International Journal of Database Theory and Application
Vol.9, No.10 (2016), pp.177-188
http://dx.doi.org/10.14257/ijdta.2016.9.10.15

Pre/post-processing of Text Mining Techniques Improved through Referring to the Trend Dictionary

Sungwook Yoon and Hyenki Kim*

Dept. of Multimedia Engineering, Andong National University,
1375 Gyeongdong-ro, Andong-City, Gyeongsangbuk-Do, Republic of Korea
*uvgotmail@nate.com, *Corresponding Author hkkim@anu.ac.kr*

Abstract

Needs for the collection of data for text mining of the case with many protolanguages or emotional words distributed in SNS and following trends, and the treatment process method of purification of improved data in a previous step are raised. In data collection for mining, the online trend dictionary based on tag was referred and semi-structured data was effectively parsing processed based on tags of dictionaries according to domains of treating languages, and data for analysis was collected. Additionally, there were the cases to show inefficiency in the text processing of the general genre or the limitation of noun extraction, however, it can be suggested as an alternative on searching trend vocabularies which requires the timeliness or the class processing for corpus work of sentiment dictionary.

Keywords*: Social Text Mining, Emotional Neologism Trend Search, Emotional Corpus, Noun Extraction, Trend Dictionary, Corpus Extension*

1. Introduction

According to the purpose of research and use of a kind of analysis technique, text mining is deducting various results to the linkage analysis, statistical analysis such as regression analysis, sensitivity analysis, opinion analysis and *etc.* And in various analysis packages, the forum form or utilization method are being utilized. The correct collection and processing of data which becomes the object of analysis according to this largely effect on the result. And if the collection of data for the object of analysis and pre-processing and post-processing operation are well-processed, the reliability of analysis can be increased. However, the flexibility of the language, various implications, and sociality are not easy practically from pre-processing of the data. In case of describing morpheme, its tag work is processed by determining the adjacent morpheme or the methods like referring to the language study dictionary through Sejong corpus using relational database. Unlike English which has the clear orders of the words, this is not easy work, either, when the work based on Korean is performed. The Korean language is the class of Ural Altai and this can communicate without any specific influence, and it can be the cause to make the analysis to be difficult due to the complexity of the language such as omitting the subject word, using various postposition, and so on. Also, in case of wrong extraction as interjection, intuitional data cleaning may not be occurred like the extractions of the vocabularies with different meanings. Hence, the reliability on the extracted nouns is affected by many usages of adopted words - place names, names, specific terms, and so on, in spite of the precise frequent analysis [1]. In addition, the circulations of new word combinations, sentimental expression coinages, abbreviations, or acronyms are the social phenomena to reflect the timeliness, however, they may interrupt the sample collection or analysis in the studies of the specific areas. In fact, it is the general practice to request pre and post processing for the data collection on the data scientists of the journalists in the recent news articles based on the text mining of many

ISSN: 2005-4270 IJDTA

Big Data. The procedure of the original data processing might be performed with very empirical and heuristic methods so that it can be difficult to provide or manipulate more accurate data as the significant data are wasted practically. For this reason, it is necessary to research the collection method using collective intelligence such as online dictionaries and encyclopedias. And in this study, through the trend dictionary reference model that refers to this, I would like to suggest the process of collecting additional trend vocabulary data into the previous internal library.

