Visualizing Character Class Collocation

Paul Bucci
pbucci@gmail.com

1.0 Introduction

The Centre for Human Evolution, Cognition, and Culture (HECC) houses the Cultural Evolution of Religion Research Consortium (CERC)[1]. One of CERC's projects involves the analysis of medieval Chinese literature. On a high level, they're trying to answer questions about the philosophical and cultural implications of the literature, such as:

- Does medieval Chinese literature exhibit a sense of dualism?
- Are there high gods? If so, is there a hierarchy? Are they morally conscious? Are they affiliated with punishment and reward?

Although these are complex and fascinating questions, CERC's chosen method of inquiry is quite simple and takes a cue from the general method of collocation analysis used in corpus linguistics:

1. Determine sets of words that constitute a common topic.
   i.e., words about 'Cognition' would include 'thought', 'imagine', 'analyze', etc.
2. Identify set members within texts.
3. Determine frequency with which set members co-occur within certain proximities of each other.
4. Determine frequency with which set members co-occur within sentence boundaries.
5. Weight the metrics of (3) and (4) to determine overall set collocation strength.

The goal of this project is to create a visualization for steps 2–5 across a corpus. By allowing researchers to explore their texts using these collocation metrics, patterns should become apparent over different genres and time periods. Although addressing the high-level concerns posed above is interesting, the hope is that this visualization exposes more sophisticated questions than provides clear answers. A secondary goal is to fine-tune word sets and question domains, but this may not prove to be possible within the time period provided by this course.

1.1 Project Scope and data

CERC is very much in the research methods development phase. Like many in the humanities, they are just beginning to understand the power of computation on large datasets. As such, there is much talk about expanding the scope and depth of research, but, for the next few months, the dataset I'm concerned with will remain static. As such it contains:

- 96 clean digitized texts
- Word sets, not mutually exclusive (see Appendix A for sample)
  - Trivials (words that aren't counted in distance calculations but need to be identified)

    - Delimiters
    - Stopwords (words that are safe to ignore such as articles)
- Non-trivials
    - Gods
    - Deities
    - Punishment
    - Reward
    - Emotion
    - Cognition
    - Morality
    - Tian
    - Di

2.1 Personal expertise

I came into this project some time last year while helping a friend debug a Python script for determining collocations. This relationship continued into last summer, when his lab at CERC hired me to re-develop and extend the scripts I had been helping out with. This turned into a roughly 80 hour project over three months where I developed a text parser that identified and counted colocations within certain ranges. As such, I know the low-level mechanics of how collocations work, but I am not as familiar with how word sets were developed. This should not prove to be a problem, as word set validation is a secondary goal, and may well not be a thing I have time to do.

3.0 Terminology

Focal set	the word set to focus our analysis on.
Comparison set	the word set to compare to a focal set.
Key	a set of n characters denoting a word or short phrase.

3.1 Text pre-analysis processing procedure

Set k as max range for collocation analysis.
Choose a Focal and Comparison set.
Identify all Focal and Comparison keys in a text.
Identify all Trivials (Stopwords and Delimiters) in a text.
Create a weighted edge from all Focal keys to all Comparison keys that lay within ±k characters of each Focal key such that:

```
    distance = abs(Position(Focal) − Position(Comparison)), distance ≤ k
    weight = distance – (Count of all Trivals between Postion(Focal) and Position(Comparison))
    Where:
        Position(x) is character index of the first character of key x from the beginning of
        the text.
        abs(x) is the absolute value function.
```

The justification for this weighting procedure is that the conceptual distance between phrases is closer than absolute word counts when you take phrase structures into account. For example, in the sentence “He rode into the sunset,” it is more meaningful to abstract the sentence into “(He) (rode) (into) (the sunset),” making the conceptual distance between “He” and “the sunset” closer than the absolute distance.

The practical implication is that the word sets are composed of n-nomial (containing n contiguous characters) keys. Most keys are 1-nomial, but enough are 2- and 3-nomial keys to warrant consideration.

3.2 Text analysis tasks

Count all Focal keys in a text.
Count all Comparison keys in a text.
Count all collocations by counting edges with weight [1..k]. Overlapping collocations are OK.
Count all collocations within each sentence of text, i.e. between each set of delimiters.

3.3 Corpus analysis tasks

Note texts in which Focal keys appear even once.
Note texts in which Comparison keys appear even once.

Count all occurrences of Focal keys across texts.
Count all occurrences of Comparison keys across texts.
Count all collocations across texts.

Bin collocations by weight ranges.
Bin collocations by text genre.

4.1 Proposed system

The proposed visualization system must support the analysis tasks as stated above by providing clear visual representations of the above in different levels of aggregation.

At the lowest level, it will be useful to see the values in (3.2) in a small summary for a single text. This visualization should expose the structure of the text, such as distribution, frequency and strength of collocations.

Direct comparisons between two texts are not particularly useful, but summaries of (3.2) across a set of selected texts will be.

Although the aggregate analysis is important, the most interesting use of the visualization will be comparing across sets of texts. These comparisons will be by genre, time period, and between Focal and Comparison set pairs.

