Web based computer-assisted text analysis

Christopher Ohge, Martin Steer

Riga Technical University, September 2019

Hathi Trust Research Centure (HTRC)

https://pro.europeana.eu/post/text-mining-2-how-scholars-can-support-digital-libraries

Hathi Trust Research Centure (HTRC)

- Promotes scholarly use of the HathiTrust Digital Library
- Custodian of Google books
  - OCR corpus – many character errors!
  - Biased – e.g C19th lacks genre metadata
- Provides
  - Data analysis infrastructure
  - “Non-consumptive” use
  - Extracted Features Dataset
  - Visualisation tools
  - Lots more…

https://pro.europeana.eu/post/text-mining-2-how-scholars-can-support-digital-libraries

HTRC Extracted Features Dataset

- Pre-extracted features
- Python/R libraries, API access
- Volume-level
  - Bibliographic metadata
    - title, pubDate, language,
    - genre, typeOfResource
    - names, imprint, rightsAttributes
  - pageCount, size

Data Stats

# of volumes	15,722,079
# of pages	5,787,519,444
# of tokens	2,449,739,213,773
# of IC volumes	9,914,509
# of IC pages	3,005,448,348
# of IC tokens	1,777,793,828,310
# of PD volumes	5,807,570
# pd pages	2,602,212,586
# pd tokens	1,197,838,539,662

https://wiki.htrc.illinois.edu/display/COM/Extracted+Features+Dataset

HTRC Extracted Features Dataset

- Page level
  - Header, Body, Footer info
  - Page sequence numbers
  - tokenCount, lineCount, sentenceCount
  - Languages (inferred per page)
  - tokenPosCount (POS tags)
  - beginCharCounts, endCharCounts


```
{
    "id": "uc1.b2617728",
    "metadata": {
        "schemaVersion": "1.3",
        "dateCreated": "2016-06-25T03:45:05.180
        "volumeIdentifier": "uc1.b2617728",
        "accessProfile": "google",
        "rightsAttributes": "ic",
        "hathitrustRecordNumber": "10281140",
        "enumerationChronology": " ",
        "sourceInstitution": "UC",
        "sourceInstitutionRecordNumber": ".b165
        "oclc": [
            "70545428"
        ],
        "isbn": [],
```

HTRC Extracted Features Dataset

- Page level
  - Header, Body, Footer info
  - Page sequence numbers
  - tokenCount, lineCount, sentenceCount
  - Languages (inferred per page)
  - tokenPosCount (POS tags)
  - beginCharCounts, endCharCounts


```
{
    "seq": "00000033",
    "tokenCount": 273,
    "lineCount": 36,
    "emptyLineCount": 0,
    "sentenceCount": 11,
    "header": {
        "tokenCount": 0,
        "lineCount": 0,
        "emptyLineCount": 0,
        "sentenceCount": 0,
        "capAlphaSeq": 0,
        "beginCharCounts": {},
        "endCharCount": {},
        "tokenPosCount": {}
    },
    "body": {
        "tokenCount": 273,
        "lineCount": 36,
        "emptyLineCount": 0,
```

