symmetry

MDPI

Article

Big Data Analysis for Personalized Health Activities: Machine Learning Processing for Automatic Keyword Extraction Approach

Jun-Ho Huh

Department of Software, Catholic University of Pusan, Geumjeong-gu, 57 Oryundae-ro, Busan 46252, Korea; 72networks@pukyong.ac.kr or 72networks@cup.ac.kr; Tel.: +82-510-0662

Received: 14 February 2018; Accepted: 30 March 2018; Published: 2 April 2018

Abstract: The obese population is increasing rapidly due to the change of lifestyle and diet habits. Obesity can cause various complications and is becoming a social disease. Nonetheless, many obese patients are unaware of the medical treatments that are right for them. Although a variety of online and offline obesity management services have been introduced, they are still not enough to attract the attention of users and are not much of help to solve the problem. Obesity healthcare and personalized health activities are the important factors. Since obesity is related to lifestyle habits, eating habits, and interests, I concluded that the big data analysis of these factors could deduce the problem. Therefore, I collected big data by applying the machine learning and crawling method to the unstructured citizen health data in Korea and the search data of Naver, which is a Korean portal company, and Google for keyword analysis for personalized health activities. It visualized the big data using text mining and word cloud. This study collected and analyzed the data concerning the interests related to obesity, change of interest on obesity, and treatment articles. The analysis showed a wide range of seasonal factors according to spring, summer, fall, and winter. It also visualized and completed the process of extracting the keywords appropriate for treatment of abdominal obesity and lower body obesity. The keyword big data analysis technique for personalized health activities proposed in this paper is based on individual's interests, level of interest, and body type. Also, the user interface (UI) that visualizes the big data compatible with Android and Apple iOS. The users can see the data on the app screen. Many graphs and pictures can be seen via menu, and the significant data values are visualized through machine learning. Therefore, I expect that the big data analysis using various keywords specific to a person will result in measures for personalized treatment and health activities.

Keywords: big data; personalized health activities; machine learning (ML); automatic keyword extraction; visualization

1. Introduction

Nowadays, there was significant development in the field of intelligent big data (IBD) analysis where a multicore platform based on a large computing cluster was used. Despite the improvement, too much complex information is still being provided for a single institute or a computing center for processing.

Especially, the number of multimedia and user population will increase continuously and exponentially due to the rapid spread of smartphones and social networking sites.

The obese population is increasing rapidly in Korea due to the change of lifestyle and diet habits. According to the Ministry of Health and Welfare, the prevalence of obesity (over 19 years old, standardization) has increased from 26.0% in 1998 to 29.2% in 2001 and 31.7% in 2007. For the last seven years, it has remained at 31~32%. During the same period, obesity in men increased from 25.1% in 1998 to 36.2% in 2007—up by 11.1% in the past nine years—and remained at around 35~38%. Obesity in women remained at around 25% from 1998 to 2014 [1].

Obesity can cause various complications, and it has become a social disease. Nonetheless, there are many obese patients who have no medical measures that are right for them. Although various online and offline obesity management services are emerging, they are not enough to attract the attention of users and have yet to be helpful in solving the problems.

The emergence of big data due to the spread of digital economy in the 21st century can provide a clue to solving some problems in our society and economy. One of the most valuable uses of big data is the health and fitness industry. The development of IT technology has given birth to a new phase of transformation in the medical field. Note, however, that the introduction of big data in the domestic medical industry has yet to be activated. It is still restricted in terms of actual use due to difficulties in the search and statistics of atypical data. Big data can produce very meaningful results depending on the collection and analysis method of a vast volume of data. The purpose of this study is to visualize big data using text mining and word clouds and to prepare measures, if any, of personalized health activities from more various perspectives.

The big data analysis can visualize a form by collected unstructured data fragments as a puzzle generates a picture by matching scattered pieces. It can show meaningful results depending on which algorithms or techniques it applies. Therefore, this study collected big data using crawling method, visualized big data using text mining and word cloud, and took a machine learning approach to prepare measures for personalized health activities in a variety of viewpoints. Figure 1 shows bird's-eye view of machine learning processing for automatic keyword extraction.

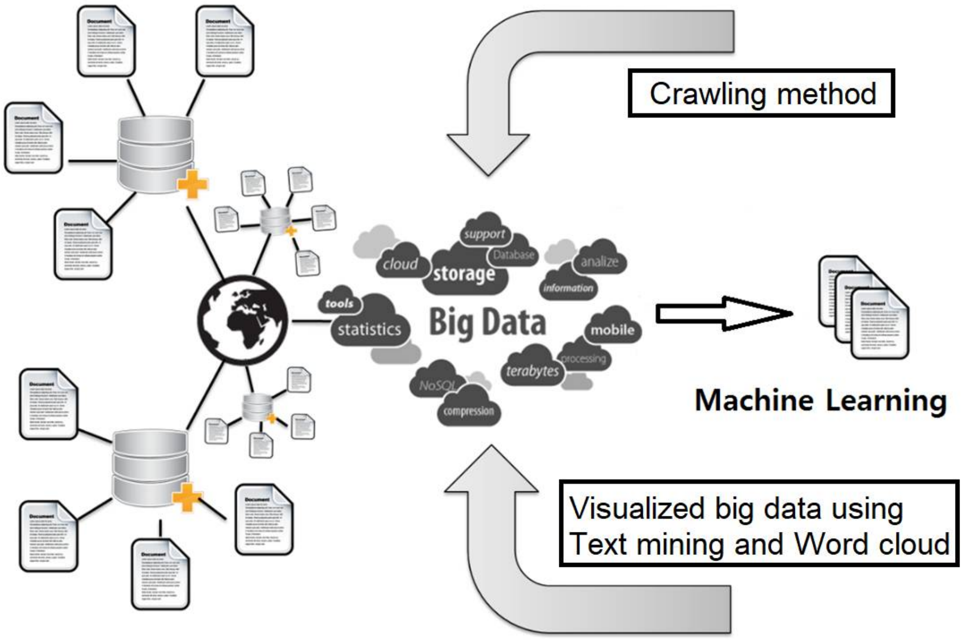


Figure 1. Bird's-eye view of machine learning processing for automatic keyword extraction.

The research team led by Professor Dev Roy at the MIT Media Lab worked with six fact-checking organizations such as PolitiFact and factcheck.org to analyze 126,000 news articles that were categorized as real or fake in 2006–2017 and confirm that the dissemination speed of fake news overwhelmed real news. It published an article in *Science* [2]. Professor Roy and his research team used AI to collect the activity data of 3 million users who directly referred or shared the categorized news and analyzed the dissemination speed and the number of sharing on the network. The analysis result showed that the number of shared fake news was 70% higher than the real news. It meant that fake news spread much more widely than real news. The research team reported that the statistical physical analysis of comments to news showed that disgusting and amazing fake news is shared much faster than real news. The paper explained that "fake news with new and exciting features can be easily transmitted on the Internet and SNS" [2]. Such studies of big data analysis can visualize a form by collected unstructured data fragments as a puzzle generates a picture by matching scattered pieces.

2. Related Works

Accordingly, the market for the big data is becoming larger over time and the data is being used in different areas of our daily lives and much information is shared by the general population. However, since the analysis of big data is very complicated and difficult that sometimes it is quite hard to recognize its meaning and direction, the visualization of big data has come into the picture. Recently, the big data analysis is shifting from SPSS/AMOS to R/TensorFlow.

Machine learning (ML) refers to studying various methods of achieving human-like learning ability through machines, and from the data analysis results, the program can learn about rules or new knowledge by extracting them automatically by itself. The techniques related to machine-learning remaining at the basic level is now becoming more sophisticated due to the emergence of new data mining techniques which can maximize their potential.

Recently, ML is one of the major areas of interest for the artificial intelligence systems, being at the intersection of informatics and statistics and closely related to the data science and knowledge discovery as well as the healthcare industry [3,4]. Especially, probabilistic ML is quite useful for the health informatics where most of the problem-solving process involves removing of uncertainties. The theoretical basis of the probabilistic ML was initially laid by Thomas Bayes (1701–1761) [5]. The probabilistic inference holds a key position in artificial intelligence and statistical learning where the inverse probability allows one to infer unknown facts, deducing them from the available data and making predictions [6,7].

Meanwhile, the scale of big data is much larger than that of the data generated from the analog environment of the past, shorter in generation cycles, and not only the numerical data but the character and image data are included in the big data as well. Since the use of PC, internet, or mobile devices has become part of people's daily routine, the volume of data left behind by them is increasing rapidly.

Along with the fact that the volume of big data has increased explosively, the types of data have been also diversified such that people's behaviors, as well as their thoughts and opinions can be anticipated through positional information and SNS services. Many countries and companies are attempting to construct and utilize the big data system now.

Accordingly, the market for the big data is becoming larger over time and the data is being used in different areas of our daily lives and much information is shared by the general population. However, since the analysis of big data is very complicated and difficult that sometimes it is quite hard to recognize its meaning and direction, the visualization of big data has come into the picture [8].

Wu et al. [9] have argued that a large volume of data (big data) can be problematic when frequent itemset mining has been used for the following reasons: (1) spatial complexity: the algorithm may not be run as the system memory deal with a large input data as well as large intermediate results and output pattern; (2) time complexity: many existing approaches depend on an exhaustive search or a complicated data structure to obtain a frequent pattern but this is not suitable for big data. Thus, they proposed an iterative sampling-based frequent itemset mining that samples the subsets instead of processing entire dataset all together and then extracting the frequent itemsets from them.

Yang Luo et al. [10] maintained that segmenting the Left Ventricle (LV) from the cardiac MRI image is essential when computing the clinical indices such as stroke volume, ejection fraction, etc. Thus, in this study, an automated LV segmenting method where the hierarchical extreme learning machine (H-ELM) is combined with a new location recognition method is proposed.

Ghadah Aldehim [11] claimed that using all the data for the feature selection, which is becoming increasingly important in big data analysis and machine-learning, may lead to a selection bias while using the partial data could lead to an underestimation the relevant features under some conditions. Thus, a research on the method with which can decide a suitable method for a specific dataset in terms of reliability and effectiveness is being introduced in this study. Also, Tri Doan et al. [12] maintained that selecting an appropriate categorization (classification) algorithm is a very important step in all the data mining procedures. The run-time is used to assess the efficiency of a categorization algorithm. In this study, a method that is helpful in finding an adequate algorithm in terms of efficiency has been studied by introducing a tool that estimates the run-time of a particular categorization algorithm used for a dataset based on the concept referred as meta-learning.

Meanwhile, Junhai et al. [13] proposed an algorithm having a higher performance level (i.e, in G-mean) than other existing ensemble algorithms in terms of speed and scalability to effectively categorize the imbalanced data into two classes.

A number of research directions are recognized [14,15]. First, sentimental classification which classifies the contents in relation to the sentiments involving the opinion targets. Second, feature (aspect)-based opinion mining that analyzes the sentiment towards certain characteristics of an object. The examples of this can be found in [16,17]. Third, comparison-based opinion mining focuses on the text where similar objects are being compared [18]. It is essential that the opinion mining methods are identified with three individual levels: document, sentence, or entity/aspect levels but most of the classification methods depend on identifying the opinion words or phrases involved. Also, their basic algorithms are categorized as (1) supervised learning of which can be found in [19,20]; (2) unsupervised learning as described in [21]; (3) partially supervised learning illustrated in [22]; and (4) other approaches using the algorithms that use some of the latent variable models such as hidden markov model (HMM) [23], conditional random fields (CRF) [24], latent semantic association (LSA) [25], or pointwise mutual information (PMI) [26]. For these varying techniques, a number of researchers had experimented them with a series of different algorithms and made comparisons [27–29].

A few research works have clearly focused on Web 2.0. In that case, while many of research had dealt with weblogs [30–33] mainly investigating the correlation between blog posts and 'real-life' situations, a few other researchers evaluated the techniques used for the opinion mining targeting the context of weblogs such that the main trend in the mining technique has not been identified or suggested. Liu et al. made a comparison between varying linguistic features when classifying the blog sentiment [34]. Some of the experimental studies were made with lexical and sentimental features using separate learning algorithms to identify the opinionated blogs [35]. It is quite interesting that there are so little research has been conducted about the opinion mining in the field of "Discussion Forums" [36,37] whereas quite a number of researchers carried out their research focusing on the microblogs (for Twitter, especially) and published the papers on them [38–42].

For the opinion mining of microblogs, the researchers primarily adopt the supervised or the semi-supervised learning technique for the microblogs. Contrary to the rapid spread of social network services led by Facebook and Twitters, the number of research concerning the opinion mining in the social networks is not enough [43,44]. Although there have been quite a number of research works published in the past decade concerning product reviews, it seems very little works for determining the most effective opinion mining technique were introduced. It is evident that most of the researchers are using the tex classification algorithms such as SVM, naïve Bayes, or a combination of different methods to enhance the reliability of opinion mining results but one of the most encouraging mining technique would be LDA [45,46]. The LDA-based model that identifies both aspects and sentiments together is proposed in [47]. Such a model also described in [48,49] stands on the premise that entire words used in a sentence are relevant to a single topic. Gerald Petz introduced his research in the paper titled "Opinion Mining on Characteristics of User Generated Content and Their Impacts" [50–59].

2.1. National Health Information Data

Among the data held by National Health Insurance Corporation according to the Korean Government 3.0 Policy, the national health information data is public data that is open to the public and which enjoys high demand from the private sector. National health information data are "medical history information" and "prescription medicine information and health examination information" of national health insurance subscribers accumulated by the corporation as the corporation serves the role of national health insurance data provider. In order to open safe data, the corporation excluded or masked personal information and sensitivity data, and the target date of data provided is from 2002 to 2016. It is planning to expand the target period continually in the future.

2.2. National Health Data Selection Criteria

Table 1 shows national health data selection criteria.

Table 1. National Health Data Selection Criteria.

Category	Contents
Extraction of sample	Randomized extraction of 1 million patients each year
Limitations on combination by resources	Individual serial number of each DB and serial number of charge differing according to sectional data
Removal of personal identifier	Resident registration number → personal serial number (8 digits)
Categorization	Age grouping, Age → Age group (by 5 years), 85 years or older categorized as "85 years or older"
Data masking	Sensitive disease D, O, P, X, and Y code (5 types and 114 kinds)
Top-level local code provision	Provided only for try codes (17 units offered) considering the recognition of samples in small areas

2.3. Excluding Personally Identifiable Information

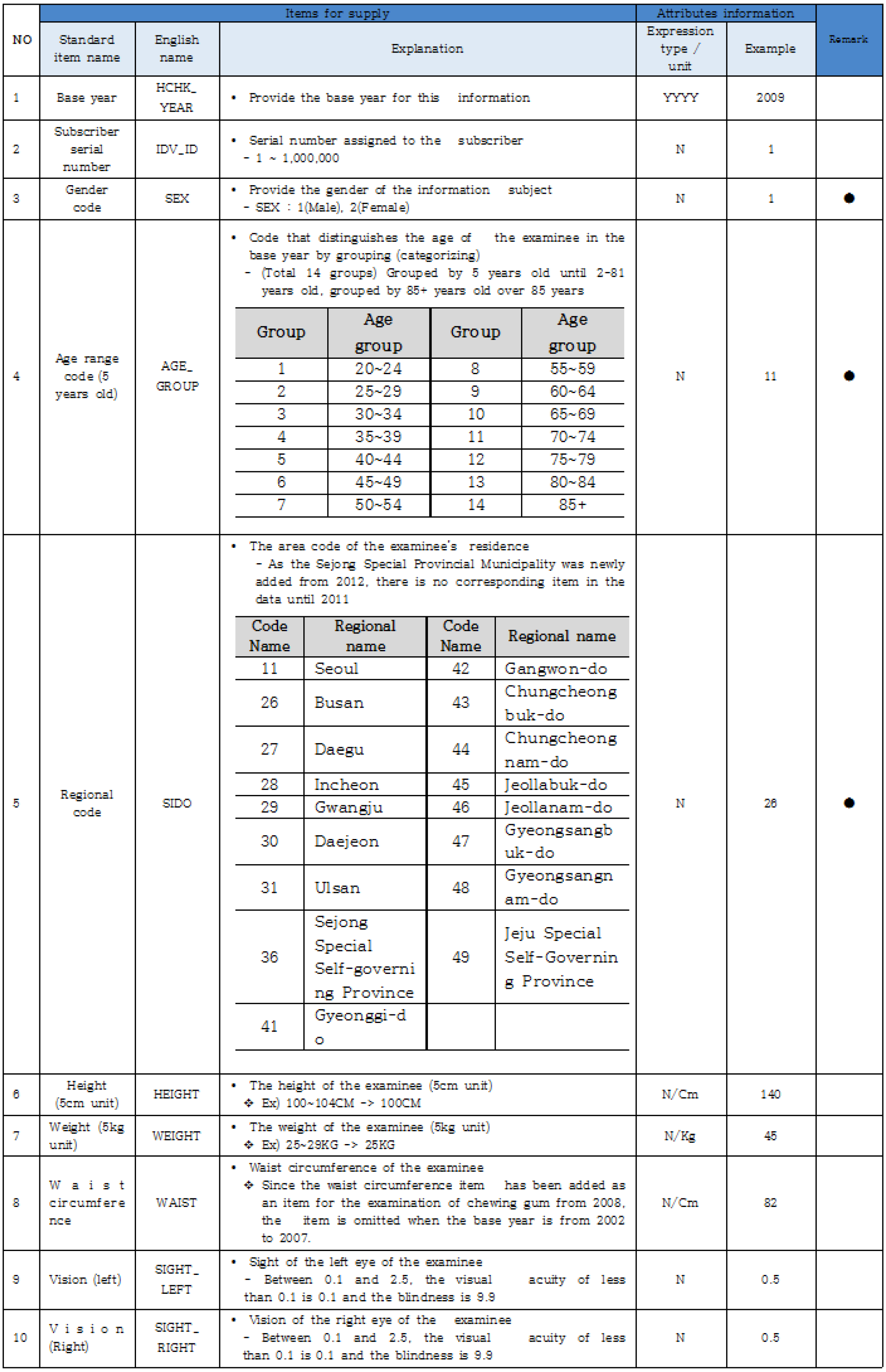
Personal identification information (resident registration number, national health insurance subscriber number, etc.) and easily identifiable information (name, telephone number, address, photograph, etc.) are excluded from the national health information data.