4.2 Annotated illustrations

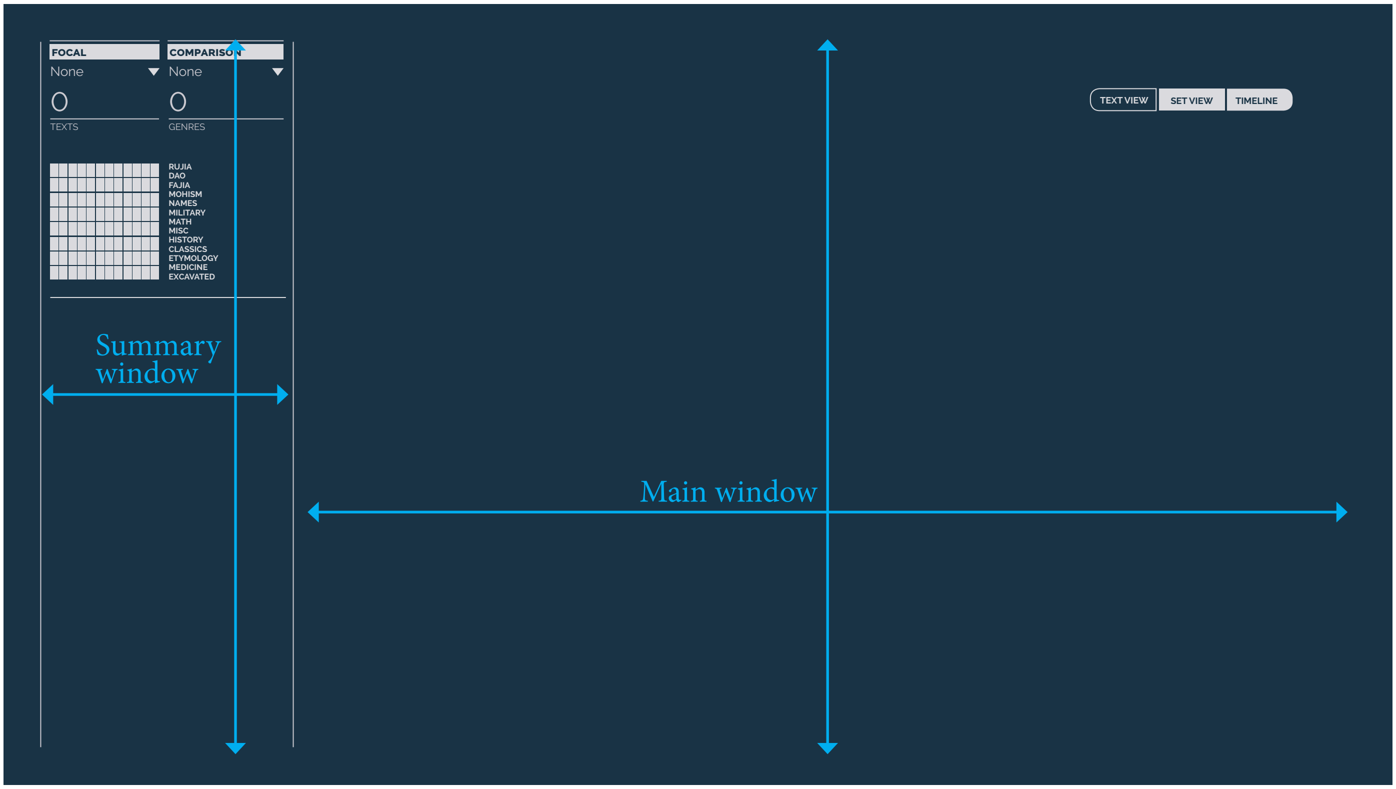


Figure 1.0 : Empty opening screen.

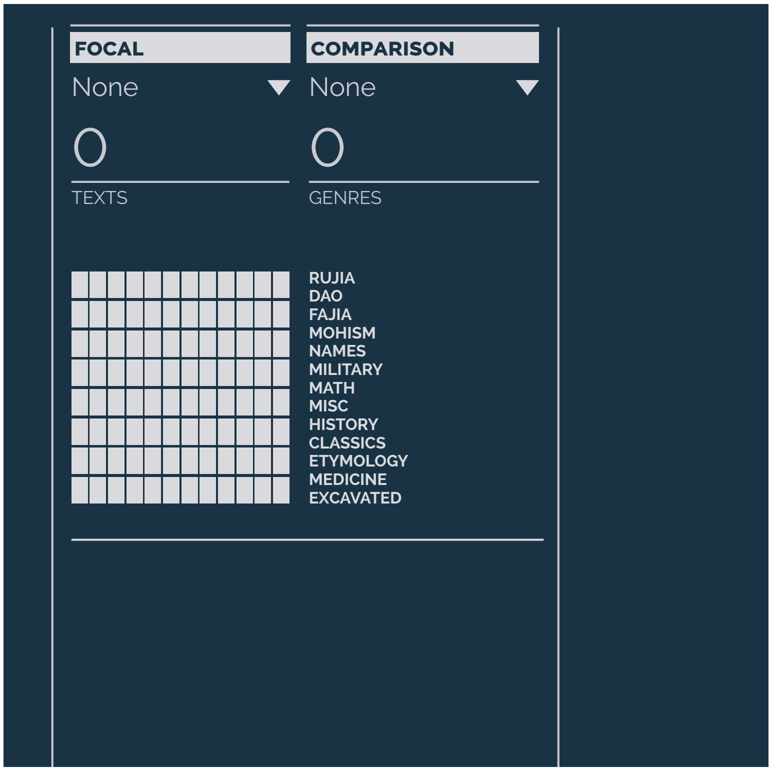


Figure 1.1 : Nothing selected.