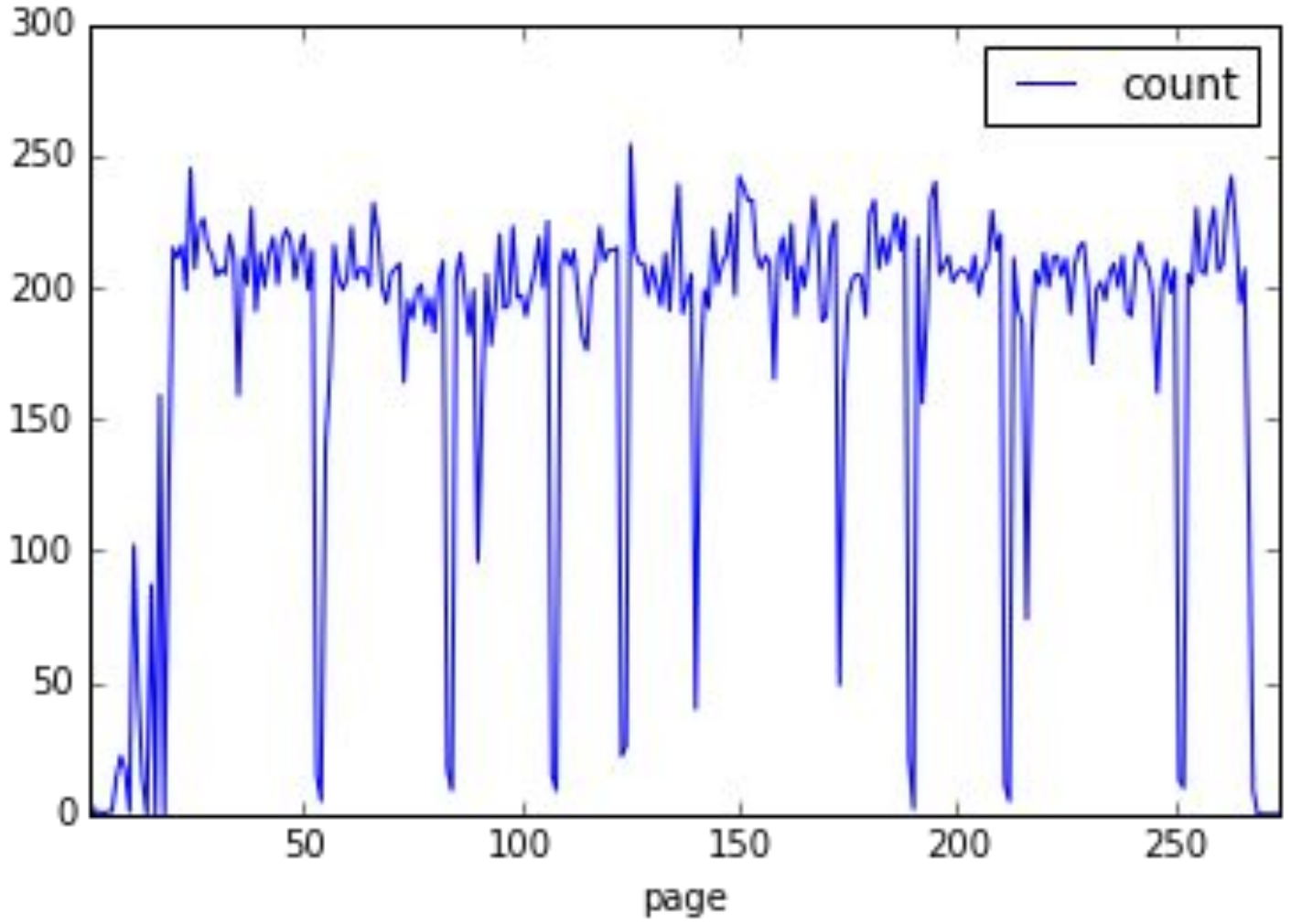
HTRC Extracted Features - Dataframe

page	section	token	pos	count
27	body	those	DT	1
		within	IN	1
28	body	a	DT	3
		be	VB	1
		deserted	VBN	1
		faintly	RB	1
		important	JJ	1