2.4. Applying the Personal Information Non-Identifying Processing Technique

Data whose re-identification is possible is excluded from the opening through the prior filtering of identifiable information by combining with other information. The possibility of identification was excluded by applying the non-identifying processing technique suitable for individual items.

2.5. Health Checkup Information

Health checkup information is obtained by randomly selecting one million Korean national health insurance subscribers who had health checkups in the year; the basic information of the selected subscribers after the item selection process and the examination result information were then extracted. The data are organized by year, with one million data and 34 attributes per year. Figure 2 shows survey of Korean national health information data.



<table>
<tr><th rowspan="2">NO</th><th colspan="3">Items for supply</th><th colspan="2">Attributes information</th><th rowspan="2">Remark</th></tr>
<tr><th>Standard item name</th><th>English name</th><th>Explanation</th><th>Expression type / unit</th><th>Example</th></tr>
<tr><td>1</td><td>Base year</td><td>HCHK_ YEAR</td><td>• Provide the base year for this information</td><td>YYYY</td><td>2009</td><td></td></tr>
<tr><td>2</td><td>Subscriber serial number</td><td>IDV_ID</td><td>• Serial number assigned to the subscriber
- 1 ~ 1,000,000</td><td>N</td><td>1</td><td></td></tr>
<tr><td>3</td><td>Gender code</td><td>SEX</td><td>• Provide the gender of the information subject
- SEX : 1(Male), 2(Female)</td><td>N</td><td>1</td><td>●</td></tr>
<tr><td>4</td><td>Age range code (5 years old)</td><td>AGE_ GROUP</td><td>• Code that distinguishes the age of the examinee in the base year by grouping (categorizing)
- (Total 14 groups) Grouped by 5 years old until 2-81 years old, grouped by 85+ years old over 85 years
<table>
<tr><th>Group</th><th>Age group</th><th>Group</th><th>Age group</th></tr>
<tr><td>1</td><td>20~24</td><td>8</td><td>55~59</td></tr>
<tr><td>2</td><td>25~29</td><td>9</td><td>60~64</td></tr>
<tr><td>3</td><td>30~34</td><td>10</td><td>65~69</td></tr>
<tr><td>4</td><td>35~39</td><td>11</td><td>70~74</td></tr>
<tr><td>5</td><td>40~44</td><td>12</td><td>75~79</td></tr>
<tr><td>6</td><td>45~49</td><td>13</td><td>80~84</td></tr>
<tr><td>7</td><td>50~54</td><td>14</td><td>85+</td></tr>
</table></td><td>N</td><td>11</td><td>●</td></tr>
<tr><td>5</td><td>Regional code</td><td>SIDO</td><td>• The area code of the examinee's residence
- As the Sejong Special Provincial Municipality was newly added from 2012, there is no corresponding item in the data until 2011
<table>
<tr><th>Code Name</th><th>Regional name</th><th>Code Name</th><th>Regional name</th></tr>
<tr><td>11</td><td>Seoul</td><td>42</td><td>Gangwon-do</td></tr>
<tr><td>26</td><td>Busan</td><td>43</td><td>Chungcheongbuk-do</td></tr>
<tr><td>27</td><td>Daegu</td><td>44</td><td>Chungcheongnam-do</td></tr>
<tr><td>28</td><td>Incheon</td><td>45</td><td>Jeollabuk-do</td></tr>
<tr><td>29</td><td>Gwangju</td><td>46</td><td>Jeollanam-do</td></tr>
<tr><td>30</td><td>Daejeon</td><td>47</td><td>Gyeongsangbuk-do</td></tr>
<tr><td>31</td><td>Ulsan</td><td>48</td><td>Gyeongsangnam-do</td></tr>
<tr><td>36</td><td>Sejong Special Self-governing Province</td><td>49</td><td>Jeju Special Self-Governing Province</td></tr>
<tr><td>41</td><td>Gyeonggi-do</td><td></td><td></td></tr>
</table></td><td>N</td><td>26</td><td>●</td></tr>
<tr><td>6</td><td>Height (5cm unit)</td><td>HEIGHT</td><td>• The height of the examinee (5cm unit)
❖ Ex) 100~104CM -> 100CM</td><td>N/Cm</td><td>140</td><td></td></tr>
<tr><td>7</td><td>Weight (5kg unit)</td><td>WEIGHT</td><td>• The weight of the examinee (5kg unit)
❖ Ex) 25~29KG -> 25KG</td><td>N/Kg</td><td>45</td><td></td></tr>
<tr><td>8</td><td>Waist circumference</td><td>WAIST</td><td>• Waist circumference of the examinee
❖ Since the waist circumference item has been added as an item for the examination of chewing gum from 2008, the item is omitted when the base year is from 2002 to 2007.</td><td>N/Cm</td><td>82</td><td></td></tr>
<tr><td>9</td><td>Vision (left)</td><td>SIGHT_ LEFT</td><td>• Sight of the left eye of the examinee
- Between 0.1 and 2.5, the visual acuity of less than 0.1 is 0.1 and the blindness is 9.9</td><td>N</td><td>0.5</td><td></td></tr>
<tr><td>10</td><td>Vision (Right)</td><td>SIGHT_ RIGHT</td><td>• Vision of the right eye of the examinee
- Between 0.1 and 2.5, the visual acuity of less than 0.1 is 0.1 and the blindness is 9.9</td><td>N</td><td>0.5</td><td></td></tr>
</table>

Figure 2. Questionnaire for Citizen Health Data in Korean [1].

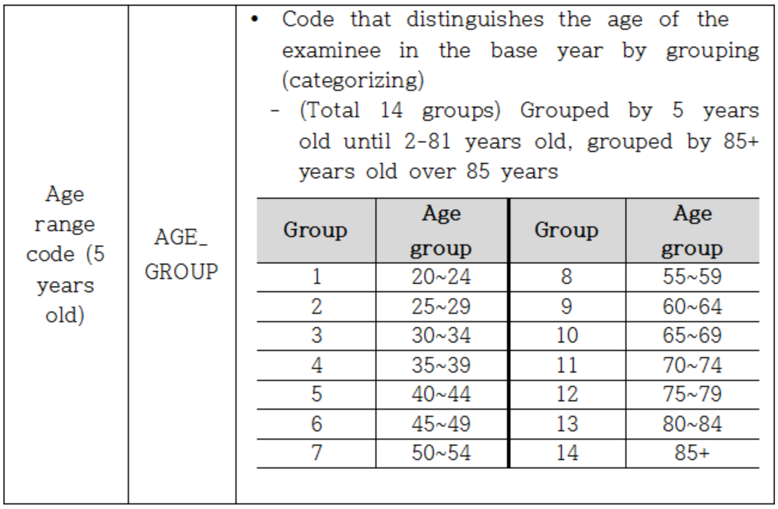
3. A Big Data Analysis Method for Personalized Health Activities

3.1. Data Analysis

Aggregating data into a frequency table can show the overall characteristics better than raw data. However, a person who is weak in numbers may not get any meaningful idea from the table. Therefore, I use a graph called histogram to show the full data more intuitively. A histogram is a bar chart with the class interval of the frequency distribution table on the horizontal axis and the frequency on the vertical axis. I used R Studio because it was the most appropriate big data analysis tool for our data.

With R Studio, there were 542,321 men and 457,679 women in the total data of 1 million people in the "Health Examination Information Data Set". First, the body information of males and females were checked, and each distribution was compared using histograms.

The health screening information data age group code is shown in Figure 3. The age group codes of the national health information data set are grouped by age group (categorized by age 5). Therefore, by using the subset () function of R Studio, the data were categorized according to age. As a result, in the analysis of height by age, the data were collected by categorizing into 6 groups—20–24 years, 25–29 years, 30–34 years, 35–39 years, 40–44 years, and 45–49 years—and the mean and standard deviation of height were determined. Table 2 shows data categorization by age and mean and standard deviation of height.



Age range code (5 years old)

AGE_GROUP

- Code that distinguishes the age of the examinee in the base year by grouping (categorizing)
  - (Total 14 groups) Grouped by 5 years old until 2-81 years old, grouped by 85+ years old over 85 years

Group	Age group	Group	Age group
1	20~24	8	55~59
2	25~29	9	60~64
3	30~34	10	65~69
4	35~39	11	70~74
5	40~44	12	75~79
6	45~49	13	80~84
7	50~54	14	85+

Figure 3. Code and Data Parsing for each Age Group of Health Examination Data.

Table 2. Data Categorization by Age and Mean and Standard Deviation of Height.

Age Group	**Male**		**Female**	
	M(cm)	**SD**	**F(cm)**	**SD**
20~24	171.2276	6.087624	159.6198	5.377917
25~29	172.1218	5.981225	159.7551	5.378908
30~34	172.2063	5.879721	159.3983	5.412011
34~39	5.879721	5.792054	158.3623	5.401406
40~44	169.9034	5.799039	157.0913	5.385961
45~49	168.4616	5.667555	155.733	5.28603

In the next place, the study compared the height of the whole age group according to gender by the histograms in Figure 4. In males, height was most distributed at 165–170 (cm), with females' height most distributed at 150~155 (cm). The histogram of females showed a symmetrical distribution of the population according to height, but the male histogram had a long tail to the left. To better understand the difference in height between male and female, the distribution of gender by combining histograms was analyzed. The distribution of male and female histograms according to height shows that the proportion of males increases with increasing height, whereas the proportion of females increases with decreasing height.

The results of big data analysis as shown as the histogram of weight in Figure 5, indicate that the portion of males increased as the weight increased while the portion of females increased when the weight decreased. It shows the similar pattern as the height histogram depicting the difference of heights of males and females, and thus the analysis results were successfully visualized.

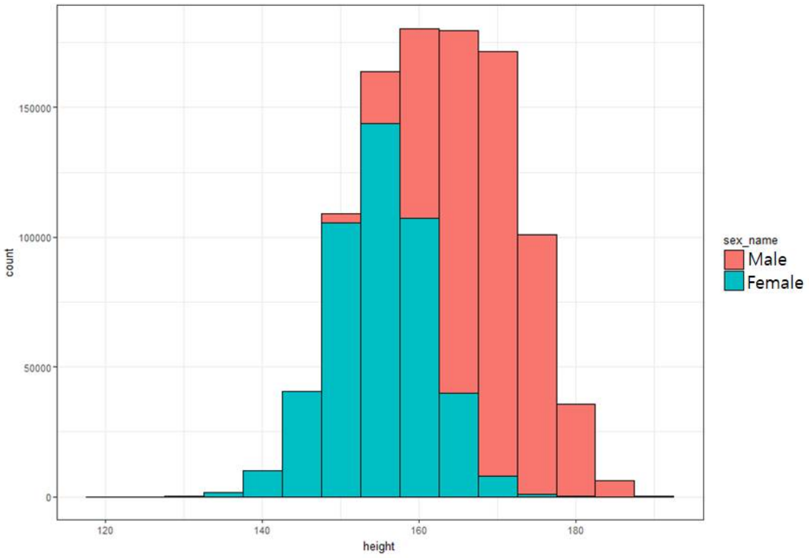


Figure 4. Male and Female's Height Histogram.

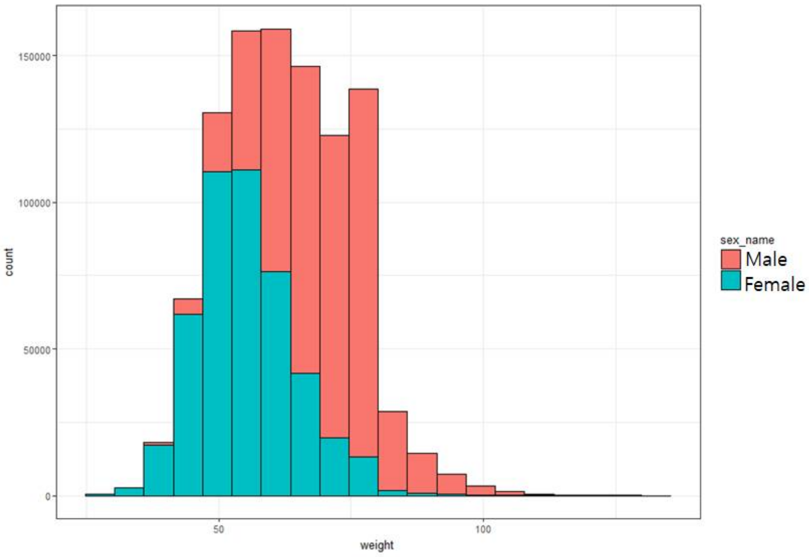


Figure 5. Male and Female's Weight Histogram.

Next, a body mass index (BMI) column is added to the data set to check for obesity. The BMI value of each row is the weight (kg) divided by the square of height (m) (body weight (kg)/height (m^2)). At this time, the classification of obesity by body index is defined as 'underweight' when BMI is less than 18.5, 'normal' when BMI is 18.5~22.9, and 'overweight' when BMI is 23~24.9. The distribution of BMI according to gender is as follows:

In this case, the obesity distribution can be seen as a kind of normal distribution as shown in Figure 6. This analysis can be influenced by the difference in the observed number of obese. Therefore, the distribution of obesity (BMI ≥ 25) ≥ on body weight was analyzed. Also, Table 3 shows data categorization by age and mean and standard deviation of height.

Table 3. Data Categorization by Age and Mean and Standard Deviation of Height.

Age Group	Male		Female	
	Weight (kg)	SD	Weight (kg)	SD
20~24	67.3603	11.01164	52.9281	9.047276
25~29	70.83247	11.73332	52.86799	8.80432
30~34	72.7386	11.64975	54.26461	9.46236
34~39	71.89527	11.0495	54.9672	9.108493
40~44	70.66983	10.39266	55.53917	8.915136
45~49	69.09484	9.692245	55.96567	8.463123

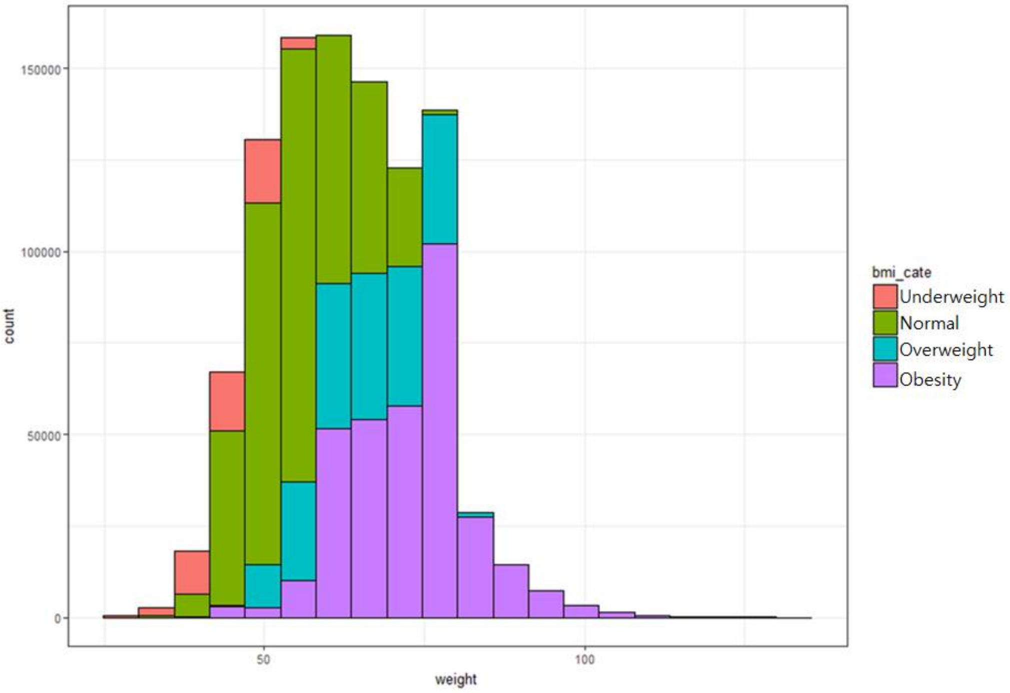


Figure 6. BMI Histogram of Male and Female.

