 doi:10.20944/preprints202105.0601.v1

Article

A study on ways to improve mobile RPG using big data text mining

DongHyun Youm [1] and JungYoon Kim [2,*]

[1] School of Game, Chungkang College of Cultural Industries, 389-94 Chungkang gachang-ro, Majang-Myeon, Icheon-si, Gyeonggi-Do, Republic of Korea; dhyoum@ck.ac.kr
[2] Department of Game Media, College of Future Industry, Gachon University, Seongnam-si, 13120, Gyeonggi-do, Republic of Korea
* Correspondence: kjyoon@gachon.ac.kr; Tel.: +82-31-750-8666

Abstract: As RPG has high sales and profits, lots of developers have supplied various RPG to market but it changed to mass production type with sensational advertising, low quality and excessive charging and similar contents which affects game market and users' game play experience. The author of this paper studied ways to improve mobile RPG by collecting and analyzing users' reviews using crawling on Google Play Store. The author of this paper used topic modeling that uses text mining technique and LDA (Latent Dirichlet Allocation) to extract meaningful information from collected big data and visualized it. Inferring users' reviews, figuring out opinions objectively and seeking ways to improve games are helpful in improving mobile RPG that can be played continuously.

Keywords: Mobile RPG; Big Data; Text Mining; Topic Modeling

1. Introduction

As internet has developed and disseminated widely, ways that consumers purchase goods have changed rapidly to on-line purchase from off-line purchase. For off-line purchase, it is possible for consumers to have an opportunity to select, touch and test goods directly. On the other hand, for on-line purchase, it is common that consumers do not have an opportunity to touch and test goods. Accordingly, in case of on-line purchase, consumers tend to depend heavily on reviews by other consumers. Such reviews provide potential customers and companies with necessary and meaningful information.

This paper studied ways to improve mobile RPG (Role Playing Game) by using big data analysis technique. To extract meaningful information from big data, the author of this paper collected users' reviews from Google Play Store through crawling and extracted meaningful text data through tokening. Results were visualized through LDA (Latent Dirichlet Allocation) and Topic Modeling. Main topics were found through interpretation based on results of visualization.

2. Relevant studies

Big data analysis aiming to get meaningful information undergoes algorithm and mathematical process according to purposes of analysis. Text mining which is a part of data mining is to find meaningful patterns based on enormous text data [1]. Collected texts allow us to extract frequency of words and includes lots of meaningless words and thus it is not easy to find main topics from set of such words. Topic modeling algorithm is statistical method which aims to find topics by analyzing words used in vast amount of texts and analyze how topics are correlated and how they change over time. Specifically [2], LDA (Latent Dirichlet Allocation) proposed by Blei is algorithm known as standard tool in the study of topic modeling [3].

LDA algorithm is production model and finds topics hidden in documents. LDA algorithm aims to reason hidden variable such as structure of documents through observed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2021 doi:10.20944/preprints202105.0601.v1

variable for example documents and words. LDA algorithm can be expressed as stochastic graph model as shown in [Figure 1].

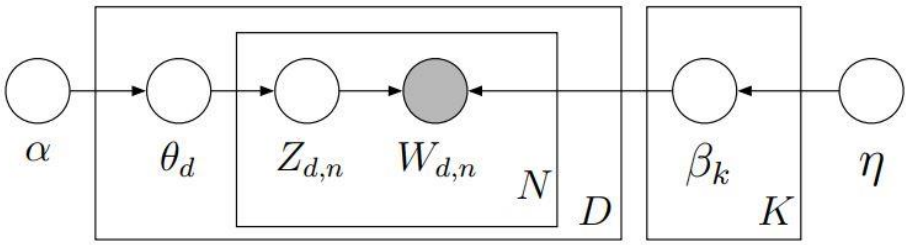

Figure 1. LDA Probabilistic Graph Model.

Topic is β1:K and each βk is distribution of words. Topic proportion of d th document is θd . θd,k is topic proportion of document d for topic k. zd is topic designation of document d . zd,n which is n th topic allocation of d can be obtained from fixed words [4]. Kai Tian used LDA to sort out software automatically [5]. Tirunillai used LDA to study online chatters' mining marketing by analyzing strategic brands of big data [6]. Bolelli used LDA to study topics and trends on text collection [7]. Somasundaran used LDA to sort out bug report automatically [8].