HTRC Extracted Features - Dataframe

section	token	count
body	,	3258
	"	1670
	the	1565
	.	1532
	and	1252

HTRC Extracted Features - Plotting



https://programminghistorian.org/en/lessons/text-mining-with-extracted-features

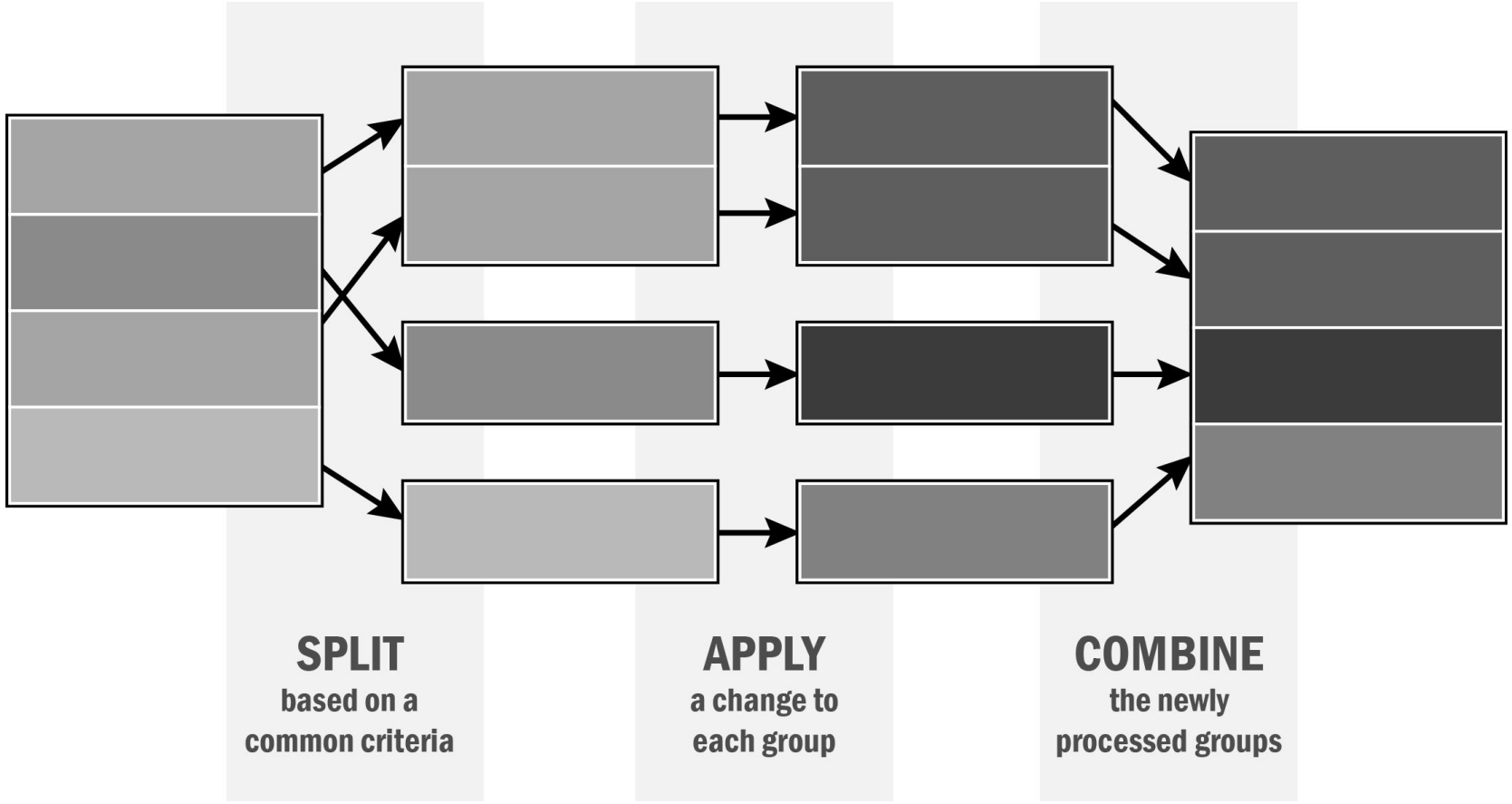
HTRC Extracted Features - Grouping

page	section	token	pos	count
19	header	CHAPTER	NNP	1
35	header	CHAPTER	NNP	1
56	header	CHAPTER	NNP	1
73	header	CHAPTER	NNP	1
91	header	CHAPTER	NNP	1
115	header	CHAPTER	NNP	1
141	header	CHAPTER	NNP	1
158	header	CHAPTER	NNP	1
174	header	CHAPTER	NNP	1
193	header	CHAPTER	NNP	1