2. Related Researches

2.1. Definition of Big Data and their Utilizations

Big Data is defined as the technique to collect large amount of typical or atypical data sets, to extract the values from these data, and to analyze the results in the data collection, storage, management, and analysis of the management tools in the existing big database [2]. Generally, the characteristics of Big Data to differentiate from the traditional data processing can be described as 3Vs including data Volume, Variety forms, and fast generation Velocity. The methodologies on these 3Vs have been emerged with the definition, the analysis, and the creation of significant new values. Gartner revised the existing definition in 2012 since the new types of processing were required to make a decision, find the insights, and enhance the process optimization by this as the data resources with large volume, fast velocity, and high level of variety. IBM defined them with the addition of Veracity, and Brian Hopkins et al did with the addition of Variability. Recently, the creation plan on the added values through these has been suggested including Complexity, additionally. Data volume means to secure the large amount of capacity, which leads to zeta-byte era with exponential growth of the digital information volume by technological advances and routinization of ICT (Information and Communication Technology) [3]. Variety forms mean to become atypical with the increase of data forms such as log, social, location based services, consumption information, real-life data, and so on. Fast generation velocity is understood as the platform to secure the timeliness due to the generation of real time data and the increase of circulation speed [4]. Complexity means the characteristics related to the increases of unstructured data, differences of data storage methods, the duplication issues, and so on, and the new technique is required on the data management and deepening of the processing complexity. The technical elements of Big Data can be considered with data collection, storage of raw data, storage of trade data, platform of real time analysis, performing allocated analysis, statistical tool of data mining, cluster management and monitoring, data continuity, semantic analysis, and so on [5].

2.2. Analysis of Big Data

Big data has put mass production of information, sharing of data of the public nature for business activities, news, information exchange of social network service, sensor network based on internet of the things as sources of collecting data [6]. The various collection methods include Big Data crawling, robots, Open API, FTP, RSS feed crawling, streaming, log collections, RDB based data collection, and so on. Additionally, opened public data for public and private information sharing which has a variety of subject classifications is the platform for public contribution [7].

Big data analysis finds a new flow by analyzing the general natural language in a lard lead and in real time. Also through the process of extracting as meaningful things by analysing this, the order of new data is found and applied to the related area.

Big Data analysis is composed of steps such as collection, storage, management, processing, analysis, using and *etc.* Collected data is filtered, transformed and cleansed. And for post-processing, it is integrated, transformed and reduced. Storage is being

processed in means of RDB, NoSQL, distributed file systems and *etc*. and formats of data to be saved are reviewed and selected in a favorable storage method for storage and management. The data format to be stored is reviewed and the advantageous type for the storage management shall be selected. Upon the review of feasibility to infringe the data security, the actions for the security management are taken such as access control, blocking, authentication, coding, de-identification, post-hoc monitoring, *etc*. by each step.

3. Analysis Design

3.1. Definition of Corpus Trends Directory

Most of current online based dictionary searching systems are reflecting trends for corpus such as the analogue, neologism, vulgarism, dialect, and *etc*., as well as words or vocabularies classified by existing morphemes [8]. This reflects online based characteristics of use by users such as the instantaneity, social nature of language and *etc*., and it is resulted from the change of user interface according to the input system of portable devices. For example, neologism and abbreviation are generated due to structural factors of typographical errors, errors in recommended vocabulary, trend of using the first letter and the used of the initial consonant language.

Coinages that are used like this have many cases of being used beyond the rapid period mostly, not going through the normal process of social agreements. And mostly, the possibility for using is decided at the boundary of inference possibilities. Coinage used like this can't confirm the meaning in regular dictionaries, but it becomes important material for the opinion analysis since there are many cases of containing feeling or meaning of emotions mostly. Thus, words that contain information which follow trends even though they are not searched by corpus of general dictionaries or words that evokes the transitional, or potential social trends by it use, such as manipulation of specific vocabulary in a new form and creed of the emotional language and *etc.* are called trend language. And corpus libraries that define vocabularies of new trend languages are defined as trend corpus dictionaries (trend dictionaries) in this study.

And if we assume that it represents newness and grasps the futuristic tendency by playing an important role from aspects of availability such as the sensitivity analysis, opinion mining and contextual analysis even though the trend vocabulary doesn't have the status of general morpheme, it is enough as a value of the material.

3.2. Trend Dictionary Reference

Once the original goal of the data which are the collection subjects is set, the data are collected by the methods like crawler, robot, and so on, and trend dictionary is referred through the collected subjects. If necessary, data are cleansed, integrated, and reduced through processing of noun extraction vectors, technical terms, identification of trend terms, and parsing of meta-information data (Figure 1).

