Mandarin Tokenizers for NLP

Lola C Manning

Department of Data Science

College of William and Mary

ABSTRACT

Tokenization is an essential step to prepare text for Natural Language Processing (NLP) algorithms; however, tokenizing Mandarin Chinese presents challenges due to the absence of clear word boundaries. This research aims to identify the optimal tokenization method for literary Chinese texts by assessing accuracy in predicting text genres. Three tokenization approaches were evaluated: 1-gram, 2-gram, and Jieba. A dataset of 32,474 literary Chinese texts, primarily predating 1911, was utilized. Results revealed that the 1-gram tokenizer outperformed both the 2-gram tokenizer and Jieba, with an average accuracy of 80.53%. The 2-gram tokenizer yielded an average accuracy of 66.89%, while Jieba achieved 72.45%. Selecting an appropriate tokenization strategy is crucial for optimizing NLP performance in processing literary Chinese texts, and these findings underscore the importance of tailoring tokenization methods to the specific characteristics of the text corpus.

1. INTRODUCTION

Tokenization of input text is one of the primary steps of Natural Language Processing (NLP). In English, this tokenization consists of breaking texts down into words by using whitespace as the delineator, and using stemming, lemmatization, and stop words. In the Chinese language, word delineation by whitespace is ineffective due to the absence of spaces after words. Chinese is

character-based, and Chinese characters do not undergo conjugation, rendering stemming and lemmatization techniques ineffective as well.

In modern Chinese, words are often made up of two or more characters which has made tokenization by character faulty. To accommodate this, text segmentation modules, such as Jieba, were created to tokenize modern Chinese into words.

While this has proved successful for modern Chinese, it falls short when it comes to literary Chinese. While modern Chinese is largely disyllabic, literary Chinese is often monosyllabic. This means that when it comes to word segmentation for NLP, Jieba may not be the best method. Despite sharing many common characters ("zi"), each era of late imperial Chinese exhibits its own distribution of words ("ci"), making character based segmentation a promising approach, across literary Chinese.

This study will focus on the analysis of individual Chinese characters, known as "zi," as the fundamental unit of analysis. While current endeavors primarily target modern Chinese for applications like machine translation and text-to-speech software, their implications are relevant to literary research as well. Character n-grams offer a feasible alternative to word-based tokenization, providing deterministic outcomes and addressing challenges associated with word segmentation.

Initial exploratory data analysis involved 1-gram, 2-gram, and 3-gram tokenization. The following visualizations demonstrate the relationship between these methods.

