Authors

Ching-Syang Jack Yue

Professor, Department of Statistics, National Chengchi University
64, Sec.2, Zhinan Rd., Wenshan District, Taipei, Taiwan 11605
Tel: (02) 2938-7695 Email: csyue@nccu.edu.tw

Li-Hsing Ho

Assistant Professor, Department of Chinese Literature, National Tsing Hua University
101, Sec. 2, Kuang-Fu Rd., Hsinchu, Taiwan 30013
Tel: (03) 516-5039 Email: lillianlhho@gmail.com

Yan-Yan Pan

Master, Department of Statistics, National Chengchi University
64, Sec.2, Zhinan Rd., Wenshan District, Taipei, Taiwan 11605
Tel: (02) 2939-3091 Ext.89020 Email: christinafun1128@163.com

Wen-Huei Cheng

Professor, Department of Chinese Literature, National Chengchi University
64, Sec.2, Zhinan Rd., Wenshan District, Taipei, Taiwan 11605
Tel: (02) 2939-3091 Ext.88051 Email: wenhuei_cheng@yahoo.com.tw

A Quantitative Study of Chinese Writing Style based on *the New Youth Magazine*

Abstract

Big Data probably is the most popular term in recent years and it also has significant influences on the study of languages. Through vast archives of digital text and new developed analysis methods, the linguists have more options to do research. In particular, now we can cross-check several materials at the same time and in greater detail than before. Digital humanity becomes a fast growing research field all around the world. However, even with the help of computers, the humans are still the core of digital humanity and we rely on experts' opinions to determine key elements for data analysis. Perhaps this is the main reason that the development of text mining is relatively slow.

One of the main difficulties in conducting text mining is the data input. Texts are a form of unstructured data and they need to be quantified first. So far, there is no standard operating procedure (SOP) for quantifying text and experts play an important role in selecting relevant information (or variables) for data analysis. In this study, our goal is to propose a SOP of text mining and use it to study the change of Chinese writing style, from classical Chinese to modern Chinese. The study material is Volumes 1~7 of *the New Youth magazine*. It is believed that, by the end of Volume 7, modern vernacular Chinese had almost completely replaced classical Chinese

We will adapt the idea of un-supervised learning from statistical learning theory to define and identify important variables. In particular, we will use the notion of Exploratory Data Analysis (or EDA, proposed by famous statistician J.W. Tukey in 1977) to evaluate potential variables which can differentiate the language styles of Volumes 1~7 in *the New Youth magazine*. Also, according to our previous study, there are quite a lot of variables and data reduction methods will be needed. We use principle component analysis to select fewer variables and apply classification methods (e.g., logistic regression) to judge whether the style of an article is close to classical Chinese or modern Chinese. Also, to avoid over parameterization (i.e., using too many unnecessary variables), we use cross validation to select feasible models. Cross validation is to separate the data into training set (or in-sample) and testing set (or out-sample). The training set is used to construct model and the constructed model is applied to the testing set for calculating model accuracy.

Our study shows that the change in writing style of articles from *the New Youth magazine* seems to be gradual from Volume 1 to Volume 7. For example, about 60% of articles in Volume 4 are classified to the group of Volume 7 (or modern Chinese), comparing to 99% of articles in Volume 1 are classified to classical Chinese and 98% of those in Volume 7 are classified to modern Chinese. The prediction accuracy of articles in Volume 4 is about 84% (cross-checked with experts of linguists in Chinese). The results of numerical analysis are promising and we should continue the study of modern Chinese writing via quantitative analysis.

Keyword: Classical and Modern Chinese, New Youth Magazine, Species Diversity,
Unstructured Data, Logistic Regression

1. Introduction

Big data is one of the most popular topics in recent years and experts in all application fields are talking about its influence. The quantitative analysis of languages and text, namely Text Mining, also becomes popular. However, the development of text mining is relatively slow and one of the main reasons is that the analysis of text requires extra efforts, such as the domain knowledge in languages. Basically, there are two types of data: structured data and unstructured data. The structured data are those with a high degree of organization, such as data stored in Excel spreadsheets, and the unstructured data are those without such organization. Most of the information recorded belongs to the unstructured data (about 80% ~ 90%), but the majority of data analyses are still for the structured data.