As a result, the obesity distribution graph according to body weight in Figure 7 shows that the normal weight, overweight, and obesity appear evenly between 60 and 70 Kg in body weight, however, the proportion of obesity increased rapidly. The proportion of BMI for the total population of the dataset is shown below. Figure 8 shows the distribution of obesity according to body weight. Table 4 shows weight of the BMI in national health information data set.

As shown in Table 5, the big data analysis of the distribution of obesity according to body weight revealed that people with a body weight of 25–40 kg tended to be underweight while those with a body weight of 40–70 kg tended to be normal. The overweight people were evenly distributed up to 60–80 kg while 100% of people weighing more than 90 kg were obese. The obesity distribution graph according to weight as shown in Figure 9 visualizes such change.

The next section is machine learning to create a suitable learning model through a large number of data of one million people and to discover useful information.

Table 4. Weight of the BMI in National Health Information Data Set.

Under Weight	Normal	Overweight	Obesity
5.1026	41.9603	19.3267	33.6104

Table 5. Distribution of Obesity According to Body Weight.

Weight	Under Weight	Normal	Overweight	Obesity
25	100.00	0.00	0.00	0.00
30	98.00	2.00	0.00	0.00
35	85.81	14.19	0.00	0.00
40	65.62	33.90	0.46	0.02
45	23.75	71.29	0.70	4.27
50	13.29	75.75	8.84	2.12
55	1.94	74.70	16.99	6.37
60	0.00	42.70	24.85	32.40
65	0.00	35.74	27.26	36.99
70	0.00	21.83	31.18	46.99
75	0.00	1.17	32.08	66.75
80	0.00	0.09	15.02	84.90
85	0.00	0.00	3.75	96.25
90	0.00	0.00	0.00	100.00
95	0.00	0.00	0.00	100.00
100	0.00	0.00	0.00	100.00
105	0.00	0.00	0.00	100.00
110	0.00	0.00	0.00	100.00
115	0.00	0.00	0.00	100.00
120	0.00	0.00	0.00	100.00
125	0.00	0.00	0.00	100.00
130	0.00	0.00	0.00	100.00

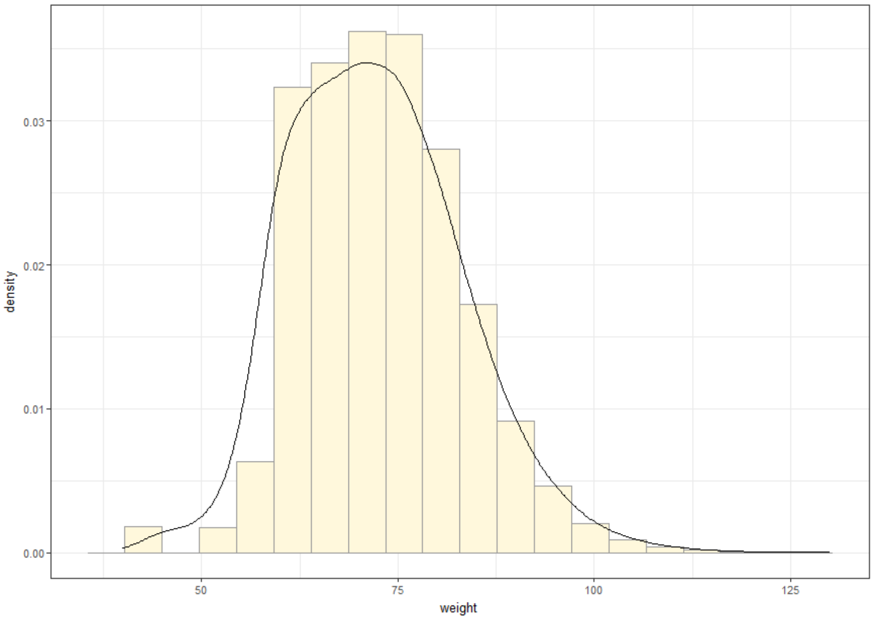
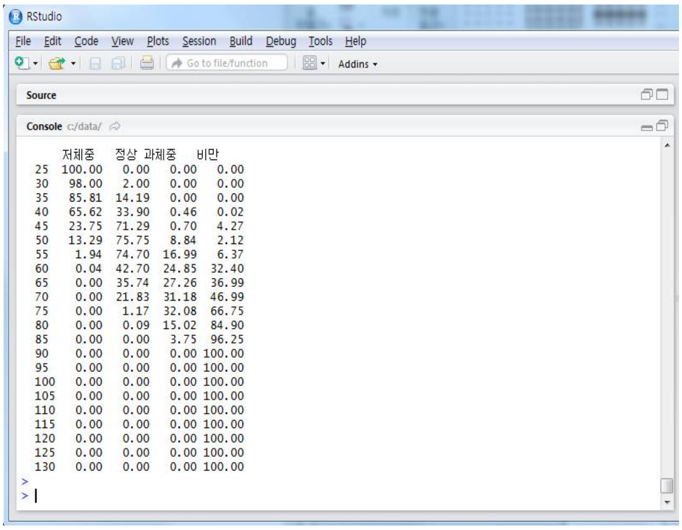


Figure 7. Obesity Distribution According to Body Weight.



```
      저체중   정상 과체중    비만
25  100.00   0.00   0.00   0.00
30   98.00   2.00   0.00   0.00
35   85.81  14.19   0.00   0.00
40   65.62  33.90   0.46   0.02
45   23.75  71.29   0.70   4.27
50   13.29  75.75   8.84   2.12
55    1.94  74.70  16.99   6.37
60    0.04  42.70  24.85  32.40
65    0.00  35.74  27.26  36.99
70    0.00  21.83  31.18  46.99
75    0.00   1.17  32.08  66.75
80    0.00   0.09  15.02  84.90
85    0.00   0.00   3.75  96.25
90    0.00   0.00   0.00 100.00
95    0.00   0.00   0.00 100.00
100   0.00   0.00   0.00 100.00
105   0.00   0.00   0.00 100.00
110   0.00   0.00   0.00 100.00
115   0.00   0.00   0.00 100.00
120   0.00   0.00   0.00 100.00
125   0.00   0.00   0.00 100.00
130   0.00   0.00   0.00 100.00
>
> |
```

Figure 8. Obesity Distribution by Body Weight Confirmed by R Studio.

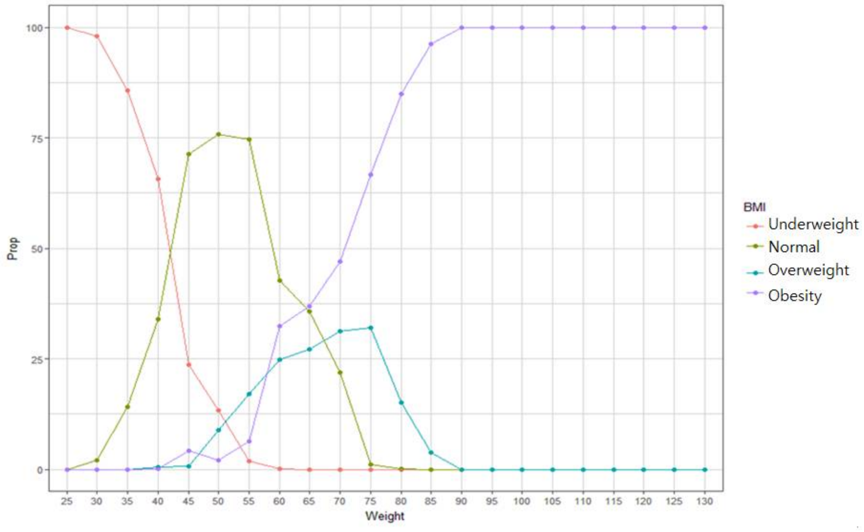


Figure 9. Obesity Distribution Graph According to Weight.

3.2. Machine Learning

Machine learning involves the study of various methods of implementing human-like learning abilities through a machine. It analyzes the given data and automatically extracts rules or new knowledge that can be learned from the analyzed results and aims to get the effect that the machine learns. Techniques related to machine learning have remained at the basic level, but they are becoming more feasible due to the emergence of a large number of big data that can maximize the potential of machine learning techniques.

In particular, the regression technique of machine learning differs from other machine learning techniques since there are a number of techniques that can be applied to a task beyond one algorithm. There are various techniques such as linear regression using one independent variable, multiple regression using two or more independent variables, and logistic regression used to model binary categorical results. The same basic principles apply to all regression techniques.

The “health checkup information data set” [1] includes various health checkup results in addition to the obesity-related variables above. To investigate the effect of these variables on blood pressure, a regression technique that predicts the numerical data is used, and an appropriate model is created to analyze the correlation.

The basic installation of R Studio does not include machine learning. In order to use the machine learning algorithm implemented in R Studio, R Weka package, class package, and stats package were installed and analyzed using the 'install.packages' () function.

To apply machine learning to data, the library () function is used to load the package. Initially, a scatter matrix of "weight", "total cholesterol", "BMI", "systolic blood pressure", and "diastolic blood pressure" variables are created to visualize the relationship between major properties.

Here, the ellipse on the scatter chart shows how strong the correlation is with the correlation ellipse. "Weight" and "BMI" mean that the correlation is strong when they are extended to an ellipse. If the circle shape such as "weight" and "total cholesterol" is strong, it means weaker correlation. In the next place, "smoke" and "drink" are generated by using the ifelse () function to compare blood pressure according to smoking status and alcohol consumption.

Meanwhile, as shown in Figures 10 and 11, the "systolic blood pressure (bp_high)" and linear regression model associated with ten variables are fitted. In this case, b_data is new data created by adding the variables required for a_data used in the preceding data analysis. After creating the model, enter the object name of the model to check the regression coefficient.

The estimated regression coefficient shown in Figure 12 suggests how much bp_high (systolic blood pressure) increases for each attribute when the other attributes remain constant. Bp_high (systolic blood pressure) increases by 0.83 when age_group (age group code) is increased by 1 with other values held constant. The values of height, weight, blds, and tot_chole (total cholesterol) showed much lower values, indicating that blood pressure is very difficult to explain. bp_lwst (diastolic blood pressure) is similar to systolic blood pressure, and women have a mean blood pressure that is 0.98 point lower than men. Likewise, BMI (Body Mass Index), drink (drinking alcohol), and smoke (smoking) were more closely related compared with other variables. Next, the performance of the model is evaluated by the summary () command.

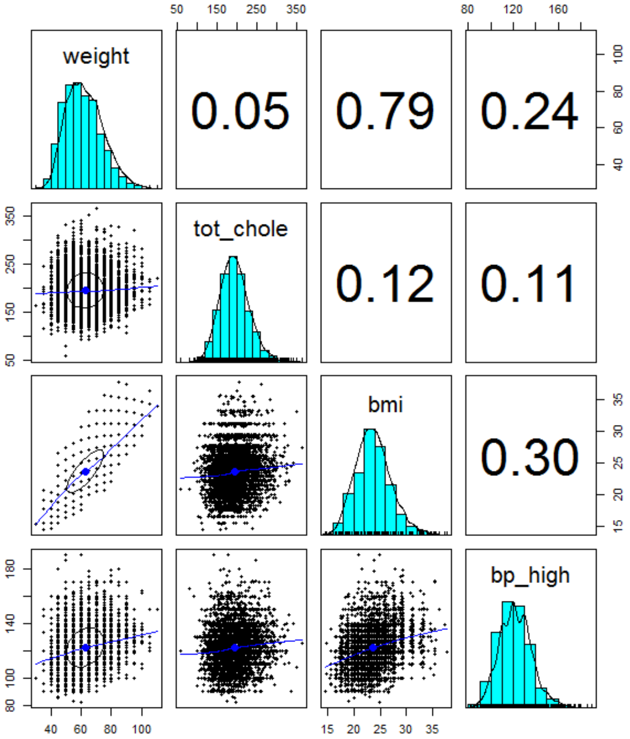


Figure 10. Systolic Blood Pressure and Dispersion of Variables.

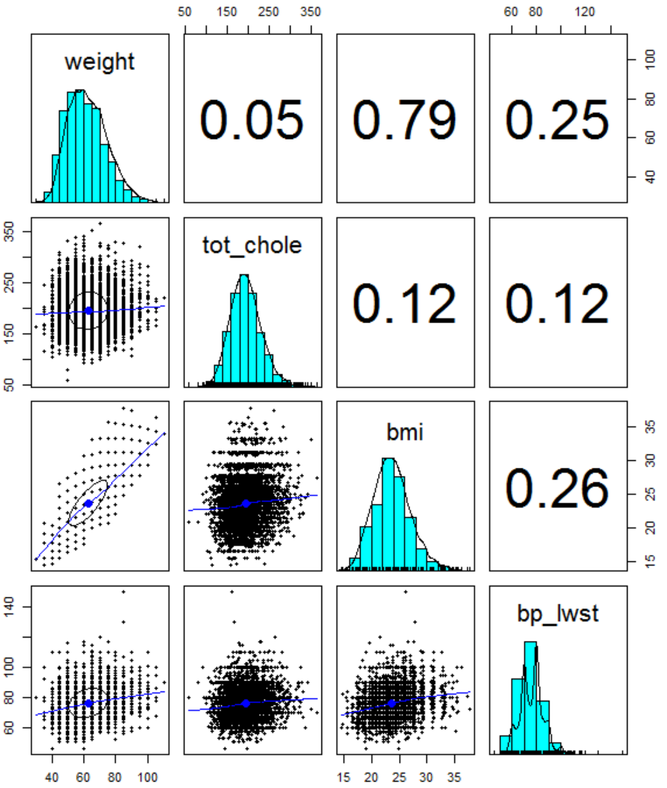
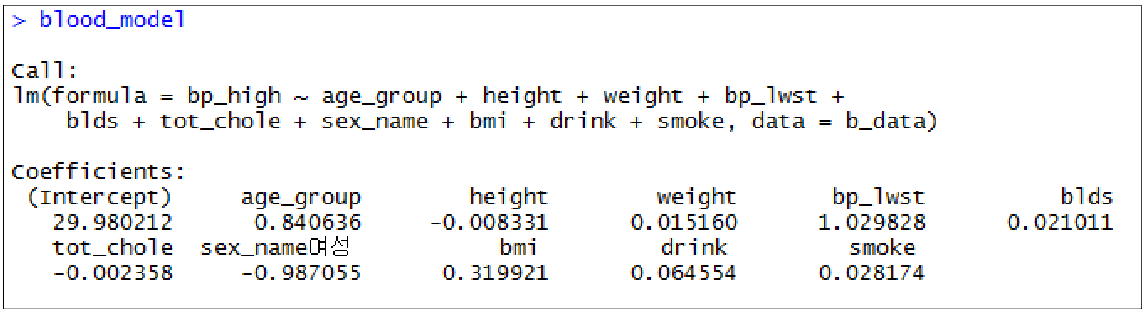


Figure 11. Diastolic Blood Pressure and Dispersion of Variables.



```
> blood_model

Call:
lm(formula = bp_high ~ age_group + height + weight + bp_lwst +
    blds + tot_chole + sex_name + bmi + drink + smoke, data = b_data)

Coefficients:
 (Intercept)      age_group         height         weight        bp_lwst           blds
   29.980212       0.840636      -0.008331       0.015160       1.029828       0.021011
   tot_chole    sex_name여성            bmi          drink          smoke
   -0.002358      -0.987055       0.319921       0.064554       0.028174
```

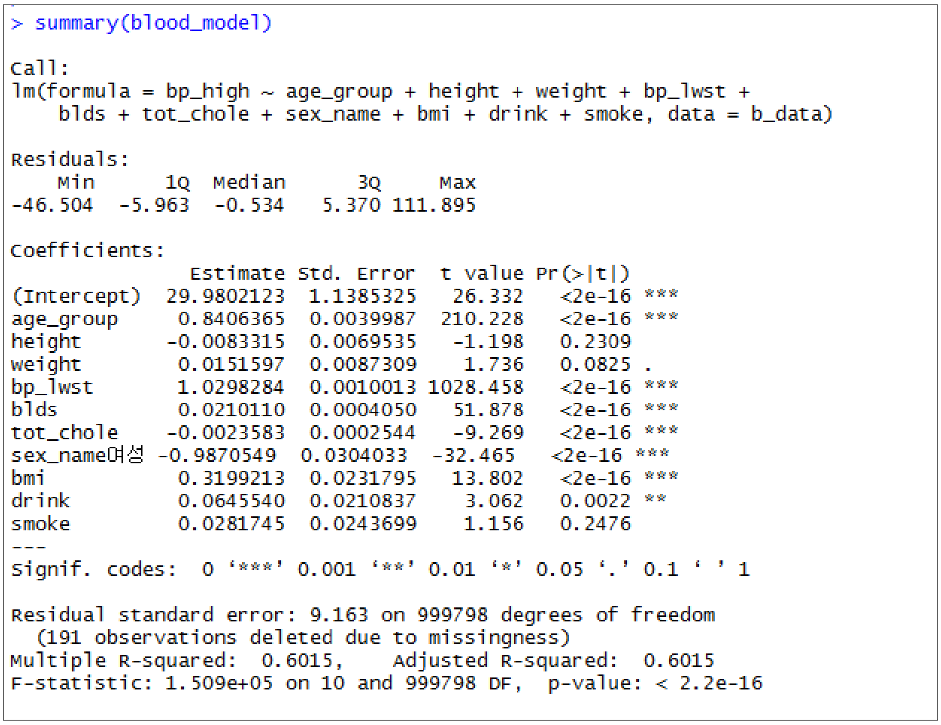
Figure 12. The Regression Coefficient of the 'Blood_model' Model Object.