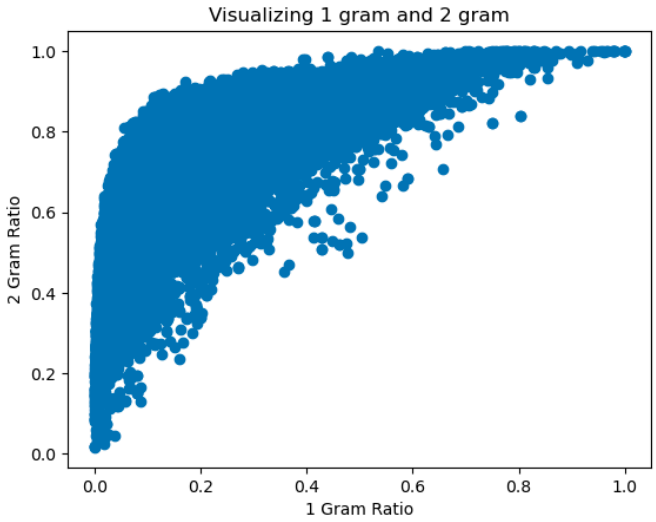

Figure 1: 3-gram Ratio and 2-gram Ratio

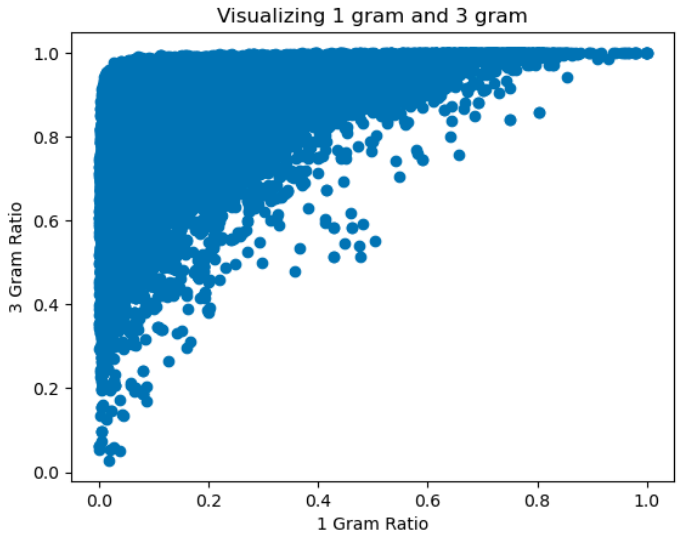

Figure 2: 1-gram Ratio and 3-gram Ratio

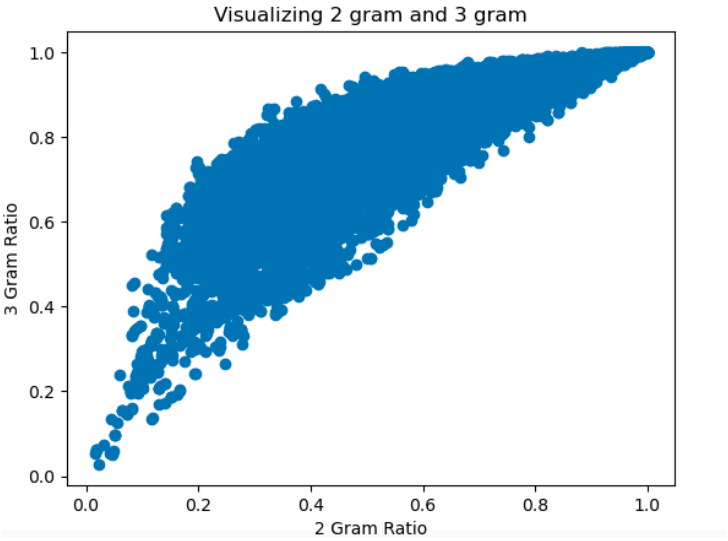

Figure 3: 2-gram Ratio and 3-gram Ratio

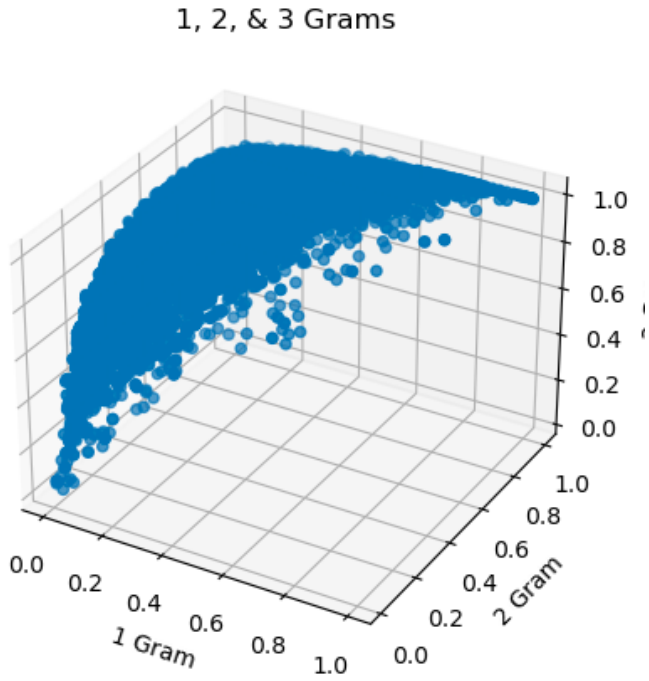

Figure 4: 1-gram Ratio, 2-gram Ratio, and 3-gram Ratio

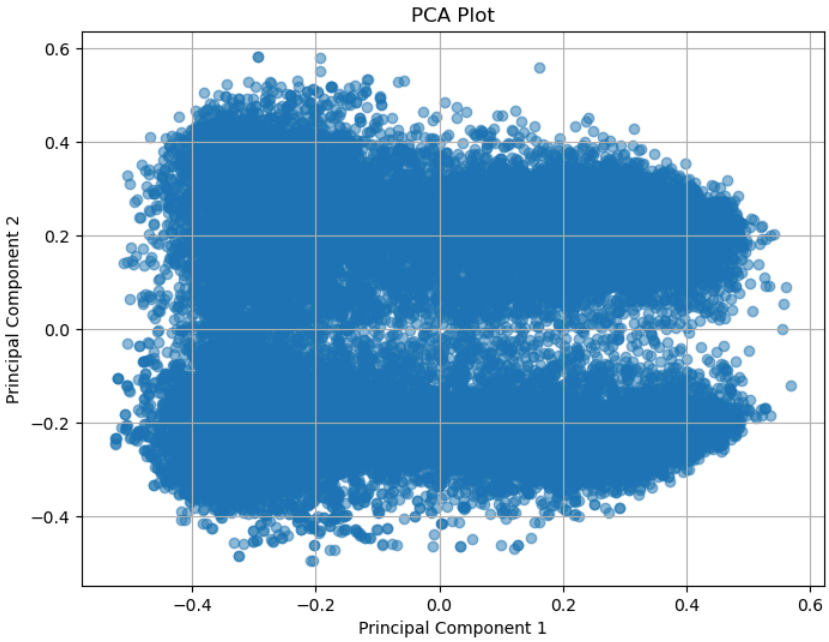

Figure 5: Principal Component Analysis of Texts

This research explores the complexities of literary Chinese text analysis, leveraging characters and n-grams to gain insights into linguistic and literary phenomena. By delving into tokenization methodologies and emerging technologies, this research contributes to the broader discourse on computational analysis in literary Chinese studies.

2. LITERATURE REVIEW

Since the early 1990s, the quest for effective methods to tokenize Chinese texts has been a focal point for statisticians and linguists alike. An initial approach was developed by Richard Sproat and Chilin Shi that employed probabilistic techniques to segment Chinese sentences. Their

approach, based on the frequency of character associations within Chinese newspapers, achieved around 90% accuracy of correct boundary recognition (Sproat & Shi, 1990).

Recent approaches have utilized machine learning techniques for word boundary detection. One such technique uses "maximum matching" algorithms, which start at the beginning of a sentence and slowly add the subsequent characters to the first character until it no longer makes a word. It leaves the most plausible word as the longest match, and continues this process to the end of the sentence (Tsai, 2014). However, while machine learning-based algorithms exhibit promise, they often necessitate extensive training datasets comprising pre-parsed texts.

Another recent development in word segmentation technology has introduced unsupervised algorithms, eliminating the need for training sets. Accessible to mainstream scholars, the Stanford Word Segmenter represents a noteworthy advancement, employing machine learning to consider lexical features and contextual cues (Tseng et al., 2005). While this method has proved accurate for modern Chinese, it leaves room for improvement when looking at literary Chinese.