In order to conduct the quantitative analysis to the unstructured data, first we need to give them a structure. However, there are no correct answers or guidelines to define the structure. Perhaps this is why analysing the unstructured data is difficult. In this study, we propose a quantitative approach to give the Chinese texts a structure and use the variables defined to analyse Chinese writing style. In specific, we are interested in comparing the classical Chinese and the modern Chinese. It is believed that the May Forth Movement in 1919 is the key event that divides the classical and modern Chinese, and *the New Youth Magazine* is the most important magazine to spot the change of Chinese writing.

There are 11 volumes in *the New Youth Magazine* and, according to our previous study, modern vernacular Chinese had almost completely replaced classical Chinese as the main written language by the end of Volume 7 (published in 1920). In other words, we can use Volumes 1~7 to study the process of modern Chinese writing transform. If we use the term “species” to resemble “keywords” then the change of writing style bears some analogy to the change of habitat, where the old species are replaced by new species. Thus, we should apply the concept of ecological habitat to explore the writing style changes over time, and use quantitative analysis to compare the differences of writing style between classical Chinese and modern Chinese.

In specific, we propose two kinds of approaches, supervised learning and un-supervised learning. The first approach is to analyse the variables assigned from expert opinions. We adapt the expert opinions from humanist for data analysis. The second approach basically is data-driven, applying the concepts and methods used in lexical analysis. In addition, we adapt the ideas in ecology and specifically we analyse the changes of writing style according to views in species diversity and evolution. We can treat the word as the species and deem individual's writing style as different ecological system.

Feature selection plays an important role in distinguish writing style. Thus, we should adapt the idea of un-supervised learning from statistical learning theory to define and identify important variables. In particular, we will use the notion of Exploratory Data Analysis (or EDA, proposed by famous statistician J.W. Tukey in 1977) to evaluate potential variables which can differentiate the language styles of articles in *New Youth Magazine*. Also, according to our previous study of writing style in Chinese, it is very likely that there are quite a lot of variables and data reduction methods will be needed. We will use principle component analysis to select fewer variables and then apply classification methods (e.g., logistic regression, classification tree) to judge whether the style of an article is close to classical Chinese or modern Chinese. Also, to avoid over parameterization (i.e., using too many unnecessary variables), we use cross validation to select feasible models. Cross validation is to separate the data into training set (or in-sample) and testing set (or out-sample). The training set is used to construct model and the constructed model is applied to the testing set for calculating model accuracy.

2. Methodology and Data

Our goal is to study the language change of the early volumes of *the New Youth Magazine*, particularly the change of writing style from classical to modern Chinese. In specific, we speculate that Volumes 1 and 2 of *the New Youth* Magazine are not influenced by the May Forth Movement in 1919 and can be deemed as the classical Chinese writing style. On the other hand, the writing style of Volume 7 is close to the modern Chinese. In other words, we can label 0 and 1 (i.e., modern and classical Chinese) for articles from Volume 7 and Volume 1 of *the New Youth Magazine*, respectively, indicating some and little influences of the May Forth Movement. Thus, we can use the logistic regression, with the target value ranging between 0 and 1, to determine whether articles are influenced by the May Forth Movement. Based on the selected variables, we construct a regression model to classify articles and apply the model to evaluate the articles closer to classical Chinese or modern Chinese.

The logistic regression is to use independent variables (or $x_i, i = 1, 2, \ldots, k$) to assign the target variable (or y) a value between 0 and 1,

$$y = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}} \tag{1}$$

or

$$\log it(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{2}$$

if we define $E(y) = p$ (Agresti, 1990). The key for applying the logistic regression usually is to select appropriate variables and their functional forms, and it is especially difficult in the case of text mining and other soft data. We should provide a possible way for selecting variables via exploratory data analysis (Tukey, 1977) in the following section.