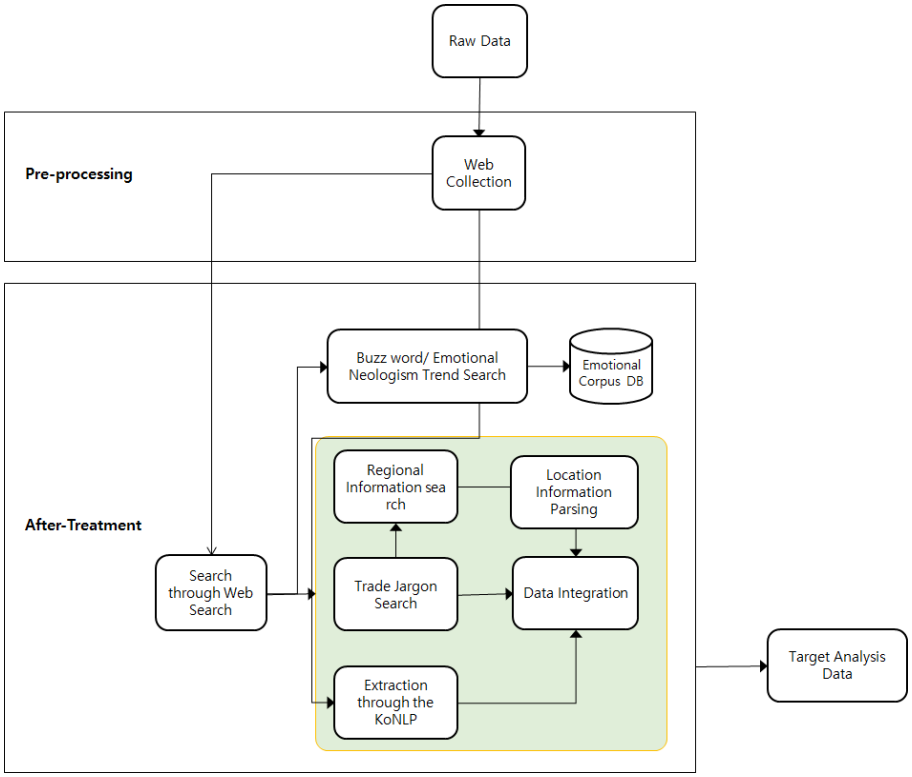

Figure 1. Keyword Extraction by Trend Dictionary Reference

The data collection processing of the website which is made of hypertext can be easily and flexibly performed by understanding of tag types in the website since the required tags or elements are defined to receive parsing through the package of Beautiful Soup. Java-based JSoup is also popular library of the web data collection. Also, data collection using Scrapy is effective which can support a variety of crawls and collection methods.

The followings Figure 2 are the overall data collection methods and pre/post processing plans in this article. The data of specific period are collected in the subject website or web-services to be processed as the queries to trend dictionary progressively dividing by each word. The results by the number of trend dictionaries are performed with parsing, and the original types of nouns and words are collected. If the part of speech is clear, the analysis by the existing dictionary library is performed. Whether the vocabulary is to be registered as the analysis subject in the dictionary corpus is decided upon the searching trend dictionary and analysis of sentimental dictionary although newly developed words, sentimental words, slangs, *etc.* are not registered in the dictionary.

