not only reaching out to others but to maintain a variety of the social bond, a shared sense of values and communal awareness, Linguists of all persuasions seem to agree that a language should be viewed as a system; a set of elements, each of which has a capacity of contributing to the workings of the whole [2].

vi. Expression of social relationships and social identity: Language pervades social life. It is the principal vehicle for the transmission of cultural knowledge, and the primary means by which we gain access to the contents of others' mind. Just as *language use* pervades social life, the elements of social life constitute an intrinsic part of the way language is used. Linguists regard language as an abstract structure that exists independently of specific instances of usage (much as the calculus is a logico-mathematical structure that is independent of its application to concrete problems), but any communicative exchange is situated in a social context that constrains the linguistic forms participants use.

4. Language Mining

The use of Information and Communication Technology is employed to simplify the mining process of the various

meanings in languages, processing the various idioms, proverbs and statement to bring out the creativity inherent in them. Mining in the field of Geosciences involves the extraction of valuable materials, mineral deposits and other valuable resources from their natural environments following the process of recovery and separation of the target substance. In Information Technology, the term *Data mining* was introduced in the 1990s, though data mining was the evolution of a field with a long history. It can also be referred to as knowledge discovery in databases, i.e. the process of discovering interesting and useful patterns and relationships in large volumes of data. The field combines tools from statistics and artificial intelligence (such as neural networks and machine learning) with database management to analyse large digital collections, known as data sets [6].

Language mining is a relatively new technology as it requires the integration of database technology and artificial intelligence. Language mining is the process of extracting the hidden facts and knowledge that people do not know in advance but potentially useful by the introduction of an artificial Intelligence system. The archive of all the Nigerian languages needs as a matter of urgency to be synchronized with a central repository. The combinatory property of Linguistics data and information Technology plays a very crucial role in the prevention of Nigerian languages and its inherent creativity from extinction.

4.1. Why Language Mining

For mining to take place, the origin of the language must be established and also the various faculties and meta-knowledge involved needs to be studied. This begs to ask important questions like:

a. Who invented language?
b. Who was the first person that spoke your language?
c. How did the different objects on earth get their name in your own language?
d. Why do people from certain cultural background behave different from others?
e. Why do people bear different names?
f. Why do we call some things evil in some certain areas?
g. Why is emotion important to creativity?
h. Does spoken words represent experiences?
i. Why should humans communicate?

Native creativity and Innovativeness stems from spoken words/languages. The analogy below makes the understanding of mining process easier. In Nigeria, the late Chief Odimegwu Ojukwu, used a technological tool called *Ogbu n'Igwe* in Igbo language during the Biafran/Nigerian war. *Ogbu n'Igwe* was the major arsenal of the Biafrans during the war. *Igwe* in Igbo language, means *multitude of people* while *Ogbu* means *Killer*. A combinatory property of the two terms gives what is called *Ogbu n'Igwe* which literarily translates to *killer of multitude* in English language. Though, the English meaning for *Ogbu n'Igwe* is *Monster Bombs*, but, translating *Monster Bombs* back to Igbo language means another thing. In this context, literal translation is not necessarily the answer. The answer is the function of the object in question. Therefore, there is need to understand where certain words stem from in order to use them.

4.2. Principal Means of Mining

i. Inductive learning method: Inductive learning methods include information methods or decision tree and set theory methods. Decision tree method is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The algorithm is non-parametric and can efficiently deal with large, complicated datasets without imposing a complicated parametric structure [7]. Using this method, it can handle the following problems, including Language simplification, Language association discovery, Language meaning evaluation and approximate analysis of Language.

ii. Imitation biotechnology method: Imitation biotechnology method includes neural networks and genetic algorithms. Artificial neural networks imitating biological neural network in structure, is a nonlinear prediction model through training, and can be used for classification, clustering, feature extraction and other operations. In Language mining, it can be expressed as a search problem with the use of powerful search capabilities of genetic algorithms to find the optimal solution in a given sample of texts [8].

iii. Statistical analyses: Using statistical analysis, various information extracted are unknown mathematical model/formula from sample analysis (Languages captured or historical data). This involves rigorous statistical procedures, judgment hypothesis and errors control.

iv. Fuzzy mathematics: Fuzzy logic is the fusion of fuzzy set and Boolean Logic. The true value of a formula cannot just be zero or one, rather inclusive of all numerical data in (0-1) count. In Language mining and knowledge discovery, fuzzy logic for Language query, sort, evidence combination and confidence calculation are often used.