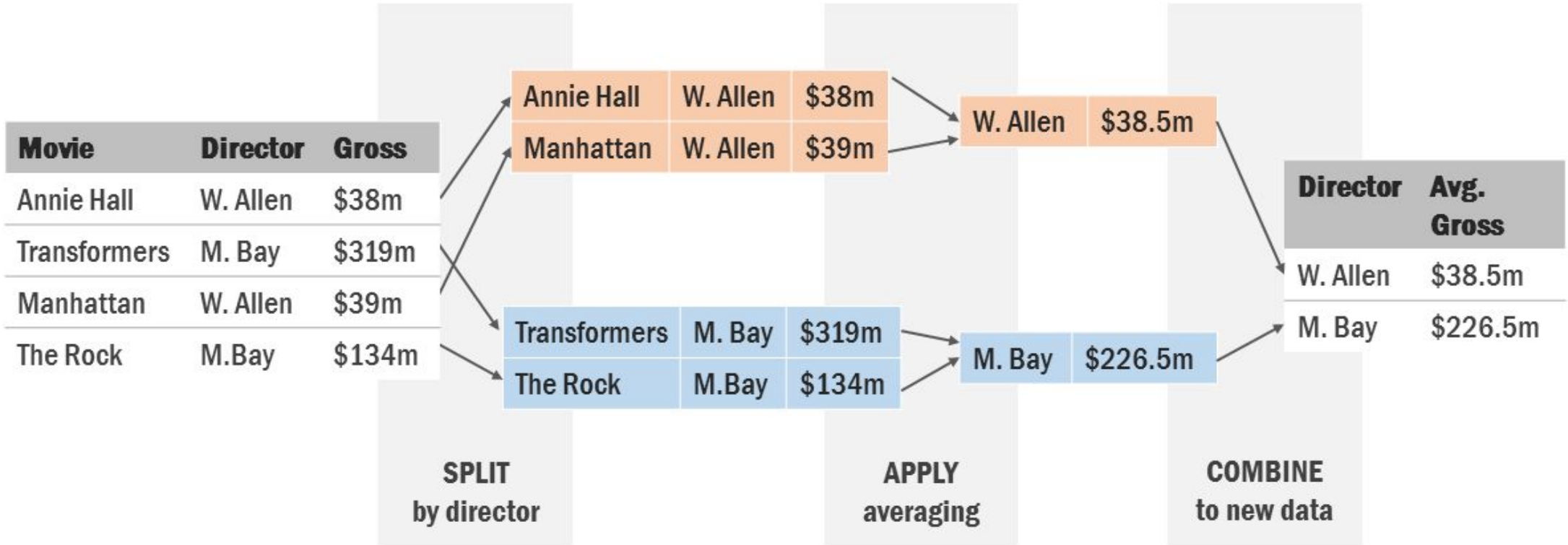
HTRC Extracted Features - Plotting

https://programminghistorian.org/en/lessons/text-mining-with-extracted-features

Split-apply-combine

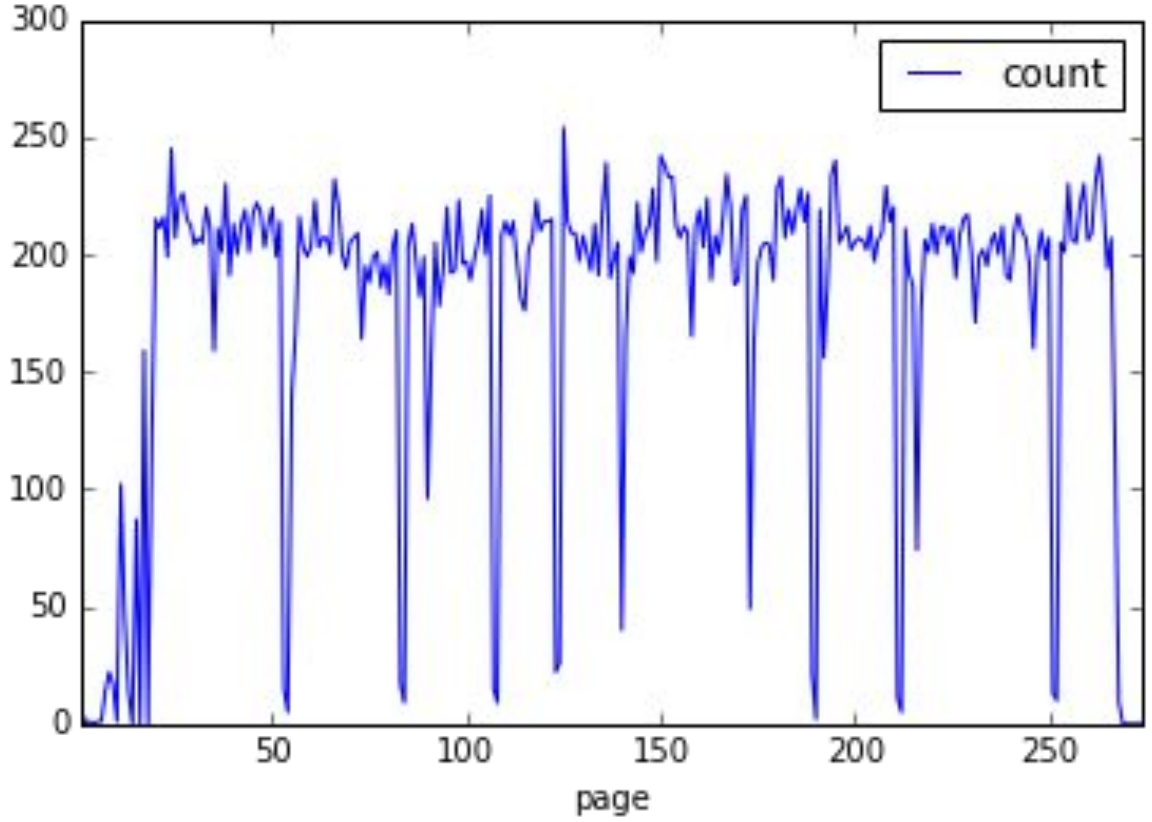


Split-apply-combine



Text analysis use cases

- Distant reading
- Word similarity
- Topic models
- Plot analysis
- Emotion analysis
- Visual structure



https://programminghistorian.org/en/lessons/text-mining-with-extracted-features

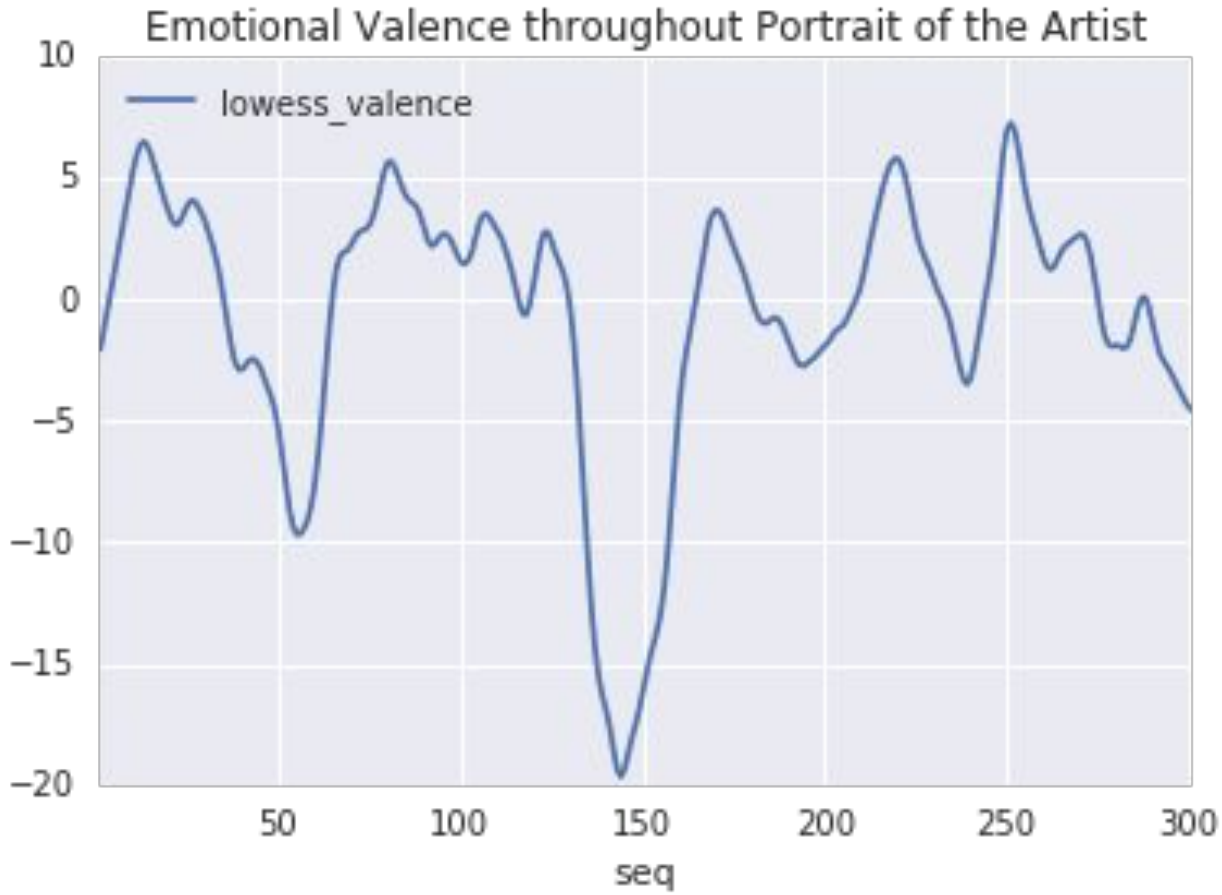
Text analysis use cases

- Distant reading
- Word similarity
- Topic models
- Plot analysis
- Emotion analysis
- Visual structure

https://wiki.htrc.illinois.edu/display/COM/Extracted+Features+in+the+Wild

Text analysis use cases

- Distant reading
- Word similarity
- Topic models
- Plot analysis
- Emotion analysis
- Visual structure