This study selects RPG genre from various mobile game genres and collects and analyzes users' thoughts and requests through text mining technique to propose an improvement plan. The most objective way of evaluating games is to figure out the voice of users. Analysis of game community allows us to figure out the voice of users most objectively. Text mining technique enables us to collect, draw and analyze users' thoughts and requests. This study selected games which many users play among RPG registered in Google Play to collect effective user's opinions. Mobile RPG that occupies higher rank in sales became the subject of this study because games with high sales are high in DAU.

3. RPG (Role Playing Game)

RPG (Role Playing Game) is one of game genres which users play most across platform such as PC, console and mobile. For mobile RPG, among mobile games registered in Google Play Store as of June 2020, RPG genre accounted for 56 out of top 100 sales [9].

Advantages and characteristics of RPG are as follows:

▶ RPG gives users role of characters and determines identity of users and direction of games through sense of unity between game characters and users and exercises a direct influence on performing roles in games [10].

▶ Users are assigned roles in games and they play a role and they are immersed in games deeply and for such sense of immersion, world in games should be alive and dynamic [11].

▶ Various forms of contents that users should perform playing a role in games exist [12].

Above mentioned advantages and characteristics along with long play time, outstanding extensibility and strong game addiction make RPG suitable for on-line platform which matches interests of game developers that should make profits continuously. This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

4. Problems by recent excessive RPG supply

Game developers' expectation for high sales and continuous profit making has led developers to supply various RPG to markets and accordingly the number of users who enjoy RPG has increased. RPG produced recently has shown problems of sensational advertising, low quality, excessive charging and similar contents which is different from

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2021 doi:10.20944/preprints202105.0601.v1

various and characteristic RPG at the beginning of service [13]. Several studies have reported that such mass production type game affects game market and users' game play experience negatively.

Yeong-joon, Jun said in his study on collective emotion and mobile game use experience that a feedback which quality of contents is not fully reviewed on platform is a serious problem which may worsen reliability on service [14]. Dai-hyun, Ki argued in his study on interaction between social network service and social network game users that keyword of 'mass production of low quality plagiarism game' may bring negative view of platforms that serve games [15]. Sung-hwa, Chung said that mass production games without improvements prevents domestic games from growing and leads to service failure [16].

DAU (Daily Active User) and ARPU (Average Revenue Per User) are very important indicators for game developers that place profits before anything else. This leads to a profit immediately and better contents are reflected in game development based on such profit which can create virtuous circle. Developing mass production type games which are characterized by similar contents for sales only is highly likely to lead to the lack of diversity in games and users' distrust. Accordingly, presenting direction which can satisfy both game developers and users is needed.

5. Using big data text mining

5.1. Text mining

This study presents a way that collects game reviews by various users and analyze them through text mining technique. Most reviews that can be collected on the internet, in other words online exist in atypical text form and thus text mining technique is used as a way to extract information from such atypical data [17]. Text mining is a part of data mining and finds a meaningful pattern based on huge text data [1].

Human languages have characteristics in terms of vocabulary and grammar. Forms of expression are so diverse and complex that it is difficult to find regularity. Human languages continue to change according to language use environment. Natural language processing analyzes and processes languages expressed as characters and understands its structure and meaning. Natural language processing allows us to convert documents to a form which can be analyzed passing through collection and preprocessing [18].

5.2. Topic modeling

Topic modeling finds main topics by analyzing words used in enormous amount of text collected and changes according to association between subjects and time [2].

Topic modeling is a study methodology which is usefully covered in the field of text mining and LDA (Latent Dirichlet Allocation) proposed by Blei is algorithm established as a standard tool in topic modeling studies.

5.3. LDA (Latent Dirichlet Allocation)

LDA is one of methods which are used most in topic modeling to process natural language. LDA shows topics through topic probability. Words with the highest probability in each topic provide a good idea of topics [19]. LDA algorithm finds topics hidden in documents as production model. LDA algorithm can figure out topics in entire document set, topic percent by documents and probability which each word is included in each topic and infers posterior probabilities based on conditions that words are produced under the assumption that words are not independent [20].

5.4. Visualization of data

Results of text mining through LDA are provided as raw number and it is very difficult to analyze them and thus visualizing data is needed to make it easy to analyze [21].

6. Mobile RPG data analysis

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2021 doi:10.20944/preprints202105.0601.v1

Meaningful topics should be extracted from mobile RPG users' reviews for this study. On this end, data are extracted through process of a few steps as shown in Figure 2 below.