With this referencing to trend dictionary, the word data sets for the analysis are organized and post-processed, if necessary, to enter the analysis step by location data searching in case of meta-data such as location names, *etc.*

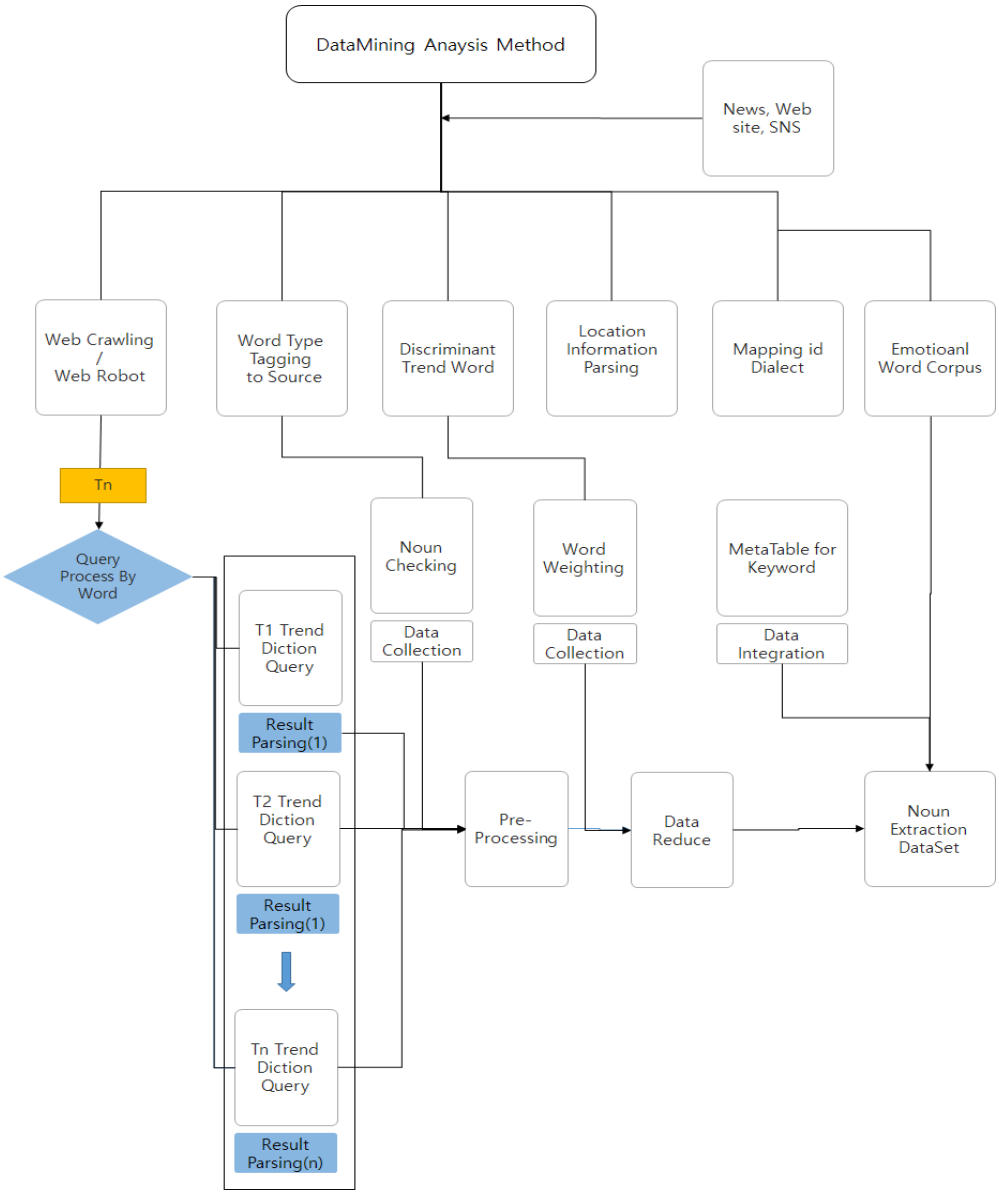

Figure 2. Diagram of the Concept for Sentimental Trend Corpus Expansion

3.3 Trend Word Extraction and Analysis

Generally, it is saved in a text form by collecting as arrangement data by distinguishing links of news, tiles, news contents and news information through differentiating tags of new articles

In order to pre-process the collected news data for each date, searching for trend dictionaries are tried by saving database in text forms gathered generally in a vector space temporarily and querying each vector sequentially. Trend dictionaries processed plural online dictionary service in steps, and was set by dividing the coefficient of classification standard in fixed amounts.

