Analysis on the book *Three Hundred Tang Poems*

Three Hundred Tang Poems is an anthology of Chinese poems from Tang Dynasty. It has been used for education purposes for a long time, so it has been a window to Tang Dynasty for modern people. In this project, I will analyze the poets and poems that are selected into this book and have an understanding about what this book has shown readers about Tang Dynasty.

The analysis will be conducted with python on 2 aspects, poets and poems. For the poets, whose poems are selected the most and among four periods of Tang Dynasty, from which period the poets are selected the most. For the poems, what is the common theme of this book and why.

1. Research steps:
    1) Data gathering and cleaning:

        From Github I got the access to an open resource text file of the book *Three Hundred Tang Poems*, which contains the name of poems and poets, the format of the poems, rhythms and the content of the poems.

        To clean this data, same as a lot of other txt files, firstly I opened this text file with python. After reading it into a variable, I removed the newlines(\n), blanks and the rhythms, which I will not be using in this study. Now the text file has already been converted into a list, in which every object is one line of text file without blanks and new lines.

        Script is shown in the python file named “txt_process”.

    2) Data processing

        Firstly, to calculate the number of poets, the script mentioned above is enlarged to create

a list of poets and a string of the content of poems. In order to do this, in every line where the word “作者” (writer) appears, from the third character to the end is the name of the writer. Similarly, we can obtain the content of the poems by selecting the characters after “詩文”(poem content). Then, we have the content written into a text file to be tokenized while the poets are counted right away with Counter function of python and saved into the file “poets_counter.rtf”.

Script is also shown in the file named “txt_process”.

Secondly, a timeline is created with a web app named “timeline maker”. From Wikipedia the four periods of Tang Dynasty are acquired and then are typed into the timeline as four time periods, they are “Chu Tang” (Early Tang), “Sheng Tang” (High Tang), “Zhong Tang” (Middle Tang), “Wan Tang” (Late Tang). To mark every poet into the timeline as well, the medians of their birth year and death year are calculated. There are altogether 77 poets including one anonymous, but 11 of them do not have an exact birth and death year. Thus, there are 66 poets in the timeline, which is shown below.

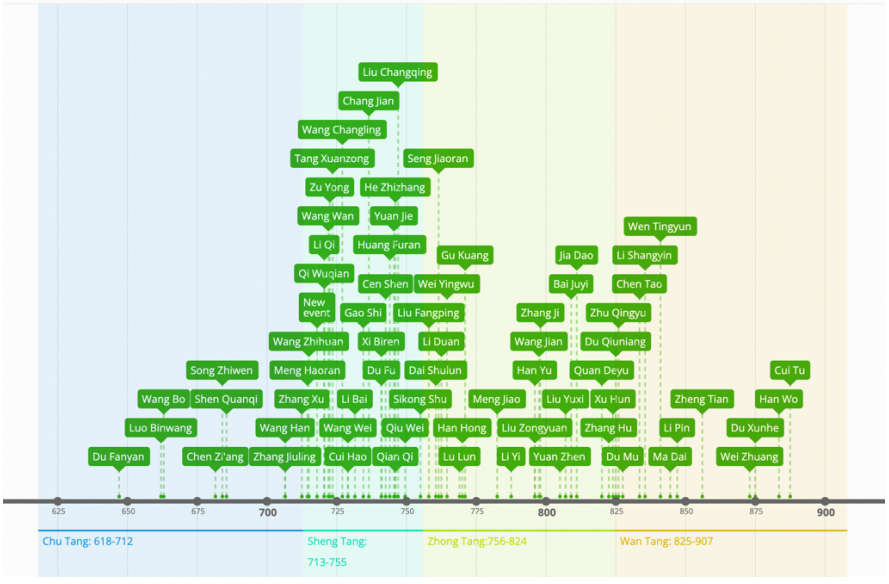

Figure 1. Timeline of poets

Then, we will turn to solve the poem content research question. To do that, my theory is that by acquiring the unigram and bigram frequencies, it is possible to summarize a vague understanding about the common theme.

So, the content text file is loaded in python and tokenized by jieba, which can parse Chinese texts. Finally the tokenized words are saved in a list, by passing this list to wordcloud function of pyhton, we can get the frequencies of the words. Also, we can get unigram and bigram(or possibly trigram) frequencies by setting a "if else" sentence to limit the length of every word in the list. During this process, I thought about the stop list a lot. On the one hand, theoretically there are a lot of non-meaningful words in poems, it is reasonable to delete them before analyzing the word frequencies. On the other hand, after testing, I found out that these non-meaningful words do not have any big influence on the word frequencies. So, I did not include any word into the stop list, only comma, period and question mark.

