
Findings from Siku Quanshu Data Mining

《四库全书》数据挖掘中得出的发现

Ashley Zheng 郑钰骐

3 Stages

01. Learn method and practice pattern recognition using Suzhou data
02. Refine 1st batch Siku data
 - indentation inconsistencies
 - docx package bugs: omission of rare characters
03. Find new solution for 2nd batch Siku data
 - color & paragraph formatting
 - tradeoff: concision vs functionality

学习苏州数据中的归类方法和规律识别

优化第一批四库数据时遇到的问题

- 段落缩进
- 语言包吞罕见字符

为第二批四库数据寻找新方案

- 颜色和段落格式
 - 代码简洁和功能性的妥协
-

Suzhou Data

- tags:

	A	B	C	D	E	F	G	H	I	J
1	#	line	juan	file		tag	content	title	ArticleTitle	ArticleContent
3	18	崇樓言言，上	東皋錄卷上	東皋錄	##	ArticleContent	崇樓言言，上下太清；鳴鐘在懸，飛鯨是名。在昔先民，制器尚形	崇樓言言，上下		
5	20	貧非樂也；樂	東皋錄卷上	東皋錄	##	ArticleContent	貧非樂也；樂惟道義，匪哲弗居。孰云細事？室如懸磬，俯仰無愧	貧非樂也；樂惟		
7	22	循彼南澗，言	東皋錄卷上	東皋錄	##	ArticleContent	循彼南澗，言采其蒲。采之何為？滫瀡是圖。彼蒲之良，利用為屨	循彼南澗，言采		
9	24	昔先佛心嘗得	東皋錄卷上	東皋錄	##	ArticleContent	昔先佛心嘗得徐兢所篆藏暉齋三大字於諸篋衍；後薦經離亂，圖籍	昔先佛心嘗得徐		

no indentation or color formatting

Siku 1 Challenges and Solutions

- indentation inconsistencies
- 缩进不一致

梅仙山奉次李古冲憲長春靄憑欄之贈撫昔人
之遺蹤感高懷於莫既悲悵自失不覺有言
春含翠篠密如城江擁晴嵐蘸水明雲護虹亭無客到

聶雙江侍御巡閩闢梵宮開五經書院聚生儒講
禮教移檄見徵晚發舟南下呈華古溪先生三
十韻

- regular pattern 2+3; 2+3+3
- irregular 2+2+2, 2+4, etc.

several causes:

- 序 → regex []

題寧都州學圖 有序
同夫鄭君修郡庠之禮殿垂成而代去其友人唐
旂王禮田觀懼君之去此無以慰其思而繫斯情
也乃命繪工圖其新廟併肖其環州山水之勝以
贈君於是南平劉楚敬贊于圖之右隅以申美之
出自西門至於泮宮我游伊何以觀民風乃觀新殿翼

- heading tagged as title/content
→ max 21 characters/line

pseudo code

Proposed adjustment: if first line is in the format []+up to 19 characters, then it's a heading; if there're 19 characters, and the next line starts with [] then it's a continuation of the same heading, and if the second line has 18 words, and the third line starts with [], then whatever left in the third line is also part of the same title

Siku 1 Challenges and Solutions

- omission of characters
- 语言包吞特殊字符

五言排律←

送陸使君入覲←

```
tag:title 五言排律
tag:heading 送君入覲
```

壽李北山先生八十

```
tag:heading 先生八十|
```

change Document (docx) to
other packages: mammoth,
textract, **py pandoc**

Siku 1 Challenges and Solutions

- inconsistent tagging (same format, different tags; author tagged as heading; heading tagged as content...)

```
tag:heading 三洲十景 有叙
tag:title    余家三洲上大海浩淼自東來西接木蘭泗華二
鉅麗橋北東舊置壘明斥埃昇平日久有老樹亭亭盤其上若飛蓋
```

```
80 tag:heading 明 劉 炳 撰 楊維禎 評
1141 tag:heading 明 于慎行 撰
```