As shown in Figure 13, the Residuals section provides summary statistics for the error. The maximum error of 111.895 indicates that the model has a difference of at least one predicted value in at least one example. The value of Multiple R-Squared indicates how well the model describes the value of the dependent variable. Like the correlation coefficient, if the value approaches 1.0, the model fully explains the data.

0.6015 as the value of R-Squared means that this 'blood_model' model can account for 60% of the dependent variable. A model with more attributes can provide higher values.

While the size of the error is part of the consideration, the regression model 'blood_model' has a value of 0.6015 and appears to work substantially well.

Next, this study analyzed not only the data analysis of the dataset but also the various keyword trends that have been publicized through the media in the meantime and examined the various problems and interest trends of national health to provide various kinds of personalized information.



```
> summary(blood_model)

Call:
lm(formula = bp_high ~ age_group + height + weight + bp_lwst +
    blds + tot_chole + sex_name + bmi + drink + smoke, data = b_data)

Residuals:
    Min      1Q  Median      3Q     Max
-46.504  -5.963  -0.534   5.370 111.895

Coefficients:
               Estimate Std. Error  t value Pr(>|t|)
(Intercept)  29.9802123  1.1385325   26.332   <2e-16 ***
age_group     0.8406365  0.0039987  210.228   <2e-16 ***
height       -0.0083315  0.0069535   -1.198   0.2309
weight        0.0151597  0.0087309    1.736   0.0825 .
bp_lwst       1.0298284  0.0010013 1028.458   <2e-16 ***
blds          0.0210110  0.0004050   51.878   <2e-16 ***
tot_chole    -0.0023583  0.0002544   -9.269   <2e-16 ***
sex_name여성 -0.9870549  0.0304033  -32.465   <2e-16 ***
bmi           0.3199213  0.0231795   13.802   <2e-16 ***
drink         0.0645540  0.0210837    3.062   0.0022 **
smoke         0.0281745  0.0243699    1.156   0.2476
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.163 on 999798 degrees of freedom
  (191 observations deleted due to missingness)
Multiple R-squared:  0.6015,    Adjusted R-squared:  0.6015
F-statistic: 1.509e+05 on 10 and 999798 DF,  p-value: < 2.2e-16
```

Figure 13. Performance Evaluation of ‘Blood_model’ Model.

3.3. Keyword Analysis Method

First, this study selected *Naver News*, an Internet newspaper, to analyze the field of interest in public health. To collect Internet news related to obesity over the last 1 year, the crawling technique was used, and the frequency of embedded words was analyzed. In addition, big data-based services such as Naver Trend and Google Trends showing the keyword trends in real time were used. As in the previous dataset analysis, web crawling for Internet news was done using statistical program R, and text mining was performed to find meaningful information of embedded words.

3.3.1. Text Mining

Text mining is a technique for extracting and processing important information such as the patterns, trends, and distributions of the text by analyzing the unstructured texts. With the recent availability of big data, interest in large-capacity text analysis technology has increased, and the importance of text mining technology is emphasized. Text mining basically expresses unstructured/semi-structured data as a simplified model.

While the purpose of a data mining is to draw useful and potential patterns from the structured data, text mining is the process of discovering new knowledge from a large unstructured textual group composed of natural language. In other words, text mining is a process of finding interesting and useful patterns from unstructured text finds new unknown knowledge or patterns with the resulting logic. Since most of the information I use is in the form of unstructured textual data, the automated analysis of textual documents in natural language is very important.

The method most commonly used in text mining is to generate a feature vector and find the new knowledge or patterns by applying various techniques such as the statistical method to the generated vector. These feature vectors extract keywords from the text and use them to categorize or summarize documents.

The text mining method combines various techniques such as the automatic classification (document clustering and text categorization), natural language processing, information extraction, and information retrieval. The automatic classification refers to a task of grouping objects with similar patterns by a classification algorithm. There are two types of automatic classification depending on the use of the preliminary classification system. The document clustering technique groups the documents having similar contents without the preliminary classification while the text categorization technique assigns the documents to the most suitable subject category classified in advance using the machine learning.

3.3.2. Word Cloud

As a technique of visualizing the key words mentioned in the study, Word cloud enables understanding intuitively the keywords and concepts of documents. For example, there is a technique that allows a word to be expressed at a glance as much as it is mentioned. It is mainly used to derive the characteristics of data when analyzing big data that deals with a huge amount of information. Big data analysis tool R Studio provides a variety of packages for crawling, text mining, and word cloud such as 'KoNLP', 'wordcloud', 'XML', 'stringr', 'httr', 'rvest', and 'dplyr'. Data was collected through crawling after installing the necessary packages. Figure 15 shows an example of word cloud text image generated by big data analysis. The users can freely change the fonts, shapes, and sizes, and there is no copyright issue.

3.3.3. Web Crawling

Web crawling is a computer program work that explores the World Wide Web in an organized, automated manner. It is used to collect certain types of information on web pages by crawling the Web using R Studio's library (httr), library (rvest), and library (dplyr) packages. After searching the news, the URL of the news web page is collected, including the URL of the news article in the web page. Then, words in the news articles are crawled to extract the text and to save. It is analyzed as shown in Figure 14.

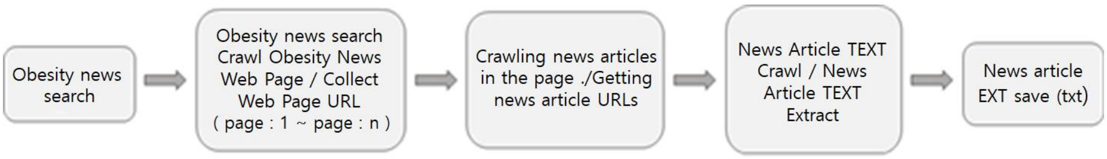

Obesity news search
Obesity news search Crawl Obesity News Web Page / Collect Web Page URL (page : 1 ~ page : n)
Crawling news articles in the page ./Getting news article URLs
News Article TEXT Crawl / News Article TEXT Extract
News article EXT save (txt)

Figure 14. Crawling Progress Step.

4. Big Data Analysis Result for Personalized Health Activities

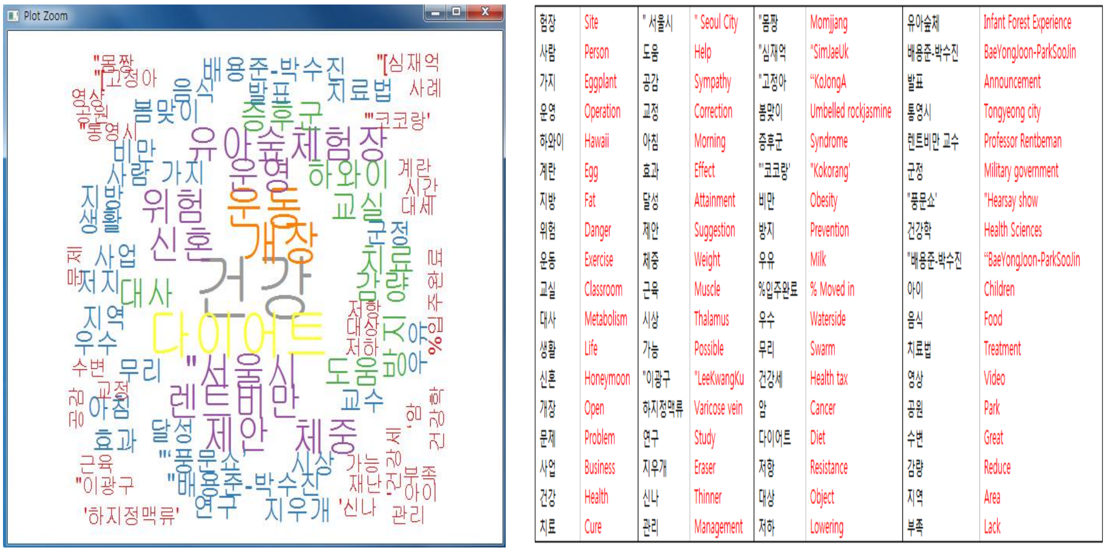
First, the study hosted periodic searches on *Naver News* Big Data in Korea to see the people's interest in obesity by season. From 1 January 2013 to 1 March 2016, the search results were 655 pages with 6545 articles, and 500 of them were crawled. Likewise, the study categorized periods for other seasons and search for news and crawled. The number of web pages and the number of articles according to each period are shown in the table below. Files crawled during the period 1 January–1 March are referred to as 'winter.txt'; files crawled from 2 March to 1 June are referred to as 'spring.txt', and files crawled from 2 June to 1 September are 'summer.txt'. Finally, files crawled from 2 September to 1 December are referred to as 'fall.txt'. Table 6 shows number of web pages and articles in obesity search by *Naver News*.

Table 6. Number of Web Pages and Articles in Obesity Search by *Naver News*.

Period (2016)	No. of Page	No. of Article	No. of Crawled Article
1.1~3.1	655	6545	500
3.2~6.1	915	9145	450
6.2~9.1	772	7716	535
9.2~12.1	887	8863	525

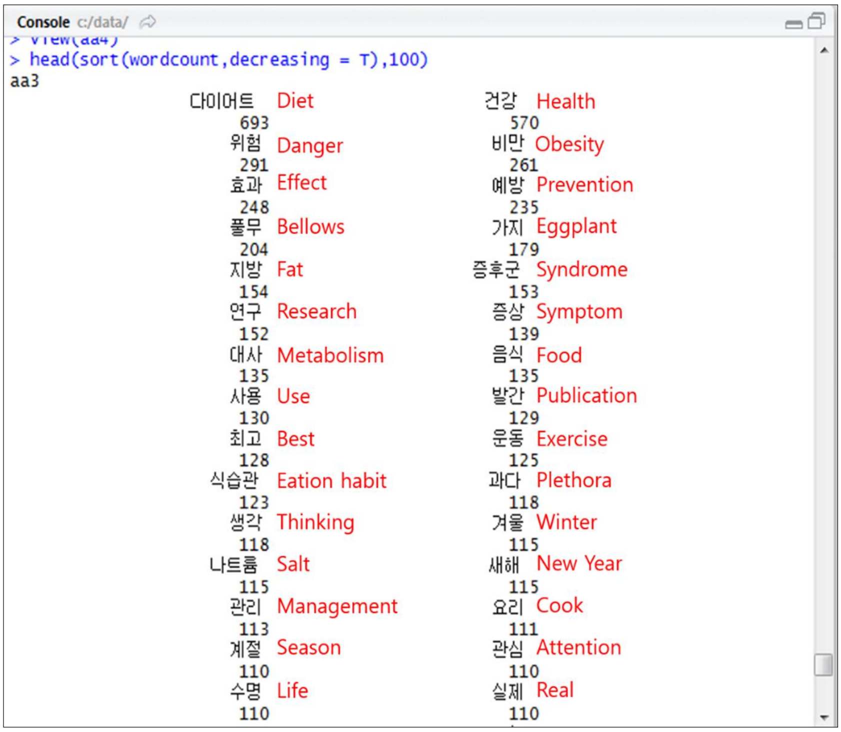
In the case of collected text files, words with “obesity” may include unnecessary words such as special characters and numbers as well as words with meaningful relationships. Therefore, the study used the gsup () function, which has a filtering function to remove characters and symbols deemed to be unnecessary during keyword extraction. The gsub () function has the function of changing the specified character or symbol to whatever character or space.

As shown in Figures 15–17, the filtered ‘winter.txt’ is shown in word cloud, and the frequency of each word is determined. Meanwhile, the five most frequently used words are selected and displayed as a graph and visualized during this period. The significance of each word is found and interpreted as follows:



험장	Site	" 서울시	" Seoul City	"몸짱	Momjjang	유아숲체	Infant Forest Experience
사람	Person	도움	Help	"심재억	"SimJaeUk	배용준-박수진	BaeYongJoon-ParkSooJin
가지	Eggplant	공감	Sympathy	"고정아	“KoJongA	발표	Announcement
운영	Operation	교정	Correction	봄맞이	Umbelled rockjasmine	통영시	Tongyeong city
하와이	Hawaii	아침	Morning	증후군	Syndrome	렌트비만 교수	Professor Rentbeman
계란	Egg	효과	Effect	"'코코랑'	"Kokorang'	군정	Military government
지방	Fat	달성	Attainment	비만	Obesity	"풍문쇼'	"Hearsay show
위험	Danger	제안	Suggestion	방지	Prevention	건강학	Health Sciences
운동	Exercise	체중	Weight	우유	Milk	"배용준-박수진	“BaeYongJoon-ParkSooJin
교실	Classroom	근육	Muscle	%입주완료	% Moved in	아이	Children
대사	Metabolism	시상	Thalamus	우수	Waterside	음식	Food
생활	Life	가능	Possible	무리	Swarm	치료법	Treatment
신혼	Honeymoon	"이광구	"LeeKwangKu	건강세	Health tax	영상	Video
개장	Open	하지정맥류	Varicose vein	암	Cancer	공원	Park
문제	Problem	연구	Study	다이어트	Diet	수변	Great
사업	Business	지우개	Eraser	저항	Resistance	감량	Reduce
건강	Health	신나	Thinner	대상	Object	지역	Area
치료	Cure	관리	Management	저하	Lowering	부족	Lack

Figure 15. Obesity Search Word Cloud in Winter.



```
> View(aa4)
> head(sort(wordcount,decreasing = T),100)
aa3
                  다이어트  Diet                        건강  Health
                      693                              570
                      위험  Danger                      비만  Obesity
                      291                              261
                      효과  Effect                      예방  Prevention
                      248                              235
                      풀무  Bellows                     가지  Eggplant
                      204                              179
                      지방  Fat                       증후군  Syndrome
                      154                              153
                      연구  Research                    증상  Symptom
                      152                              139
                      대사  Metabolism                  음식  Food
                      135                              135
                      사용  Use                         발간  Publication
                      130                              129
                      최고  Best                        운동  Exercise
                      128                              125
                    식습관  Eation habit                과다  Plethora
                      123                              118
                      생각  Thinking                    겨울  Winter
                      118                              115
                    나트륨  Salt                        새해  New Year
                      115                              115
                      관리  Management                  요리  Cook
                      113                              111
                      계절  Season                      관심  Attention
                      110                              110
                      수명  Life                        실제  Real
                      110                              110
```

Figure 16. Word Frequency Graph of ‘winter.txt’.