Logistic regression is a popular tool for binary classification, with value 0 and 1 indicating one group and the other group, respectively. Usually, we assign the observations with fitted values smaller than 0.5 (which is also threshold value) to the "0" group and those larger than 0.5 to the "1" group. The accuracy of logistic regression can be measured by $\frac{a+d}{a+b+c+d}$, where the values *a*, *b*, c, and *d* are defined in Table 1. The observations associate to *a* and *d* are correctly classified, while those associate to *b* and *c* are not. Note that, for the empirical consideration, if the sizes of "0" and "1" groups differ a lot, the threshold value can be set to a value other than 0.5.

Table 1. Evaluating Fitting Accuracy

		Fit	
		0	1
True	0	*a*	*b*
	1	*c*	*d*

If there are many independent variables, we usually apply variable (or data) reduction techniques, such as PCA (Principal Component Analysis), for variable selection. We can employ the PCA to determine the maximal number of independent variables required in regression analysis. Of course, if our goal is the accuracy of logistic regression, we might not need to consider the PCA. Note that the PCA is a famous method of multivariate analysis (Johnson and Wichern, 2007) and, in addition to data reduction, it can also be used to interpret the regression result. However, after applying the PCA, the original independent variables are transformed and the interpretation of regression model might change as well.

Since a logistic regression model for the data from Volumes 1 and 7 is constructed first and then this model is applied to data from Volume 4, we need to evaluate the stability of regression model. The evaluation can be achieved by a cross validation process, in order to make sure that the model is not data-sensitive (or data-dependent). Usually, we divide the data into k equal parts (also called k-fold). Then, we use k-1 parts of data (i.e., training set) to construct the model and apply the model to the remaining part of data (i.e., testing set). The model is deemed to be stable if the model accuracies (Table 1) of training and testing sets are about the same. The k-fold evaluation can also be done by repeating randomly separating the data for a few times. We prefer repeating the k-fold evaluation to avoid how we separate the data.

The data considered in this study are articles from *New Youth Magazine*. The *New Youth Magazine*, *or La Jeunesse,* is an important Chinese magazine in 1910's and 1920's and can be used to study the spreading of Modern Chinese since the May Forth Movement in 1919. The influence of May Forth Movement can well be observed from articles of first 7 volumes (Lo et al., 2014). The *New Youth Magazine* openly supported the communism starting from Volume 8, influenced by the 1917 Russian October Revolution, and became an official journal of Chinese Communist Party from Volume 10. We choose the articles from Volumes 1, 4, and 7 for data analysis in this study.

3. Exploratory Data Analysis

Roughly speaking, there are two types of big data: one is structured data and the other is unstructured data, which are also called hard data and soft data, respectively. The structured data are those with a high degree of organization. Examples of structured data include library catalogues of books (such as published date, authors, and place) and population census records (such as birth date, income, and address). This type of data usually can be quantified without difficulty and are easy to be input, stored, and analysed.

On the other hand, the unstructured data usually are lacking structure and quantifying them often requires certain knowledge about the application domain. Thus, converting the structure data into numeric values somehow is subjective and the values can be very different depending on the problem/users. Most of the textual data are unstructured and analysing them, or Text mining, needs a lot of extra work. We need to create a relational structure for the textual data before plugging them into logistic regression. So far, there are no standard operating procedures for structuring the textual data. We think that the exploratory data analysis (EDA), proposed by Tukey (1977), is a feasible approach and we should demonstrate how we structuring the textual data.

The EDA is to discover the data properties through basic statistical analysis, such as computing the sample average and sample variance. For the textual data, the basics of EDA include the number of words, the number of different words (or vocabularies) and their distribution. Table 2 shows the numbers of words and different words for 11 volumes of the *New Youth Magazine*. In general, there are more words for the later volumes but the number of vocabularies is not. This property can also be well described by species diversity indices, such as Simpson index and entropy (or Shannon index), which are defined as $\theta_S = \sum_i p_i^2$ and $\theta_E = -\sum_i p_i \ln(p_i)$, respectively, where p_i is the proportion of vocabulary i. Note that the larger the entropy is, the larger the species (or vocabulary) diversity. On the contrary, a smaller value of Simpson index indicates larger diversity. In other words, the statistics in Table 2 imply smaller diversity in later volumes.

