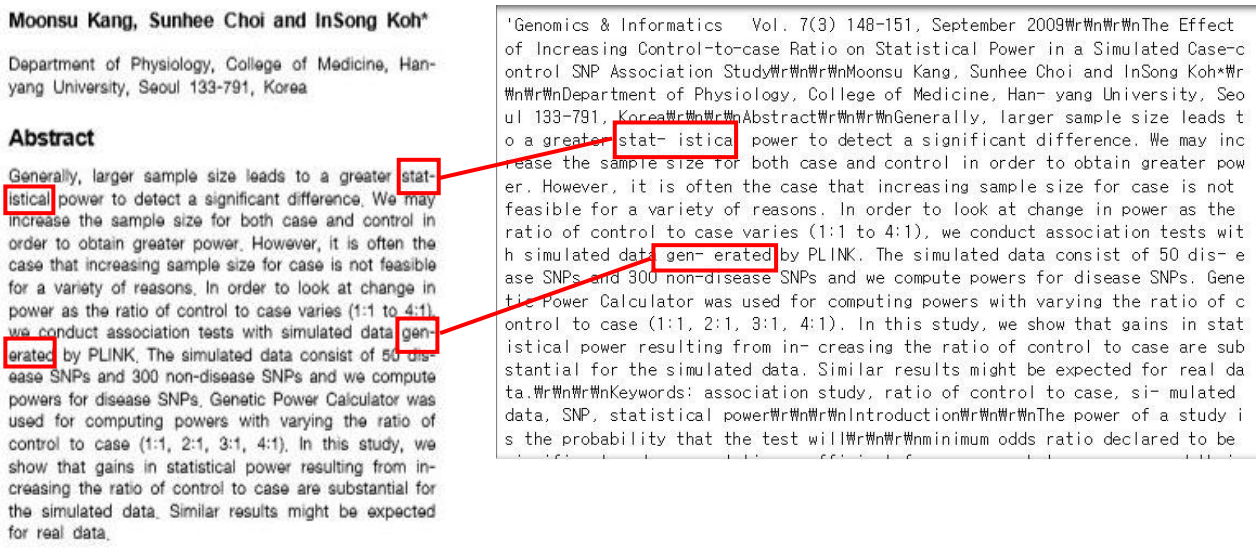
GNI Corpus Preprocessing

1829008 김민영

(먼저, 이전에 다운받았던 raw_text2 버전입니다.)

문제1 : 단어 사이가 -로 끊어짐

PlaintextCorpusReader로 읽어들인 후 raw_text2 데이터를 살펴보면, 줄이 바뀌는 부분에 단어가 걸쳐있는 경우, 아래그림의stat- istical이나 gen- erated 처럼 끊긴 단어의 대부분은 줄끝에서– 기호로 이어진 경우로, "-"기호와" "기호가 붙어있는 경우이다. 이때, re모듈을 이용하여 자동수정이 가능하다-단어사이에 '–'이 있는 경우 중 정상적인 단어의 연결도 많으므로 (예:school-age등) 모든 경우를 수정하는 것보다는 이방법으로 하는 것이 안전할 것이다.



Moonsu Kang, Sunhee Choi and InSong Koh*

Department of Physiology, College of Medicine, Han-yang University, Seoul 133-791, Korea

Abstract

Generally, larger sample size leads to a greater stat-istical power to detect a significant difference. We may increase the sample size for both case and control in order to obtain greater power. However, it is often the case that increasing sample size for case is not feasible for a variety of reasons. In order to look at change in power as the ratio of control to case varies (1:1 to 4:1), we conduct association tests with simulated data gen-erated by PLINK. The simulated data consist of 50 dis-ease SNPs and 300 non-disease SNPs and we compute powers for disease SNPs. Genetic Power Calculator was used for computing powers with varying the ratio of control to case (1:1, 2:1, 3:1, 4:1). In this study, we show that gains in statistical power resulting from in-creasing the ratio of control to case are substantial for the simulated data. Similar results might be expected for real data.