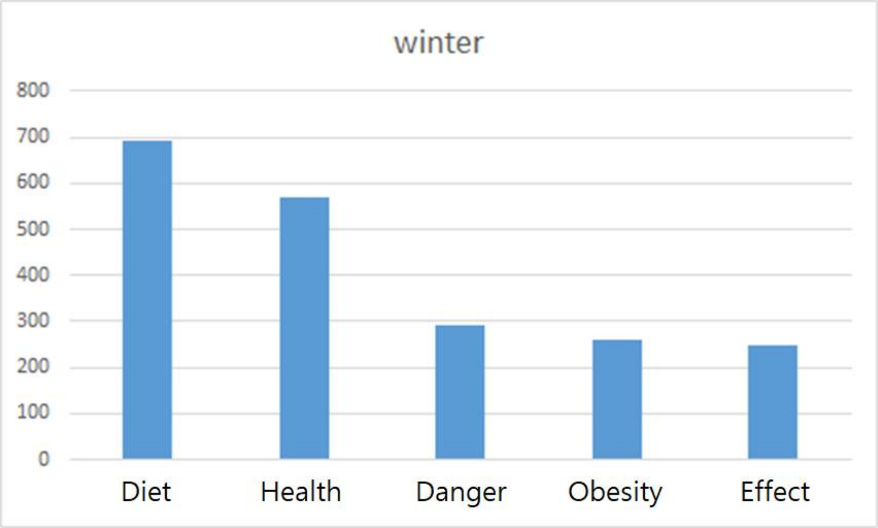
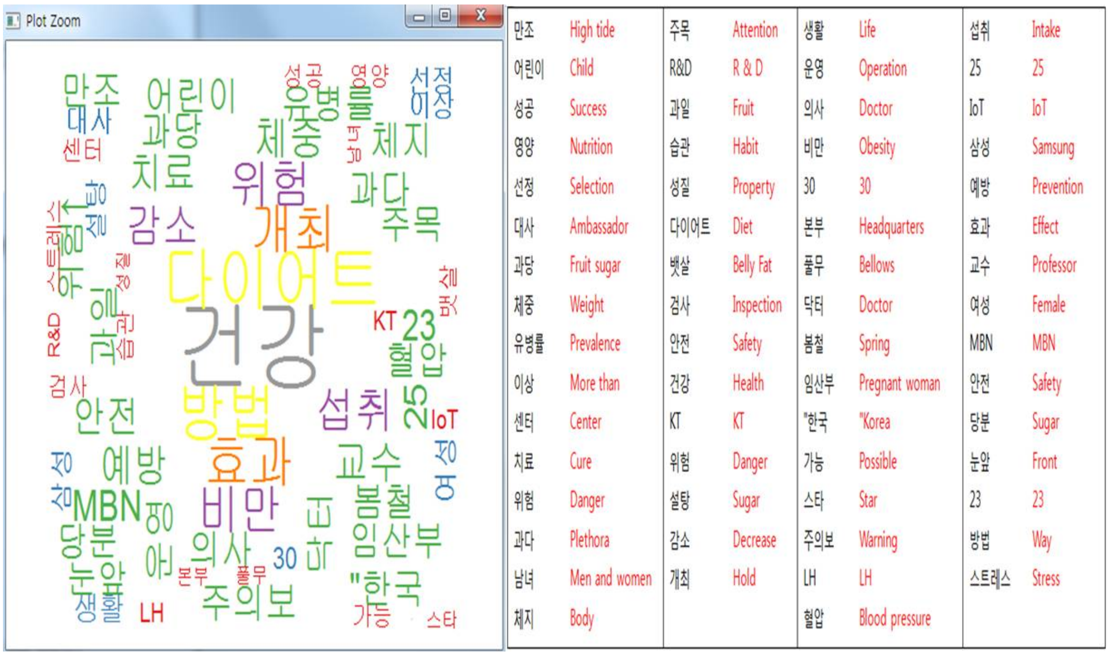


Figure 17. Word Frequency Graph 'winter.txt'.

The most commonly used words related to obesity in winter were "diet", "health", "risk", "obesity", and "effect". The analysis shows that obese people during this period are most interested in diet and health, and that they are also paying attention to the dangers of obesity. It was followed by the analysis of the crawled text file for another season.

As shown in Figures 18 and 19, the most frequently used words related to "obesity" in spring were "health", "diet", "method", "effect", and "hosting" in order of frequency. There was no significant difference compared to the winter period, but people had more interest in health and diet methods and effects than the risk of obesity. In particular, the frequency of the word "hosting" increased, indicating that various events related to obesity were hosted. In fact, during this period, various events such as the "Healthy Living Practice Contest" hosted by the Korea Health Association, "Diet Recipe Contest" hosted by the obesity professional treatment center, and a "Health Lecture" held at the Northern Health Center were hosted.



만조	High tide	주목	Attention	생활	Life	섭취	Intake
어린이	Child	R&D	R & D	운영	Operation	25	25
성공	Success	과일	Fruit	의사	Doctor	IoT	IoT
영양	Nutrition	습관	Habit	비만	Obesity	삼성	Samsung
선정	Selection	성질	Property	30	30	예방	Prevention
대사	Ambassador	다이어트	Diet	본부	Headquarters	효과	Effect
과당	Fruit sugar	뱃살	Belly Fat	풀무	Bellows	교수	Professor
체중	Weight	검사	Inspection	닥터	Doctor	여성	Female
유병률	Prevalence	안전	Safety	봄철	Spring	MBN	MBN
이상	More than	건강	Health	임산부	Pregnant woman	안전	Safety
센터	Center	KT	KT	"한국	"Korea	당분	Sugar
치료	Cure	위험	Danger	가능	Possible	눈앞	Front
위험	Danger	설탕	Sugar	스타	Star	23	23
과다	Plethora	감소	Decrease	주의보	Warning	방법	Way
남녀	Men and women	개최	Hold	LH	LH	스트레스	Stress
체지	Body			혈압	Blood pressure		

Figure 18. Obesity Search Word Cloud in Spring.

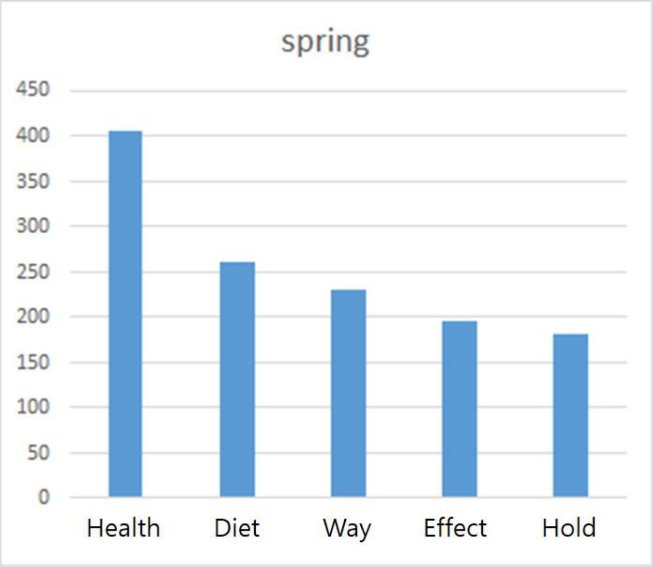
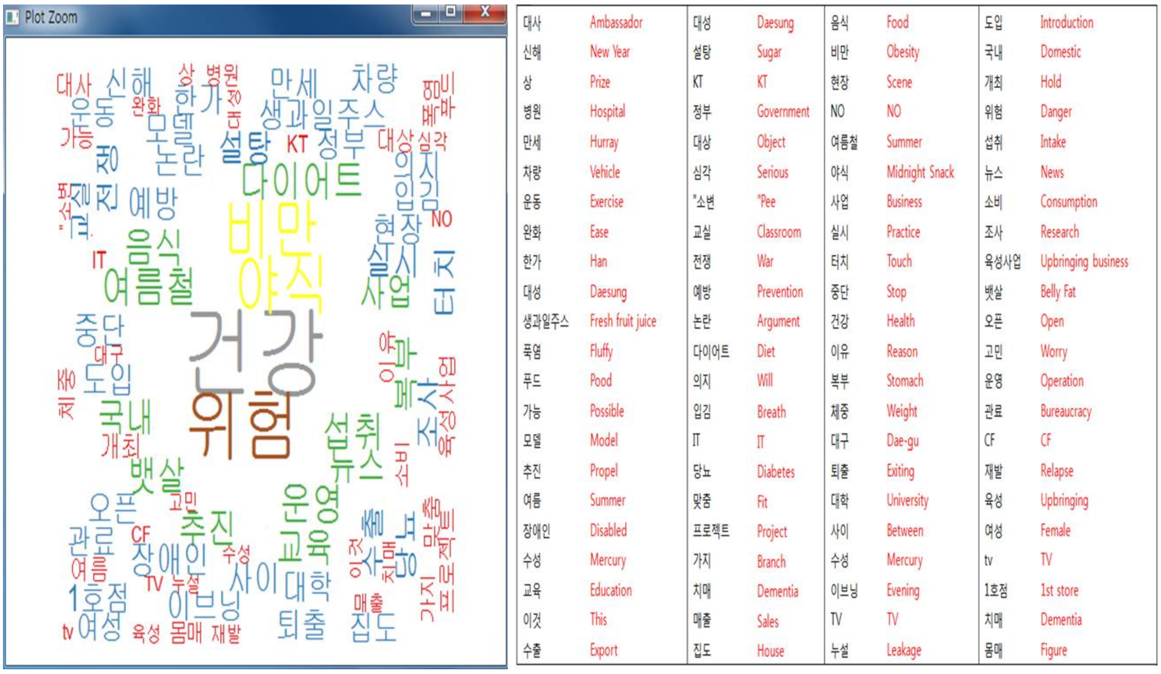


Figure 19. Word Frequency Graph of ‘spring.txt’.

As shown in Figures 20 and 21, the most commonly used words in summer were “health”, “risk”, “night snack”, “obesity”, and “summer”. During this period, interest in food and body shape is higher than in other seasons. It can be seen that they had more interest in diet habits for body fat management such as “summer”, “intake”, “food”, and “abdominal muscle”. In particular, due to the characteristics of the summer season, this result is attributable to the fact that the body shape was most noticeable and more prominent than in other seasons. In the next place, the study analyzed ‘fall.txt’, which crawled obesity news during the fall period.



대사	Ambassador	대성	Daesung	음식	Food	도입	Introduction
신해	New Year	설탕	Sugar	비만	Obesity	국내	Domestic
상	Prize	KT	KT	현장	Scene	개최	Hold
병원	Hospital	정부	Government	NO	NO	위험	Danger
만세	Hurray	대상	Object	여름철	Summer	섭취	Intake
차량	Vehicle	심각	Serious	야식	Midnight Snack	뉴스	News
운동	Exercise	"소변	"Pee	사업	Business	소비	Consumption
완화	Ease	교실	Classroom	실시	Practice	조사	Research
한가	Han	전쟁	War	터치	Touch	육성사업	Upbringing business
대성	Daesung	예방	Prevention	중단	Stop	뱃살	Belly Fat
생과일주스	Fresh fruit juice	논란	Argument	건강	Health	오픈	Open
폭염	Fluffy	다이어트	Diet	이유	Reason	고민	Worry
푸드	Pood	의지	Will	복부	Stomach	운영	Operation
가능	Possible	입김	Breath	체중	Weight	관료	Bureaucracy
모델	Model	IT	IT	대구	Dae-gu	CF	CF
추진	Propel	당뇨	Diabetes	퇴출	Exiting	재발	Relapse
여름	Summer	맞춤	Fit	대학	University	육성	Upbringing
장애인	Disabled	프로젝트	Project	사이	Between	여성	Female
수성	Mercury	가지	Branch	수성	Mercury	tv	TV
교육	Education	치매	Dementia	이브닝	Evening	1호점	1st store
이것	This	매출	Sales	TV	TV	치매	Dementia
수출	Export	집도	House	누설	Leakage	몸매	Figure

Figure 20. Obesity Search Word Cloud in Summer.

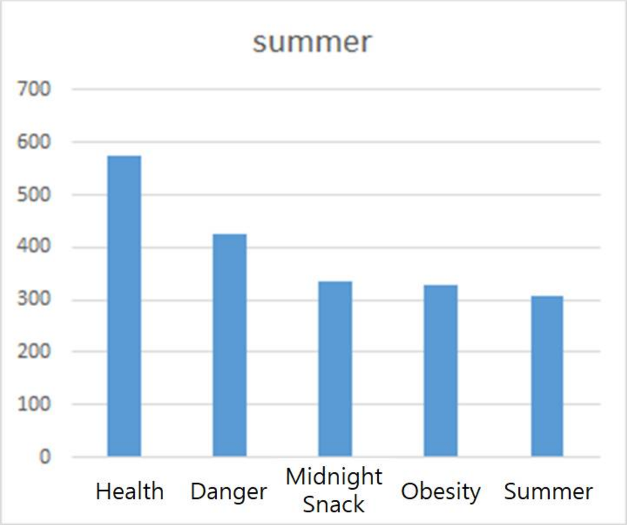


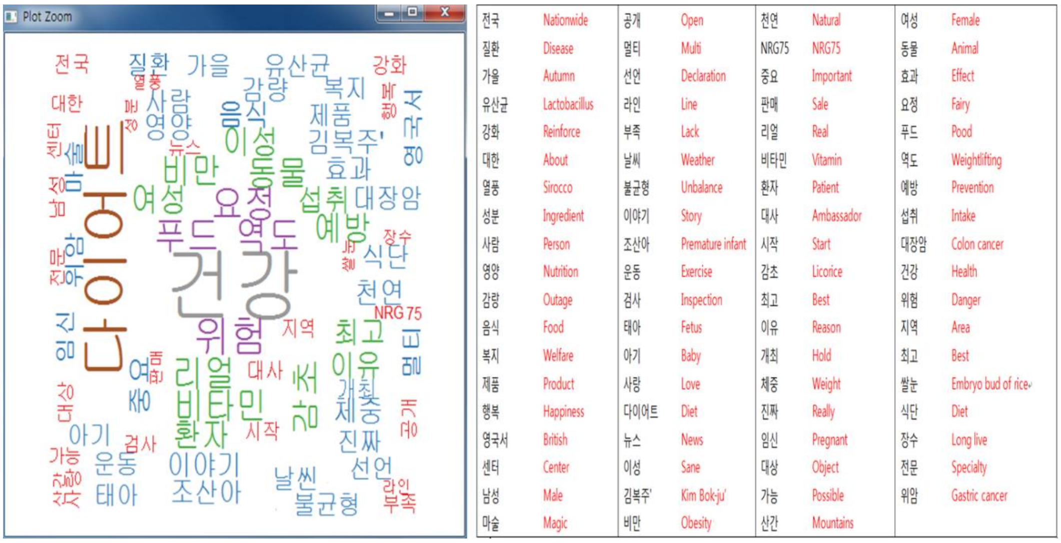
Figure 21. Word Frequency Graph of 'summer.txt'.

As shown in Figures 22 and 23, the three most frequently used words of "obesity news" in the fall period were "health", "diet", and "risk", which did not show much difference compared to other seasons. Note, however, that the frequency of the two words "weight loss" and "fairy", which have a low relationship with obesity, increased greatly. It is analyzed that "Kim Bok-joo, weightlifting fairy", a drama aired during fall~winter 2016, has had great effect. It indicates that, in the crawling and word cloud processes, more caution is required regarding the filtering process that excludes unnecessary information. In addition to considering the exclusion of simple special characters, English alphabet, etc., it is also necessary to pay attention to the general public's social interest.

In the next place, the study tried to visualize the change in the interest of the Korean people in obesity using the "biggest data-based service", Google Trends, and check the relevant search terms to study the customized countermeasures according to obesity.

Meanwhile, as shown in Figure 24, when we observe the graph of "obesity" change of interest of Google Trends over the last year, the interest was high in January and May compared to other periods but showed a sharp increase in December.

The influence of the drama as confirmed in the previous crawling process is analyzed to be large.