Table 2. Words Count of the *New Youth Magazine*

Volume	# of Words	# of Vocabularies	Simpson Index	Entropy
1	248,833	4,379	0.004568	6.654036
2	291,848	4,344	0.004500	6.649539
3	290,038	4,227	0.004954	6.541824
4	305,020	4,298	0.004172	6.539378
5	343,519	4,125	0.004672	6.461579
6	389,407	3,848	0.005749	6.348547
7	586,942	3,850	0.006053	6.328604
8	461,731	3,753	0.006035	6.320355
9	437,748	3,745	0.005574	6.322103

10	342,778	2,980	0.005700	6.177278
11	489,223	3,093	0.005712	6.212699

In addition to the counts of words and vocabularies, we can also use other attributes of the Chinese writing for style classification. The early volumes of the *New Youth Magazine* are deemed to be of the style of classical Chinese, while the later volumes are of modern Chinese. Two of the main differences between these two writing styles are the sentence length and the common function words. For the sentence length, the punctuations “ ， 。 ； ！ ？ ” (comma, period, semicolon, exclamation mark, and question mark) are used to separate sentences. The proportion of sentences with 4, 5, & 6 words is about 50% in volume 1 and the proportion of sentences with 4, 5, 6, 7, & 8 words is about 50% in volume 7, showing the message that later volumes have longer sentences and larger variances. We can also compute the average numbers of words (Figure 1) and their standard deviations to demonstrate the sentence length in a volume, and they are 7.07 (1.07) and 9.27 (1.99) for volumes 1 and 7, respectively. It is obvious that more words are used in the later volumes and it matches to our consensus that the classical Chinese are simple and concise.

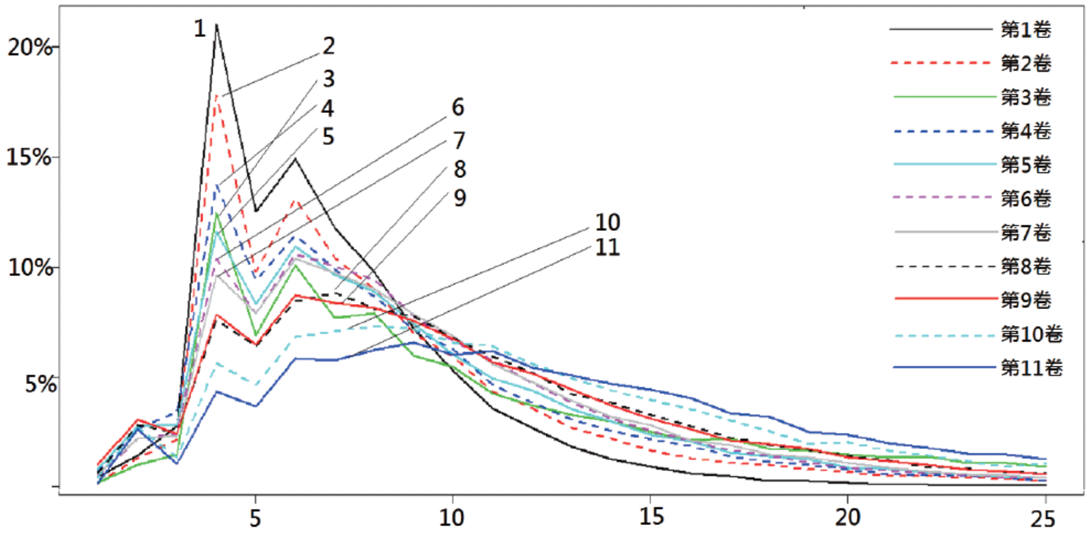

Figure 1. Average Number of Words in a Sentence

Other variables showing the species richness are also considered in this study, in addition to those introduced above, including the number of new vocabularies every 1,000 words in each volume, 10 most common single words, and 10 most common two-word keywords. The idea behind the number of new vocabularies every

1,000 words is similar to that of type-token ratio (TTR). Since more words usually are with smaller TTR, we use 1,000 words as a form of normalization.