'Genomics & Informatics Vol. 7(3) 148-151, September 2009₩r₩n₩r₩nThe Effect of Increasing Control-to-case Ratio on Statistical Power in a Simulated Case-control SNP Association Study₩r₩n₩r₩nMoonsu Kang, Sunhee Choi and InSong Koh*₩r₩n₩r₩nDepartment of Physiology, College of Medicine, Han- yang University, Seoul 133-791, Korea₩r₩n₩r₩nAbstract₩r₩n₩r₩nGenerally, larger sample size leads to a greater stat- istical power to detect a significant difference. We may increase the sample size for both case and control in order to obtain greater power. However, it is often the case that increasing sample size for case is not feasible for a variety of reasons. In order to look at change in power as the ratio of control to case varies (1:1 to 4:1), we conduct association tests with simulated data gen- erated by PLINK. The simulated data consist of 50 dis- ease SNPs and 300 non-disease SNPs and we compute powers for disease SNPs. Genetic Power Calculator was used for computing powers with varying the ratio of control to case (1:1, 2:1, 3:1, 4:1). In this study, we show that gains in statistical power resulting from in- creasing the ratio of control to case are substantial for the simulated data. Similar results might be expected for real data.₩r₩n₩r₩nKeywords: association study, ratio of control to case, si- mulated data, SNP, statistical power₩r₩n₩r₩nIntroduction₩r₩n₩r₩nThe power of a study is the probability that the test will₩r₩n₩r₩nminimum odds ratio declared to be

해결

정규표현식 '[A-Za-z]+- [a-z]+' 로 re.sub로 해결할 수 있다.

```
In [11]: pattern1 = re.findall(r'[A-za-z]+- [a-z]+',giRaw)
         pattern1

Out[11]: ['Han- yang',
          'stat- istical',
          'gen- erated',
          'dis- ease',
          'in- creasing',
          'si- mulated',
          'cor- relation',
          'val- ues',
          'ep- idemiology',
          'Ambro- sius',
          'in- crease',
          'ra- tio',
          'neg- ative',
```

```
In [12]: [re.sub(r'- ','',p)for p in pattern1]

Out[12]: ['Hanyang',
          'statistical',
          'generated',
          'disease',
          'increasing',
          'simulated',
          'correlation',
          'values',
          'epidemiology',
          'Ambrosius',
          'increase',
          'ratio',
          'negative',
          'statistical',
          'prevalence',
          'statistical',
          'simulated',
          'disease',
          'model',
```

문제2: 아래첨자 띄어쓰기 문제

아래첨자에서 띄어쓰기가 발생한 것을 확인했다. 예를들어, 아래 그림처럼 귀무가설 대립가설 기호가 띄어쓰기로 잘못읽혀있다. (H0 -> H 0, H1-> H 1)

모두 찾기는 어려우므로, 위 기호와 같은 것들은 먼저 처리할 수 있을 것 같다.

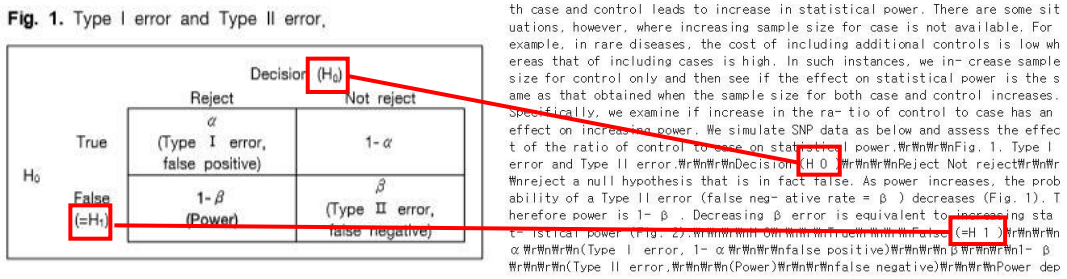


Fig. 1. Type I error and Type II error.