4.3. Steps in Language Mining

Language mining is a systemic process used for the extraction of facts unknown, effective and practical information from large databank of languages. Language mining process includes four steps, data/language storage, data preparation, mining process, results.

i Data/Language Storage: The data collected at this stage are stored in a single repository that houses various entities e.g. Igbo, Hausa, Yoruba etc… The data housed are raw translation of the words, proverbs or sentences. The database implementation needs to be relational in nature because one of the criterion and key purpose of this paper is to outline the possibility of building bridges/relationships across all languages.

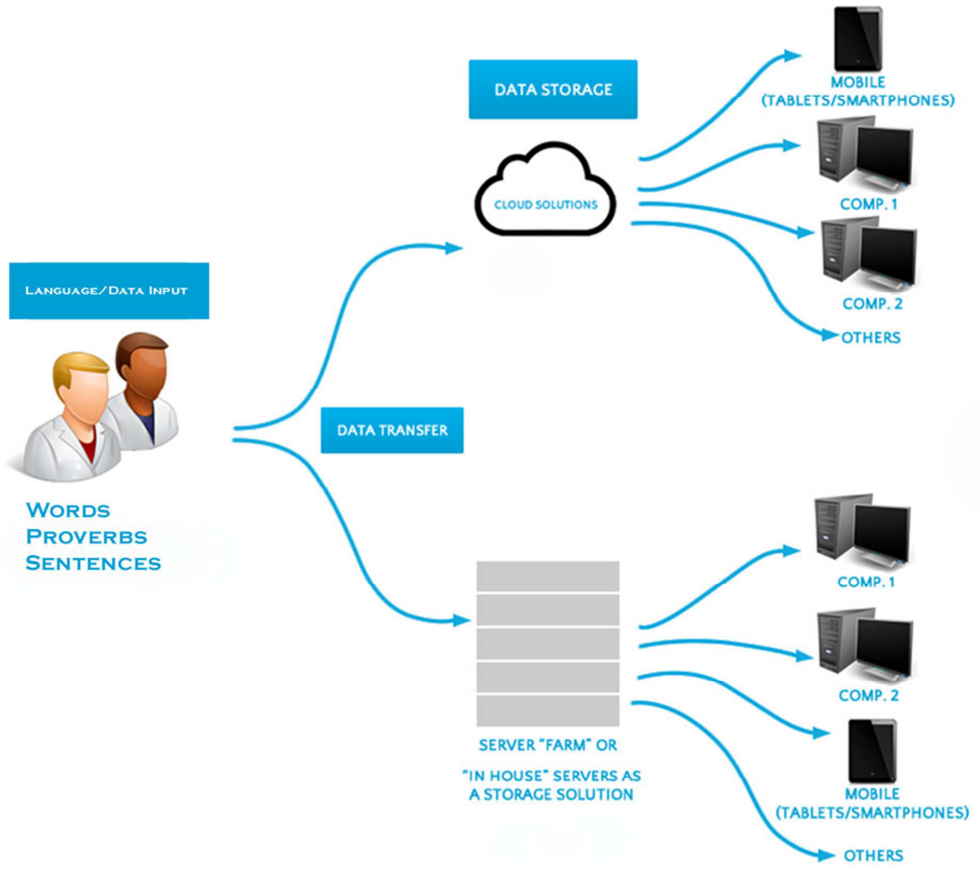

Figure 1. *Data storage mechanism from a user.*

ii Data Preparation: The data stored at this stage is not quantitative in nature, thus needs to be structured in a way that makes it meaningful to read. Here the ETL (Extraction, Translation and Loading) process is applied in order to galvanize data. Also, we apply meta-knowledge to the already provided data at this stage. The final result at this stage is a well-structured and articulate knowledge base.