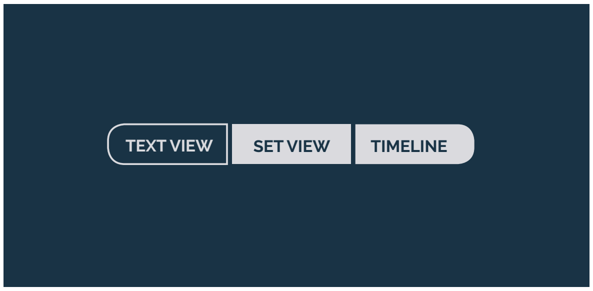
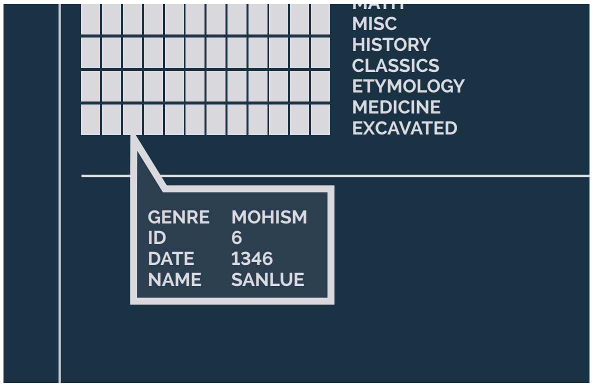


Figure 1.2 : Detail, view navigation pills.



GENRE	MOHISM
ID	6
DATE	1346
NAME	SANLUE

Figure 1.3 : Text detail on hover over text.

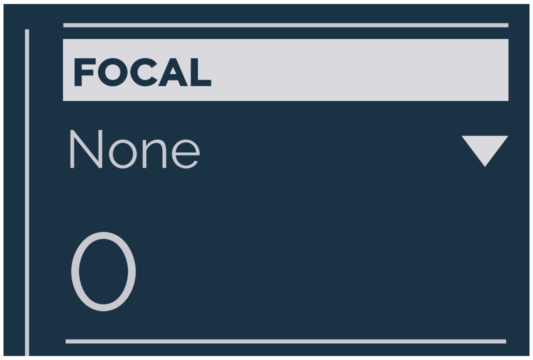


Figure 1.4 : Dropdown menu for focal set.

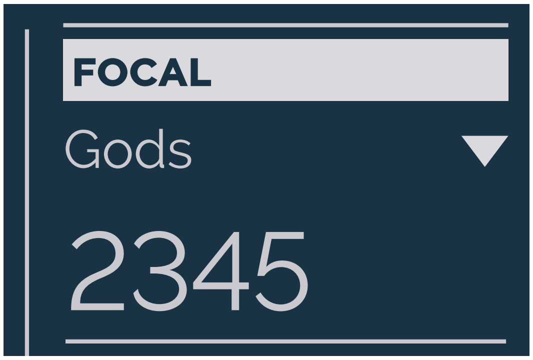


Figure 1.5 : Count across selected texts apears below.

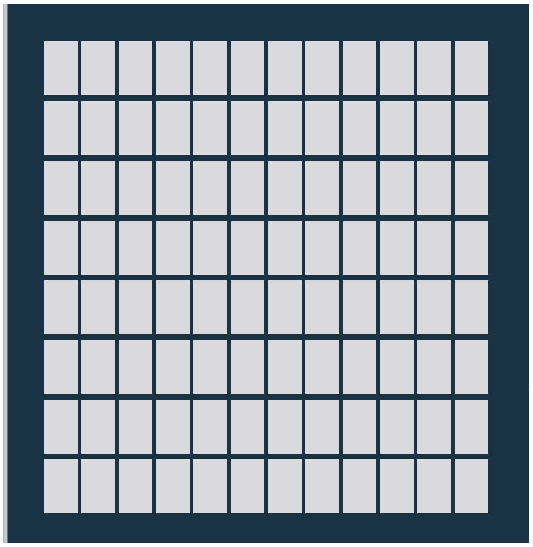


Figure 1.6 : Each cell represents a text.

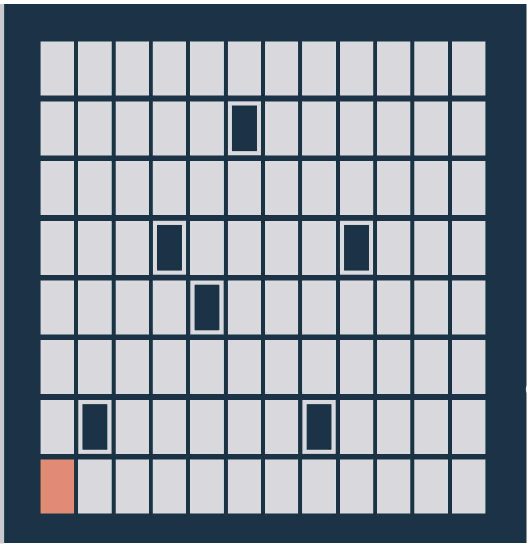


Figure 1.6 : If a Focal set is selected, texts in which the Focal set does not appear appear with no fill. The pink represents a selected set of texts.