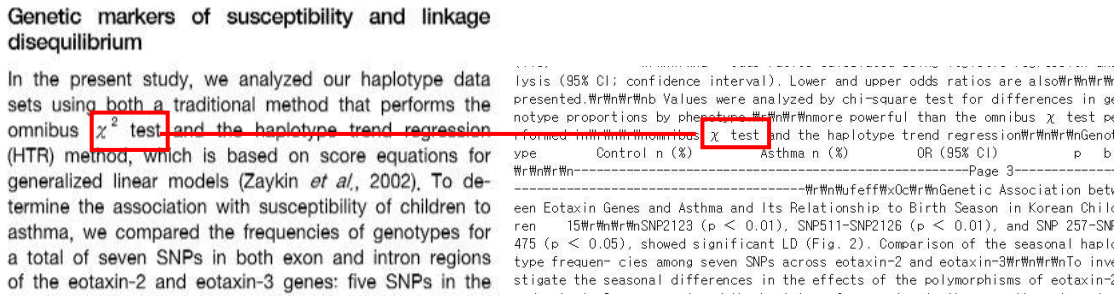
		Decision (H_0)	
		Reject	Not reject
H_0	True	α (Type I error, false positive)	$1-\alpha$
	False ($=H_1$)	$1-\beta$ (Power)	β (Type II error, false negative)

th case and control leads to increase in statistical power. There are some sit uations, however, where increasing sample size for case is not available. For example, in rare diseases, the cost of including additional controls is low wh ereas that of including cases is high. In such instances, we in- crease sample size for control only and then see if the effect on statistical power is the s ame as that obtained when the sample size for both case and control increases. Specifically, we examine if increase in the ra- tio of control to case has an effect on increasing power. We simulate SNP data as below and assess the effec t of the ratio of control to case on statistical power.₩r₩n₩r₩nFig. 1. Type I error and Type II error.₩r₩n₩r₩nDecision (H 0)₩r₩n₩r₩nReject Not reject₩r₩n₩r ₩nreject a null hypothesis that is in fact false. As power increases, the prob ability of a Type II error (false neg- ative rate = β) decreases (Fig. 1). T herefore power is 1- β . Decreasing β error is equivalent to increasing sta t- istical power (Fig. 2).₩r₩n₩r₩n1 Of₩r₩n₩r₩nFalse (=H 1)₩r₩n₩r₩n α ₩r₩n₩r₩n(Type I error, 1- α ₩r₩n₩r₩nfalse positive)₩r₩n₩r₩n β ₩r₩n₩r₩n1- β ₩r₩n₩r₩n(Type II error,₩r₩n₩r₩n(Power)₩r₩n₩r₩nfalse negative)₩r₩n₩r₩nPower dep

해결

```
In [15]: pattern2 = re.findall(r'[H]+ [0-1]+',giRaw)
         pattern2

Out[15]: ['H 0', 'H 0', 'H 1']

In [16]: [re.sub(r' ','',p)for p in pattern2]

Out[16]: ['H0', 'H0', 'H1']
```

문제3 : 카이스퀘어 검정에서 제곱 기호가 누락됨



Genetic markers of susceptibility and linkage disequilibrium

In the present study, we analyzed our haplotype data sets using both a traditional method that performs the omnibus χ^2 test and the haplotype trend regression (HTR) method, which is based on score equations for generalized linear models (Zaykin *et al*., 2002). To determine the association with susceptibility of children to asthma, we compared the frequencies of genotypes for a total of seven SNPs in both exon and intron regions of the eotaxin-2 and eotaxin-3 genes: five SNPs in the

lysis (95% CI; confidence interval). Lower and upper odds ratios are also
presented.
b Values were analyzed by chi-square test for differences in g
notype proportions by phe
more powerful than the omnibus χ test p
formed
χ test
and the haplotype trend regression
Genot
ype Control n (%) Asthma n (%) OR (95% CI) p b
Page 3
Genetic Association bet
een Eotaxin Genes and Asthma and Its Relationship to Birth Season in Korean Chil
ren 15
SNP2123 ($p < 0.01$), SNP511-SNP2126 ($p < 0.01$), and SNP 257-SN
475 ($p < 0.05$), showed significant LD (Fig. 2). Comparison of the seasonal hapl
type frequen- cies among seven SNPs across eotaxin-2 and eotaxin-3
To inv
stigate the seasonal differences in the effects of the polymorphisms of eotaxin-

