Authorship of Dream of the Red Chamber: A Topic Modeling Approach

Keli Du
keli.du@stud-mail.uni-wuerzburg.de
University of Würzburg, Germany

Dream of the Red Chamber (DRC, 红楼梦) is one of the most famous Chinese classic novels, written by Cao Xueqin (曹雪芹) during the 18th century. The original version of DRC had 80 chapters. But in 1791, Gao E (高鹗) and Cheng Weiyuan (程伟元) claimed that they had found more manuscripts of Cao and published another edition with 120 chapters. Since then, there has been a lot of discussions regarding the number of authors of DRC. Many scholars see the last 40 chapters as a later addition. Currently Hu Shih's (胡适) (Hu, 1988) research is most widely accepted, where he argued these last 40 chapters were written by Gao E. According to some modern research approaches, the first 80 and the last 40 chapters are written by two authors. Evidence also suggest that Chapters 64 and 67 may not be written by Cao (Hu, Wang, & Wu, 2014; Tu, & Hsiang, 2013).

Using Delta (Burrows, 2002), a measure of difference between two texts, the same conclusion has been obtained (Du, 2016). The 120 chapters are written by two authors. Red texts are the first 80 chapters and green texts are the last 40 chapters. Delta also suggest that chapter 6, 10, 11 and 67 might be written by the second author (see Fig. 1, 2, 3).

Although most the results obtained are within expectations, the presence of the four chapters in the second group certainly deserves further investigation. Three hypotheses are proposed in this paper as the cause for this situation:

- This test result from the Delta method is not 100% accurate.
- These four chapters shares many names or plot related terms with the last 40 chapters.
- Stylistic difference. Compared to the other chapters, the use of some less plot related terms indicates that the second author wrote Chapter 6, 10, 11 and 67.

Topic Modeling was used to test DRC on this regard. I used the version of the DRC that Tu (2013) deemed "the closest to the earliest editions" for this study. Topic Modeling can automatically discover the contents of a large collection of documents, and is often used as an alternative method to explore the documents. A topic is a probabilistic distribution over words appearing in the corpus. The model finds groups of related words, and words that occur frequently together will be clustered in the same group. If some words tend to co-occur in the last 40 chapters of DRC, Topic Modeling would be effective in highlighting their presence. LDA (Latent Dirichlet Allocation) (Blei et al., 2003) was used to model my corpus and do my test, and MAchine Learning for LanguagE Toolkit (MALLET) was used as the topic-modeling tool.

At the preprocessing step, tokenizing and chunking were performed. Tokenizing is required to process texts written in the Chinese Language to spilt texts into words, as there are no spaces to mark word boundaries. Tools like Stanford Chinese Word Segmenter can only be used on modern Chinese texts, as the segmentation standards are not suitable for classic Chinese. Character bigrams were hence selected as the "word". Breaking the texts up allows the relationship among words to be explored more thoroughly, hence the chapters were split, where each document for the test contains 500 bigrams. Stopwords present another issue. A test run was first performed with MALLET to acquire some topics. The results of the test run show the bigram words assigned to the topics. In these topics the correct bigram words and person names were observed. Besides, a significant amount of meaningless bigrams in the topics were also found, for example: 著一 (writings, one), 的干 (and so on, do), 云笑 (cloud, smile). Thereby a stopword list with the following was compiled: a collection of meaningless bigrams and person names and function words like 一個 (a), 這個 (this), 只得 (have to) and so on.