전국	Nationwide	공개	Open	천연	Natural	여성	Female
질환	Disease	멀티	Multi	NRG75	NRG75	동물	Animal
가을	Autumn	선언	Declaration	중요	Important	효과	Effect
유산균	Lactobacillus	라인	Line	판매	Sale	요정	Fairy
강화	Reinforce	부족	Lack	리얼	Real	푸드	Pood
대한	About	날씨	Weather	비타민	Vitamin	역도	Weightlifting
열풍	Sirocco	불균형	Unbalance	환자	Patient	예방	Prevention
성분	Ingredient	이야기	Story	대사	Ambassador	섭취	Intake
사람	Person	조산아	Premature infant	시작	Start	대장암	Colon cancer
영양	Nutrition	운동	Exercise	감초	Licorice	건강	Health
감량	Outage	검사	Inspection	최고	Best	위험	Danger
음식	Food	태아	Fetus	이유	Reason	지역	Area
복지	Welfare	아기	Baby	개최	Hold	최고	Best
제품	Product	사랑	Love	체중	Weight	쌀눈	Embryo bud of rice
행복	Happiness	다이어트	Diet	진짜	Really	식단	Diet
영국서	British	뉴스	News	임신	Pregnant	장수	Long live
센터	Center	이성	Sane	대상	Object	전문	Specialty
남성	Male	김복주'	Kim Bok-ju'	가능	Possible	위암	Gastric cancer
마술	Magic	비만	Obesity	산간	Mountains		

Figure 22. Obesity Search Word Cloud in Fall.

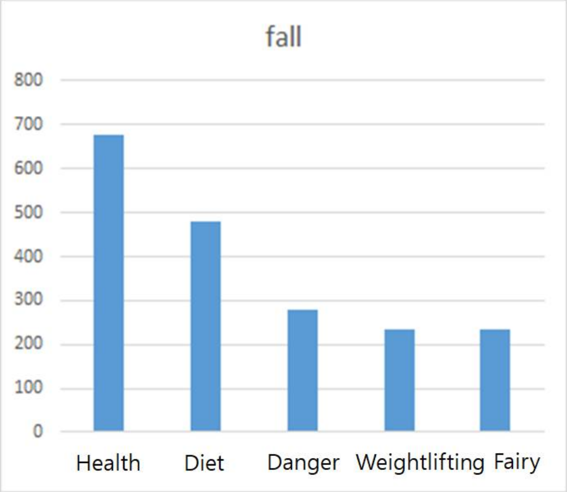
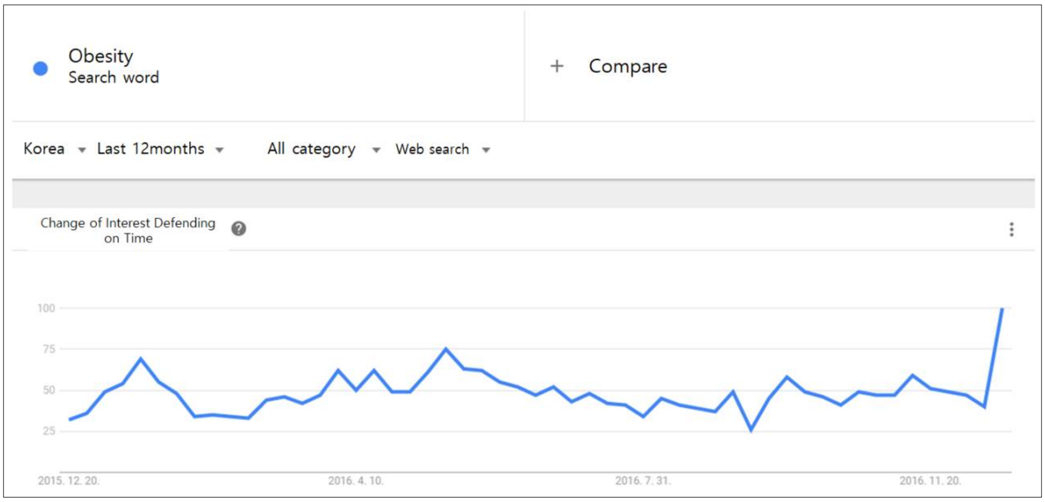


Figure 23. Word Frequency Graph of 'fall.txt'.

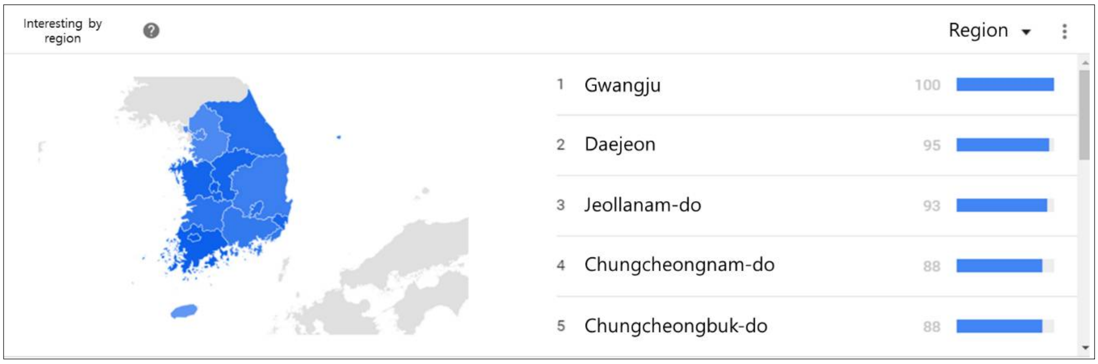

● Obesity
Search word

+ Compare

Korea ▾ Last 12months ▾ All category ▾ Web search ▾

Figure 24. The Change of Korea's Interest in Obesity over the Last Year in Google Trends.

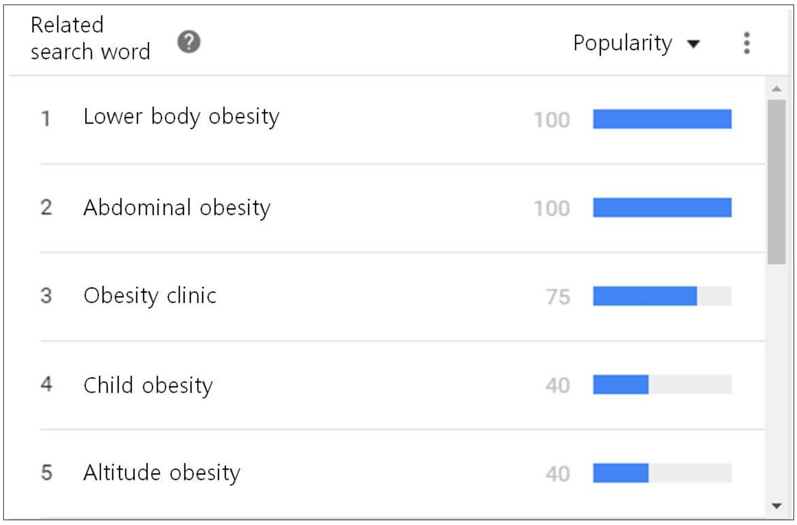
As shown in Figure 25, Gwangju, Daejeon, and Jeollanam-do recorded 100, 95, and 93, respectively, indicating that they were most interested in obesity. Jeju Island recorded 45, which is the lowest. In this case, since the value of interest indicates the percentage of the total search words instead of the absolute search number, the residents of Jeju Island show that the interest in obesity is about half of that of Gwangju and Daejeon.



1	Gwangju	100
2	Daejeon	95
3	Jeollanam-do	93
4	Chungcheongnam-do	88
5	Chungcheongbuk-do	88

Figure 25. Obesity Interest by Region of Google Trends.

As shown in Figure 26, interest was high in order of "lower body obesity", "abdominal obesity", "obesity clinic", "childhood obesity", and "high degree obesity". Obese people are less concerned about the causes of obesity and are more likely to have obesity and abdominal obesity.

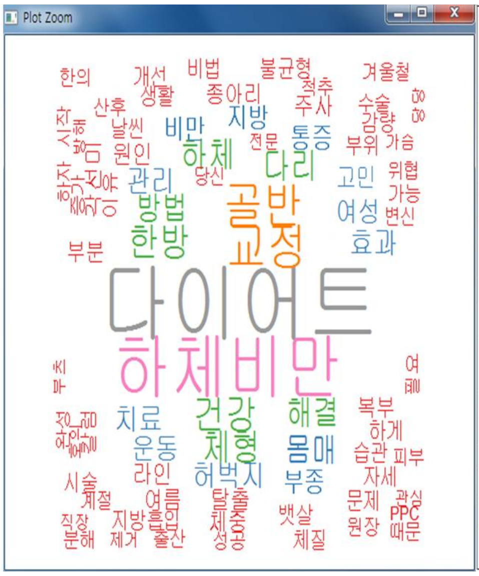


Related search word		Popularity
1	Lower body obesity	100
2	Abdominal obesity	100
3	Obesity clinic	75
4	Child obesity	40
5	Altitude obesity	40

Figure 26. Search Term Ranking Related to Obesity in Google Trends.

Next, this study analyzed crawl and word cloud for lower body obesity and abdominal obesity and customized measures. In *Naver News*, 126 pages and 1276 articles were formed as a result of the search for lower body obesity treatment, and 200 of them were crawled to generate a 'lower.txt' file. The search results for abdominal obesity treatment constituted 311 pages and 3120 articles, and 250 articles were crawled and stored as 'ob.txt'.

As shown in Figures 27–29, words mostly related to lower body obesity treatment in *Naver News* were diet, correction, and pelvis. For people with lower body obesity, exercise such as pelvic correction is helpful. In addition, considering the high frequency of the word "herbal", oriental medicine such as herbal diet is considered effective. Besides, it seems that attention should be paid to words or complications that cause only lower incomes such as "swelling" and "pain". Liposuction using injections is also mentioned as one of the treatment methods.



하체	Lower body	탈출	Escape	비법	Secret	칼럼	Column
다리	Leg	체중	Weight	종아리	Calf	시술	Surgery
방법	Way	성공	Success	불균형	Unbalance	계절	Season
한방	Oriental	뱃살	Belly Fat	적추	Chilli	직장	Rectal
건강	Health	체질	Constitution	주사	Injection	분해	Decomposition
해결	Solution	필요	Need	겨울철	Winter	지방	Fat
체형	Body type	복부	Stomach	수술	Operation	제거	Remove
관리	Management	하게	To do	감량	Outage	라인	Line
비만	Obesity	습관	Habit	부위	Part	여름	Summer
지방	Fat	피부	Skin	가슴	Chest	흡입	Inhale
통증	Ache	자세	Posture	당당	Sacred	출산	Childbirth
고민	Worry	문제	Problem	위협	Threat	증가	Increase
여성	Female	관심	Attention	가능	Possible	방해	Hindrance
효과	Effect	PPC	PPC	변신	Transform	각선미	A nice leg line
치료	Cure	원장	Ledger	부분	Part	이유	Reason
운동	Exercise	때문	Because	단신	Short	산후	Postpartum
허벅지	Thigh	한의	Oriental doctor	전문	Specialty	날씨	Weather
몸매	Figure	하자	Lets do it	부츠	Boots	원인	Cause
부족	Lack	시작	Start	완성	Complete	개선	Improving
				호의	Favor	생활	Life

Figure 27. Word Cloud in Lower Body Treatment Search.

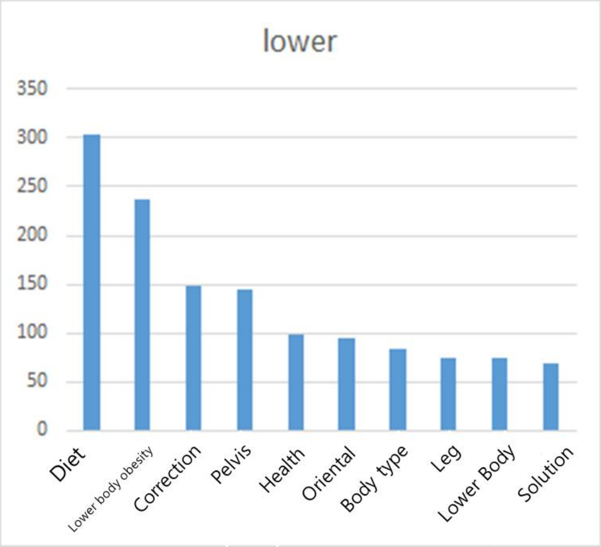
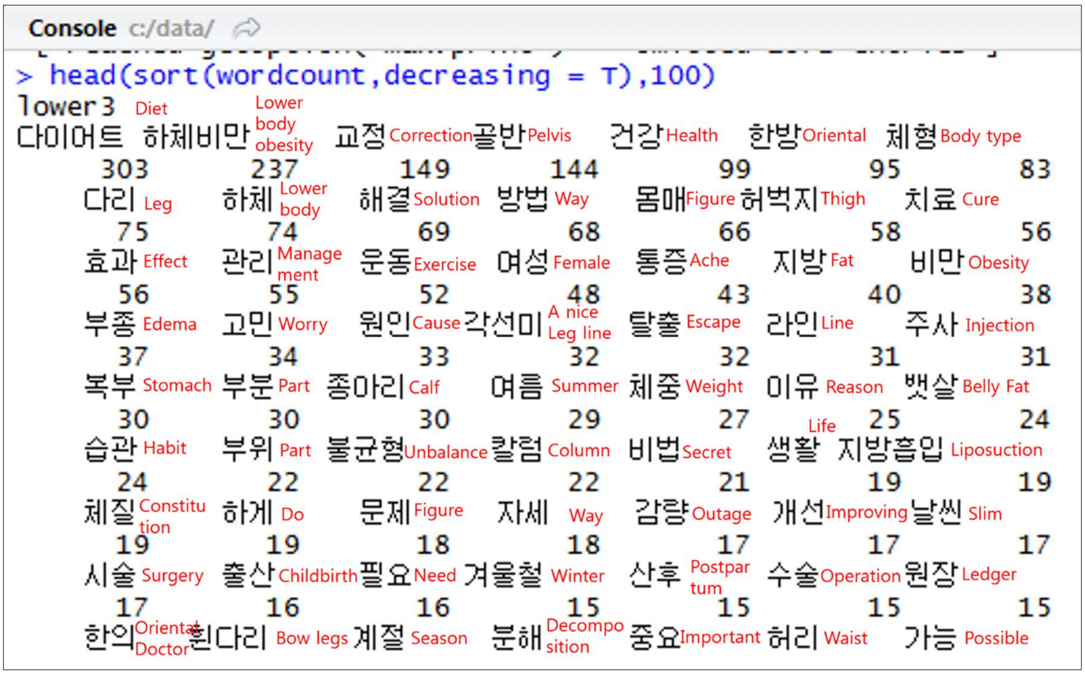


Figure 28. Frequency Graph of Word 'lower.txt'.



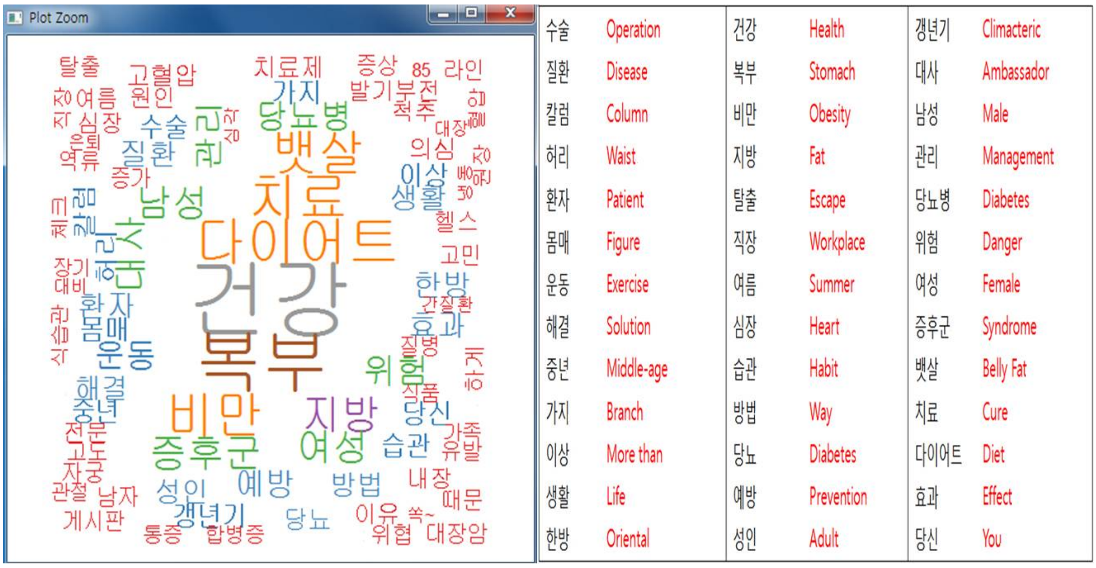
```
> head(sort(wordcount,decreasing = T),100)
lower3
다이어트 Diet   하체비만 Lower body obesity   교정 Correction   골반 Pelvis   건강 Health   한방 Oriental   체형 Body type
      303          237          149          144           99           95           83
  다리 Leg      하체 Lower body   해결 Solution   방법 Way   몸매 Figure   허벅지 Thigh   치료 Cure
       75           74           69           68           66           58           56
  효과 Effect   관리 Management   운동 Exercise   여성 Female   통증 Ache   지방 Fat   비만 Obesity
       56           55           52           48           43           40           38
  부종 Edema   고민 Worry   원인 Cause   각선미 A nice Leg line   탈출 Escape   라인 Line   주사 Injection
       37           34           33           32           32           31           31
  복부 Stomach   부분 Part   종아리 Calf   여름 Summer   체중 Weight   이유 Reason   뱃살 Belly Fat
       30           30           30           29           27           25           24
  습관 Habit   부위 Part   불균형 Unbalance   칼럼 Column   비법 Secret   생활 Life   지방흡입 Liposuction
       24           22           22           22           21           19           19
  체질 Constitution   하게 Do   문제 Figure   자세 Way   감량 Outage   개선 Improving   날씬 Slim
       19           19           18           18           17           17           17
  시술 Surgery   출산 Childbirth   필요 Need   겨울철 Winter   산후 Postpartum   수술 Operation   원장 Ledger
       17           16           16           15           15           15           15
  한의 Oriental Doctor   휜다리 Bow legs   계절 Season   분해 Decomposition   중요 Important   허리 Waist   가능 Possible
```

Figure 29. The Number of Frequency of 'lower.txt' Word Cloud.