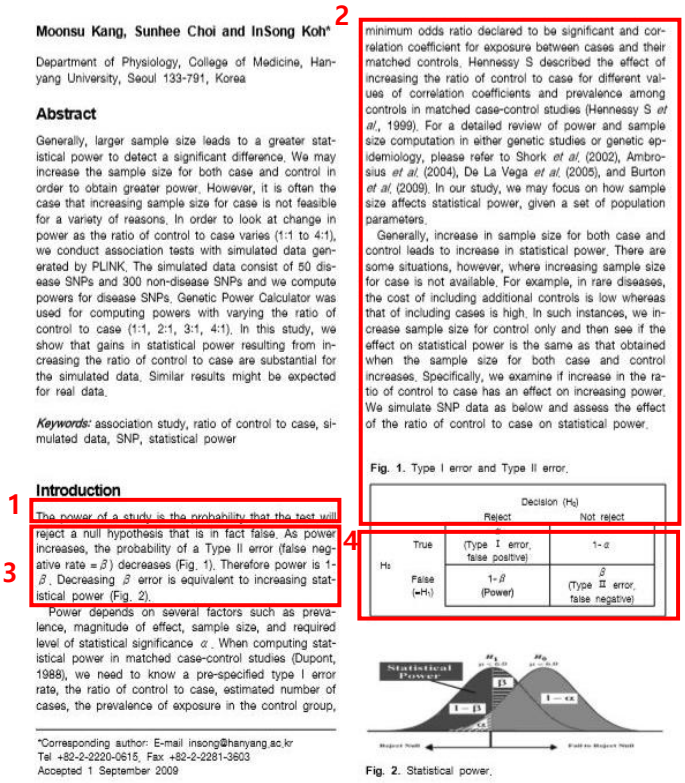
해결

아래와 같이 정규표현식을 사용하여 해결할 수 있다.

```
In [37]: pattern3= re.findall(r'χ test',giRaw3)
         pattern4

Out[37]: ['χ test', 'χ test']

In [38]: [re.sub(r'χ test','χ2 test',p)for p in pattern3]

Out[38]: ['χ2 test', 'χ2 test']
```

문제4 : 순서 문제

그림이 포함되어 있을 때, 아래와 같이 순서가 뒤죽박죽 되는 것을 확인할 수 있었다.



Moonsu Kang, Sunhee Choi and InSong Koh*

Department of Physiology, College of Medicine, Hanyang University, Seoul 133-791, Korea

Abstract

Generally, larger sample size leads to a greater statistical power to detect a significant difference. We may increase the sample size for both case and control in order to obtain greater power. However, it is often the case that increasing sample size for case is not feasible for a variety of reasons. In order to look at change in power as the ratio of control to case varies (1:1 to 4:1), we conduct association tests with simulated data generated by PLINK. The simulated data consist of 50 disease SNPs and 300 non-disease SNPs and we compute powers for disease SNPs. Genetic Power Calculator was used for computing powers with varying the ratio of control to case (1:1, 2:1, 3:1, 4:1). In this study, we show that gains in statistical power resulting from increasing the ratio of control to case are substantial for the simulated data. Similar results might be expected for real data.

Keywords: association study, ratio of control to case, simulated data, SNP, statistical power

Introduction

The power of a study is the probability that the test will reject a null hypothesis that is in fact false. As power increases, the probability of a Type II error (false negative rate = β) decreases (Fig. 1). Therefore power is 1-β. Decreasing β error is equivalent to increasing statistical power (Fig. 2).