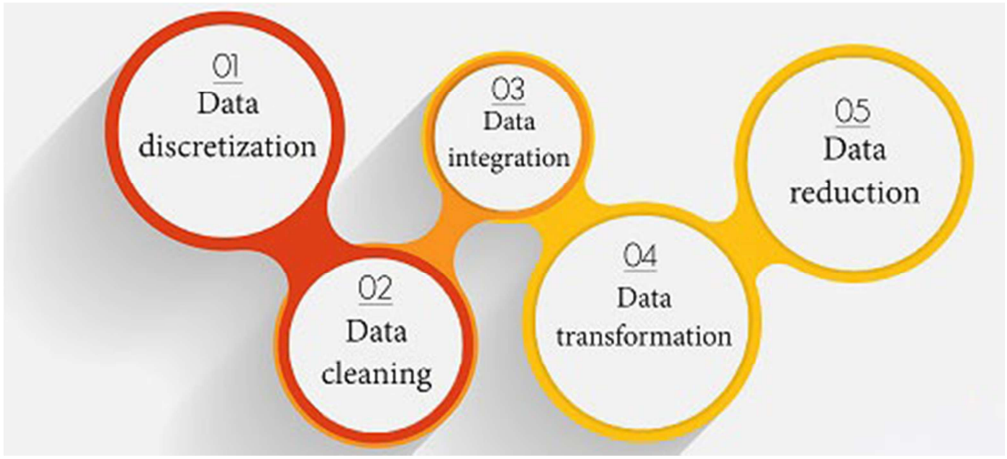

Figure 2. *Data preparation steps.*

During the preparation stage, the various language data that forms a relationship with others are not yet established and thus forms a scattered nonlinear plots found in Figure 3.

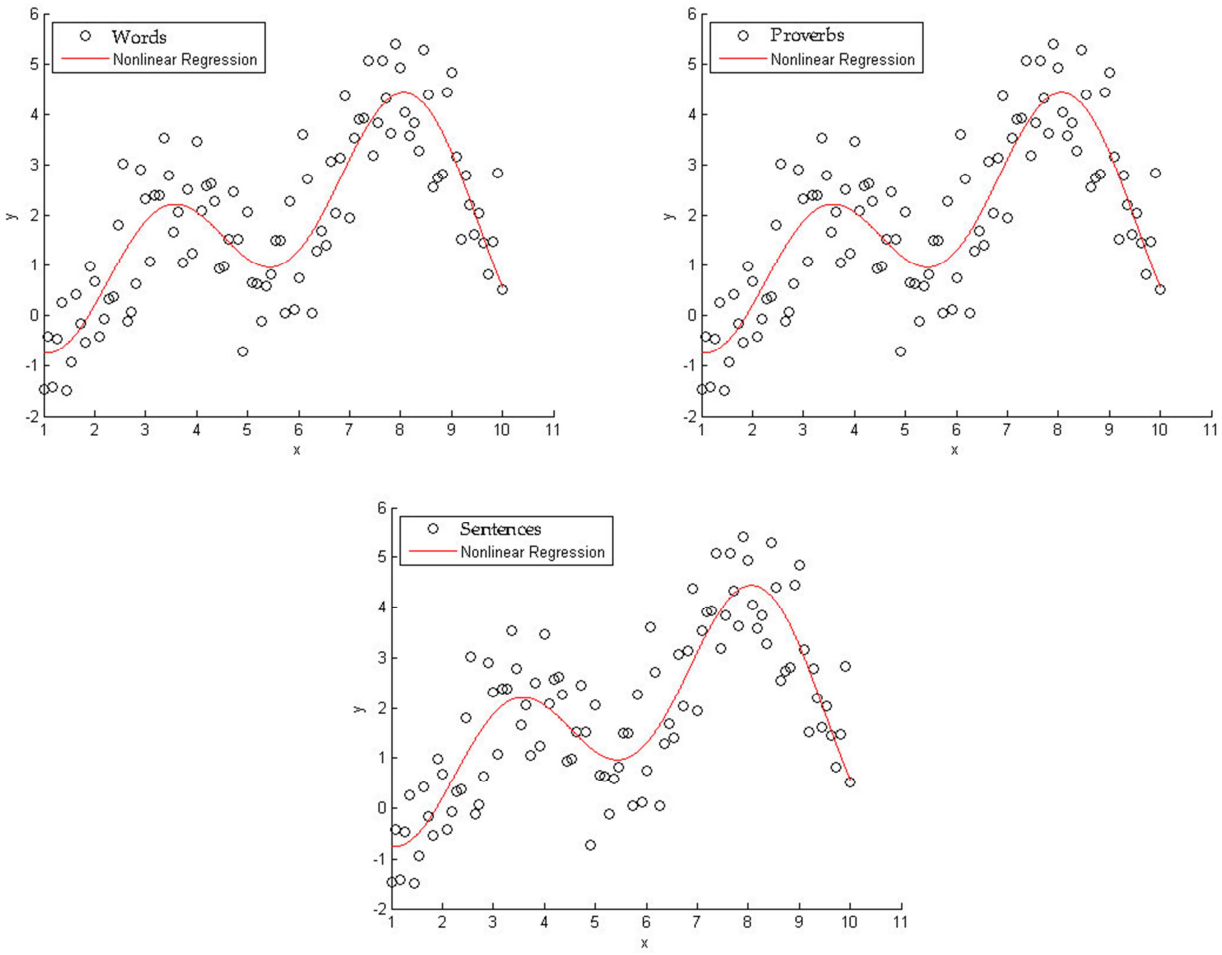

Figure 3. *Individual data plots of various language construct without relationship.*

iii Mining Process: After data preparation, a rough knowledge is given but on individual basis not relating to other entities i.e. the words, sentences and proverbs translations prepared are first in a scattered form without correlation with each other. The application of the mining mechanism works in a spiral form, meaning that it uses unsupervised neural network to loop through each correlation. The mining property is applied to map the various meanings given to a word as it relates to either sentences or proverbs. On the same verge, same is done on proverbs to sentences mapping as well to get specific correlations. This stage gives various deciphered meaning from inputted data, though, multiple loops are performed to gather more knowledge.