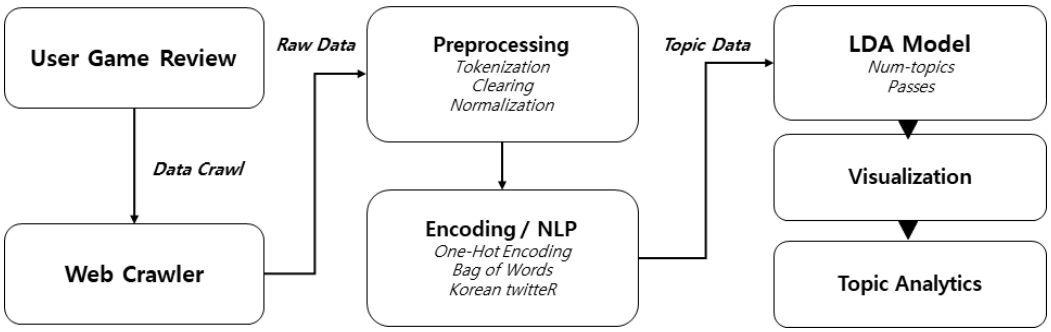

Figure 2. Mobile RPG data analysis.

Firstly, collect users' reviews. For this, method of web crawling is used.

Secondly, convert text data collected through web crawling to numerical data that computers can analyze. Preprocessing which filters unnecessary words for converting text data to figure data is needed. Preprocessing makes only meaningful letters left in text data.

Thirdly, configure topic models by using LDA algorithm and find optimum topic models by conducting performance evaluation.

Fourthly, visualize result of topic models so that it can be interpreted easily and seek direction of improvement of games by reasoning subjects of users.

This study used python.

6.1. Collectiong data by using Python

Web is basically expressed as HTML and it is managed in a typical form within HTML. Technique that brings typical data on the internet, parses and extracts only data needed is called crawling [22] which is conducted by using python.

6.2. Preprocessing by using Python

Corpus data obtained from crawling is preprocessed such as tokenization, clearing and normalization to one's needs. Tokenization is a process [23] that classifies and sorts out a series of input text section and separates language that cannot be divided any more in terms of grammar in other words token [24]. Clearing means removing unnecessary data in the process of tokenization. Special symbols are representative data that should be removed. Normalization means binding words that are different but have the same meaning together.

Table 1. Example of preprocessing results using Python.

Before tokenization	I am a boy, You are a girl!
After tokenization	"I", "am", "a", "boy", "You", "are", "a", "girl"!
Before tabletting	I am a boy, You are a girl!
After tableting	I am a boy You are a girl
Before normalization	You, Thou, Thee, Thy
After normalization	You

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2021 doi:10.20944/preprints202105.0601.v1

6.3. Encoding

Encoding is performed through One-Hot Encoding while a computer converts characters into numbers to process letters [25]. In case of One-Hot Encoding which is collection of words that do not allow duplication when making word set, it has weakness that as the number of words increases, storage space increases and thus data are digitized by focusing on word appearance frequency without considering BoW (Bag of Words) model- order of words.

6.4. Natural Language Processing

Natural language processing is a series of technical set that analyzes, extracts and understands meaningful information from text. This paper used twitter package among python packages for processing Korean information [26].

7. Topic modeling by using Python

7.1. Marking word dictionary

Gensim, library for topic modeling implemented as python provides LDA algorithm.

7.2. LDA model training

The following two parameters were selected as important parameters to make LDA model in Gensim.

- passes – LDA model learning recall
- passes – LDA model learning collection

Two parameters are adjusted and confusion score and consistency score are measured and parameters with best evaluation are determined to complete the final LDA model.

7.3. Finding optimum passes

Testing passes with steps classified from 1 to 50 by multiples of 5 assuming that num_topics is 10 among LDA model parameters and measuring confusion score and consistency score allows us to get graphs in Figure 3. Passes with best score by analyzing graphs are designated as final LDA model parameter values.

