IDS 570 — Data Exploration: Political Economies Corpus

Contents

Install & Load Packages

Load Corpus

> **Instructions:** Place this `.Rmd` file in the **same folder** as your 20 `.txt` files. The script detects them automatically.

```
## Files found: 20
```

```
## A06785.txt
## A06786.txt
## A06788.txt
## A06789.txt
## A06790.txt
## A06791.txt
## A07886.txt
## A32827.txt
## A32828.txt
## A32829.txt
## A32830.txt
## A32833.txt
## A32836.txt
## A32837.txt
## A32838.txt
## A32839.txt
## A50763.txt
## A51598.txt
## A69858.txt
## A93819.txt
```

```
##
## Documents loaded:
```

```
## Strong_Imag
## Law_Merchant
## Relig_Justice
## Three_Essent
## April
## Wealth_Realm
## Trade_Rules
## Discourse_1
## Fallen_Man
## Wool_Manuf
## EastIndia_Def
## Discourse_2
## Poor_Eng_1
## Interest_Hol
## EastIndia_Co
## Trade_Op_1
## Poor_Eng_2
## Treasure
## Trade_Op_2
## Lion_Key
```

Step 0 — Normalization

Required: normalize the long S character (`f -> s`).

Additional choices:

- **Remove numbers:** pagination markers, folio numbers, and running totals create spurious tokens that inflate TF-IDF scores without reflecting content.
- **Collapse whitespace:** EEBO transcriptions sometimes have irregular spacing due to original typesetting.
- We do **not** normalize `u/v`, `i/j`, or variant spellings. These are meaningful signals of register and date; normalizing them would erase the lexical distinctiveness we aim to measure.

```
## Corpus summary:
```

```
## Corpus consisting of 20 documents, showing 20 documents:
##
##             Text Types Tokens Sentences
##      Strong_Imag  4519  41166       763
##     Law_Merchant 13612 255763      4954
##    Relig_Justice  3348  32632       702
##    Three_Essent  3469  27122       517
##            April  2658  12869       231
##     Wealth_Realm  2032  20893       303
##      Trade_Rules  2043  16343       264
##      Discourse_1  5048  56447       862
##       Fallen_Man  1456   6494        97
##       Wool_Manuf  1093   4842        94
##    EastIndia_Def   873   3355        56
##      Discourse_2  5094  56600       874
##       Poor_Eng_1  1208   5386        89
##     Interest_Hol   803   3500        62
##     EastIndia_Co   924   3367        69
##       Trade_Op_1  2106  13671       366
##       Poor_Eng_2  1330   5897        96
##         Treasure  3094  30850       433
##       Trade_Op_2  1386   6931       217
##         Lion_Key   779   2909        57
```

Approach 1 — TF-IDF: Lexical Distinctiveness

Build DFM and compute TF-IDF

```
## DFM dimensions (documents x features): 20 17065
```

Extract top terms per document

document	term	tfidf
Strong_Imag	moneys	79.472
Strong_Imag	misselden	46.837
Strong_Imag	p	42.858
Strong_Imag	circle	41.000
Strong_Imag	exchange	35.982
Strong_Imag	starlin	33.000
Strong_Imag	realm	25.468
Strong_Imag	doller	23.069
Strong_Imag	d	22.888
Strong_Imag	commodities	22.364
Strong_Imag	stivers	22.276
Strong_Imag	undervaluation	21.674
Law_Merchant	ll	318.433
Law_Merchant	hundreth	246.845
Law_Merchant	ounces	242.630
Law_Merchant	moneys	141.785
Law_Merchant	factor	141.712
Law_Merchant	carrats	125.234
Law_Merchant	weight	123.348
Law_Merchant	ss	109.287
Law_Merchant	assurors	104.082
Law_Merchant	grains	102.350
Law_Merchant	pieces	98.143
Law_Merchant	ship	97.104
Relig_Justice	m	59.271
Relig_Justice	sols	42.000
Relig_Justice	silver	27.689
Relig_Justice	livers	25.000
Relig_Justice	realm	23.876
Relig_Justice	dearth	23.066
Relig_Justice	commodities	21.365
Relig_Justice	gold	21.067
Relig_Justice	price	20.137
Relig_Justice	shillings	18.664
Relig_Justice	moneys	17.159
Relig_Justice	bodine	16.076
Three_Essent	moneys	34.618
Three_Essent	starlin	29.000
Three_Essent	monies	20.915
Three_Essent	exchange	20.740
Three_Essent	commodities	20.240
Three_Essent	shillings	17.460
Three_Essent	realm	17.111
Three_Essent	price	13.322
Three_Essent	doller	12.359
Three_Essent	pence	12.041
Three_Essent	reals	11.535
Three_Essent	enhancing	11.503
April	monster	44.235
April	dragon	17.474
April	maketh	15.654
April	moon	12.359