A critical consideration in tokenization approaches is the use of character n-grams versus "word" tokens or word n-grams. N-gram tokenization offers deterministic outcomes, contrasting with the probabilistic outputs of many parsing algorithms. Furthermore, the lack of a distinct boundary between characters (*zi*) and words (*ci*) in Chinese underscores the utility of n-gram analysis, mitigating challenges posed by the absence of robust parsers. However, even in literary Chinese, in which words are frequently single characters, compound-character words exist, leaving room for ambiguity with 1-gram tokenization (Xue, 2003). 2 and 3-gram analyses contribute valuable contextual information, yet as the value of n increases, sparse data becomes increasingly problematic, leaving the extension of the length of the n-gram relatively insignificant.

Ongoing research suggests that for specific downstream tasks, particularly those pivotal to this study, n-gram analysis demonstrates comparable, if not superior, performance compared to word-based analysis. However, the emergence of cutting-edge technologies, such as multilingual transformer-based models like Google's BERT, Baidu's ERNIE, and OpenAI's GPT3 (alongside the impending release of GPT4), presents significant potential for advancing word segmentation in literary Chinese. Through training or fine-tuning on premodern Chinese corpora, these models have the capacity to unveil novel avenues for sophisticated analytics, potentially reshaping the landscape of literary Chinese natural language processing.

3. METHODOLOGY/DATA SET

The dataset comprises a Corpus containing essential information about each text, including title, author, genre, time period, and other attributes. Additionally, a folder houses each text listed in the Corpus, sourced from various repositories such as The Kanseki repository, the Daizhige repository, Chinese Wikisource, open lit, and Gutenberg. In total, the dataset encompasses 32,474 literary Chinese texts primarily dated before 1911.

To preprocess the text data, all non-Chinese characters were eliminated. Subsequently, the total length of each text was computed and appended to the Corpus dataset. Following thorough data cleaning and preparation for analysis, an n-gram function was developed to segment the text into n-character sequences. Ratios based on 1-gram, 2-gram, and the Jieba tokenizer were calculated and integrated into the Corpus dataset.