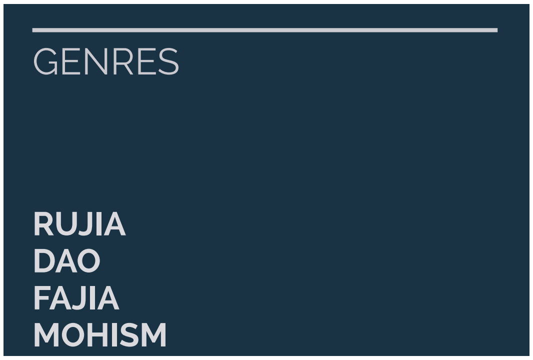
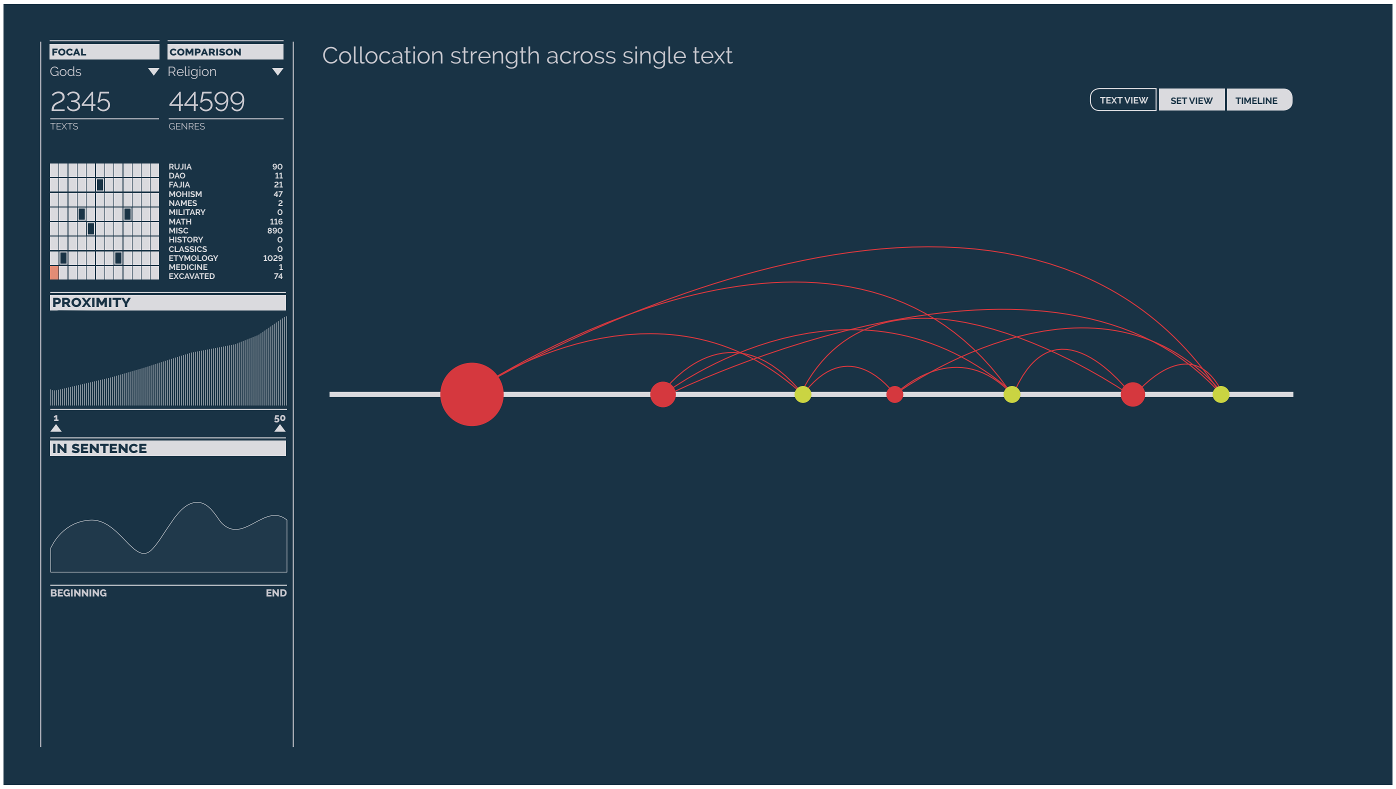


Figure 1.7 : a clickable list of genres appears on the side of the texts for quick selection.