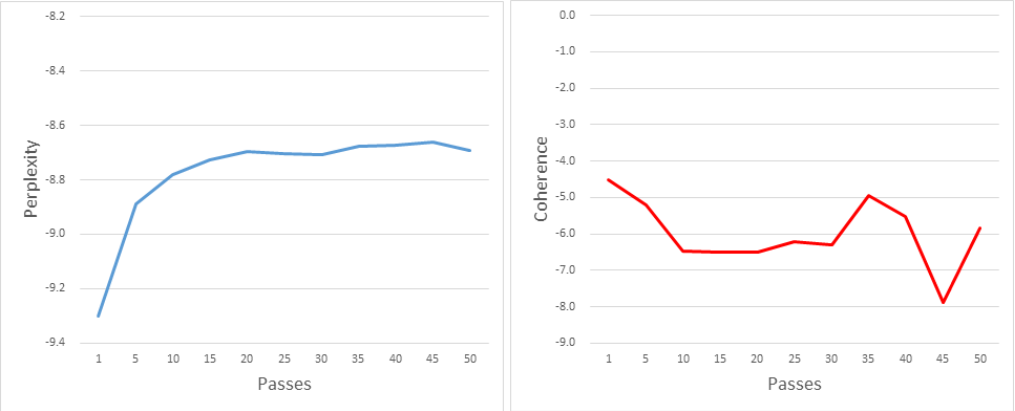

Figure 3. Example of graph of confusion score and consistency score according to Passes.

7.4. Finding num_topocs

Apply values of passes obtained from 7.3 and find num_topics among LDA model parameters. Testing num_topics with steps classified from 2 to 20 by multiples of 2 and measuring confusion score and consistency score allows us to get graphs in Figure 4.

num_topics with best scores by analyzing graphs are designated as final LDA model parameter values.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 May 2021 doi:10.20944/preprints202105.0601.v1

In case that the number of topics is twenty or fewer, confusion score and consistency score continue to worsen and thus inspecting topics of twenty or more is meaningless.

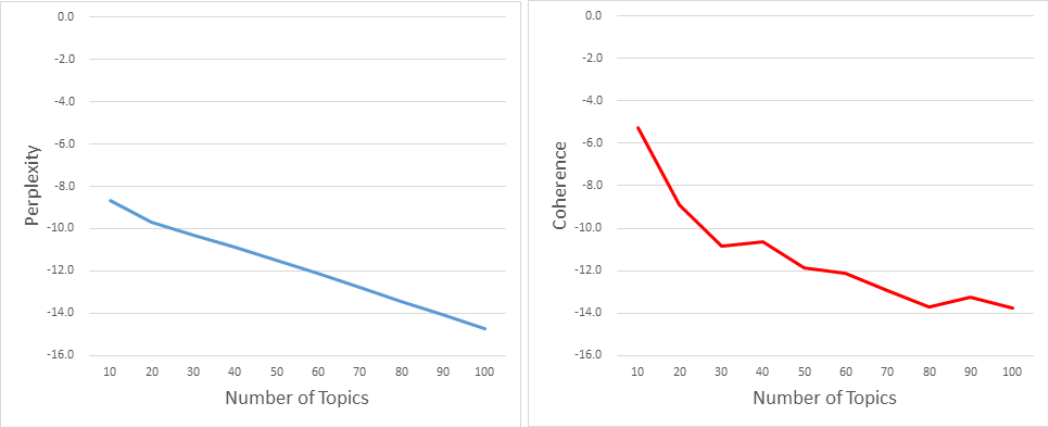

Figure 4. Example of graph of confusion score and consistency score according to the number of topics.

8. Visualization by using Python

Produce bag of words through pyLDAvis to visualize bag of words topic model and after producing LDA, schematize LDA model as shown in Figure 5 below.