Power depends on several factors such as prevalence, magnitude of effect, sample size, and required level of statistical significance α. When computing statistical power in matched case-control studies (Dupont, 1988), we need to know a pre-specified type I error rate, the ratio of control to case, estimated number of cases, the prevalence of exposure in the control group, minimum odds ratio declared to be significant and correlation coefficient for exposure between cases and their matched controls. Hennessy S described the effect of increasing the ratio of control to case for different values of correlation coefficients and prevalence among controls in matched case-control studies (Hennessy S *et al*., 1999). For a detailed review of power and sample size computation in either genetic studies or genetic epidemiology, please refer to Shork *et al*. (2002), Ambrosius *et al*. (2004), De La Vega *et al*. (2005), and Burton *et al*. (2009). In our study, we may focus on how sample size affects statistical power, given a set of population parameters.

Generally, increase in sample size for both case and control leads to increase in statistical power. There are some situations, however, where increasing sample size for case is not available. For example, in rare diseases, the cost of including additional controls is low whereas that of including cases is high. In such instances, we increase sample size for control only and then see if the effect on statistical power is the same as that obtained when the sample size for both case and control increases. Specifically, we examine if increase in the ratio of control to case has an effect on increasing power. We simulate SNP data as below and assess the effect of the ratio of control to case on statistical power.

*Corresponding author: E-mail insong@hanyang.ac.kr
Tel +82-2-2220-0615, Fax +82-2-2281-3603
Accepted 1 September 2009

Fig. 1. Type I error and Type II error.

		Decision (H_0)	
		Reject	Not reject
H_0	True	α (Type I error, false positive)	$1-\alpha$
	False (=H_1)	$1-\beta$ (Power)	β (Type II error, false negative)

Fig. 2. Statistical power.

1 The power of a study is the probability that the test will

2 minimum odds ratio declared to be significant and cor- relation coefficient for exposure between cases and their matched controls. Hennessy S described the effect of increasing the ratio of control to case for different val- ues of correlation coefficients and prevalence among controls in matched case-control studies (Hennessy S et al ., 1999). For a detailed review of power and sample size computation in either genetic studies or genetic ep- idemiology, please refer to Shork et al . (2002), Ambro- sius et al . (2004), De La Vega et al . (2005), and Burton et al . (2009). In our study, we may focus on how sample size affects statistical power, given a set of population parameters.

Generally, increase in sample size for both case and control leads to increase in statistical power. There are some situations, however, where increasing sample size for case is not available. For example, in rare diseases, the cost of including additional controls is low whereas that of including cases is high. In such instances, we in- crease sample size for control only and then see if the effect on statistical power is the same as that obtained when the sample size for both case and control increases. Specifically, we examine if increase in the ra- tio of control to case has an effect on increasing power. We simulate SNP data as below and assess the effect of the ratio of control to case on statistical power.

3 Fig. 1. Type I error and Type II error.

Decision (H 0)

Reject Not reject

4 reject a null hypothesis that is in fact false. As power increases, the probability of a Type II error (false neg- ative rate = β) decreases (Fig. 1). Therefore power is 1- β . Decreasing β error is equivalent to increasing stat- istical power (Fig. 2).

H 0

True

False (=H 1)

α

문제 5: 표 인식 문제

위에서의 순서문제와 마찬가지로 표의 중간에 다른 문장에 끼어들었으며, 표를 인식하는 순서도 뒤죽박죽인 것을 확인할 수 있었다.

Fig. 1. Type I error and Type II error.

H_0	Decision (H_0)	
	Reject	Not reject
True	α (Type I error, false positive)	$1-\alpha$
False (=H_1)	$1-\beta$ (Power)	β (Type II error, false negative)

Decision (H 0)

Reject Not reject

reject a null hypothesis that is in fact t
rate = β) decreases (Fig. 1). Therefore
power (Fig. 2).