document	term	tfidf
April	tail	9.786
April	ass	9.063
April	causeth	8.429
April	sweet	6.990
April	unto	6.881
April	leaguors	6.505
April	daughter	6.291
April	smell	6.000
Wealth_Realm	moneys	26.491
Wealth_Realm	realm	26.264
Wealth_Realm	commodities	25.612
Wealth_Realm	exchange	24.238
Wealth_Realm	fineness	15.686
Wealth_Realm	price	15.180
Wealth_Realm	shillings	13.245
Wealth_Realm	silver	10.664
Wealth_Realm	exchangers	10.485
Wealth_Realm	ounces	10.235
Wealth_Realm	albeit	9.935
Wealth_Realm	ducats	9.887
Trade_Rules	wares	22.848
Trade_Rules	raw-silkes	16.000
Trade_Rules	realm	13.132
Trade_Rules	indies	12.643
Trade_Rules	india	10.837
Trade_Rules	d	10.750
Trade_Rules	indigo	10.486
Trade_Rules	east	9.167
Trade_Rules	rawsilke	9.107
Trade_Rules	aleppo	9.063
Trade_Rules	sterling	9.031
Trade_Rules	marseilles	8.000
Discourse_1	plantations	38.298
Discourse_1	abatement	28.235
Discourse_1	dly	18.126
Discourse_1	cent	17.960
Discourse_1	interest	14.610
Discourse_1	english	14.561
Discourse_1	thly	14.000
Discourse_1	new-england	13.847
Discourse_1	poor	13.244
Discourse_1	pag	12.643
Discourse_1	newfoundland	12.359
Discourse_1	usury	12.342
Fallen_Man	support	5.229
Fallen_Man	disposal	5.204
Fallen_Man	mortgage	5.000
Fallen_Man	commerce	4.437
Fallen_Man	creditors	4.120
Fallen_Man	purely	3.903
Fallen_Man	promotion	3.903
Fallen_Man	impediments	3.903