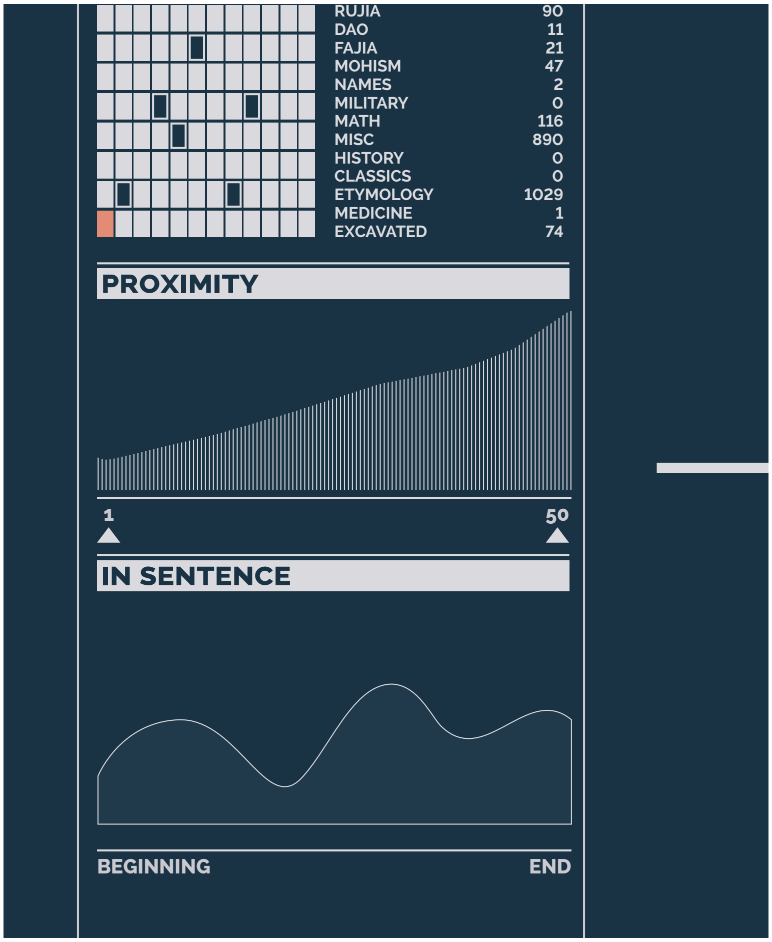


Collocation strength across single text

FOCAL	COMPARISON
Gods	Religion
2345	44599
TEXTS	GENRES

RUJIA	90
DAO	11
FAJIA	21
MOHISM	47
NAMES	2
MILITARY	0
MATH	116
MISC	890
HISTORY	0
CLASSICS	0
ETYMOLOGY	1029
MEDICINE	1
EXCAVATED	74

Figure 2.0 : A Focal set, Comparison set, and single text are selected. Graph of Focal→Comparison edges appears in Main window.



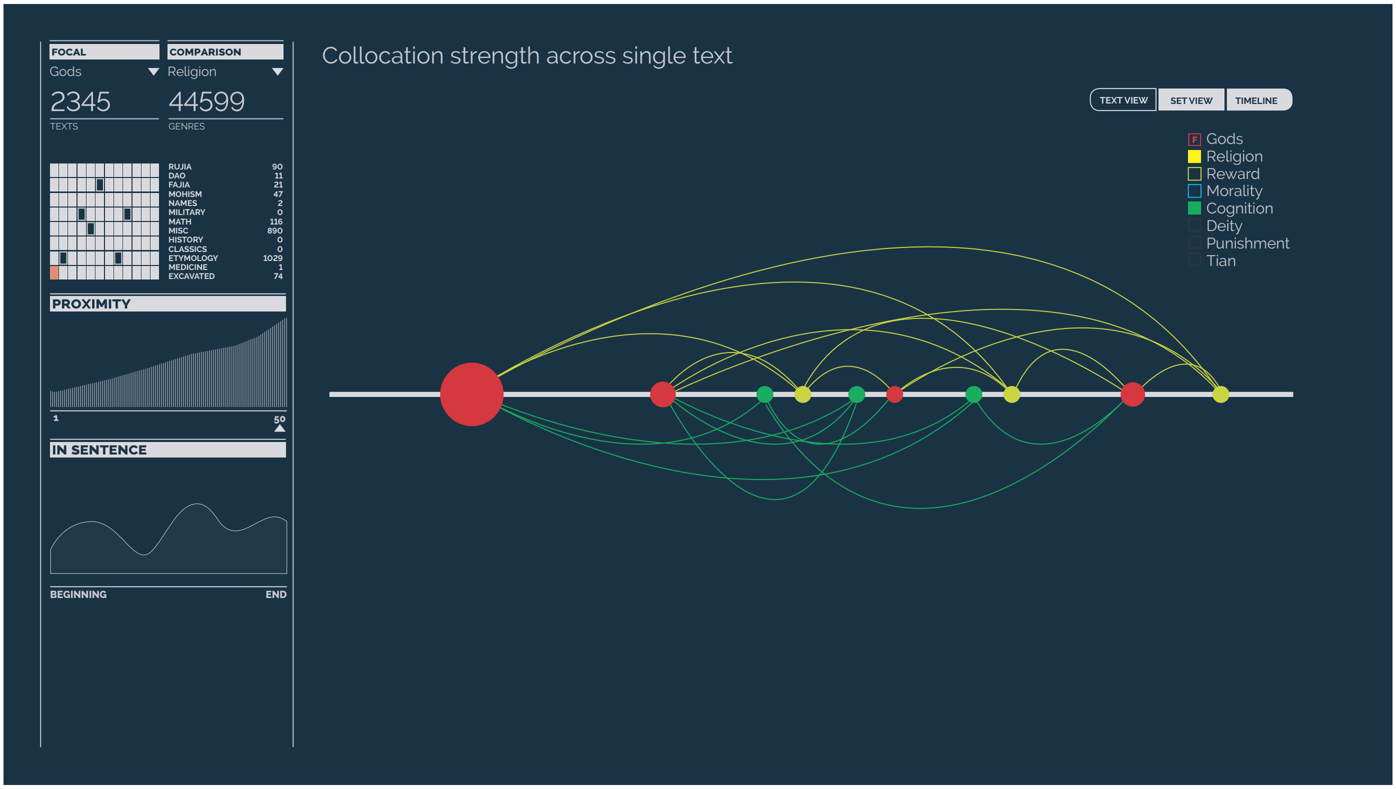
RUJIA	90
DAO	11
FAJIA	21
MOHISM	47
NAMES	2
MILITARY	0
MATH	116
MISC	890
HISTORY	0
CLASSICS	0
ETYMOLOGY	1029
MEDICINE	1
EXCAVATED	74

Figure 2.1

Top: genres update with collocation counts.