regular expression: identify patterns
and translate to python in between txt
and csv conversion

```
# translate regex to remove tag:heading with authors
cleaned_text = re.sub(r"tag:heading[^\n]+\s撰\n", "", cleaned_text)
cleaned_text = re.sub(r"tag:heading[^\d\n]*撰[^\d\n]*評\n", "", cleaned_text)
# translate regex to remove tag:title with authors but keep volumn
cleaned_text = re.sub(
    r"(tag:title\D{7,15})明\D{4,}撰", # Keep up to 明, drop everything from 明 to 撰 inclusive
    r"\1",
    cleaned_text
)
# clean tag:vol only keep the volume, delete author in between 明 and 撰
cleaned_text = re.sub(
    r"(tag:vol[^\n]*?)明.*撰", # Keep up to 明, drop everything from 明 to 撰 inclusive
    r"\1",
    cleaned_text
)
# clean tag:vol only keep the volume, delete author in between 元 and 撰
cleaned_text = re.sub(
    r"(tag:vol[^\n]*?)元.*撰", # Keep up to 明, drop everything from 元 to 撰 inclusive
    r"\1",
    cleaned_text
)
```

Limitations

- tradeoff between code reusability and specificity
- 代码可复用性和处理特例能力之间的权衡
- programs can save time, but manual checking for irregularities are still necessary
- 程序可以节省时间，但仍然需要人工检查异常情况

槽頭醪醕倒不盡池上梅花開欲齊今日好懷良足慰
穀 百年高興若為題銀燈絢席口瑚爛 敏 綵筆聯詩錦
繡迷細雨作花簷際落 堪 清歌如縷拍中低渾忘司馬
青衫濕 穀 似聽宮鶯綠樹啼公子金猊燒篤耨 敏 美人
犀筋勸駝啼也知北海能豪傑 堪 況有東方善滑稽腰
細舞低翻弱柳 穀 手纖盃重壓柔荑先判五夜同傾倒
敏 未信三邊尚鼓鼙倏忽年華梁旦暮 堪 風流人物晉
東西江湖空濶鴻偏到 穀 梧竹高低鳳獨棲乘興放船
來笠澤 敏 散愁投轄憶潯溪襟期久矣交如漆 堪 笑謔
頽然醉似泥潮漲空灘移舴艋 穀 雨晴何處看鳬鷺會
須發達攄憂憤 敏 慎勿登臨動慘悽失喜今宵是何夕
酒闌起舞忽聞雞 堪

咸 B 行為楊璽改葬乃祖作

曾有感 據陳本增二二四七二二四七二二四七二二四八二二四八二二四八
-二二五一二二五二樓頭 據陳本增王戌除夕中秋有感四憶詩附錄》增據
二二五二二二五二二二五三羅洪先集 卷三十二五言排律：：：：：：：
三月十三日駕車太學陪祀孔廟聽講彝倫堂 據四庫本增：：：重晤劉孔

Siku 2 Challenges and Solutions

- color and style formatting
- 以顏色和樣式做的標記

▪ 御題劉宗周黃道周集有序[←]
比因彙輯《四庫全書》，各省傳訪遺編以進，
再懇侍養疏萬歷乙巳三月上[←]
行人司行人臣劉宗周謹奏為遵例再懇天恩終賜侍[←]

- style names different from what appears in word: .e.g. 四庫11，內文小註，案語
- different styles across documents
- packages cannot recognize RGB

- add print statements to check unknown attributes (run vs. paragraph style)
 - keep tag:content but put purple and blue characters (annotations by editors) in []
 - pattern not always obvious
-

Siku 2 Challenges and Solutions

- one paragraph separated into several tag: content lines
 - 同一段文字分成几行
 - must combine lines after converting to txt because line breaks are required for identifying different styles
 - compromise: not the most elegant/succinct solution
-

Pattern Recognition Tips

1

- what's the function of color/style/indentation?

2

- is it really an irregularity or an unnoticed pattern?
-