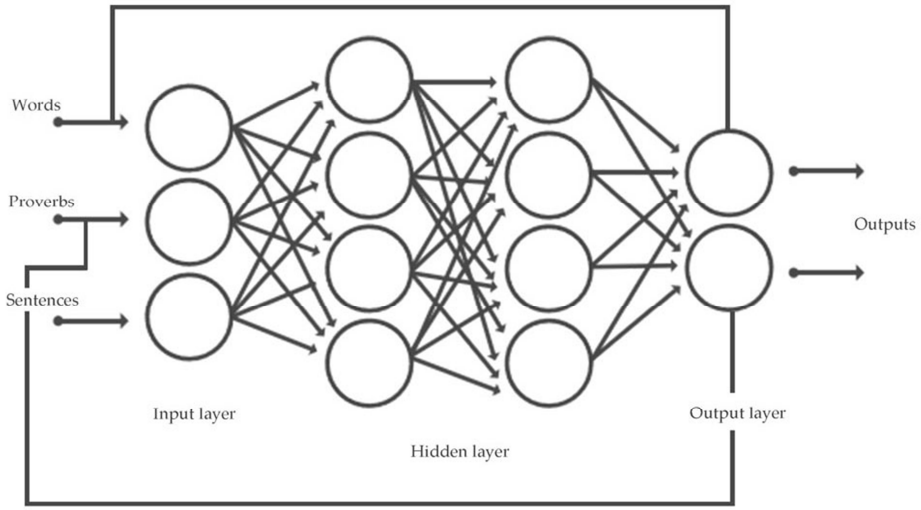

Figure 4. *The mining process utilizing unsupervised neural network.*

iv Result: The result at this stage are refined terms from various languages: After the ETL process, the output is a refined information that can be used as inference for practical purposes [9]. In our case we get various scientific terms and hidden meanings behind most languages and why they are very unique. The ETL process uses a user defined mining algorithm. Algorithm research and testing is often used in standard Language sets. However, in reality, Language sources of users are often various, which have a lot of missing Language, noise Language and non-standardized Language. So in the Language mining process, it is very important for Language pre-processing and transformation, and it is an important guarantee for Language mining to improve the efficiency and get better result.

5. Conclusion

Nigeria is a Nation that's not only rich in oil but rich in culture. Our diverse cultural heritage can be harnessed to our advantage, just as countries like China and India who changed their economic climate with their uniqueness. It's time to turn the table and export our rich heritage through a sync with Information Technology.

6. Recommendations

There is need for collaboration both Nationally and internationally in terms of giving this project the necessary publicity that it needs. The orthography of all the major languages are synchronized into a single large repository for future reference, and with that, researchers can tap from hidden knowledge extracted from the system.

References

[1] Pulse Newspaper. (2018). Why are Nigerians shying away from their mother tongue?

[2] Beaugrande and Dressler (1992). Nigeria and the role of English language in the 21st century Retrieved from https://eujournal.org/index.php/esj/article/download/1153/1169

[3] Nhlapo, T., Arogundade, E., & Garuba, H. (2014). Things fall apart? reflections on the legacy of Chinua Achebe. Zhao, Z. A., & Liu, H. (2011). Spectral feature selection for data mining. Chapman and Hall/CRC.

[4] Danladi, S. S. (2013). Language policy: Nigeria and the role of English language in the 21st century. European Scientific Journal, ESJ, 9(17).

[5] Yusuf, O. (2010). Basic Linguistics for Nigerian Languages. Ijebu-Ode: Shebiotimo Publications.

[6] Chuvakin, A., Schmidt, K., & Phillips, C. (2012). Logging and log management: the authoritative guide to understanding the concepts surrounding logging and log management. Newnes.

[7] Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

[8] Agatonovic-Kustrin S1, Beresford R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10815714

[9] Khan, A., Baharudin, B., & Lee, L. H. (2010). Khairullah khan, (2010). A Review of Machine Learning Algorithms for Text-Documents Classification, journal of advances in information technology, 1(1).