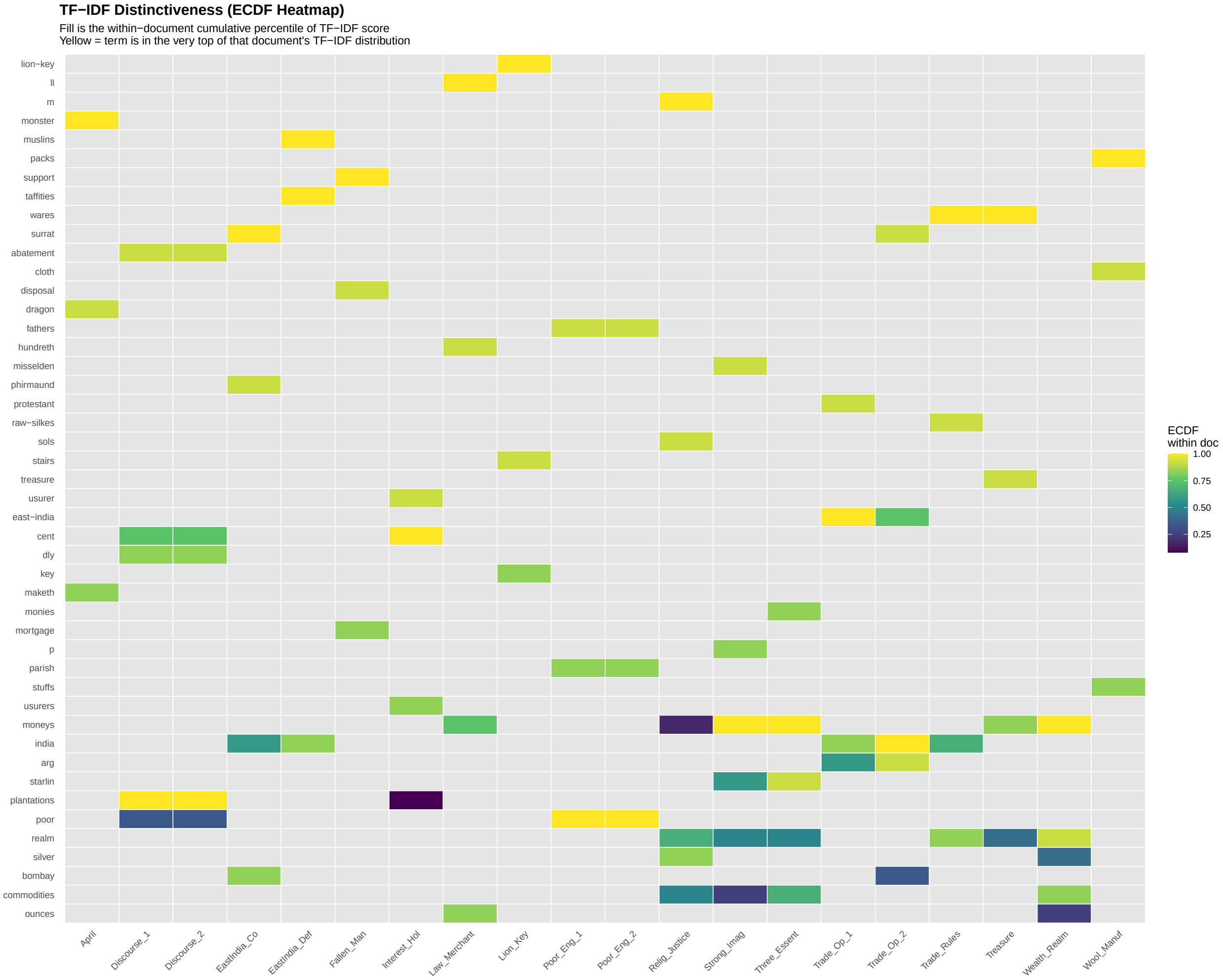
document	term	tfidf
Fallen_Man	advantages	3.647
Fallen_Man	produce	3.311
Fallen_Man	various	3.296
Fallen_Man	earth	3.121
Wool_Manuf	packs	9.000
Wool_Manuf	cloth	6.922
Wool_Manuf	stuffs	5.720
Wool_Manuf	merchant-adventurers	5.204
Wool_Manuf	wool	4.659
Wool_Manuf	broad-cloths	4.120
Wool_Manuf	manufactors	3.903
Wool_Manuf	growers	3.903
Wool_Manuf	african	3.903
Wool_Manuf	thousand	3.873
Wool_Manuf	clothiers	3.815
Wool_Manuf	price	3.718
EastIndia_Def	muslins	3.903
EastIndia_Def	taffities	3.903
EastIndia_Def	india	2.709
EastIndia_Def	em	2.472
EastIndia_Def	product	2.408
EastIndia_Def	prime	2.000
EastIndia_Def	mony	2.000
EastIndia_Def	etc	1.704
EastIndia_Def	expended	1.648
EastIndia_Def	scotch	1.648
EastIndia_Def	undeniable	1.648
EastIndia_Def	worsted	1.648
Discourse_2	plantations	40.122
Discourse_2	abatement	27.713
Discourse_2	dly	18.126
Discourse_2	cent	17.586
Discourse_2	new-england	15.051
Discourse_2	english	14.561
Discourse_2	interest	14.540
Discourse_2	thly	14.000
Discourse_2	poor	13.368
Discourse_2	newfoundland	13.183
Discourse_2	pag	12.643
Discourse_2	usury	12.342
Poor_Eng_1	poor	8.621
Poor_Eng_1	fathers	5.549
Poor_Eng_1	parish	5.229
Poor_Eng_1	propose	4.214
Poor_Eng_1	question	3.328
Poor_Eng_1	assembly	3.192
Poor_Eng_1	hospitals	3.137
Poor_Eng_1	pious	3.010
Poor_Eng_1	quest	2.796
Poor_Eng_1	defect	2.736
Poor_Eng_1	elected	2.408
Poor_Eng_1	election	2.408