H 0

True

False (=H 1)

α

(Type I error, 1- α

false positive)

β

1- β

(Type II error,

(Power)

false negative)

해결 : 문장이 아니므로 제거하는 것이 좋을 것 같음

문제 6 : 아래 사진과 같이 수식이 제대로 읽히지 않는다.

metry of the data of the probability distribution:

$$skewness = \frac{m_3}{\sigma^3}, \quad m_3 = \frac{\sum_{i=1}^{n}(x_i - X)^3}{n} \qquad (1)$$

skewness = m 3 =

,

σ (1)

n

n

∑ (x i − X) 3

m 3

i = 1

3

To determine the classification model, we adopted the logistic regression method with these features:

$$y = \alpha + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}, \quad i = 1, \ldots, n \qquad (2)$$

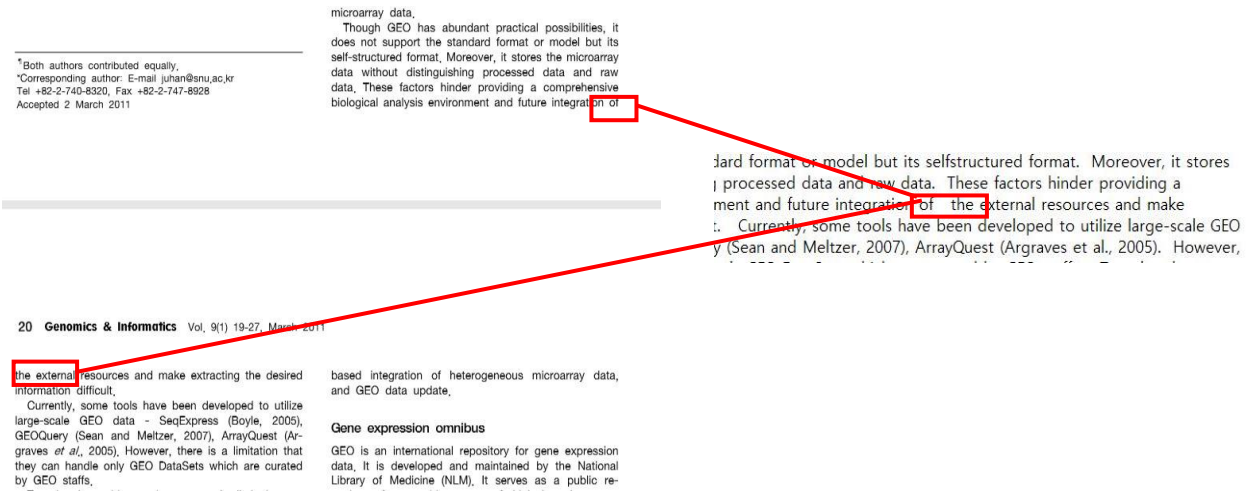
y = α + β 1 x 1 , i + L + β k x k , i ,
i = 1 , K , n

(2)

해결 : 문장이 아니므로 제거하는 것이 좋을 것 같다.

(여기서부터는 업데이트된 raw_text2 버전입니다.)