This script is saved in "content_wordcloud.py".

3) Analyzing results

Until now, the data processing part is done and it's time to analyze the results returned by python scripts.

2. Analysis of the results

The book *Three Hundred Tang Poems* has selected altogether 321 Tang poems, with the file "poets_counter.rtf" we can rank the poets according to the number of their poets selected. The first place is taken by Du fu, with 39 poems. In the second place is Li bai, 35, and follows him is Wang wei, 29. The top three have already contributed about 33%

of the poems in this book. As expected, they are very famous poets even until now with their talent in writing and unique personalities. Du fu is also known as the greatest poet of Tang Dynasty and “poet-historian”. Apart from his own talent, he also had certain concerns about the country and the people shown in his poems, although he didn’t have the chance to lead the country onto a right track with his abilities. Li bai is known for his romance style and he was acclaimed to be a poet genius since he didn’t need time to draft a poem, but he wrote it directly and perfectly. His poems are regarded as one of the “Three wonders” in Tang Dynasty. Wang wei was especially talented in writing poems and painting, a lot of his poems are preserved until now. He is known for his poems about beauty of nature.

In general, this ranking of poets is in accord with our expectation. This book has thus selected most poems from the poets, who were already famous and unique.

The timeline(figure1) of poets has shown us that most poets lived during Sheng Tang and Zhong Tang, which needs to be analyzed combining with historical facts. Sheng Tang is a period when there was no war and the national power of the Tang Dynasty grew steadily, which means that the national economy kept developing and culture began to flourish. During this period, poets started to innovate and create art works since they didn’t have much problems around their lives. In the contrary, during Zhong Tang, the national power of Tang Dynasty started to decrease and wars took place. People still managed to live, but poets were very concerned about the country, so they wrote poems to express their feelings or bad living conditions of people. During Chu Tang and Wan Tang, the wars had done severe damages on people’s lives. Nearly everyone was worrying about their food or accommodations, so it’s reasonable that there were fewer good poets and poems during those periods.

In the terms of the poems content, we have to look at the word clouds generated by python.

Figure 2. Unigram word cloud

In Chinese poetries, it is always bigrams that carry important meanings and unigrams themselves can be interpreted as different meanings. According to Figure 2, the most frequent unigrams are “长”(long), “月”(moon), “不”(no), “无”(without), “在”(at, existence), from which we cannot tell anything about the themes but the main feeling is rather negative, because each one has negative meaning except the last one.

Figure 3. Bigram (and above) word cloud

With the bigram word cloud, the analysis and inference of common theme is possible. According to Figure 3, the most frequent bigrams are “明月”(the bright moon), “昨夜”(last night) , “不知”(do not know), “何处”(where), “不得”(shall not), “琵琶”(traditional Chinese instrument), “故人”(old acquaintance), “万里”(500km). Respectively, the bright moon in Chinese poems has always been the symbol of missing someone; “last night” does not always mean last night literally, but the reminiscence of the past days; The expression of “do not know”, “where” and “shall not” show the bewilderment and helplessness of the poets; “琵琶”, as the traditional Chinese instrument, is always regarded as symbol of nostalgia, sorrow or farewell; The “old acquaintance” can represent that the poet missed old friends or also the old days; and finally, “500km” is just one expression of a great distance.

In general, if we put these expressions together, we can already make an inference that

the most common theme might be this: the poet was sad because he was lonely and he misses the great old days with friends, and his feelings of bewilderment and helplessness were expressed as well.

All in all, this book has selected most its poems from famous Tang poems written by famous poets and the most common theme is reminiscence.

3. Possible biases and problems

To begin with, the data I got from github is obviously pre-cleaned and processed, is it cleaned in an objective way, is there any important information deleted? To answer this question, we need to check all the poems and information of the text file.

Secondly, the main focus of this project is vague. The main purpose is to conduct a computational study on this book, but it is too general. Every individual research question can be a complete study, but due to the limitation of my python skills, it is not very possible to go that deep, so my study is conducted to answer several questions shallowly instead of one question deeply.

Last but not least, the meanings of the most common bigrams are my interpretation on the basis of my poetry knowledge, but it is obviously not enough for a study, results of articles or projects should be used instead.