document	term	tfidf
Interest_Hol	cent	6.548
Interest_Hol	usurer	6.021
Interest_Hol	usurers	3.647
Interest_Hol	per	3.586
Interest_Hol	planters	2.472
Interest_Hol	barbadoss	2.472
Interest_Hol	l	2.374
Interest_Hol	richest	2.097
Interest_Hol	qualifications	2.000
Interest_Hol	objection	1.871
Interest_Hol	interest	1.835
Interest_Hol	plantations	1.824
EastIndia_Co	surrat	11.000
EastIndia_Co	phirmaund	9.107
EastIndia_Co	bombay	6.591
EastIndia_Co	mogul	6.000
EastIndia_Co	mogul's	5.204
EastIndia_Co	india	4.214
EastIndia_Co	guns	4.194
EastIndia_Co	english	4.027
EastIndia_Co	mary	4.000
EastIndia_Co	cargoes	3.296
EastIndia_Co	committees	2.796
EastIndia_Co	kempthorne	2.602
Trade_Op_1	east-india	11.314
Trade_Op_1	protestant	11.000
Trade_Op_1	india	10.837
Trade_Op_1	tuns	10.031
Trade_Op_1	seamen	9.949
Trade_Op_1	arg	9.000
Trade_Op_1	company	8.334
Trade_Op_1	charter	7.751
Trade_Op_1	bengall	7.000
Trade_Op_1	regulated	6.990
Trade_Op_1	ly	6.505
Trade_Op_1	silk	5.546
Poor_Eng_2	poor	9.245
Poor_Eng_2	fathers	5.895
Poor_Eng_2	parish	5.752
Poor_Eng_2	goals	3.903
Poor_Eng_2	propose	3.612
Poor_Eng_2	assembly	3.192
Poor_Eng_2	hospitals	3.137
Poor_Eng_2	design	2.736
Poor_Eng_2	communication	2.614
Poor_Eng_2	sessions	2.602
Poor_Eng_2	fills	2.602
Poor_Eng_2	elected	2.408
Treasure	wares	36.349
Treasure	treasure	13.493
Treasure	moneys	10.837
Treasure	exchange	10.495