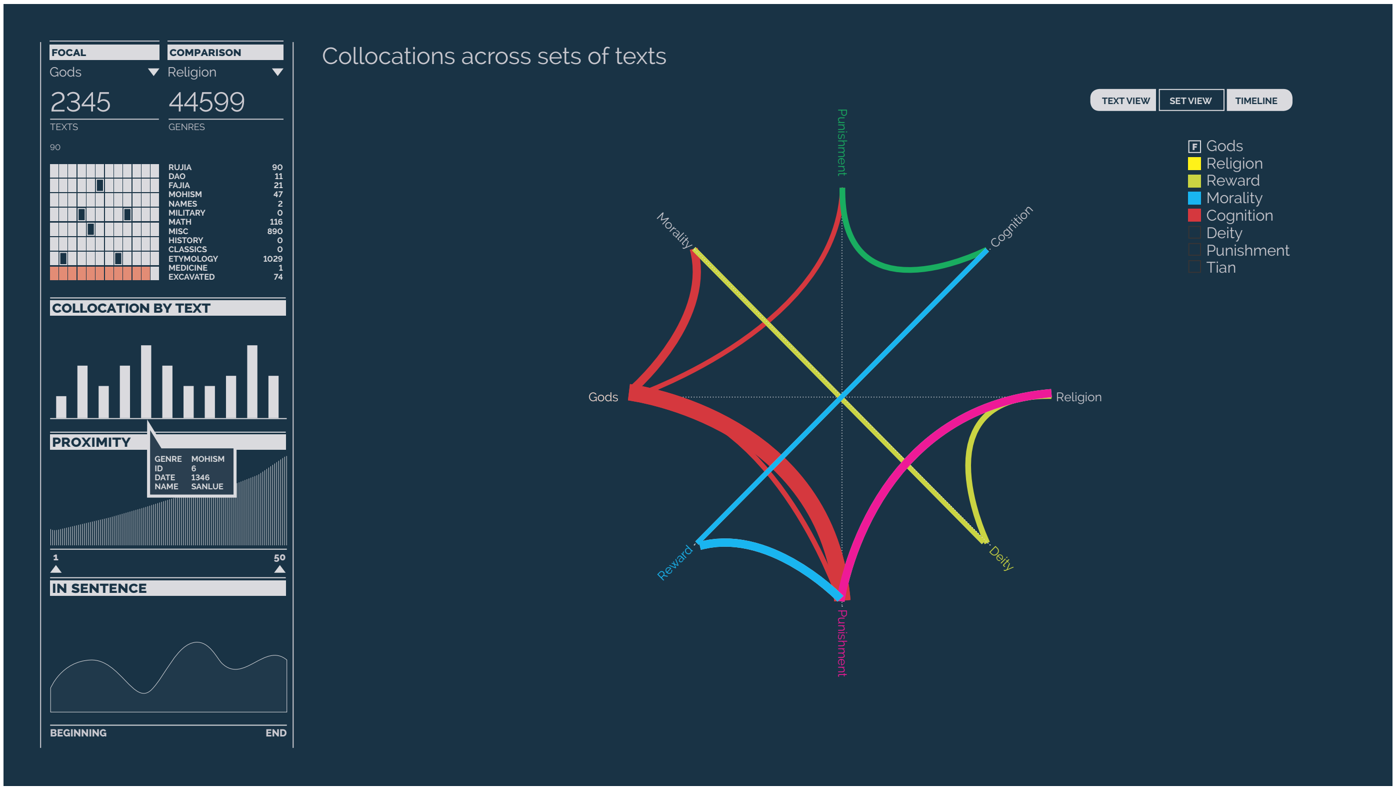
Middle: A cumulative proximity graph appears, counting collocations along the range of the slider.

Bottom: An in-sentence distribution graph appears, counting the number of times across the entire text that an in-sentence collocation occurs.



RUJIA	90
DAO	11
FAJIA	21
MOHISM	47
NAMES	2
MILITARY	0
MATH	116
MISC	890
HISTORY	0
CLASSICS	0
ETYMOLOGY	1029
MEDICINE	1
EXCAVATED	74

Figure 2.2 : With multiple comparison sets selected, the single text view shows Focal→Comparison edges for both Comparison sets.



RUJIA	90
DAO	11
FAJIA	21
MOHISM	47
NAMES	2
MILITARY	0
MATH	116
MISC	890
HISTORY	0
CLASSICS	0
ETYMOLOGY	1029
MEDICINE	1
EXCAVATED	74

GENRE	MOHISM
ID	6
DATE	1346
NAME	SANLUE

Figure 3.0 : With multiple comparison sets selected as well as multiple texts selected, a set view is needed.