문제 : 페이지가 넘어갈 때, 문장이 걸쳐있는 경우, 띄어쓰기가 추가로 발생

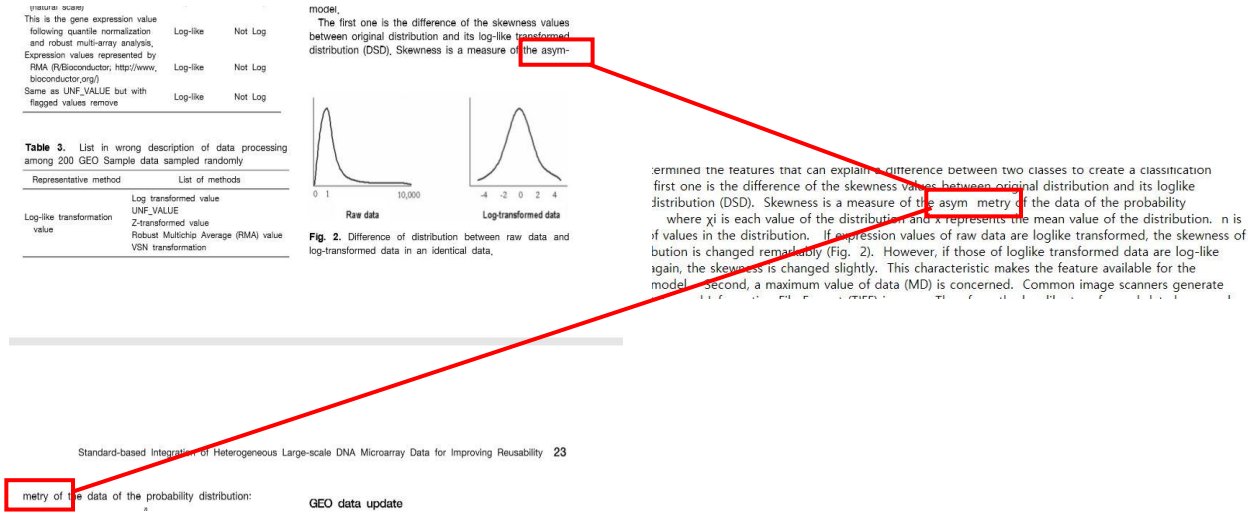


해결

한 파일에서 같은 문제가 총 4번 발생했음을 발견했고, 이를 정규표현식을 활용하여 아래와 같이 해결할 수 있었다.

```
In [54]: p1 = re.findall(r'[A-za-z][^.]+   [a-z]+', raw)
         p1

Out[54]: ['These factors hinder providing a comprehensive biological analysis environment and
         future integration of   the',
          'Each row in the data table corresponds to a single element, and includes sequence a
         nnotation and tracking information as   provided',
          'On the other hand, the fields of dual channel   in',
          'The training data set currently consists of 190   examples']

In [59]: [re.sub(r'   ',' ',p)for p in p1]

Out[59]: ['These factors hinder providing a comprehensive biological analysis environment and
         future integration of the',
          'Each row in the data table corresponds to a single element, and includes sequence a
         nnotation and tracking information as provided',
          'On the other hand, the fields of dual channel in',
          'The training data set currently consists of 190 examples']
```

문제 : 페이지가 넘어갈 때, 단어가 걸쳐있는 경우, 띄어쓰기가 추가로 발생



(natural scale)
This is the gene expression value following quantile normalization and robust multi-array analysis, Log-like Not Log
Expression values represented by RMA (R/Bioconductor; http://www.bioconductor.org/) Log-like Not Log
Same as UNF_VALUE but with flagged values remove Log-like Not Log

Table 3. List in wrong description of data processing among 200 GEO Sample data sampled randomly

Representative method	List of methods
Log-like transformation value	Log transformed value UNF_VALUE Z-transformed value Robust Multichip Average (RMA) value VSN transformation

model.
The first one is the difference of the skewness values between original distribution and its log-like transformed distribution (DSD). Skewness is a measure of the asym-

Fig. 2. Difference of distribution between raw data and log-transformed data in an identical data.

ermined the features that can explain a difference between two classes to create a classification
first one is the difference of the skewness values between original distribution and its loglike
distribution (DSD). Skewness is a measure of the asym metry of the data of the probability
where χi is each value of the distribution and x represents the mean value of the distribution. n is
f values in the distribution. If expression values of raw data are loglike transformed, the skewness of
bution is changed remarkably (Fig. 2). However, if those of loglike transformed data are log-like
again, the skewness is changed slightly. This characteristic makes the feature available for the
model. Second, a maximum value of data (MD) is concerned. Common image scanners generate

Standard-based Integration of Heterogeneous Large-scale DNA Microarray Data for Improving Reusability 23

metry of the data of the probability distribution:

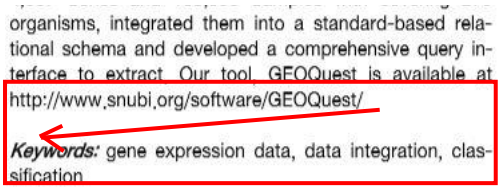
GEO data update

해결

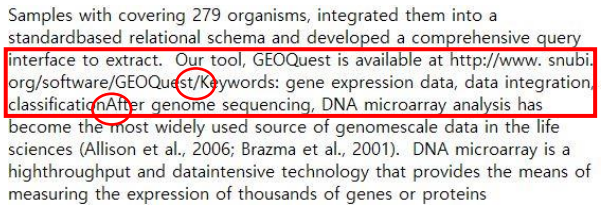
아래와 같이, 정상적으로 분리되어있던 단어들도 있었다. 따라서 경계에 해당하는 단어 (asymmetry) 만 따로 변경해주어야한다.

```
In [37]: p1 = re.findall(r'[^.][A-z0-9]+  [A-z0-9]+', raw)
         p1
Out[37]: [' Display  accession', ' number  was', ' distribution  In', 'asym  metry']

In [36]: p1 = [re.sub(r'asym  metry','asymmetry',p)for p in p1]
         p1 = [re.sub(r'Display  accession','Display-accession',p)for p in p1]
         p1
Out[36]: ['Display-accession', 'number  was', 'distribution  In', 'asymmetry']
```

문제 : Keywords 문제


organisms, integrated them into a standard-based relational schema and developed a comprehensive query interface to extract. Our tool, GEOQuest is available at http://www.snubi.org/software/GEOQuest/

Keywords: gene expression data, data integration, classification


Samples with covering 279 organisms, integrated them into a standardbased relational schema and developed a comprehensive query interface to extract. Our tool, GEOQuest is available at http://www. snubi.org/software/GEOQuest/Keywords: gene expression data, data integration, classificationAfter genome sequencing, DNA microarray analysis has become the most widely used source of genomescale data in the life sciences (Allison et al., 2006; Brazma et al., 2001). DNA microarray is a highthroughput and dataintensive technology that provides the means of measuring the expression of thousands of genes or proteins

GeoQuest/뒤에 키워드가 붙어버렸다.

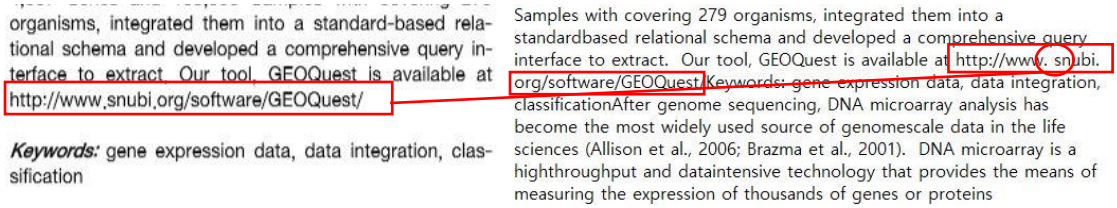
해결 키워드는 예외적으로 있는 거니까 키워드 앞에 공백문자를 넣어주면 된다.

```
In [113]: p4 = re.findall(r'[A-za-z][^.]+Keywords:', raw)
          p4
Out[113]: ['org/software/GEOQuest/Keywords:']

In [114]: [re.sub(r'Keywords:',' Keywords:',p)for p in p4]
Out[114]: ['org/software/GEOQuest/ Keywords:']
```

다만, Keywords 문장이 끝나는 부분에서도 다음문장과 붙어버리는(classificationAfter) 문제도 있는데, 이는 이 파일에서 찾기 어려우므로, Keywords 가 들어간 문장을 찾아 공백문자를 넣어주어 수작업으로 진행해야한다.

문제 : 링크 띄어쓰기 문제

아래 사진과 같이 www.snubi.org 로 읽혀야하는데, www. snubi. org 이런식으로 띄어쓰기가 들어갔다. 경우에 따라, 띄어쓰기가 두번 들어가기도 했고, 띄어쓰기가 한번 들어가기도 했다.



해결

아래와 같이 정규표현식으로 url주소를 먼저 찾은 후, 공백을 제거해주면 된다.

```
In [104]: p3 = re.findall(r'http://[A-z]*₩. [A-z]*₩.(?:  | )[A-z]*', raw)
          p3

Out[104]: ['http://www. snubi. org',
           'http://mged. sourceforge  net',
           'http://geo. snubi. org']

In [105]: [re.sub(r' ','',p)for p in p3]

Out[105]: ['http://www.snubi.org', 'http://mged.sourceforge.net', 'http://geo.snubi.org']
```