As shown in Figures 30–32, words most related to abdominal obesity treatment were health, diet, treatment, obesity, fat, and syndrome. The incidence of syndrome is higher than lower body obesity treatment, and it can be seen that various syndromes can be caused by abdominal obesity. As with lower body obesity treatment, abdominal obesity treatment is also considered to be highly effective in herbal diet, and caution is required because it can cause adult diseases such as diabetes and hypertension. Finally, this study searched the keyword "obesity" in the US search site *About.com* and conducted a web crawl. *About.com* is a US online information site founded in 1997, and it continues to operate to date. It has a lot of information in various fields such as food information and recipe, health, economy, and travel information since it has a long history of operation.

The search period for the news to crawl was 1 year from 1 January 2016 to 31 December 2016. Similar to previous data analysis, crawled words were extracted, filtered, and saved as 'obesity.txt' and visualized using word cloud. As shown in Figures 33 and 34, the ten most commonly used words in the search results for "obesity" in *About.com* are "Weight", "Obesity", "Loss", "Fat", "Overweight", "Health", "Kids", "Body", "Childhood", and "Children". Unlike Korea, in the US, the words "Kids", "Childhood", "Children", and so on were used for words related to obesity. Since the US has a higher rate of childhood obesity than other countries, it focuses more on childhood life and diet habits related to obesity.

The analysis showed a wide range of seasonal factors according to spring, summer, fall, and winter. Its significance is that it completed visualization of the process of extracting the keywords appropriate for treatment of abdominal obesity and lower body obesity. In other words, this study collected big data by applying the machine learning and crawling methods to unstructured national health information data and search data of *Naver News* and Google and then visualized them using text mining and word cloud.



수술	Operation	건강	Health	갱년기	Climacteric
질환	Disease	복부	Stomach	대사	Ambassador
칼럼	Column	비만	Obesity	남성	Male
허리	Waist	지방	Fat	관리	Management
환자	Patient	탈출	Escape	당뇨병	Diabetes
몸매	Figure	직장	Workplace	위험	Danger
운동	Exercise	여름	Summer	여성	Female
해결	Solution	심장	Heart	증후군	Syndrome
중년	Middle-age	습관	Habit	뱃살	Belly Fat
가지	Branch	방법	Way	치료	Cure
이상	More than	당뇨	Diabetes	다이어트	Diet
생활	Life	예방	Prevention	효과	Effect
한방	Oriental	성인	Adult	당신	You

Figure 30. Word Cloud for Abdominal Obesity Treatment Search.

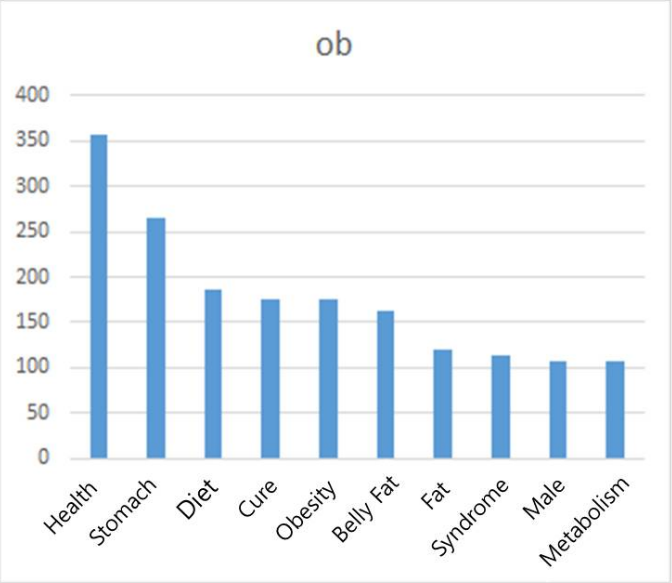
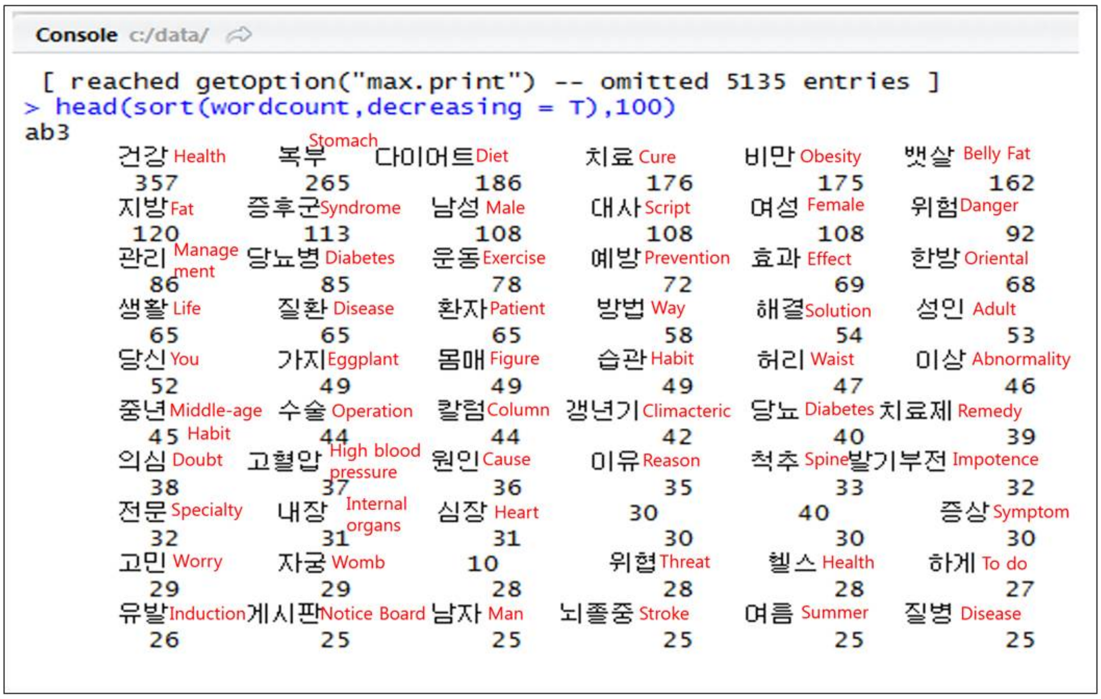


Figure 31. Frequency Graph of Word 'ob.txt'.



```
Console c:/data/
 [ reached getOption("max.print") -- omitted 5135 entries ]
> head(sort(wordcount,decreasing = T),100)
ab3
       건강 Health   복부 Stomach   다이어트 Diet   치료 Cure      비만 Obesity    뱃살 Belly Fat
        357           265           186            176            175            162
       지방 Fat      증후군 Syndrome  남성 Male     대사 Script    여성 Female     위험 Danger
        120           113           108            108            108             92
       관리 Management  당뇨병 Diabetes  운동 Exercise  예방 Prevention  효과 Effect  한방 Oriental
         86            85            78             72             69             68
       생활 Life     질환 Disease    환자 Patient   방법 Way       해결 Solution   성인 Adult
         65            65            65             58             54             53
       당신 You      가지 Eggplant   몸매 Figure    습관 Habit     허리 Waist      이상 Abnormality
         52            49            49             49             47             46
       중년 Middle-age  수술 Operation  칼럼 Column  갱년기 Climacteric  당뇨 Diabetes  치료제 Remedy
         45 Habit      44            44             42             40             39
       의심 Doubt    고혈압 High blood pressure  원인 Cause  이유 Reason  척추 Spine  발기부전 Impotence
         38            37            36             35             33             32
       전문 Specialty  내장 Internal organs  심장 Heart   30          40          증상 Symptom
         32            31            31             30             30             30
       고민 Worry    자궁 Womb      10            위협 Threat    헬스 Health     하게 To do
         29            29            28             28             28             27
       유발 Induction  게시판 Notice Board  남자 Man  뇌졸중 Stroke  여름 Summer   질병 Disease
         26            25            25             25             25             25
```

Figure 32. The Number of Frequency of 'ob.txt' Word Cloud.

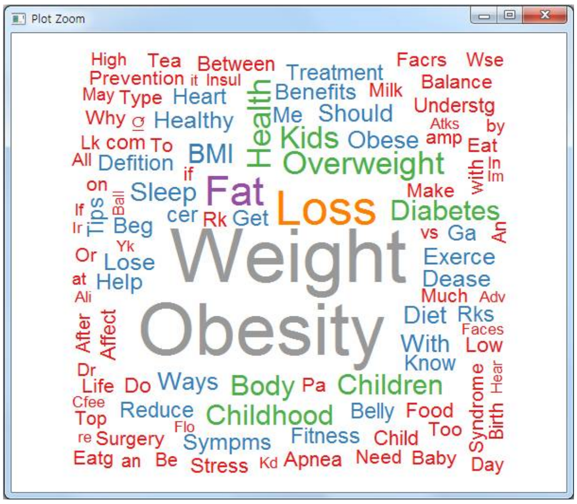


Figure 33. Obesity Search Word Cloud.

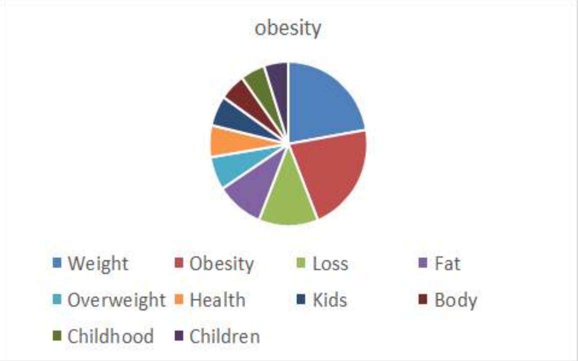


Figure 34. Frequency Graph of Word 'obesity.txt'.

5. Android App and iOS App and their Implementation to Present the Significant Analysis Results

Figure 35 shows the user interface (UI) that visualizes the big data compatible with Android and Apple iOS. The users can see the data on the app screen. Many graphs and pictures can be seen via menu, and the significant data values are visualized through machine learning. The toolbar is divided into four areas including the menu, message, share, and help; the sub-directories are displayed when a user taps the menu.

The previous section described the parameters related to the age, season, region, and obesity. The sub-menu shows four buttons for these parameters. The user can select the menu and tap the button to display the graph of the data that are intelligently computed.

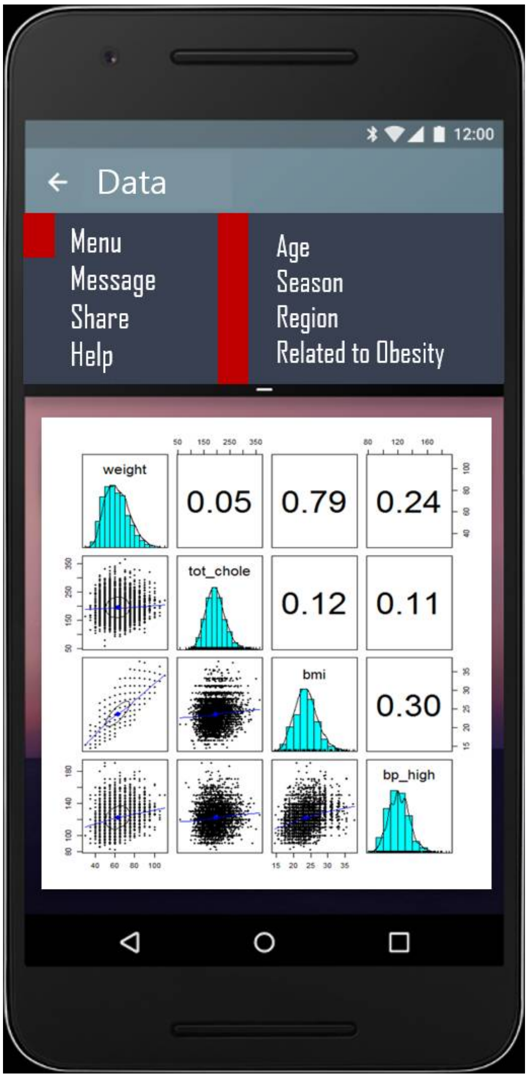


Figure 35. User Interface to Visualize Big Data Compatible with Android App and Apple iOS App.

Figure 36 shows the interest in obesity according to the visualized region of the big data for seasonal health search. It used the bar graph that looks like stairs to help the users understand it at a glance. The data progresses from left to right.

Also, the UI is organized to show the randomly arranged words, show the graph that represents them, and identify the most frequently used word. The analysis result indicates that the word "health" was used most frequently among the words displayed in the graph of four seasons. The number signifies the number of times the word "health" was used. Figure 37 visualizes the interest in obesity according to the region. The regions can be ranked for users' convenience. The big data analysis presents the meaningful result like a puzzle that shows a big picture when all pieces are put together.

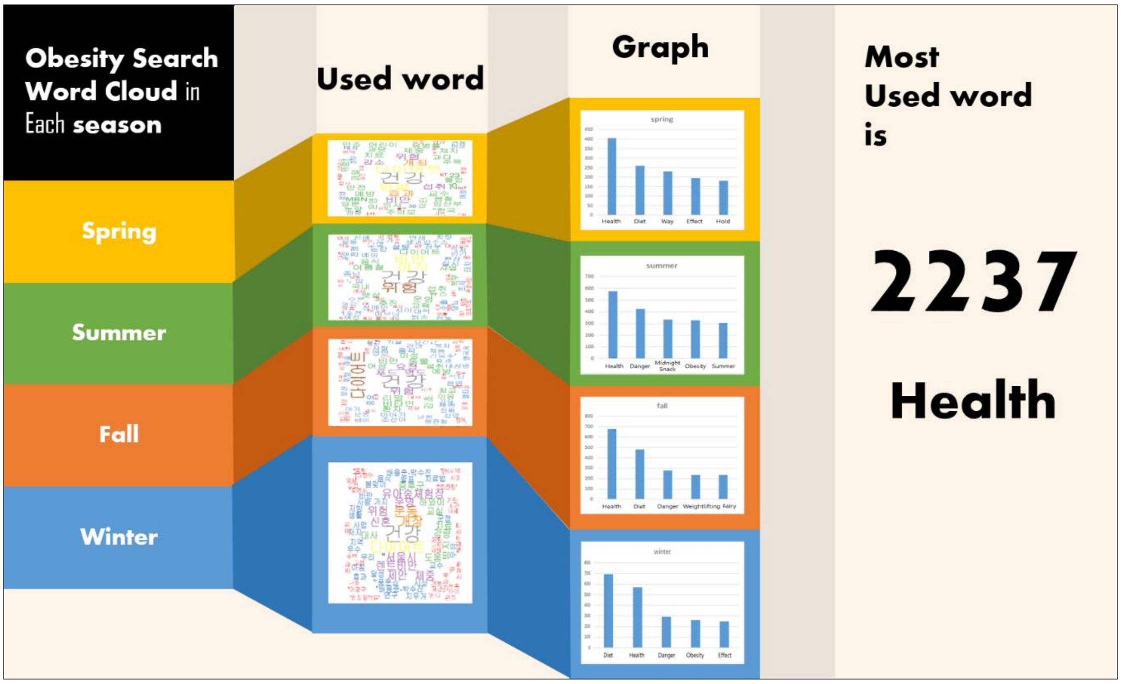


Figure 36. Interest in Obesity According to Visualized Region of Big Data for Seasonal Health Search.