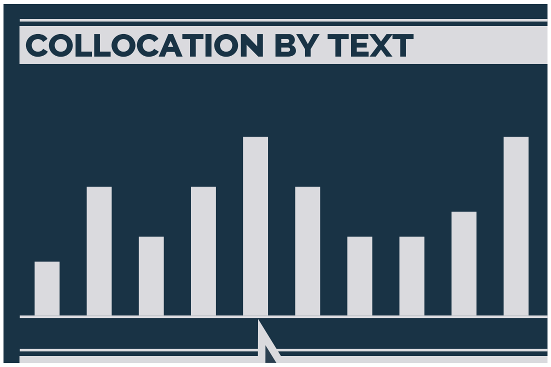
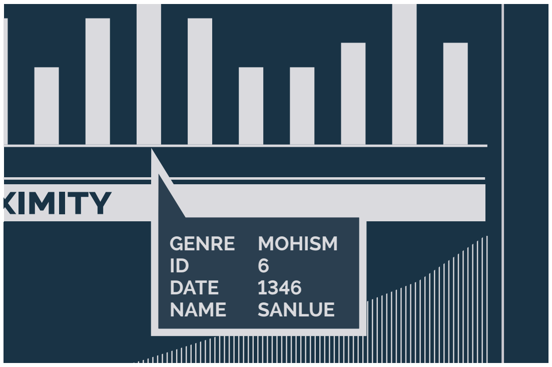
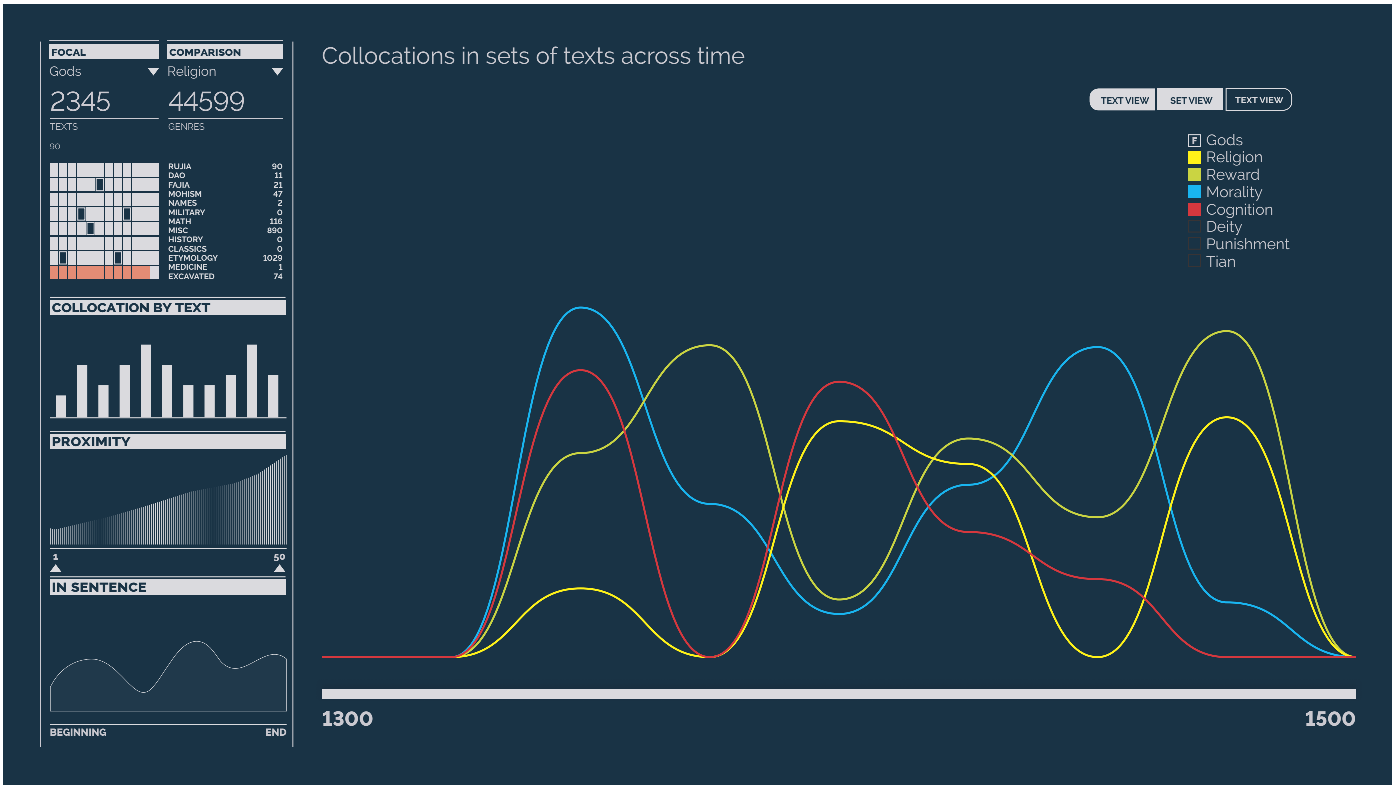


Figure 3.1 : A bar chart appears when multiple texts are selected, showing collocations strength by text. This is normalized over text size.



GENRE MOHISM
ID 6
DATE 1346
NAME SANLUE

Figure 3.2 : As with above, hovering over a bar shows text information.



RUJIA	90
DAO	11
FAJIA	21
MOHISM	47
NAMES	2
MILITARY	0
MATH	116
MISC	890
HISTORY	0
CLASSICS	0
ETYMOLOGY	1029
MEDICINE	1
EXCAVATED	74

Figure 4.0 : The timeline view plots collocation strength over time for all selected texts.

5.0 Scenario of use

Carson is a researcher for CERC. He is interested in examining whether Gods can be said to be morally conscious, and, if they are, whether they more often dole out punishment or reward. To begin, he selects the “Gods” Focal set, and “Punishment” comparison set, and the entire range of texts. A summary of all texts is displayed in the summary view window. Interested in exploring his data, he hovers over individual texts to see their statistics, which update on hover in the summary view window.

Carson decides to inspect a single text by selecting it from the sidebar. The main view window updates to show the distribution of Focal→Comparison characters. He repeats this for a few texts.

Noticing a pattern in the frequency of collocations in the “History” genre, he selects all of the texts from “History” and the summary view updates accordingly. The main view shows a distribution of Gods→Punishment across the range of [n..k] for all selected texts. Carson update n and k with a slider to pinpoint the range of collocations he wants to explore.