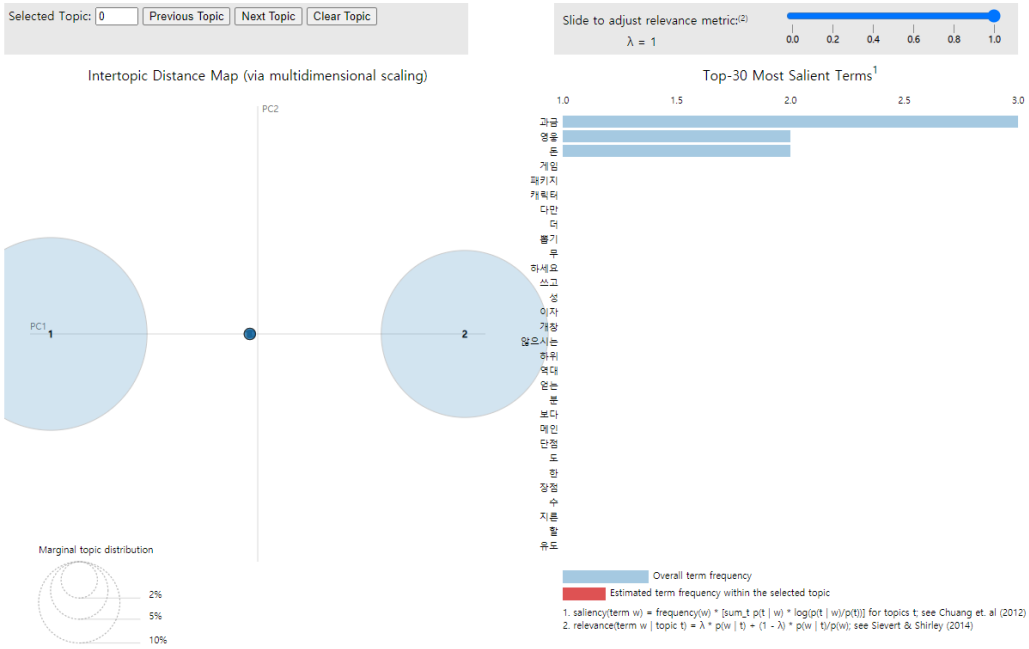

Figure 5. Visualization example using pyLDAvis.

9. Topic analysis

Find topics based on LDA model schematization by combining and inferring words with high salient values. In case that there is no knowledge of documents and it may be difficult to find topics. There is a considerable concern that analyzer's subjective thoughts is involved in analysis.

10. Conclusion

RPG is better than games of other genres in terms of DAU and ARPU based on users' continuous play. Therefore, RPG has advantages in securing profits stably in terms of game developers. In case of mobile RPG, games are developed in a manufacturing manner

rather than characteristic RPG by each developer is developed. Such problem causes users' complaints which may affect profits of game developers and lower reliability of mobile RPG games.

This paper studies a way to analyze users' opinions and cope with it through text mining technique. In order to obtain and analyze data, user reviews are crawled and tokened by using python and open-source modules and users' opinions are figured out by extracting meaningful words. In addition, LDA topic modeling technique was used to grasp accurate topic (subject) and find meaningful words. This study found optimum performance of LDA model by comparing confusion score and consistency score to evaluate performance of LDA model. Data analyzed by LDA model show correlations between topics by schematizing. This study can analyze and organize relevant topics through topics analyzed by LDA model and obtain main words composing topics. It is necessary to enhance accuracy by improving sophistication in the process of tokening through crawling. In addition, it is necessary to make a comparative analysis of studies based on analysis models. This study is expected to be used in developing and complementing games by extracting meaningful data applying findings of this study to games of other genres.

Author Contributions: Conceptualization and methodology, D.H.Y.; data curation, D.H.Y.; funding acquisition, J.Y.K.; investigation, D.H.Y.; supervision, J.Y.K.; visualization, D.H.Y.; writing—original draft, D.H.Y.; writing—review and editing, J.Y.K. All authors have read and agreed to the published version of the manuscript.

Author Contributions: Conceptualization and methodology, D.H.Y.; data curation, D.H.Y.; funding acquisition, J.Y.K.; investigation, D.H.Y.; supervision, J.Y.K.; visualization, D.H.Y.; writing—original draft, D.H.Y.; writing—review and editing, J.Y.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by Ministry of Culture, Sports and Tourism and Korea Creative Content Agency(Project Number: R2020040243).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: This research is supported by Ministry of Culture, Sports and Tourism and Korea Creative Con-tent Agency(Project Number: R2020040243), Excerpt submitted thesis by DongHyun, Youm for Ph.D., University of Gachon, South Korea, 2020.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Navathe, S.B.; Ramez A.E. *Fundamentals of Database Systems with Cdrom and Book*. Addison-Wesley Longman Publishing Inc, : Boston, United States, 2001.
2. Kang, B.I.; Song, M.; Jho, H.S. A Study on Opinion Mining of Newspaper Texts based on Topic Modeling. *Korean Society For Library And Information Science*. 2013, *47(4)*, 315-334. doi.org/10.4275/KSLIS.2013.47.4.315
3. Blei, D.M.; Andrew Y.N.; Michael I.J. Latent dirichlet allocation. *Advances in neural information processing systems*. 2001, *14(1)*, 601-608
4. Blei, D.M. Probabilistic topic models. *IEEE signal processing magazine*. 2010, *27(6)*, 55-65. doi.org/10.1109/MSP.2010.938079
5. Tian, K.; Meghan, R.; Denys, P. Using latent dirichlet allocation for automatic categorization of software. *Mining Software Repositories, 2009. MSR `09. 6th IEEE International Working Conference on 2009 May*. 2009, 163-166.
6. Tirunillai, S.; Gerard J.T. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of marketing research*. 2014, *51(4)*, 463-479. doi.org/10.1509/jmr.12.0106
7. Bolelli, L.; Ertekin, S.; Giles, C.L. Topic and trend detection in text collections using latent dirichlet allocation. *Lecture notes in computer science*. 2009, *5478*, 776-780. doi.org/10.1007/978-3-642-00958-7_84
8. Somasundaram, K.; Gail, C.M. Automatic categorization of bug reports using latent dirichlet allocation. *Proceedings of the 5th India software engineering conference*. 2012, 125-130. doi.org/10.1145/2134254.2134276
9. Google Play. Available online: https://play.google.com(accessed on 18 July 2020).