Table 3. 10 Common Function Words

	Classical Chinese	Modern Chinese
Words	矣乎焉歟哉耳豈之乃無	的是們個了和麼著嗎吧
Volume 1 Proportion	3.6%	0.7%
Volume 7 Proportion	0.5%	8.8%
Total Proportion	2.4%	7.3%

The variables mentioned above belong the type of un-supervised learning (or data driven) and we should also consider supervised learning variables, i.e., suggestions from expert opinions. Function words are one of the popular choices for the study of writing style. We adapt the idea of Ho et al. (2014) and choose 10 common function words in classical Chinese and 10 in modern Chinese. Again, we use the statistics of volumes 1 and 7 as a demonstration. As expected, there are more classical function words for volume 1 and more modern function words for volume 7 (Table 3). It seems that the pattern of function words used is very different.

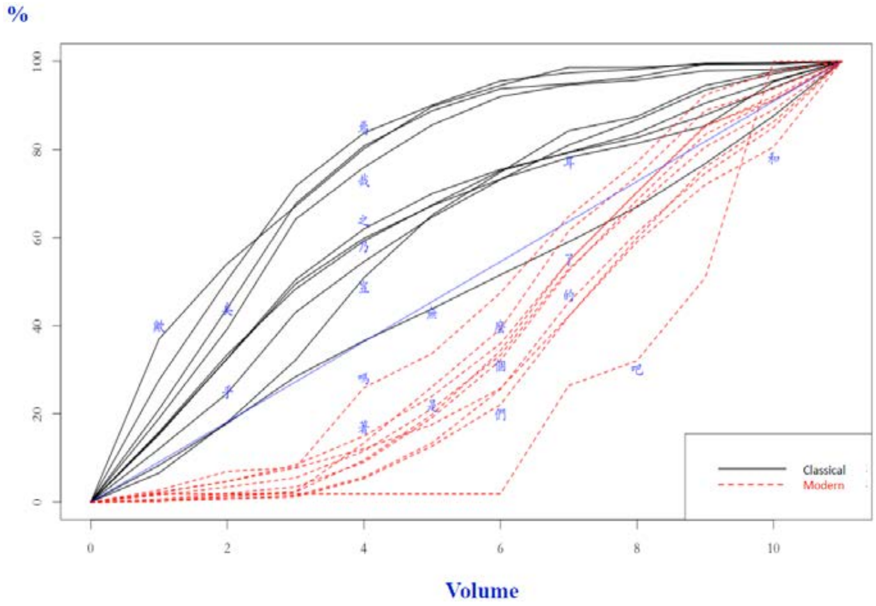

Figure 2. Time Trends of 20 Function Words

We also use graphs to demonstrate the time trends of these 20 function words. The time trend of each word is calculated similar to the idea of Gini's index, via cumulating the percentage of word used, as shown in Figure 2. The lines above the blue line indicate words used more in the early volumes, while the lines below the blue line are the words used more in the later volumes. Except the word "無", all lines of classical Chinese function words are above the blue line and all lines of modern Chinese function words are below the blue line. It seems these 20 function words are fine candidates to differentiate the Chinese writing style.

Table 4. Variables in the Logistic Regression

Types	Variables
Words & Vocabulary	Total numbers of words and vocabulary
	New vocabulary every 1,000 words
	The number of cumulative vocabulary
	Simpson Index
	Entropy
Sentence	Average sentence length
	Variance of sentence length
Function Words	10 Classical and 10 modern Chinese
Common Words	10 most common single words and 10 most common two-word keywords

All variables considered are summarized in Table 4 and they will be used to differentiate the writing styles of articles for the *New Youth magazine*. In the next section, we should apply these variables to construct logistic regression model.

4. Data Analysis of *the New Youth*

We will apply the variables selected from the previous section into the logistic regression in this section. First, we focus on verifying if the regression model can provide accurate and stable estimation. As mentioned in the previous section, the textual data from the *New Youth Magazine* are used to construct the regression model. In specific, we label "0" and "1" to the articles if they belong to modern Chinese and classical Chinese, respectively. And then we build a logistic regression model and check if we can distinguish articles are from modern Chinese or classical Chinese. Also, we also evaluate if the constructed model is stable and reliable by cross validation.