For the classification of buzzwords of dictionary that reflect the partial trend, the coefficient value in each trend dictionary is added in the case of having nouns that are derived normally by applying lambda value in half. And in the case of buzzword and *etc*., the less value is applied. It was added as 0 when there is no value that was searched. When accumulated value exceeded the coefficient value of classification standard, stop

the query and save the noun value of extracted words and then process the next query value.

In addition, in cases of searching for words that are popular recently, confirm the characteristic tag, apply the name of the class when it doesn't exist and apply to the process.

In case of trend word, the morpheme such as noun and *etc.* is not indicated, but when its meaning is deducted or when the cooperated meaning of collective intelligence is found, it is searched as words and it is processed by receiving parsing by confirming the property. The search result of buzzwords in trend dictionaries is reflecting the trend language, but it is not treated as formal nouns. In cases of parsing trend words, the reference of trend dictionary of plurality is proceeded and saving of recommended words are possible for the possible comparison through parsing and saving of words.

The model proposed in this study is treated with query in online trend dictionary by extracting data for each syntactic words saved in vector spaces with n-th noun words respectively. Trend dictionaries refer to the prior database reflecting the trend, and if contents such as Wiki, Daum, Naver dictionary, Google and *etc.* exist, they are calculated by including in threshold value of trend dictionaries.

The threshold value is determined by whether or not words that exist in dictionaries data that are treated in various trend dictionaries present or not. And the frequency status of popularity of words is determined by looking at the referred ratio. The query is terminated according to the threshold value, and if it is not the case, questions for pre-set site are performed repeatedly. If it is expressed as a formula, it is as follows Figure 3.

Lk : a sequential itemset with k items
Ck : a candidate itemset with k items
n: number of trend dictionary
c: number of test data
Tp : processing of query for prototype vector
Tr : check trend dictionary morpheme parsing after query
λ : Classification standard value (threshold)
Tn : serial number of trend dictionary

begin
build L1
for(k=1; Lk $\neq \phi$; k++) do
C_{k+1}= candidate which are generated from Lk
for each $c \in C_{k+1}$
for each contents d in database do
if (n is contained Lk) n.count++
$Tr_{k+1} = c \in T_{k+1}$
if($\frac{t}{n} >$ λ) select Tr
end
return $\bigcup_k L_k$
end

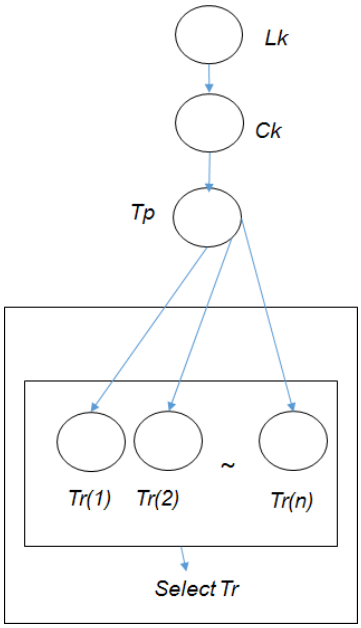

Figure 3. How Reference at Trend Dictionary

The coefficient of classification standard value means to accept the noun extraction if the reply rate of serial trend dictionary exceeds the level. By adjusting coefficient value, it may affect the difference of the time to take searching query.

3.4 Expanded Algorithm of Sentimental Corpus Dictionary

Corpus is the text group with big structure. It is used for the test of rule occurrence within the certain language domain. To analyze this, the effectiveness of corpus or n-gram [9] is investigated by the analyses of morpheme, adjacent relation, array of word order, and so on.