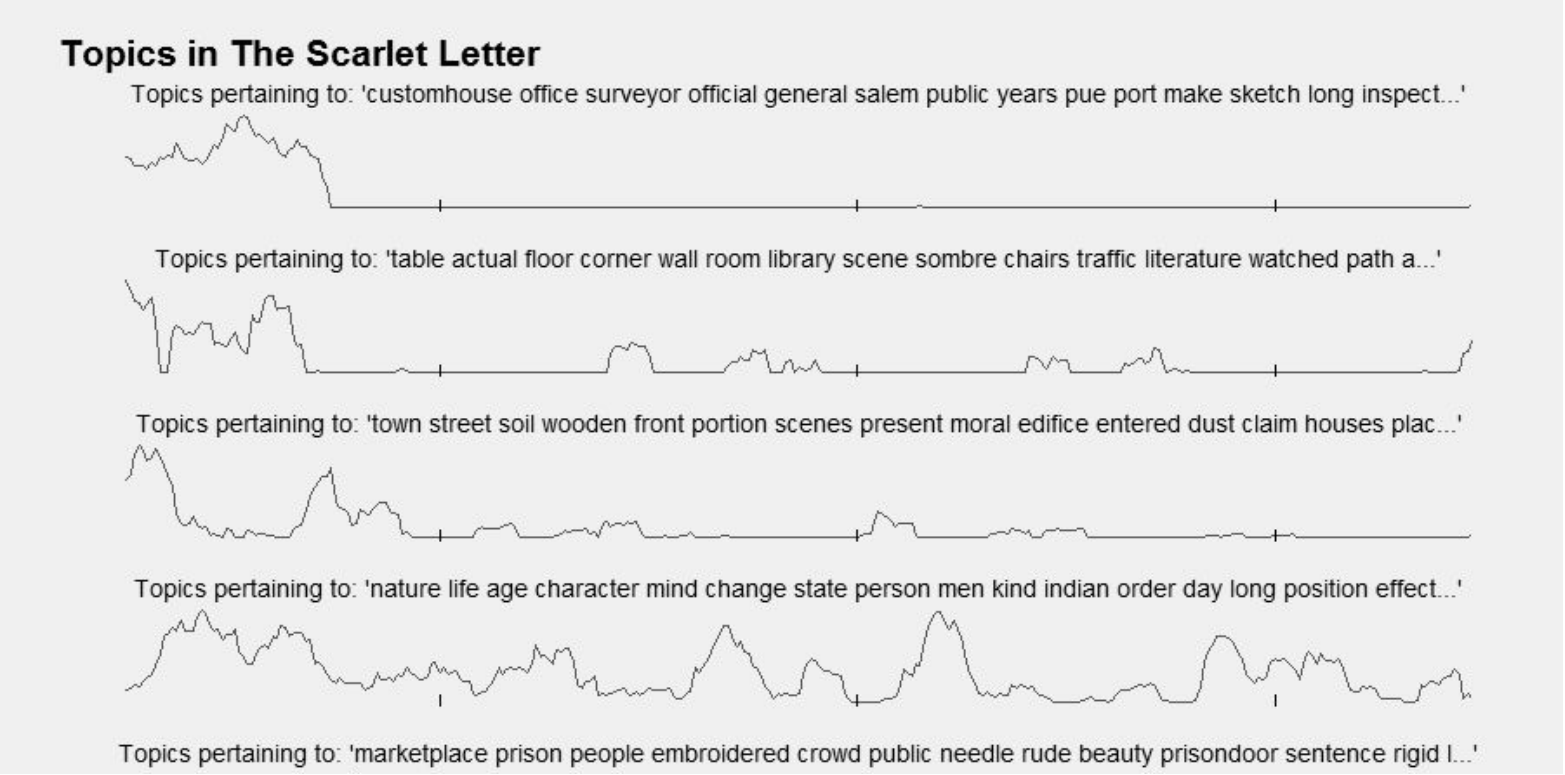


https://programminghistorian.org/en/lessons/text-mining-with-extracted-features

Text analysis use cases

- Distant reading
- Word similarity
- Topic models
- Plot analysis
- Emotion analysis
- Visual structure