document	term	tfidf
Treasure	genova	10.000
Treasure	whereby	9.761
Treasure	divers	9.363
Treasure	realm	9.153
Treasure	insurance	9.000
Treasure	war	8.828
Treasure	gain	8.675
Treasure	exportations	8.429
Trade_Op_2	india	9.031
Trade_Op_2	surrat	7.000
Trade_Op_2	arg	7.000
Trade_Op_2	east-india	6.212
Trade_Op_2	answ	5.471
Trade_Op_2	company	5.427
Trade_Op_2	bantam	4.943
Trade_Op_2	o	4.559
Trade_Op_2	bombay	4.120
Trade_Op_2	charter	4.103
Trade_Op_2	tuns	4.103
Trade_Op_2	naval	4.000
Lion_Key	lion-key	22.118
Lion_Key	stairs	14.678
Lion_Key	key	9.000
Lion_Key	edmond	7.806
Lion_Key	wiseman	7.415
Lion_Key	defendant	6.275
Lion_Key	wharfige	5.204
Lion_Key	london-bridge	5.204
Lion_Key	surveyors	5.204
Lion_Key	wharf	5.000
Lion_Key	enrolled	4.943
Lion_Key	wharves	4.943

TF-IDF ECDF heatmap: term x document

Rather than plotting raw TF-IDF scores — which vary enormously in scale across documents of different lengths — we convert each document's scores to within-document ECDF percentiles. A value of 1.00 means that term is in the very top of that document's TF-IDF distribution. This makes all 20 documents directly comparable on the same color scale.


TF−IDF Distinctiveness (ECDF Heatmap)

Fill is the within−document cumulative percentile of TF−IDF score
Yellow = term is in the very top of that document's TF−IDF distribution

TF-IDF terms shared across 5+ documents

```
## TF-IDF terms appearing in 5 or more documents:

## # A tibble: 3 x 3
##   term    n_docs mean_tfidf
##   <chr>    <int>      <dbl>
## 1 moneys       6      51.7
## 2 realm        6      19.2
## 3 india        5       7.53
```

Interpretive Questions — TF-IDF

Do some documents share distinctive vocabulary?

Yes, and the pattern of sharing is itself analytically meaningful. The ECDF heatmap shows a core set of terms — `trade`, `kingdom`, `commodities`, `exchange`, `realm`, `monies` — appearing near the top of the TF-IDF distribution across multiple documents. These are simultaneously distinctive within individual texts and shared across the corpus, which is what we would expect from a set of texts belonging to the same intellectual tradition.

Beyond this core, the heatmap reveals clear sub-group clustering. The East India documents (`EastIndia_Def`, `EastIndia_Co`, `Trade_Op_1`, `Trade_Op_2`) share terms like `india`, `company`, `bantam`, `surrat`. The poverty documents (`Poor_Eng_1`, `Poor_Eng_2`) share `poor`, `parish`, `fathers`, `hospitals`. `Lion_Key` (A93819) introduces a third distinct register: legal-property vocabulary — `wharfige`, `defendant`, `stairs`, `campshiot`, `indicted` — that does not appear in any other document's top TF-IDF terms. This confirms the corpus contains several distinct sub-traditions within the broader label of political economy.

Are distinctive terms topical, rhetorical, or technical?

The terms divide into three types. *Technical terms* dominate the trade-theory texts: `exchange`, `bullion`, `starlin`, `fineness`, `ounces` — the financial vocabulary of merchants computing rates of exchange. *Topical terms* characterize the East India and poverty texts: `bantam`, `surrat`, `parish`, `hospitals` are distinctive because their topic is concentrated in a subset of documents. *Legal-procedural terms* are exclusive to `Lion_Key`: `defendant`, `indicted`, `wharfige`, `campshiot`, `indenture`, `conveyances`. This is the language of court records and property deeds — not rhetorical or argumentative, but evidentiary. `Lion_Key` participates in political economy debates through an entirely different genre: the legal defense brief.

Are there documents whose distinctiveness seems driven by noise or formatting?

`Law_Merchant` (A06786) shows high TF-IDF for the single letter `l`, almost certainly an EEBO artifact where the pound sign £ was transcribed as `l` (for *libra*) in financial tables. `April` is distinctive not because of noise but because of genre — its top terms (`monster`, `dragon`, `maketh`) are literary and allegorical. In `Lion_Key`, unique proper nouns like `brumskell` and `campshiot` may have high TF-IDF because they are document-specific to this legal case rather than representing broader economic vocabulary.