10. Byeon, H.S. The Impacts of Artistic Creativity, Scientific Creativity, General Creativity on Perceived Enjoyment and Intention to Reuse : Focused on Role-Playing Game Player. *Journal of Korea Game Society*. 2011, *11(1)*, 59-67. doi.org/10.7583/JKGS.2011.11.1.059
11. Role-playing game. Available online: https://en.wikipedia.org/wiki/Role-playing_game(accessed on 18 July 2020).
12. Kim, G.H.; Lee, N.Y. The Quality Evaluation Model for Mobile RPG. *The Korea Institute of Information and Commucation Engineering*. 2014, 457-460.
13. Column - Why is a Chinese-made mass-produced game a problem?. Acrofan. Available online: https://url.kr/qt6WZw(accessed on 18 July 2020).
14. Cheon, Y.J.; Kwak, K.T. Collective Sentiments and Users' Feedback to Game Contents : Analysis of Mobile Game UX based on Social Big Data Mining. *Journal of Korea Game Society*. 2015, *15(4)*, 145-156. doi.org/10.7583/JKGS.2015.15.4.145
15. Ki, D.H.; Park.Ch.H. Two-way interaction between social network service (SNS) and social network game (SNG) users. Asia-pacific journal of multimedia services convergent with art, humanities, and sociology. 2019, 9(12), 1321-1329. doi.org/10.35873/ajmahs.2019.9.12.115
16. Jeong, S.H.; Kyung, B.P.; Lee, D.L.; Lee, W.B.; Ryu, S.H. Study for the Transformation and Growth of MMORPGs : TIME FLOW Scenario Design. *Journal of Korea Game Society*. 2015, *15(4)*, 79-92. doi.org/10.7583/JKGS.2015.15.4.79
17. Kim, J.Y.; Kim, D.S. A study on the method for extracting the purpose-specific customized information from online product reviews based on text mining. *Journal of society for e-business studies*. 2016, *21(2)*, 151-161.
18. Hong, W.E.; Kim, U.H.; Cho, S.H.; Kim, S.S.; Yi, M.Y.; Shin, D.H.; Export Control System based on Case Based Reasoning: Design and Evaluation. *Journal of intelligence and information systems*. 2014, *20(3)*, 109-131. doi.org/10.13088/jiis.2014.20.3.109
19. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; Zhao, L. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimedia Tools and Applications*. 2019, *78(11)*, 15169-15211. doi.org/10.1007/s11042-018-6894-4
20. Park, J.H.; Song. M. A Study on the Research Trends in Library & Information Science in Korea using Topic Modeling. Journal of the Korean society for information management. 2013, 30(1), 7-32. doi.org/10.3743/KOSIM.2013.30.1.007
21. Ali, S.M.; Gupta, N.; Nayak, G.K.; Lenka, R.K. Big data visualization: Tools and challenges. *Contemporary Computing and Informatics (IC3I), 2016 2nd International Conference on 2016*. 2016, 656-660. doi: 10.1109/IC3I.2016.7918044
22. Olston, C.; Marc, N. *Web crawling*. Now Publishers Inc, : Boston, United States, 2010. doi.org/10.1561/1500000017
23. Lexical analysis. Available online: https://en.wikipedia.org/wiki/Lexical_analysis#Tokenization(accessed on 18 July 2020).
24. Text Tokenization. Available online : https://url.kr/nZJKua(accessed on 18 July 2020).
25. Cerda, P.; Gaël, V.; Balázs, K. Similarity encoding for learning with dirty categorical variables. *Machine Learning*. 2018, 107(8/10), 1477-1494. doi.org/10.1007/s10994-018-5724-2
26. Morphological analysis and POS tagging. KoNLPy. Available online : https://url.kr/LGra6k(accessed on 18 July 2020).