Within-book topic modelling, Peter Organisciak



https://github.com/organisciak/htrc-book-models

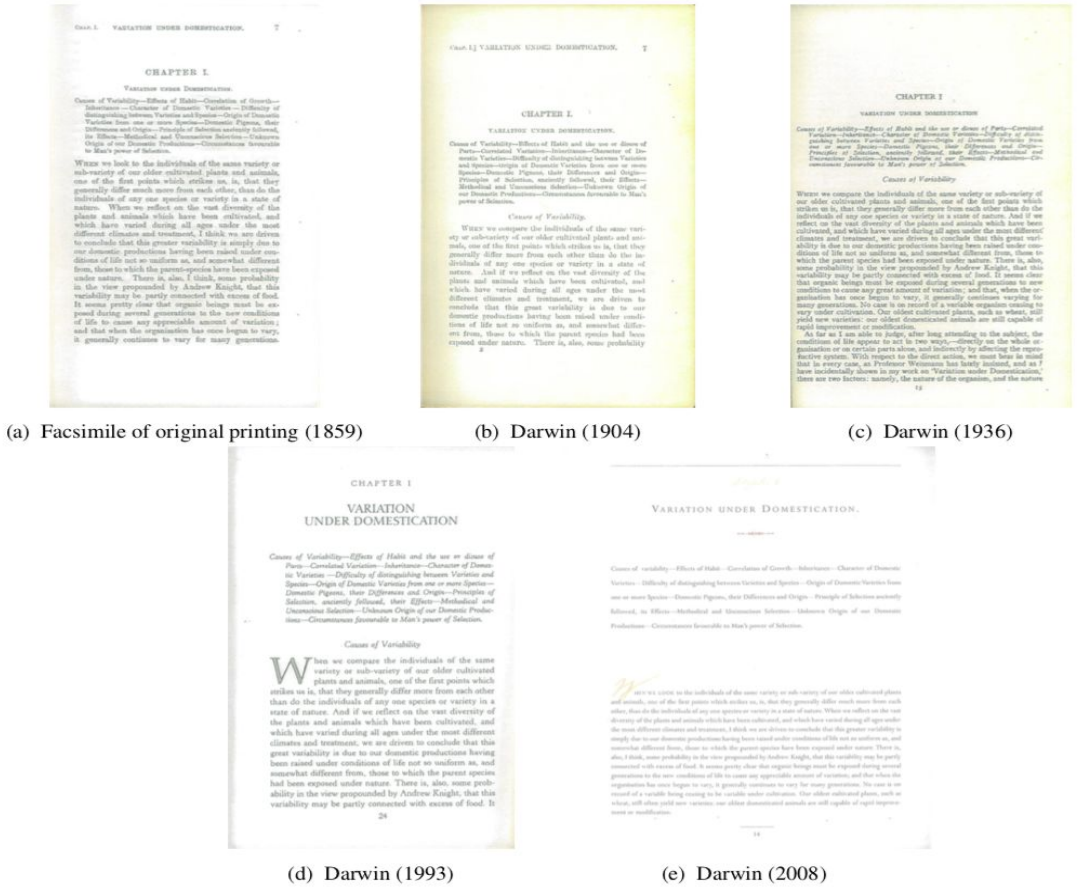
Interpretation of mediums

5 copies of Origin of Species in HTC corpora

Two trends:

- 1929 and after are on average about one quarter of an inch taller and one eighth of an inch wider than books published before

- the font size (as determined by the median line height) also appears to be increasing (by a quarter of a point)



(a) Facsimile of original printing (1859) (b) Darwin (1904) (c) Darwin (1936)

(d) Darwin (1993) (e) Darwin (2008)

Figure 1: Five versions of Darwin's *Origin of Species*, illustrating different design choice in page layout.

Bamman, David, and Björn Hartmann. 'Modeling the History of Book Design HTRC Whitepaper: Summary of Activities', n.d., 8.

Insight driven investigation

Distant Horizons, Ted Underwood (2019)

Frequency of colour terms in random fiction increases post 1800 – why?

- Does it represent the decline in 3rd person narration? A shift to description?
- We need to look for correlations

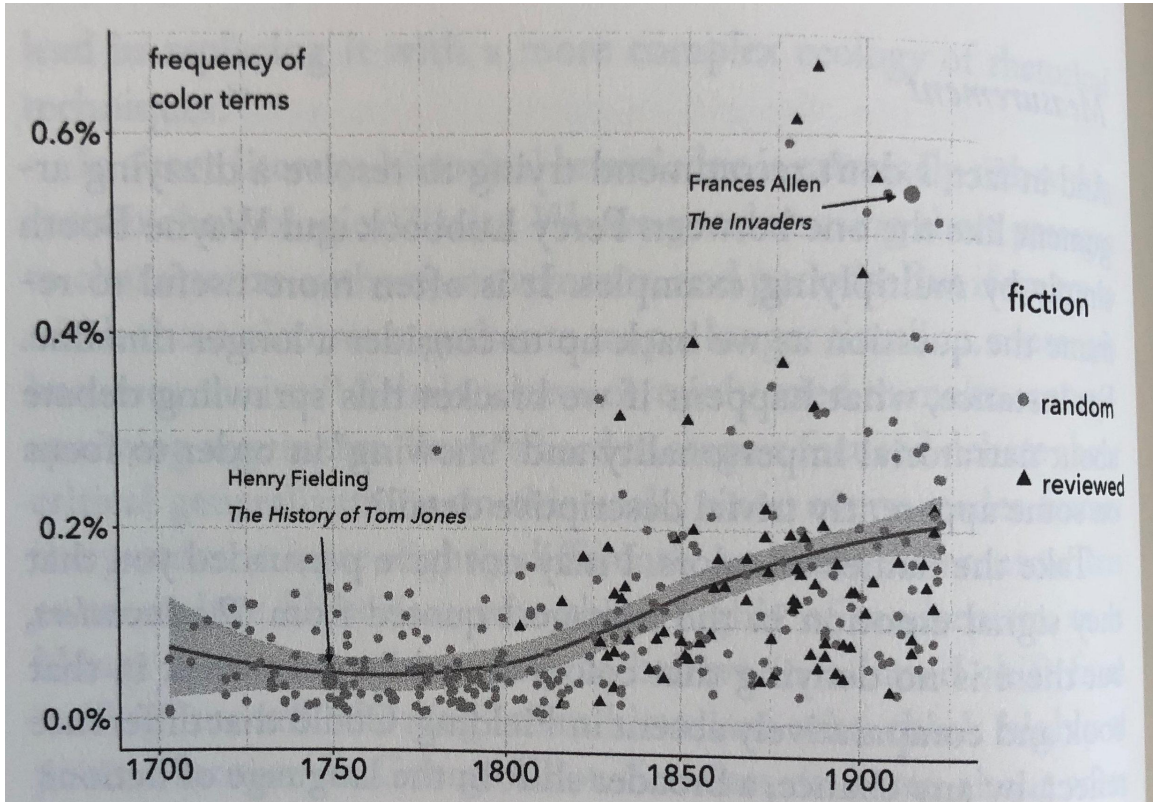


FIGURE 1.1. Frequency of color terms in a random sample of fiction, 1700–1922. Points represent 347 individual volumes. After 1800, volumes reviewed in a sample of prestigious magazines are distinguished as black triangles.