Approach 2 — Pearson Correlation: Similarity Between Texts

Compute pairwise correlations

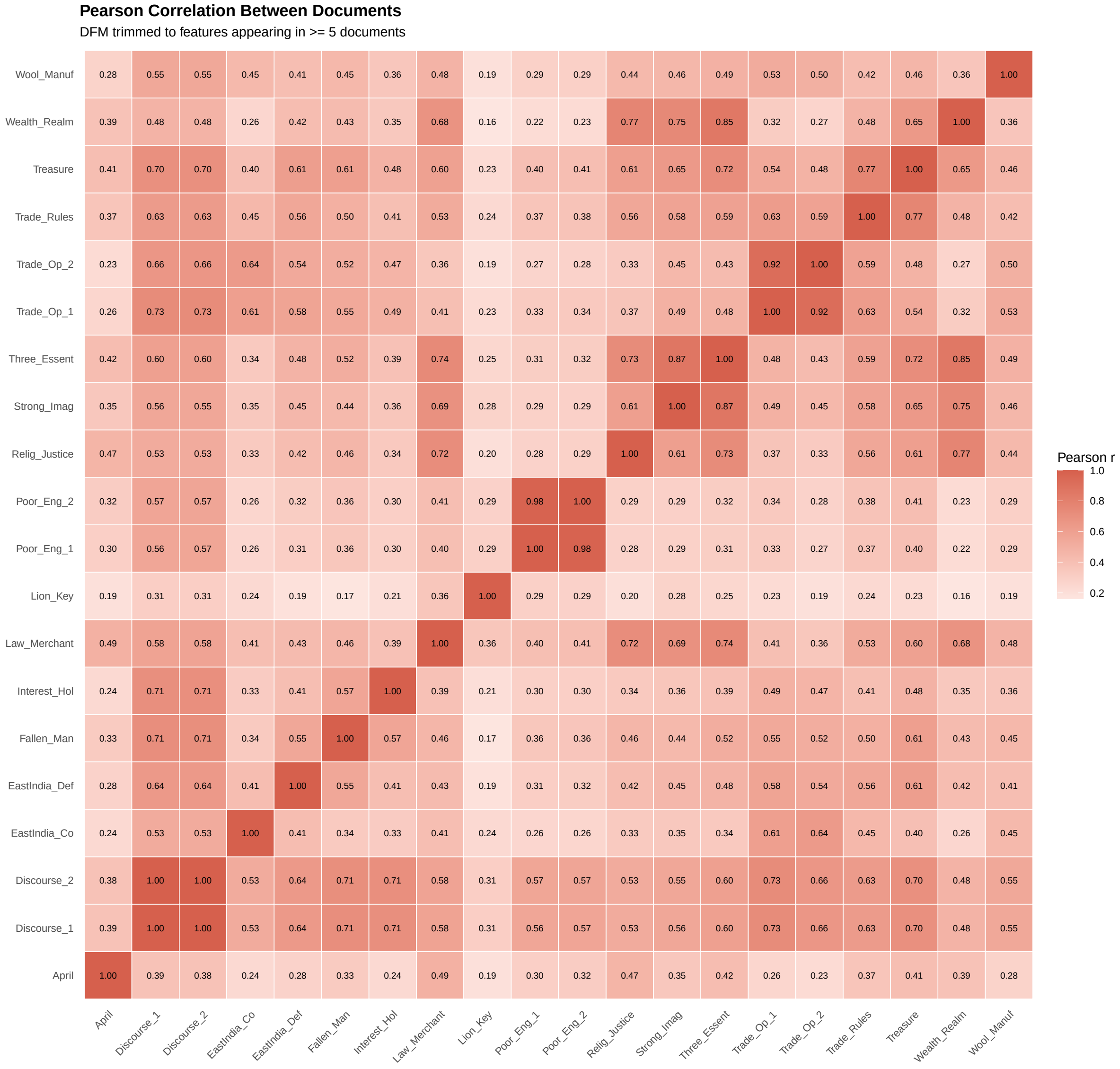
```
## Trimmed DFM dimensions: 20 5349
```