He switches his Comparison set from “Punishment” to “Reward”, and performs similar explorations.

Now wanting to compare “Gods→Punishment” to “Gods→Reward”, he selects multiple Comparison sets. The main view updates to show the strengths of collocations between sets by displaying collocations as number of edges between sets.

Carson suspects that the “Religion” Focal set may be better to explore than “Gods”, but to make sure, he selects both focal sets. The main view now shows all connections from Gods→Punishment, Gods→Reward, Religion→Punishment, Religion→Reward, colour-coded by Focal set.

Seeing no great difference, Carson selects all Comparison sets and all Focal sets, then chooses some sets to decides to drill down into again.

Carson now wants to see how collocation strength changes over time. Keeping the full set of “History” selected, he switches the main view to a timeline, and sees a bump around 1350 CE in Gods→Religion, and wants to see whether this bump holds true for all genres. Selecting all texts again, he sees that, indeed, all texts exhibit a rise in collocations between Gods→Religion, and starts the process again.

6.0 Implementation approach

The CERC lab is used to using Python, and the Collocation scripts I wrote over the summer are in Python, as such, I plan to use Python. Bokeh (*http://bokeh.pydata.org/*) seems to be powerful enough library to support the types of visualizations I need, which are 2D and relatively simple in expression.

7.0 Milestones/timeline

Oct 31–7	Confirm approach validity, explore tools
Oct 31	Send proposal for review by CERC member
Nov 2	Meet in person for conceptual design revisions
Nov 7	Conceptual design revisions finished, tools confirmed
Nov 8–14	Implement Phase I
Nov 14	Update due, send Phase I to CERC for review
Nov 15–21	Implement changes to Phase I, update timeline, start on Phase II
Nov 22–28	Finish Phase II, contingent on timeline update
Nov 28	Send Phase I and II to CERC for review
Nov 29–Dec 6	Implement changes to Phase I and II, update timeline, start on Phase III
Dec 6	Send Phase, I, II, III to CERC for review
Dec 712	Implement changes to Phase I, II, III, conduct further user studies
Dec 12	Final presentation
Dec 13–15	Write final report
Dec 15	Submit final report

As it stands, Phases I and II are the most important to complete. Phase III will be optional, as the time range within current texts is limited, but future uses of the tool may need to support larger time ranges.

Phase I	Implement single text views for Summary and Main view.
Phase II	Implement Focal/Comparison set oriented Main view.
Phase III	Implement timeline oriented view.

8.0 Previous work

This project takes a cue from text analysis and visualization tools such as PNNL's In-Spire[2], IBM's Many Eyes text analysis tools[3], as well as more general visualization approaches such as HivePlot[4]. In-Spire especially shares a conceptual similarity in that it is for exploring a set of documents, uses multiple linked views, and supports an iterative workflow for analysis. This project does not take the same bag-of-words approach that some text analysis tools do, however, since word sets have already been developed, tuned, and prioritized. In this way, this project has more in common with a specialized network analysis tool, as it is designed to capture the meaning of connections between items in sets, rather items alone. Since this project takes the the linear nature of text into account, a solution such as HivePlot provides good inspiration.

References

1. CERC homepage. http://www.hecc.ubc.ca/cerc/
2. Pacific Northwest National Laboratory In-Spire website. http://in-spire.pnnl.gov/
3. IBM Many Eyes hompage. http://www-01.ibm.com/software/analytics/many-eyes/
4. HivePlot information portal. http://www.hiveplot.net/

Appendix A

delimiters = {。}

stopwords = {#@、「」：；『』《》，、「」：；『』之 不 也 以 而 其 為 曰
者 子 有 於 十 則 無 所 故 三 二 一 是 與 夫 可 五 將 使 何 至 四 矣 自 此 太 謂 如 乃 百 皆
乎 于 在 非 六 諸 必 然 若 及 未 萬 吾 焉 我 復 千 亦 九 七 方 元 正 多 西 足 又 高 內 當 去
北 來 氏 外 同 受 反 少 常 過 后 作 因 雖 始 里 請 女 右 敢 前 易 求 說 左 起 會 定 通 對 哉
難 稱 屬 宜 聽 終 遠 盡 異 進 初 甚 本 止 興 耳 廣 益 應 還 絕 往 己 邪 固 首 由 共 徒 任 更
惠 少 文 景 武 昭 宣 元 成 哀 平 更始 光武 章 和 殤 安 順 沖 質 桓 獻 昭烈}

emotion = {安 欲 情 乐 喜 急 忙 恐 惊 爱 怒 恶 苦 怕 欢 感 忧 萧 恨 惧 怨 患 悦 哀 虑 惜 痛
悲 怜 烦 畏 愁 慌 慕 戚 忌 困 悄 悔 嫌 羞 闷 欣 恼 厌 恤 寂 匆 恚 愧 愤 惨 惶 恋 骇 惭 恍 羡
悠 怼 惮 恳 忿 疼 闵 畅 僖 慨 妒 悼 恸 忻 栗 愕 禧 讶 兢 嫉 怅 嗔 歆 憾 懊 怡 娱 瞿 竦 厄 恺
忝 诧 逍 诟 怖 寞 怏 憎 窘 尬 恻 悯 顼 愉 怿 惕 悖 愠 觊 慑 懔 悚 惴 忏 怆 愦 缱 怔 恹}

reduced_gods ={天 帝}