MALLET was run after the preprocessing stage to generate 50 topics. The topic-document distribution output was aggregated as the chapters were split into chunks at the preprocessing step. The average of the topic shares associated with each chapter were then computed and the shares were transformed into a document-topic matrix. Then the visualization of the topic proportions associated with the chapters was created (see Fig. 4). The x-axis are the topics and the y-axis are the documents. The red line divides the figure

into two parts, the first 80 chapters (above the line) and the last 40 chapters (under the line). Topic #26 became the main focus from this result: the last 40 chapters are all strongly associated with it, while the first 80 chapters (except Chapter 11 and 67) are associated to this topic with much lower probability. The topic-word assignments were then computed. The top 20 words and their unnormalized weights of the topic #26 are:

> 357.0: 太太 (lady), 246.0: 回來 (back), 233.0: 來了 (coming), 218.0: 丫頭 (slave girl), 190.0: 過來 (come over), 184.0: 答應 (answer), 146.0: 只見 (seen), 125.0: 去罷 (go), 122.0: 時候 (moment), 114.0: 叫人 (call someone), 108.0: 出來 (come out), 101.0: 言語 (speech), 100.0: 告訴 (tell), 98.0: 今日 (today), 98.0: 話儿 (talk), 95.0: 看見 (see), 92.0: 回道 (answer), 91.0: 那邊 (over there), 91.0: 連忙 (promptly), 85.0: 想起 (think of)

All the words are not person names or plot-related. The result obtained from the Topic Modelling shows that Topic #26 is a "style"-Topic. Most of them can be presented in a scene: an interaction (or a conversation) between the lady and the slave girl. DRC is an observation of the Chinese society in 18th-century. The story is about the life of two large, wealthy family compounds in the capital of China. Such scenes or plot are hence very common in the whole novel.

In conclusion, Topic Modeling was used in this paper to find the specific topic, which represents the difference between the first 80 and the last 40 chapters of DRC. The test results indicate that both hypotheses, a) the Delta test result is not 100% accurate, b) The four chapters share many names or plot related terms with the last 40 chapters, are not true. The first 80 chapters (except Chapter 11 and 67) are stylistically different from the last 40 chapters. According to the results of Delta and Topic Modeling, both Chapter 11 and 67 are definitely not written by the first author. They might be written. or at least edited by the second author of DRC.

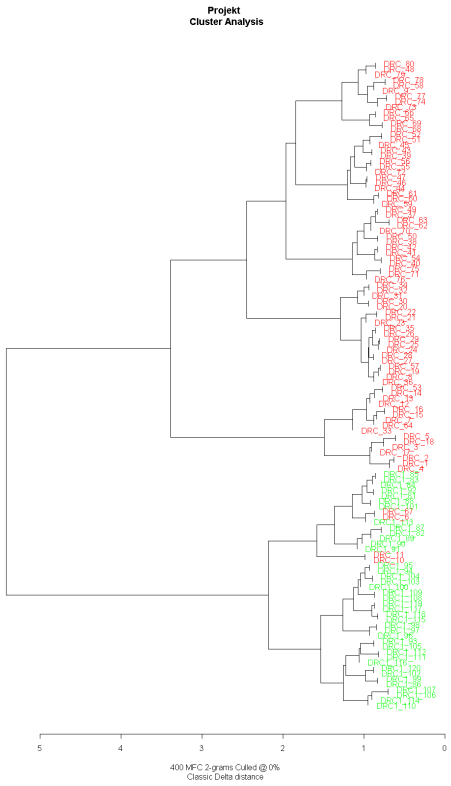

Figure 1. Delta test results of DRC, (300 MFC, 2-grams)

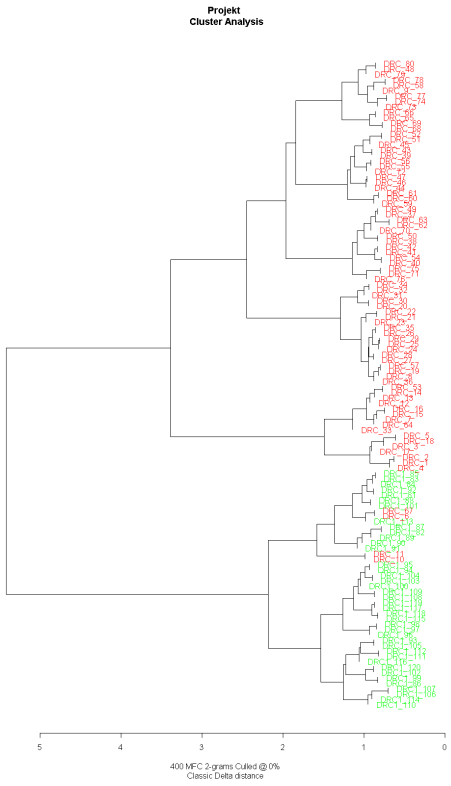

Figure 2, Delta test results of DRC, (400 MFC, 2-grams)