Particularly, sentimental word or text analysis has been positioned as the crucial element in the opinion mining, hence, the establishment of sentimental corpus dictionary to represent positivity and negativity became the important reference model among corpus dictionaries. Sentimental corpus continues to be changed upon the confirmation of the current usages by trendy fashion according to the time, changes of vocabularies, expansion power of word morpheme, *etc.* In case of newly developed words, it is necessary to identify whether they are sentimental words. If that is the case, it is required to add them into the expanded sentimental corpus database by analyzing them based on the evaluation of the existing vocabularies as well as the group intelligence. It enables of the precise analysis in the courses of understanding context and mining by this.

To distinct newly developed sentimental words, they should not have tags as morpheme in the dictionary of the trend dictionary model. With respect to newly developed words, they are confirmed with the sentiment index to perform parsing the relevant corpus or sentence primarily by the services like open Korean language, and secondarily by the queries the parsing sentences to the services of sentiment dictionary to refer to the established existing database (Figure 4).

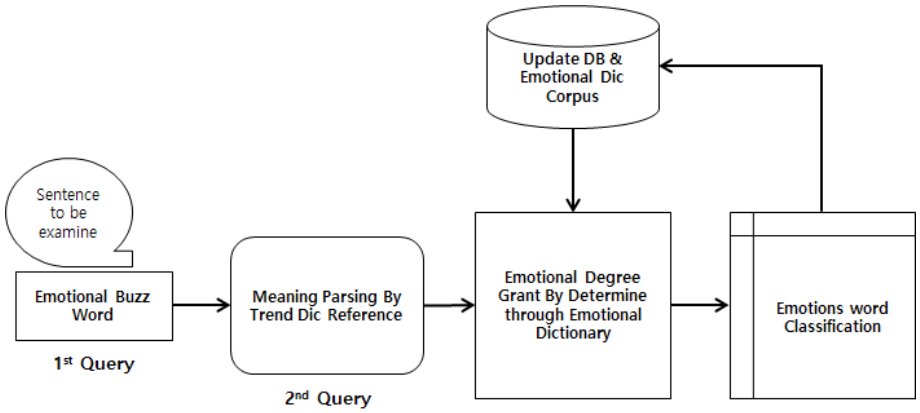

Figure 4. Emotional Corpus Expansion Algorithm

4. Experiments and Evaluation

4.1 Extraction Analysis

Overall, they are stored as the text forms by discriminations of the tags in the news articles to differentiate from news link, title, contents, and news information so as to collect them as the arrayed data. To pre-process the news data collected by dates, the whole text-type database was temporarily stored in the vector space, and tried to search the trend dictionary by queries of each vector with phrases, step by step. Trend dictionary was processed with four steps including Naver Korean language dictionary, Daum dictionary, Korean Wikipedia, and Google dictionary, and the coefficient value of classification standard was set constantly with 0.25~0.75. With respect to the classification of buzzwords which implicate some trends applying half value of lambda, 0.25 of coefficient value was added in each trend dictionary in case of normal parsing nouns, 0.125 in case of buzzwords, and zero in no searching result. If the accumulated coefficient value exceeded the classification standard, the query was hold, the extracted words were stored with noun values, and the next queries were processed.

As the examples of trend dictionary references, website collection was performed on the trendy buzzword which was not registered in the dictionary but fashioned by analyzing the searching website of Naver Korean dictionary with html5 tag. If we search the specific word selecting Naver Korean dictionary as the trend dictionary, we will use the address system of the website searching.

4.2 Evaluation Method of Sentimental Words

Sentimental analysis will be performed through three steps including data collection by opinion mining using machine learning algorithms such as Naive Bayes and Support Vector Machine, Subjectivity Detection, and Polarity Detection. Subjectivity Detection is the process of de-identifications of personal information and name, and additionally the one to exclude the non-related parts with sentiment. Objectivity is excluded upon discriminations as four-area analytics including positivity, negativity, neutrality, and objectivity. In the process of Polarity Detection, the analysis for positivity and negativity is performed to add the weighted value of sentiment on positivity and negativity so as to understand the sentence or overall context. Multi-perspective Question Answering (MPQA) [10] which is sentimental analysis corpus established in the English subjects was referenced using the note language that could show the meanings in sentimental expressions with approximately 10,000 sentences. The selection method of the index for the sentiment words to be occurred additionally in the existing sentence analytic corpus

was mainly heuristic while it is required with the effects of various group intelligence method in case of trendy sentimental language in the practice.