Distant Horizons, Ted Underwood (2019), p10

Frames of context

Standord 2012 Heuser & Le-Khac:

- Rising frequency of concrete adjectives in 19th century novels (colours, names, body parts, action verbs – used for physical descriptions)

Concluded fictional discourse was transitioning ”from telling to showing”

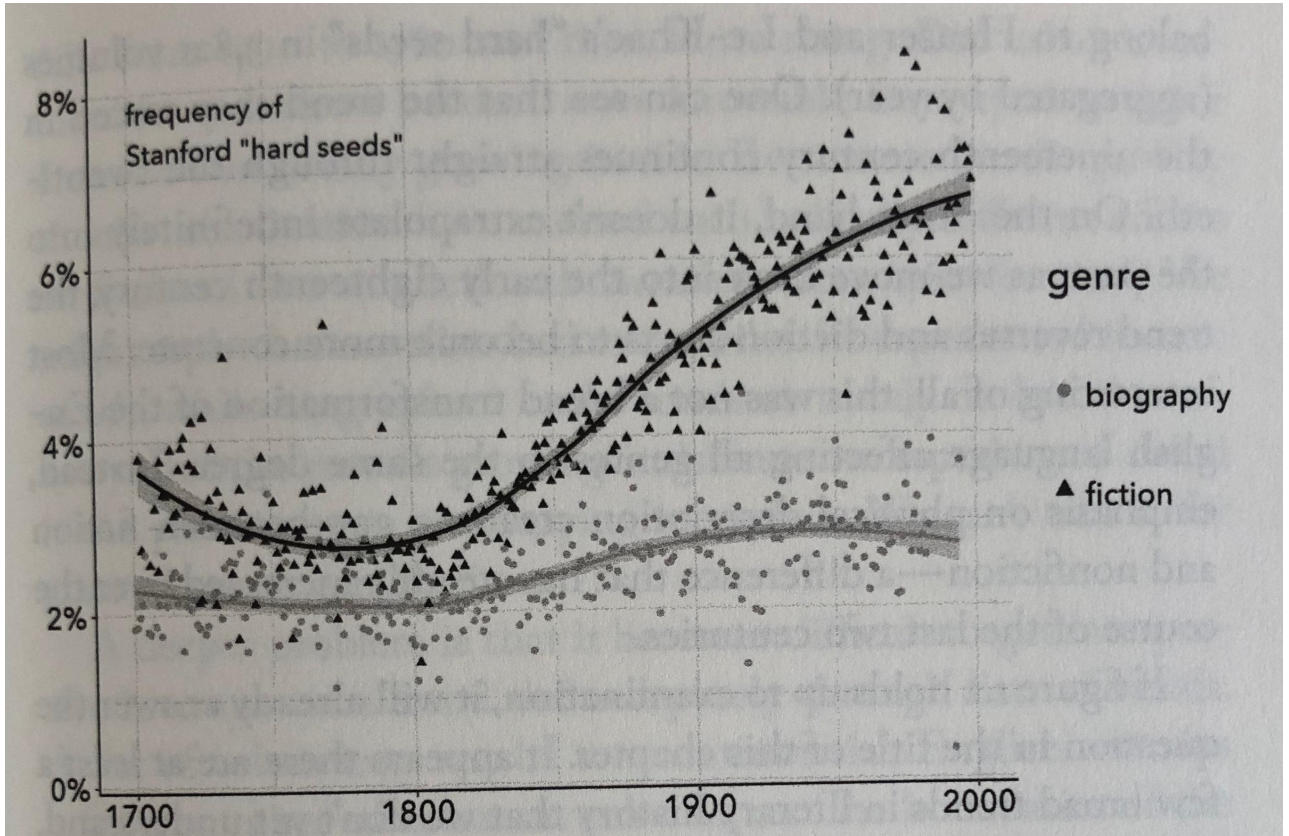
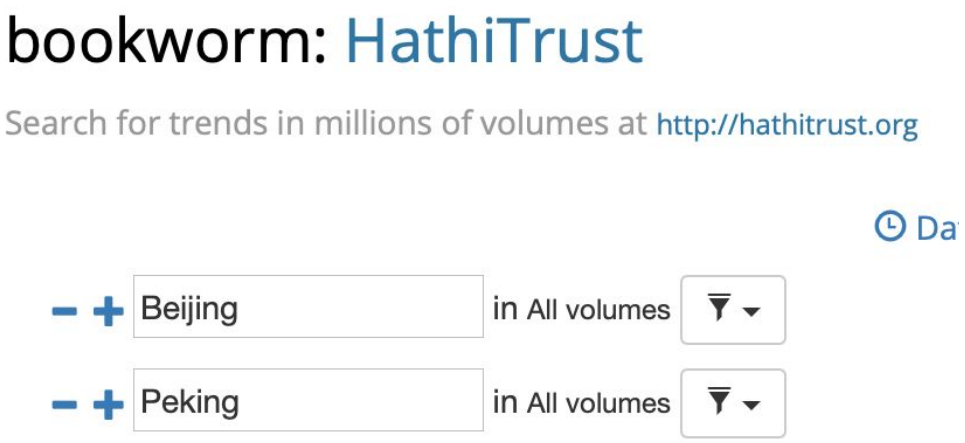


FIGURE 1.2. Frequency of Stanford “hard seeds” as percentage of fiction and biography. Data points represent yearly averages for each genre.