Table 5 shows the classification results of articles from volumes 1 and 7. There are 162 and 132 articles in volumes 1 and 7, respectively. Among all 294 articles, 289 of them are classified correctly, or 98.3% classification accuracy. Further, in addition to the accuracy, we want to check the stability of the regression model via cross validation. The regression model is first built based on training data and then applied to the testing data. The fitting accuracies of training data and testing data are recorded separately, and these two numbers of accuracy should be close if the model is stable. For every simulation run, we randomly separate the training data and testing data into proportions of 90% and 10%, respectively. Table 6 shows the averages and their standard errors of the fitting accuracy for training and testing data from 100 simulation runs. Apparently, the regression model is fairly stable since it has very similar fitting accuracy (and small standard errors) for training and testing data. Note that the size of testing data is smaller, and thus its standard error is larger.

Table 5. Classification Results of Volumes 1 and 7

		Fit	
		Volume 1 (Classical)	Volume 7 (Modern)
True	Volume 1 (Classical)	160	2
	Volume 7 (Modern)	3	129

Table 6. Cross Validation Results of Volumes 1 and 7

	Fitting Accuracy	
	Average	Standard Error
Training	96.10%	0.07%
Testing	95.95%	0.31%

Note that, since there are quite a lot of variables (Table 4), we can also employ the PCA to reduce the number of variables in the regression model, without sacrificing the accuracy of classification. The number of principal components (or variables) is 3 or 4, much smaller than the list in Table 4. However, the variables selected via the PCA are linear combinations of the original variables and it is usually difficult to give proper interpretation for these linear combinations. If our goal is to provide interpretation of classification with respect to the original variables, then we suggest not using the PCA for variable reduction.

Table 7. Classifications of articles in Volume 4

		Fit	
		Classical	Modern
True	Classical	34	0
	Modern	13	32

Next step, we apply the constructed regression model to the articles in Volume 4. Among 79 articles, 66 articles are correctly classified (83.54% accuracy), shown in Table 7. All articles of classical Chinese are perfectly classified, while the fitted accuracy of modern Chinese articles is about 72.73%. If we look at the results from the other way, the articles labelled as modern Chinese by logistic regression are truly modern Chinese. These numbers are interesting and probably this implies that the variables chosen have systematic deficiency and are insufficient to differentiate classical and modern Chinese.

Figure 3 shows the detailed results of classification, after the kernel smoothing. Fitted values closer to 0 and 1 (modern and classical Chinese) indicate stronger belief of logistic model, and values closer to 0.5 indicate ambiguous decisions. The fitted results of Volumes 1 and 7 are all close to 0 or 1, but those of Volume 4 have a moderate chance being around 0.5. This coincides with the results of classification accuracy (Tables 5, 6, and 7).

