Applications of Natural Language Processing in Classical Chinese Lyric Poetry

Tianchu Shu
Harris School of Public Policy
University of Chicago
Chicago, IL USA
tshu@uchicago.edu

ABSTRACT

The growing need for Ancient Chinese natural language processing (NLP) is largely in a range of research and commercial applications. This project aims to use modern NLP toolkits to study and analyze Classical Chinese lyric poetry, in order to understand different SongCi style, popular words etc.

KEYWORDS

Natural Language Processing(NLP), SongCi style, Classical Chinese Lyrics Poetry, ancient Chinese language;

1 INTRODUCTION

NLP Research in English literature and poetry has undergone major developments in recent years. While ancient Chinese NLP research, however, has not evolved significantly despite the fact ancient Chinese language is one of the oldest existing ancient languages in this world. This project aims to use modern NLP toolkits to study and analyze Classical Chinese lyric poetry.

2 BACKGROUND

Cí (詞) is a type of lyric poetry in the tradition of Classical Chinese poetry. Cí use a set of poetic meters derived from a base set of certain patterns, in fixed-rhythm, fixed-tone, and variable line-length formal types, or model examples. The rhythmic and tonal pattern of the ci are based upon certain, definitive musical song tunes. Cí became very a popular poetic form of Classical Chinese poetry in Song Dynasty. And it most often express feelings of desire. In terms of style,

cí can also be classified as either wǎnyuē 婉約 (grace) or háofàng 豪放 (bold).

3 DATASET

I am using a dataset of 21,047 Cí provided by the Jackey Gao on Github. The Cí are written by 1,468 different poets in Song Dynasty, using 1,420 different Cípái, which is the name of various formation of Cí. The formations of Cí are complicated, in different names of Cípái, the number of characters, syntactical structure, tones and rhyme are also different.

Because Cípái is only the name of the Cí, when performing the NLP. I am only looking at the formation of Cí itself.

4 METHODS

In summary of the various libraries I used to do this project, it is clearly Jieba is very useful to tokenize Chinese. NLTK, Gensim, Word Cloud, Matplotlib, Collections and Spacy. Right now, Chinese is not fully supported by SpaCy. I am very looking forward to see how SpaCy will improve that in the future.

5 RESULTS

Figure 1: WordCloud

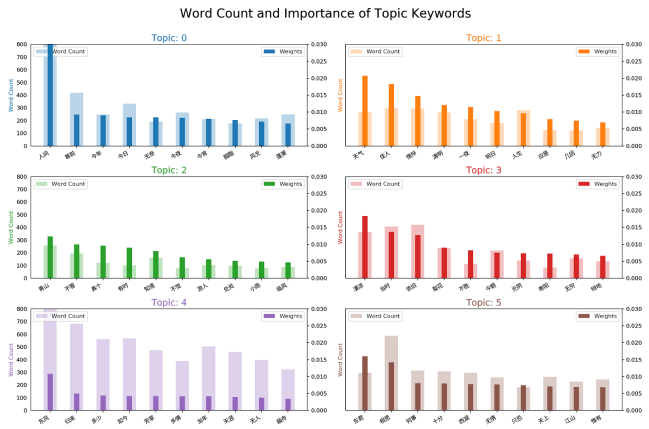

Figure 2: Word Count and Importance of Topic Keywords

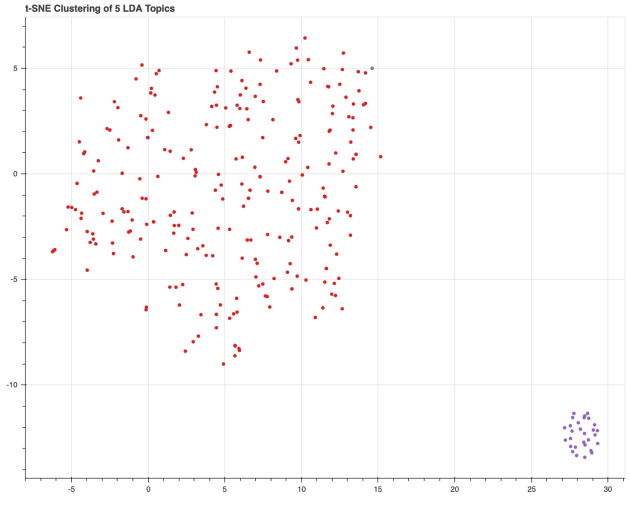

Figure 3: t-SNE Clustering of 5 LDA Topics

6 CONCLUSIONS AND FUTURE WORKS

With the aim of understanding different SongCi style and popular words were used by poets different NLP toolkits were used.

From the Figure 3, I can clearly see that the dominant topic clusters are only two, which I guess as mentioned before, cí can also be classified as either wǎnyuē 婉約 (grace) or háofàng 豪放 (bold). Since most poets wrote about their sadness and misfortune in life and career, I suspect the bigger red cluster represents wǎnyuē 婉約 (grace), while the smaller blue cluster represents háofàng 豪放 (bold).

The other interesting finding is related to word count, from figure 2 we can clearly see rénjiān 人间 (world) was popularly used by poets. I suspect they are using rénjiān 人间 to write poems to questioning the meaning of life.

Last but not least, for similar words used in Cí: dōngfēng 东风 (east wind), xīfēng 西风 (west wind), chūnfēng 春风 (spring wind) and huánghūn 黄昏 (dusk) are all considered 99% similar by Word2Vec, which are quite absurd. since they are not that much related in modern life while NLP tools think they are the same.

ACKNOWLEDGMENTS

I thank Jeffrey Tharsen, and many classmates from the natural language processing class for their valuable feedback on this work.