Analysis of “The Tale of the Rabbit(Tokki-jeon)” Versions Using Text Mining

Jeongrim Lee, MLIS
Librarian, Midwest University

Introduction

- AI technologies, driven by big data, are fueling a 4th Industrial Revolution.
- Text mining is the process of exploring and analyzing large amounts of unstructured text data aided by software that can identify concepts, topics, patterns, keywords and other attributes in the data.
- Previous studies on the classification of classic novels : Qualitative method, requires a lot of time and effort from the researcher
- Traditional Text Materials + New Technology Methodology (Text mining)

Research Questions

1. What services can Artificial Intelligence (AI) provide to library users with digitized rare books?
2. Can text mining effectively classify different versions of classic novels?

Objectives

1. Attempting to classify 64 versions of “The Tale of the Rabbit (Tokki-jeon)” using text mining
2. Exploring new possibilities in version classification research

Methods

- Materials : 64 versions include in “Series of the Tale of the Rabbit (Tokki-joen Jeonjip)”
- Data analytics software : R
- Korean text data preprocessing : KoNLP (Korean Natural Language Processing)
- Text data analysis : Word frequency analysis, Euclidean distance analysis (Text similarity), Hierarchical cluster analysis
- Comparison with the previous classification study

Results

1. Word Frequency Analysis
   - Top 100 most frequent words, with a high proportion of character nouns (44.5%)
   - High frequency of nouns related to government positions (12.5%) and places (6.1%)
   - In the case of ‘Pansori (Korean traditional narrative song)’ versions, a high frequency of words related to Pansori’s unique style or rhythm
   - The version’s writing system can be identified (Korean-only, Korean-Chinese mixed)

Figure 1. WordClouds of “The Tale of the Rabbit”

2. Euclidean Distance Analysis (Text Similarity)
   - A shorter Euclidean distance between the texts indicates a higher similarity
   - The average value is 158.5
   - The minimum distance value is 74.3 (Text 0201 -Text 0211 <Gyeongpan-bon Tosaeng-jeon> -<Togong-jeon of Kim Dong-wook collection)
   - The maximum distance value is 246 (Text 0303 – Text 0505 <Sujung Byeoljubu Sanjung Tocheosa-jeon of Park Soon-ho Collection> - <Togong-jeon of Lim Hyeong-taek Collection)

3. Hierarchical Cluster Analysis
   - Function : hclust / Method : Ward’s method / Number of groups(k) : 7

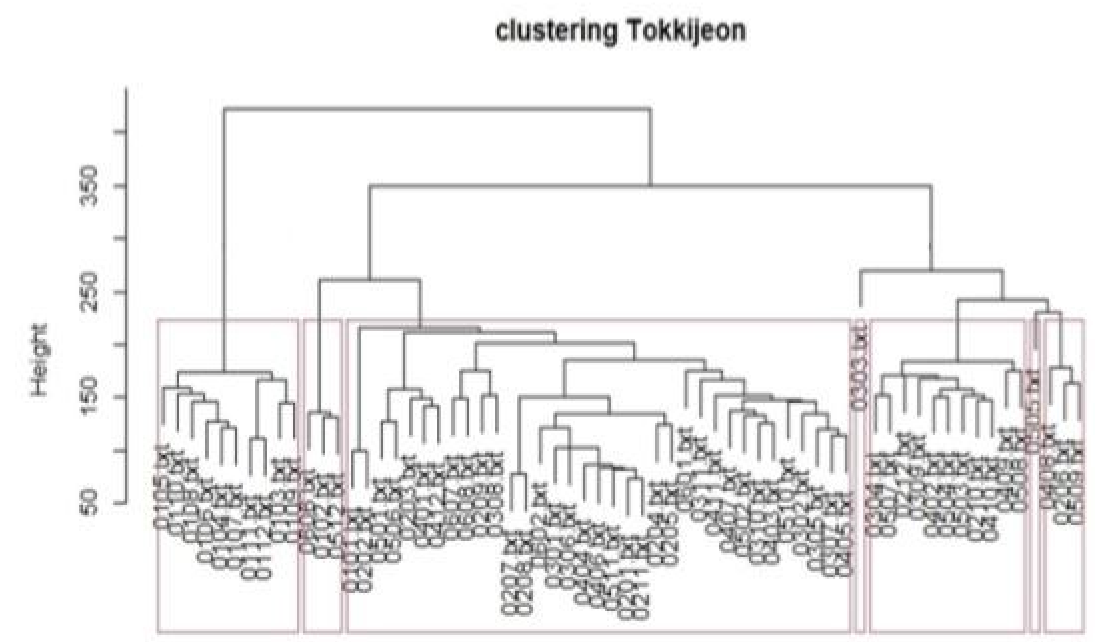

Figure 2. Clustering of “The Tale of the Rabbit”

4. Comparison with the previous classification study
   - Comparison with Kim Dong-gun’s study
   - 3 groups match completely, and 4 groups partially match

Table 1. Comparison with cluster analysis and Kim Dong-gun’s study

Cluster Analysis	Kim Dong-gun’s Study	Matching
Group 1	Pansori 1	Complete
Group 2	Pansori 2	Complete
Group 3	Print 1 + Print 2 + Print 3	Partial
Group 4	Miscellaneous	Partial
Group 5	Print 1	Partial
Group 6	Print 4	Complete
Group 7	Print 2	Partial

Conclusions

- Propose a new methodology to version classification study by using text mining
- The advantages of classification of classic novels using text mining are speed and efficiency
- Can be utilized to support previous qualitative classification research
- Utilizing text mining technologies on a digitized rare book collection, it is possible to provide various textual information to library users

- Discussion
  - Limitation : Using limited classification method
  - Recommendations for future research: Evaluating which classification method is consistent with the qualitative method

For more information, please contacts Jeongrim Lee (MLIS), Librarian at the Midwest University, jrlee@midwest.edu

References

1. Baek, Young Min (2017) Text-mining using R. Seoul: Hanul Academy.
2. Choi, Woon Ho & Kim, Dong Kun (2018). A computational approach to the classification and clustering of Tokkijeon through pairwise comparison of its narrative elements. The Studies of Korean Literature, 58, 123-154. https://doi.org/10.20864/skl.2018.04.58.123
3. In, Kwon Hwan (1991). A study on the changing aspects and the meaning of Tokkijeon’s ending. Korean Studies Quarterly, 14(3), 163-185.
4. Jo, Yun Je (2020). A study on the reproducing keywords of Park Wan-seo’s literature using Text Mining: Focusing on domestic papers and short stories. Master thesis, Graduate School of Myongji University, Department of Creative Writing.
5. Kim, Dong Kun (2001). A study of《Toggi-jun》. Ph.D. dissertation, Graduate School of Kyunghee University, Department of Korean Language and Literature.
6. Kim, Gye Soo (2017). R big data analysis for value creation. Seoul: Hannarae Academy.
7. Kim, Jin Young, Kim, Hyun Joo, Kim, Dong Kun, Lee, Sung Hee & Kim, Pil Rae ed. (1997-2003). Series of Tokkijeon(Vols. 1–6). Seoul: Pagijong Press.
8. Lee, Jee Young (1998). A study on the digitizing and indexing method for Korean old books. Master thesis, Graduate School of Ewha Womans University, Department of Library and Information Science.
9. Min, Chan (1995). A study on the novel of animal fable in the latter period of Chosun dynasty. Seoul: Thaehaksa.
10. Yoo, Min (2019) A Text Mining study on『The Dwarf』: Focused on an interpretation of subject discussion by hierarchical clustering. Master thesis, Graduate School of Education, Yonsei University.