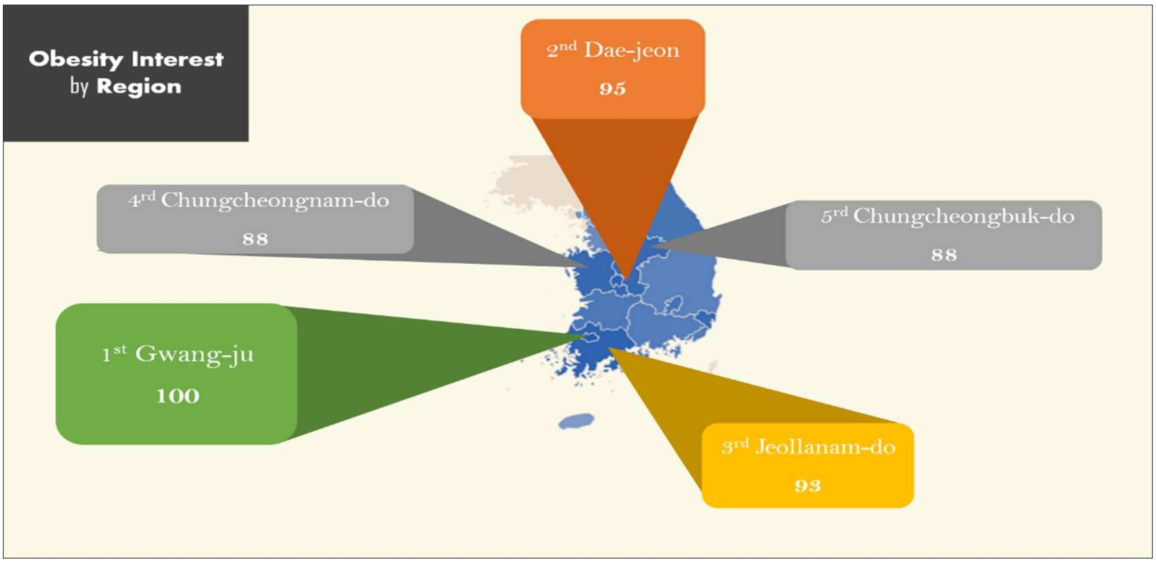


Figure 37. Interest in Obesity According to Region.

6. Conclusions and Future Work

In this paper, I analyzed the unstructured health data of one million Korean citizens using the datasets provided by the National Health Insurance Service using the machine learning and applied the text mining to the big data services such as Google Trends, *Naver News*, and *About.com* to analyzing the keyword big data for personalized health activities. It visualized the big data using text mining and word cloud. This study collected and analyzed the data concerning the interests related to obesity, change of interest on obesity, and treatment articles. The analysis showed a wide range of seasonal factors according to spring, summer, fall, and winter. Its significance is that it completed visualization of the process of extracting the keywords appropriate for treatment of abdominal obesity and lower body obesity.

As a result of analyzing the health examination information data set using the big data analysis tool R Studio, the distribution of obesity degree such as height and weight according to gender can be determined, including the obesity degree distribution according to body weight. Care should be taken when the weight exceeds 85 kg since the overweight and obese populations are high in that level. In addition to the various attributes used in this study, the health examination information data set contains more variables, so it is possible to analyze data from more diverse perspectives.

In the next place, data schematization such as crawling and word cloud can facilitate the analysis by clearly and concisely dividing the information. Nonetheless, careful attention is required because it can cause unintended and distorted results in the user's data classification or at the schematization stage.

Meanwhile, seasonal obesity did not show a significant difference; the dramatic change in the interest rate of obesity in Google Trends in December is analyzed to have been influenced by the recent drama. The degree of interest in obesity by region was also significantly different. In particular, interest in obesity between Jeju residents and Gwangju and Daejeon residents had more than double the difference. Thus, future analysis on this issue would also have a significant effect on obesity. Abdominal obesity and lower body obesity were categorized as *Naver News* crawling, with lower body obesity showing that exercise such as pelvic correction was helpful and abdominal obesity showing a higher risk of obesity-related syndrome and adult disease. Herbal diet had an effect on both abdominal obesity and lower body obesity but higher frequency in the lower body obesity treatment.

The study included data collection and analysis of obesity-related areas of interest, changes in obesity interest, and treatment articles. It was a process of extracting keywords for the treatment of abdominal obesity and lower body obesity. Since each subject has different interests, interest level, and physical constitution, however, there is a need to select a variety of keywords suitable for the individual and perform big data analysis in order to prepare a countermeasure for more personalized treatment methods and health activities. Likewise, if big data is collected, processed, and analyzed using various techniques, it is expected to be able to prevent and treat various diseases including obesity.

Acknowledgments: The first draft part of this paper was presented in at The 12th KIPS International Conference on Ubiquitous Information Technologies and Applications (CUTE 2017) [60], 18–20 December 2017, Taichung, Taiwan. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2017R1C1B5077157).

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

References

1. Ministry of Health and Welfare. *2014 National Health Statistics I*; Ministry of Health and Welfare: Sejong City, Korea, 2015. (In Korean)
2. Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. *Science* **2018**, *359*, 1146–1151. [CrossRef] [PubMed]
3. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]
4. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
5. Bayes, T. An essay towards solving a problem in the doctrine of chances. *Stud. Hist. Stat. Probab.* **1970**, *1*, 134–153. [CrossRef]
6. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*; Springer: Berlin, Germany, 2008.

7. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
8. Huh, J.H.; Kim, H.B.; Seo, K. A preliminary analysis model of big data for prevention of bioaccumulation of heavy metal-based pollutants: Focusing on the atmospheric data analyses. *Adv. Sci. Technol. Lett. SERSC* **2016**, *129*, 159–164.
9. Wu, X.; Fan, W.; Peng, J.; Zhang, K.; Yu, Y. Iterative sampling based frequent itemset mining for big data. *Int. J. Mach. Learn. Cybern.* **2015**, *6*, 875–882. [CrossRef]
10. Luo, Y.; Yang, B.; Xu, L.; Hao, L.; Liu, J.; Yao, Y.; Van de Vosse, F. Segmentation of the left ventricle in cardiac MRI using a hierarchical extreme learning machine model. *Int. J. Mach. Learn. Cybern.* **2017**. [CrossRef]
11. Aldehim, G.; Wang, W. Determining appropriate approaches for using data in feature selection. *Int. J. Mach. Learn. Cybern.* **2017**, *8*, 915–928. [CrossRef]
12. Doan, T.; Kalita, J. Predicting run time of classification algorithms using meta-learning. *Int. J. Mach. Learn. Cybern.* **2017**, *8*, 1929–1943. [CrossRef]
13. Zhai, J.; Zhang, S.; Wang, C. The classification of imbalanced large data sets based on mapreduce and ensemble of elm classifiers. *Int. J. Mach. Learn. Cybern.* **2017**, *8*, 1009–1017. [CrossRef]
14. Pang, B.; Lee, L. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 1–135. [CrossRef]
15. Kaiser, C. Opinion Mining im Web 2.0—Konzept und Fallbeispiel. *HMD Prax. Wirtsch.* **2009**, *46*, 90–99. [CrossRef]
16. Hu, M.; Liu, B. Mining Opinion Features in Customer Reviews. In Proceedings of the 19th National Conference on Artifical Intelligence, San Jose, CA, USA, 25–29 July 2004; pp. 755–760.
17. Liu, B.; Hu, M.; Cheng, J. Opinion Observer: Analyzing and Comparing Opinions on the Web. In Proceedings of the 14th International Conference on World Wide Web, New York, NY, USA, 10–14 May 2005; pp. 342–351.
18. Jindal, N.; Liu, B. Identifying Comparative Sentences in Text Documents. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery (ACM), Seattle, WA, USA, 6–11 August 2006; pp. 244–251.
19. Zhang, T. Fundamental Statistical Techniques. In *Handbook of Natural Language Processing*, 2nd ed.; Indurkhya, N., Damerau, F.J., Eds.; Chapman & Hall/CRC: Boca Raton, FL, USA, 2010; pp. 189–204.
20. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In Proceedings of the ACL-2002 Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA, USA, 6–7 July 2002; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 79–86.
21. Liu, B. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*; Springer: Berlin, Germany, 2007.
22. Dasgupta, S.; Ng, V. Mine the Easy, Classify the Hard: A Semi-Supervised Approach to Automatic Sentiment Classification. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, 2–7 August 2009; Volume 2, pp. 701–709.
23. Wong, T.-L.; Bing, L.; Lam, W. Normalizing Web Product Attributes and Discovering Domain Ontology with Minimal Effort. In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, Hong Kong, China, 9–12 February 2011; pp. 805–814.
24. Choi, Y.; Cardie, C. Hierarchical Sequential Learning for Extracting Opinions and their Attributes. In Proceedings of the ACL 2010 Conference Short Papers, Uppsala, Sweden, 11–16 July 2010; pp. 269–274.
25. Guo, H.; Zhu, H.; Guo, Z.; Su, Z. Domain Customization for Aspect-oriented Opinion Analysis with Multi-level Latent Sentiment Clues. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management, Glasgow, UK, 24–28 October 2011; pp. 2493–2496.
26. Holzinger, A.; Simonic, K.-M.; Yildirim, P. Disease-Disease Relationships for Rheumatic Diseases. In Web-Based Biomedical Textmining and Knowledge Discovery to Assist Medical Decision Making. In Proceedings of the IEEE 36th International Conference on Computer Software and Applications, Izmir, Turkey, 16–20 July 2012; pp. 573–580.
27. Cui, H.; Mittal, V.; Datar, M. Comparative Experiments on Sentiment Classification for Online Product Reviews. In Proceedings of the AAAI-2006, Boston, MA, USA, 16–20 July 2006; pp. 1265–1270.
28. Chaovalit, P.; Zhou, L. Movie Review Mining: A Comparison between Supervised and Unsupervised Classification Approaches. In Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, 6 January 2005; pp. 112–121.

29. Moghaddam, S.; Ester, M. On the Design of LDA Models for Aspect-based Opinion Mining. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, Maui, HI, USA, 29 October–2 November 2012; pp. 803–812.
30. Mishne, G.; Glance, N.S. Predicting Movie Sales from Blogger Sentiment. In Proceedings of the 21st National Conference on Artificial Intelligence, Boston, MA, USA, 16–20 July 2006; pp. 11–14.
31. Sik Kim, Y.; Lee, K.; Ryu, J.-H. Algorithm for Extrapolating Blogger's Interests through Library Classification Systems. In Proceedings of the IEEE International Conference on Web Services, Beijing, China, 23–26 September 2008; pp. 481–488.
32. Liu, Y.; Huang, X.; An, A.; Yu, X. ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, 23–27 July 2007; ACM Press: New York, NY, USA, 2007; pp. 607–614.
33. Sadikov, E.; Parameswaran, A.; Venetis, P. *Blogs as Predictors of Movie Success*; AAAI Press: Boston, MA, USA, 2009; pp. 304–307.
34. Liu, F.; Wang, D.; Li, B.; Liu, Y. Improving Blog Polarity Classification via Topic Analysis and Adaptive Methods. In Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, CA, USA, 2–4 June 2010; pp. 309–312.
35. Liu, F.; Li, B.; Liu, Y. Finding Opinionated Blogs Using Statistical Classifiers and Lexical Features. In Proceedings of the 3rd International ICWSM Conference, San Jose, CA, USA, 17–20 May 2009; AAAI Press: Boston, MA, USA, 2009; pp. 254–257.
36. Chmiel, A.; Sobkowicz, P.; Sienkiewicz, J.; Paltoglou, G.; Buckley, K.; Thelwall, M.; Hołyst, J.A. Negative emotions boost user activity at BBC forum. *Phys. A* **2011**, *390*, 2936–2944. [CrossRef]
37. Softic, S.; Hausenblas, M. Towards Opinion Mining through Tracing Discussions on the Web. In Proceedings of the 7th International Semantic Web Conference; Karlsruhe, Germany, 26–30 October 2008.
38. Go, A.; Bhayani, R.; Huang, L. Twitter Sentiment Classification using Distant Supervision. In *CS224N Project Report*; Stanford University: Stanford, CA, USA, 2009.
39. Pak, A.; Paroubek, P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC), Valletta, Malta, 17–23 May 2010; pp. 1320–1326.
40. Barbosa, L.; Feng, J. Robust Sentiment Detection on Twitter from Biased and Noisy Data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Beijing, China, 23–27 August 2010; pp. 36–44.
41. Bollen, J.; Mao, H.; Zeng, X. Twitter mood predicts the stockmarket. *J. Comput. Sci.* **2011**, *2*, 1–8. [CrossRef]
42. Derczynski, L.; Maynard, D.; Aswani, N.; Bontcheva, K. Microblog-Genre Noise and Impact on Semantic Annotation Accuracy. In Proceedings of the 24th ACM Conference on Hypertext and Social Media, Paris, France, 1–3 May 2013.
43. Thelwall, M.; Wilkinson, D.; Uppal, S. Data Mining Emotion in Social Network Communication: Gender differences in MySpace. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 190–199. [CrossRef]
44. Bermingham, A.; Conway, M.; McInerney, L.; O'Hare, N.; Smeaton, A.F. Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation. In Proceedings of the International Conference on Advances in Social Network Analysis and Mining, Athens, Greece, 20–22 July 2009; pp. 231–236.
45. Titov, I.; McDonald, R. A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In Proceedings of the ACL-2008, HLT, Columbus, OH, USA, 15–20 June 2008; ACL: Stroudsburg, PA, USA, 2008; pp. 308–316.
46. Titov, I.; McDonald, R. Modeling Online Reviews with Multi-grain Topic Models. In Proceedings of the 17th International Conference on World Wide Web, Beijing, China, 21–25 April 2008; pp. 111–120.
47. Zhao, W.X.; Jiang, J.; Yan, H.; Li, X. Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, USA, 9–11 October 2010; pp. 56–65.
48. Brody, S.; Elhadad, N. An Unsupervised Aspect-Sentiment Model for Online Reviews. In Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, CA, USA, 2–4 June 2010; pp. 804–812.

49. Jo, Y.; Oh, A. Aspect and Sentiment Unification Model for Online Review Analysis. In Proceedings of the 4th ACM International Conference on Web Search and Data Mining, Hong Kong, China, 9–12 February 2011; pp. 815–824.
50. Petz, G.; Karpowicz, M.; Fürschuß, H.; Auinger, A.; Stříteský, V.; Holzinger, A. Opinion mining on the web 2.0—Characteristics of user generated content and their impacts. In *Lecture Notes in Computer Science, LNCS*; Springer: Berlin/Heidelberg, Germany, 2013.
51. Liu, B. Sentiment analysis and opinion mining. In *Synthesis Lectures on Human Language Technologies*; Morgan & Claypool: San Rafael, CA, USA, 2012; Volume 5, pp. 1–167.
52. Huh, J.H. PLC-based design of monitoring system for ICT-integrated vertical fish farm. In *Human-Centric Computing and Information Sciences*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 7, pp. 1–20.
53. Sharma, P.K.; Moon, S.Y.; Park, J.H. Block-VN: A Distributed Blockchain Based Vehicular Network Architecture in Smart City. *J. Inf. Process. Syst.* **2017**, *13*, 184–195.
54. Garcia Lopez, P.; Montresor, A.; Epema, D.; Datta, A.; Higashino, T.; Iamnitchi, A.; Barcellos, M.; Felber, P.; Riviere, E. Edge-centric Computing: Vision and Challenges. *ACM SIGCOMM Comput. Commun. Rev.* **2015**, *45*, 37–42. [CrossRef]
55. Sun, L.; Ma, J.; Wang, H.; Zhang, Y. Cloud Service Description Model: An Extension of USDL for Cloud Services. *IEEE Trans. Serv. Comput.* **2015**, *99*, 1–14. [CrossRef]
56. Li, M.; Sun, X.; Wang, H.; Zhang, Y.; Zhang, J. Privacy-aware Access Control with trust management in Web Service. *World Wide Web* **2011**, *14*, 407–430. [CrossRef]
57. Wang, H.; Cao, J.; Zhang, Y. A Flexible Payment Scheme and its Role based Access Control. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 425–436. [CrossRef]
58. Holzinger, A. Interactivemachine learning for health informatics: When do we need the human-in-the-loop? *Brain Inform.* **2016**, *3*, 119–131. [CrossRef] [PubMed]
59. Gong, H.-S.; Weon, S.; Huh, J.-H. A Study on the Design of Humane Animal Care System and Java Implementation. 2018; unpublished.
60. Lee, S.; Le, H.-S.; Huh, J.-H. A Keyword-Based Big Data Analysis for Individualized Health Activity Using Keyword Analysis Technique: A Methodological Approach Using National Health Data. In *Advances in Computer Science and Ubiquitous Computing*; Springer: Singapore, 2017; pp. 1237–1243.