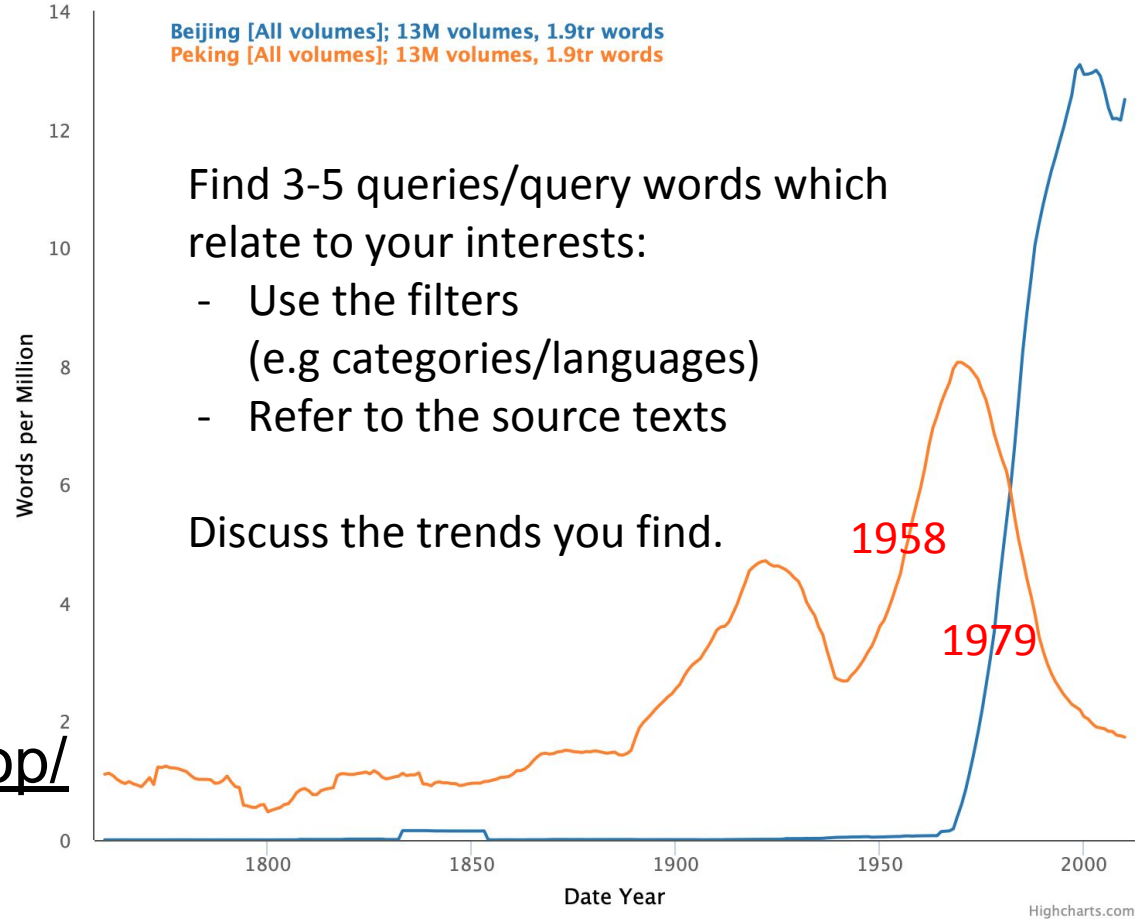
Distant Horizons, Ted Underwood (2019), p13

Practical – Finding Trends



https://bookworm.htrc.illinois.edu/develop/

https://books.google.com/ngrams



Find 3-5 queries/query words which relate to your interests:

- Use the filters (e.g categories/languages)
- Refer to the source texts

Discuss the trends you find.

Examples

Literary trends

- Red
- White

(Narrow class filter: fiction/biography)

Meaning change

- Hysteria
- Gay
- legacy/heritage

Parts of speech

Book_VERB
Book_NOUN

Geopolitical

- Persia
- Iran
- إيران

- Beijing
- China

Practical – Interpreting Topics

HathiTrust Fiction (1920-1922) topic-browser

https://jgoodwin.net/htb/#/model/grid

Pick 3 topics in each group:

- Use the GUI to investigate
- Why does the GUI help?
- Refer to the source texts

Discuss your findings.

French literature cluster?
Label = French Women or French Gender?

Contains = Fiction with a lot of gender descriptions?
A lot of polite conversation with women characters?

Computer-assisted tools and Hermeneutics

Descartes, *Discourse on Method*: “[T]hings made up of different elements and produced by the hands of several master craftsmen are often less perfect than those on which only one person has worked.”

Individual versus communal-participatory inquiry; dialogic relationship between human and human and machine, criticism and code; augmentation of criticism.

Teamwork changes the single-interpreter model that Descartes describes: “analytical tools *are* instantiations of interpretive methods that can be woven closely into other hermeneutical things” (Rockwell and Sinclair).