As the methods of sentiment token analysis from developing vector data by phrases or sentences in the subject texts and of automation for positivity and negativity indexes, two methods are possible to use Open Hangeul System Query and R analysis using KoNLP package. It is referred to the database for sentimental analysis such as http://openhangul.com/senti_text and the result will be parsing by queries. To do so, it is to make queries to Open Hangeul with the results from parsing of recommendation by Open Hangeul in trend dictionary. The following Figure 5, python example shows the process of query for trend language which is not classified in the sentiment dictionary but is started to be communicated.

Through the token analysis of the whole sentences, each ratio of positivity and negativity can be found and it can be processed quantitatively with the classifications of neutrality, positivity, or negativity of the word. The analyzed word can be added to the library upon the feedback after multiple confirmations on the existing data.

As an example of trend dictionary reference, the examination analysis on Naver Korean dictionary searching web site was conducted to proceed the web collection regarding HTML5 tag, the generally input vocabulary and trendy vocabulary that is not registered on dictionary but is in trend. The trend dictionary reference also utilized Beautiful Soup library and used python and Java, and also used import.io service when needed. The HTML 5 structure of collection subject sites were examined to collect and process the needed data in accordance with tag element, and because the collected Korean data uses overlap code set with the Greek, the environment is saved in UTF-8 text to adjust the character set.

In order to compare the trend dictionary reference, it was processed with word cloud in qualitative approach and verified the intuitive difference. Also, the frequency analysis verified effectiveness of noun extraction. Then the frequency analysis of extracted noun, independent word and trend vocabulary extracted by inner library and one trend dictionary reference from the Martin Luther King Jr's speech and 180 twitters regarding the 2016 general election was compared with word cloud and verified the order of frequency analysis to conduct experiment.

As a result, twitter comparatively had more colloquial than the speech script by Martin Luther King Jr., along with trend vocabulary, initial sound word, sensibility words. There were more vocabularies extracted from the inner library reference and trend dictionary reference in some level.

```
import requests
from bs4 import BeautifulSoup
f = open("example_Kor.txt", 'r') # Example Query Data
data = f.read()
data = data.split() # Word Separation
n=0
while data[n]!= -1:
q = str(data[n])
n=n+1
# Daum Dic Search : Reference to Trend Dic
url='http://dic.daum.net/search.do?q=' + q + '&dic=kor'
source_code=requests.get(url)
plain_text=source_code.text
soup = BeautifulSoup(plain_text,'lxml')
try:
link = soup.select('.search_word')
link = str(link).replace('<span class="search_word">', "")
link = link.replace('</span>', "")
except IndexError:
pass
print(link)
```

Figure 5. Python Example for Trend Data Reference

Due to varied word order, easy creation of sensitivity words, varied combination of onomatopoeia, and mimetic word in Korean language has led to some loss of noun extract result in dictionary library reference and trend dictionary reference. However, in case of trend dictionary, it showed equal searching effect to Sejong dictionary with the general portal dictionary, while in case of sensitivity word and typing error search, it proposed the revised and recommended word.

The popular trend vocabulary, which has consistently increasing usability, does not simultaneously deduces agreement through the collective intelligence of collaboration, while publishing the tendency explanation. Therefore, the status of analysis subject is granted by referencing the numeral trend dictionary, user value judging the part of it or setting the threshold value.