Upon establishment of the Corpus dataset, individual vectorizers were constructed for each tokenization method. These vectorizers were then fitted to the text, and their accuracy scores in predicting genre were assessed.

4. RESULTS

In our analysis, we observed varying levels of accuracy across different tokenization methods. The 1-gram tokenizer emerged as the most effective, achieving an average accuracy of approximately 80.53%. Following closely behind, the Jieba tokenizer yielded an average accuracy of 72.45%.The 2-gram tokenizer lagged behind with an average accuracy of about 66.89%. These results highlight the importance of selecting an appropriate tokenization strategy, as it significantly impacts the accuracy of predictive models in our dataset.

5. DISCUSSION

The results of our analysis shed light on the effectiveness of different tokenization methods for processing literary Chinese texts. Tokenization is a crucial step in natural language processing, as it directly impacts the performance of downstream tasks. In our study, we evaluated three tokenization approaches: 1-gram, 2-gram, and Jieba.

Our findings indicate that the 1-gram tokenizer outperformed both the 2-gram tokenizer and Jieba in terms of accuracy, averaging around 80.53%. This result suggests that for our dataset of literary Chinese texts, the 1-gram approach was the most suitable for segmenting the texts into meaningful units. The superior performance of the 1-gram tokenizer could be attributed to its ability to capture the individual characters' semantic and syntactic information effectively, which

is particularly relevant in literary Chinese, where characters are more likely to carry their own meaning than they are in modern Chinese.

Contrastingly, the 2-gram tokenizer yielded lower accuracy, averaging about 66.89%. This lower performance could be attributed to the increased complexity introduced by considering pairs of consecutive characters. While 2-gram analysis adds contextual information, it also encounters challenges with sparse data and increased computational complexity. In the context of literary Chinese, the lower performance of the 2-gram tokenizer may also be due to the often monosyllabic nature of the texts. The 2-gram tokenizer could provide a better fit for the largely disyllabic modern Chinese.

Interestingly, the Jieba tokenizer, which is widely used for Chinese text segmentation, returned an accuracy of 72.45%. While Jieba leverages a pre-existing lexicon and employs sophisticated algorithms for word segmentation, its performance in our study was not as high as the 1-gram tokenizer. This discrepancy could be attributed to the fact that Jieba tokenization was developed for modern Chinese, meaning it may not always align optimally with the characteristics of literary Chinese texts.

Overall, our study highlights the importance of considering the specific characteristics of the text corpus when selecting a tokenization method. While the 1-gram tokenizer demonstrated superior performance for our dataset of literary Chinese texts, future research could explore fine-tuning parameters to further optimize tokenization accuracy. It is also important to note that a shortcoming of the 1-gram method is its room for ambiguity as, even in literary Chinese, multi-character words exist and the 1-gram method may sacrifice some of the context and clarity that a higher n-gram method might provide. Additionally, the generalizability of these findings to

other corpora of Chinese texts, especially those from different time periods or genres, warrants further investigation. By refining tokenization techniques, researchers can enhance the effectiveness of natural language processing applications for literary Chinese texts and facilitate deeper insights into this rich cultural heritage.

6. REFERENCES

Chih-Hao Tsai, "MMSEG: A word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm," accessed March 26, 2014, http://technology.chtsai.org/mmseg/.

Garychowcmu. (n.d.). *GitHub - garychowcmu/daizhigev20:* 殆知阁古代文献. GitHub. https://github.com/garychowcmu/daizhigev20

Kanripo 漢籍リポジトリ*:* (n.d.). https://www.kanripo.org/

Project Gutenberg. (n.d.). Project Gutenberg. https://www.gutenberg.org/

Sproat, R., & Shih, C. (1990). A Statistical Method for Finding Word Boundaries in Chinese Text. Computer Processing of Chinese & Oriental Languages, 4(4), 337.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., & Manning, C. (2005). A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.

Xue, N. (2003). Chinese Word Segmentation as Character Tagging. Computational Linguistics and Chinese Language Processing, 8(1), 30.

维基文库,自由的图书馆. (n.d.). https://zh.wikisource.org/wiki/Wikisource:%E9%A6%96%E9%A1%B5

朱邦復工作室. (n.d.). 開放文學. http://open-lit.com/