[10] Bohanec, M., Moyle, S., & Wettschereck, D. (2001). A software architecture for data pre-processing using data mining and decision support models.

[11] Vaidya, J., & Clifton, C. W. (2009). Privacy-Preserving Kth Element Score over Vertically Partitioned Data. IEEE Trans. Knowl. Data Eng., 21(2), 253-258.

[12] Cook, J. E., & Wolf, A. L. (1998). Discovering models of software processes from event-based data. ACM Transactions on Software Engineering and Methodology (TOSEM), 7(3), 215-249.

[13] Cortadella, J., Kishinevsky, M., Kondratyev, A., Lavagno, L., & Yakovlev, A. (1997). Petrify: a tool for manipulating concurrent specifications and synthesis of asynchronous controllers. IEICE Transactions on information and Systems, 80(3), 315-325.

[14] Gao, F., Xing, C., Du, X., & Wang, S. (2007). Personalized service system based on hybrid filtering for digital library. Tsinghua Science and Technology, 12(1), 1-8.

[15] Santesteban, M., Pickering, M. J., Laka, I., & Branigan, H. P. (2015). Effects of case-marking and head position on language production? Evidence from an ergative OV language. Language, Cognition and Neuroscience, 30(9), 1175-1186.

[16] Hollmann, J., Ardö, A., & Stenström, P. (2007). Effectiveness of caching in a distributed digital library system. Journal of Systems Architecture, 53(7), 403-416.

[17] Khenn Adatan (2013). Transcript of Language as a Key Aspect of Culture Relationship of Language and Culture. Retrieved from https://prezi.com/fkodedzsqps1/language-as-a-key-aspect-of-culture/

[18] Yang, L., Shin, S., Choi, Y., Choi, M., & Lee, Y. (2007, April). A surrogate variable-based data mining method using CFS and RSM. In Proceedings of the 6th WSEAS International Conference on Applied Computer Science (pp. 651-657).

[19] Lorenz, R., Bergenthum, R., Desel, J., & Mauser, S. (2007, July). Synthesis of petri nets from finite partial languages. In Application of Concurrency to System Design, 2007. ACSD 2007. Seventh International Conference on (pp. 157-166). IEEE.

[20] Ren, L., Song, M., & Song, J. (2004, May). A novel data type for the protocol of data synchronization. In Computer Supported Cooperative Work in Design, 2004. Proceedings. The 8th International Conference on (Vol. 1, pp. 532-535). IEEE.

[21] Parekh, R., & Honavar, V. (2000). Grammar inference, automata induction, and language acquisition. Handbook of natural language processing, 727-764.

[22] Agarwal, R. C., Aggarwal, C. C., & Prasad, V. V. V. (2001). A tree projection algorithm for generation of frequent item sets. Journal of parallel and Distributed Computing, 61(3), 350-371.

[23] Blench, R. M. (2014). The origins of nominal affixes in MSEA languages: convergence, contact and some African parallels. Languages of Mainland Southeast Asia: The State of the Art, 550-577.

[24] Rufai, A. (1977). The Question of a National Language in Nigeria. Language and Linguistic Problems in Africa Columbia, South Carolina, Hornbean.

[25] Schrijver, A. (1998). Theory of linear and integer programming. John Wiley & Sons.

[26] Tung, A. K., Hou, J., & Han, J. (2001). Spatial clustering in the presence of obstacles. In Data Engineering, 2001. Proceedings. 17th International Conference on (pp. 359-367). IEEE.

[27] Van Dongen, B. F., Busi, N., Pinna, G., & van der Aalst, W. M. (2007). An iterative algorithm for applying the theory of regions in process mining. In Proceedings of the workshop on formal approaches to business processes and web services (FABPWS'07) (pp. 36-55). Publishing House of University of Podlasie, Siedlce, Poland.

[28] Van der Aalst, W. M., Rubin, V., van Dongen, B. F., Kindler, E., & Günther, C. W. (2006). Process mining: A two-step approach using transition systems and regions. BPM Center Report BPM-06-30, BPMcenter. org, 6.

[29] Van der Aalst, W. M., van Dongen, B. F., Herbst, J., Maruster, L., Schimm, G., & Weijters, A. J. (2003). Workflow mining: A survey of issues and approaches. Data & knowledge engineering, 47(2), 237-267.

[30] Maniatty, W. A., & Zaki, M. J. (2000). Systems support for scalable data mining. ACM SIGKDD Explorations Newsletter, 2(2), 56-65.

[31] Yermack, D. (2015). Is Bitcoin a real currency? An economic appraisal. In Handbook of digital currency (pp. 31-43).