5. Conclusions

Sentimental analysis which is an opinion mining became more important considering more generated data types in the ubiquitous SNS era. Usually, when we process the texts for Big Data analysis, they are processed by morpheme based on the dictionary which is relatively stable library and by classification of the typical language structure, or by the investigations of similar types. Most of these processes are performed with noun-focused word extractions by the estimation of study results with sematic concept. Based on the Big Data tools which are widely utilized and given dictionary library, they are performed with some value intervention of the investigator. However, the processed data without precise meaning deduction because the required data are incorrectly screened in the processes of the cleaning, transition, and mapping of the analysis subjects; the update is delayed; they are processed based on the dictionary library with different interpretation from the other dictionaries; or newly developed words like sentimental words are emerged so frequently may result in lowering the statistical reliability with the problem claiming certain extent although they deduct the outcomes later based on the precise data analysis and statistics. In this study, the process measures of data collections for mining in

the case of having many emotional language or coined words distributed in SNS or following trends, and the filtering work of improved data in a preprocessor phase were suggested and their functions were compared. They are the complementary thing of the morpheme tag of regular text based on library and the extracting method of independent language. By referring to online trend dictionaries in collecting mining data and effectively parsing processing semi-structured data based on tags of dictionaries according to treated languages, improved data was treated for the analysis. Additionally, there were the cases to show inefficiency in the text processing of the general genre or the limitation of the noun extraction, however, they can be suggested as the alternatives in the class processing on corpus work of sentiment dictionary or trend words which require timeliness.

Acknowledgements

This work was supported by a grant from 2016 Research Funds of Andong National University.

References

[1] S. Yoon, J. Lee and H. Kim, "Social Trend Text Extracting Methods by Trend Dictionary", Asia-pacific Proceedings of Applied Science and Engineering for Better Human Life., vol. 4, **(2016)**, pp. 23-26.
[2] "Bigdata", https://en.wikipedia.org/wiki/bigdata, **(2015)**.
[3] "Information and communications technology", https://en.wikipedia.org/ wiki/Information and communications technology, **(2016)**.
[4] HanKyung Academy, "Management Bigdata Analysis", Society of Digital Policy & Management, **(2014)**.
[5] S. C. Yun, G. H. Nam, S. G. Yang and H. G. Kim, "Big Data analysis and use case using Semantic Technology", The Korea Society of Management information Systems, vol. 29, no. 11, **(2012)**, pp. 524-530.
[6] B. S. Kim, "Analyzing SNS Users' Knowledge Sharing Behaviors in a Big Data Era: A Privacy Calculus Model Perspective", Global e-Business Association, e-Business Studies, vol. 15, no. 1, **(2014)**, pp. 297-315.
[7] K-ICT Big Data Center, "Big data utilizing step-by-step manual business processes and use of technology (Version 1.0)", National Information Society Agency, **(2015)**.
[8] N. H. E. Khalili, B. Haddad and H E. Ghalayini., "Language Engineering for Creating Relevance Corpus", International Journal of Software Engineering and Its Applications, IJSEIA, vol. 9, no. 3, **(2015)**, pp. 107-116.
[9] ITworld, "Read the emotion in the article", Understanding the sensitivity analysis, IDG Korea, **(2014)**.
[10] K. S. Sim, "Syllable-based Korean Morphological Analysis using n-grams extracted from POS Tagged Corpus", Journal of KIISE, Korea Institute of Information Scientists and Engineers, vol. 40, no. 12, **(2013)**, pp. 869-876.

Authors

Sungwook Yoon, received his Doctor of Philosophy in in Information & Communication engineering from Andong National University, Korea. His research area includes interests include Multimedia Contents, Data Mining, Internet of Things.

Hyenki Kim, He received the B.S. and M.S. degree in electronics engineering from Kyungpook National University, Korea, 1986 and 1988 respectively. He received Ph. D. in electronics engineering from Kyungpook National University, Korea, 2000. He joined Andong National University in 2002, where he is currently a professor at Dept. of multimedia engineering in Korea. His research interests include Big Data, mobile app. and Multimedia system and communication.