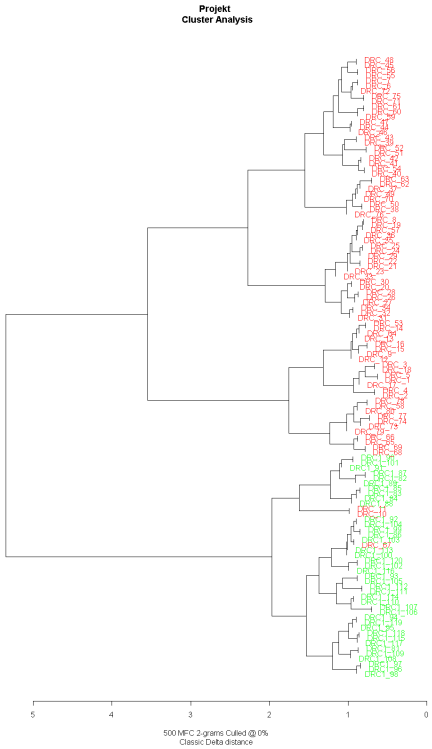

Figure 3. Delta test results of DRC, (500 MFC, 2-grams)

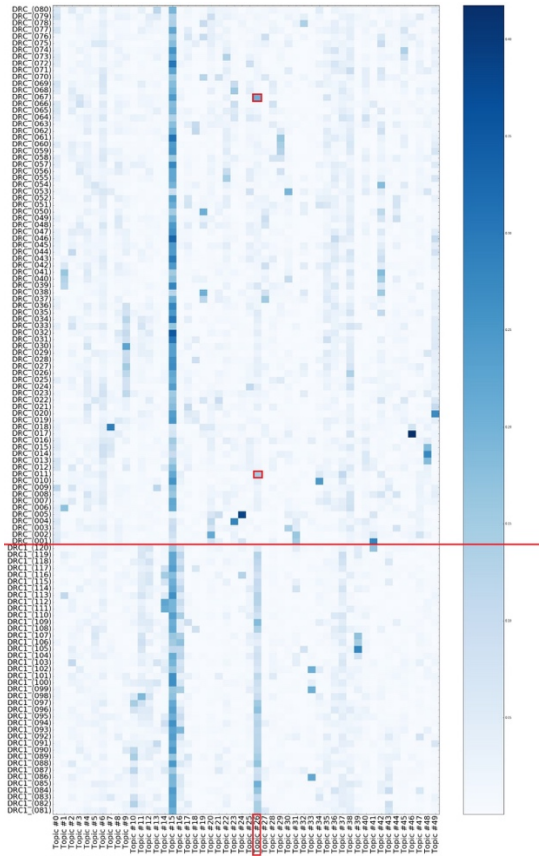

Figure 4. Topic-chapter distribution of DRC (50 topics, 120 documents)

Bibliography

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 993-1022.

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. In: Literary and Linguistic Computing, 17(3), (pp. 267-287).

Du, K. (2016). Testing Delta on Chinese Texts. In Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University, Kraków, pp. 781-783.

Hu Shhi (1988). 《胡适红楼梦研究论述全编》 [Hu Shihs Analysis of Dream of Red Chamber], Shanghai Guji Chubanshe (Shanghai Classics Publishing House)

Hu, X., Wang, Y., & Wu, Q. (2014). Multiple Authors Detection: A Quantitative Analysis of Dream of the Red Chamber. Advances in Adaptive Data Analysis, 1450012.

Tu, H. C., & Hsiang, J. (2013). A Text-Mining Approach to the Authorship Attribution Problem of Dream of the Red

Chamber. In: Digital Humanities 2013: Conference Abstracts (pp. 441-443)