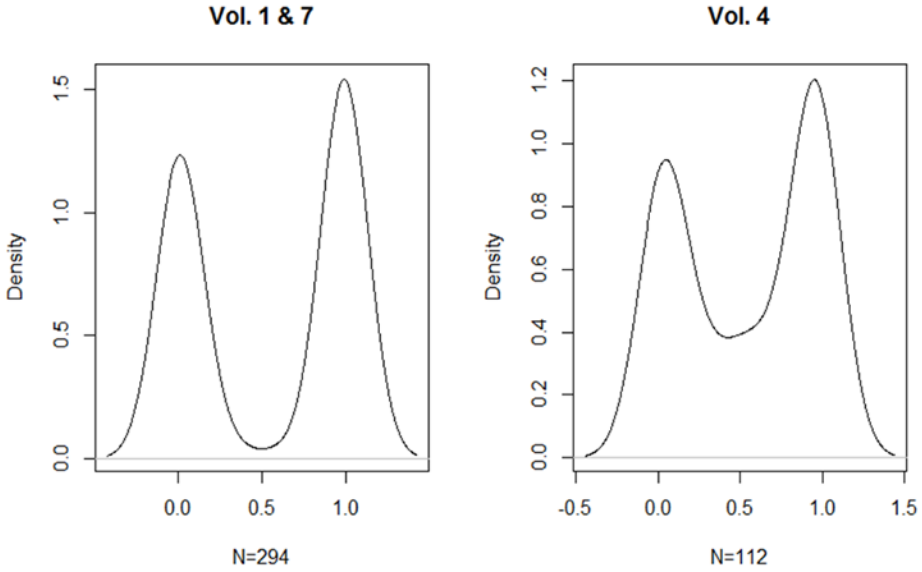

Figure 3. Classification results of Volumes 1, 4, and 7

5. Conclusion and Discussions

2015 marks the 100th anniversary of *The New Youth* magazine. By employing the digital humanity methods and focusing on the language change of the earlier volumes, we hope to make some new contribution to the study of both *the New Youth magazine* and the writing style of modern Chinese language. Our fitting results are stable and their variances are fairly reliable via cross validation. This suggests that the writing styles of Volumes 1 and 7 are very different and using logistic model and the variables can easily differentiate the articles from Volumes 1 and 7. This supports that the numerical analysis is a feasible approach.

In addition, we learn that the writing styles of articles (identified by logistic regression) in *The New Youth magazine* gradually turn from classical Chinese to modern Chinese, from Volume 1 to Volume 7. If we apply the constructed model to Volumes 2 to 6, about 68% and 12% of articles in Volume 2 and 6 are classified as classical Chinese (or 32% and 88% of articles) are classified as modern Chinese. It seems that the results of quantitative analysis coincide with the chronical change of *the New Youth Magazine*. The results of our study encourage further studies of Chinese writing styles via the notion of big data/digital humanity. The numerical approach does provide a new possibility of research in social and humanity science.

As for the variables for classification analysis, only the numbers of words/vocabularies and the length of sentences are included in this study. Of course, the information related to the articles' content should be used for classification, but it would require more work. We suggest the EDA-type procedure used in this study to select content related variables, to avoid researchers' subjective for choosing words. We can first summarize the most common words/keywords, and then decide which words/keywords can represent the meaning of an article and be used for classification. However, using different words does not necessary indicate different styles or meanings. For example, the synonyms and antonyms should be applied with care, and a negative mood can turn antonyms into synonyms.

Experts estimated that about 80% ~ 90% of the data in any organization are unstructured. Unstructured data, according to Wikipedia, are the information does not have a pre-defined data model or are not organized. Converting text data into numerical data (or giving unstructured data a structure, namely structurization) is the first and probably the most important step for analysing text data. Once the texts are transformed into numerical data, we can apply the well-developed datamining techniques to the text data. However, although past studies tried to give a standard operating procedure for structuring unstructured data, no consensus is reached on converting the text data and there are a lot of rooms for data transformation.

Text mining often requires multi-discipline knowledge and probably it is one of the reasons why the data transformation is difficult. In fact, team work is essential in handling big data, not only for text data. It is a difficult task if only one person or experts from one area are involved. For example, identification of keywords/key-phrases needs the domain knowledge of application field, as well as the considerations in quantitative analysis and data structure. In the future, we expect to see more collaboration between different fields of experts.

References

- Agresti, A., Categorical Data Analysis. New York: Wiley, 1990.
- Hastie, T., R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *2nd Edition,* Springer Series in Statistics, 2009.
- Ho, L., C.J. Yue, and W. Cheng, "From Classical Chinese to Modern Chinese: A Study of Function Words from New Youth Magazine." *Journal of the History of Ideas in East Asia* 7 (2014): 427-454.
- Johnson, R.A. and D.W., Wichern, Applied Multivariate Statistical Analysis (6th edition). Pearson, 2007.
- Karlgren, B., "New Excursions in Chinese Grammar." *Bulletin of the Museum of Far Eastern Antiquities* (Stockholm) 24 (1952): 51-80.
- Mosteller, F. and D. Wallace, Inference and Disputed Authorship: the Federalist. Addison-Wesley, 1964.
- Shannon, C.E. and W. Weaver, A Mathematical Theory of Communication. *The Bell System Technical Journal* 27 (1948): 379–423 & 623–656.
- Thisted, R. and B. Efron, "Did Shakespeare Write a Newly-discovered Poem?" *Biometrika* 74, no. 3 (1986): 445-455.
- Tukey, J.W., Exploratory Data Analysis. Addison-Wesley, 1977.