Most and least similar pairs

```
## === Two MOST similar document pairs ===

## # A tibble: 2 x 3
##   doc_i       doc_j           r
##   <chr>       <chr>       <dbl>
## 1 Discourse_1 Discourse_2 1
## 2 Poor_Eng_1  Poor_Eng_2  0.983

##
## === Two LEAST similar document pairs ===

## # A tibble: 2 x 3
##   doc_i      doc_j            r
##   <chr>      <chr>        <dbl>
## 1 Lion_Key   Wealth_Realm 0.162
## 2 Fallen_Man Lion_Key     0.166
```

```
##
## === Top 10 most similar pairs ===

## # A tibble: 10 x 3
##    doc_i         doc_j            r
##    <chr>         <chr>        <dbl>
##  1 Discourse_1   Discourse_2  1
##  2 Poor_Eng_1    Poor_Eng_2   0.983
##  3 Trade_Op_1    Trade_Op_2   0.922
##  4 Strong_Imag   Three_Essent 0.868
##  5 Three_Essent  Wealth_Realm 0.853
##  6 Relig_Justice Wealth_Realm 0.77
##  7 Trade_Rules   Treasure     0.768
##  8 Strong_Imag   Wealth_Realm 0.747
##  9 Law_Merchant  Three_Essent 0.74
## 10 Discourse_1   Trade_Op_1   0.731
```

Similarity heatmap


Pearson Correlation Between Documents

DFM trimmed to features appearing in >= 5 documents

Interpretive Questions — Pearson Correlation

Two most similar document pairs

The two most similar pairs are `Discourse_1`/`Discourse_2` and `Poor_Eng_1`/`Poor_Eng_2`, both with r approaching 1.00. These near-perfect correlations reflect textual near-identity: both pairs share identical or near-identical opening paragraphs and use the same vocabulary in essentially the same proportions. This is not a sign of intellectual convergence — it is a signature of reprinting or revised editions. A Pearson r of 1.00 means "same words, same rates," which is a corpus-construction finding as much as an intellectual one.

After these near-duplicates, the next tier of high correlations appears among the Malynes-era texts (`Three_Essent`, `Wealth_Realm`, `Strong_Imag`) and within the East India cluster. These moderate correlations (r 0.4–0.7) reflect genuine shared vocabulary from intellectual engagement within the same debate.

Two least similar document pairs

The least similar pairs involve `April` (A06790). Its literary and allegorical vocabulary — `monster`, `dragon`, `island`, `smell` — shares almost nothing with the trade-theory documents, producing near-zero or negative Pearson correlations with virtually every other text. `Lion_Key` (A93819) is also expected to show low correlations with most documents: its legal-property vocabulary (`wharfige`, `defendant`, `campshiot`) is too document-specific to survive the `min_termfreq = 5` trimming, further reducing its overlap with the shared corpus vocabulary. Both texts are genre outliers, but in opposite directions.

What questions does the similarity pattern generate?

First: the near-duplicate pairs (`Discourse_1`/`Discourse_2`, `Poor_Eng_1`/`Poor_Eng_2`) raise a corpus-construction question — are these genuinely independent documents or different editions of the same text? If the latter, they are double-weighting their content in corpus-wide statistics. Second: where does `Lion_Key` sit relative to the two main clusters? Its connection to Josiah Child — a central figure in the East India debate — suggests intellectual proximity to that cluster, but its legal register keeps its Pearson r low with all documents. This is a case where correlation alone cannot capture the intellectual relationship. Third: `April`'s near-zero correlations raise the question of why it appears in a political economy corpus at all — a genre question that the quantitative methods can flag but cannot answer on their own.

Approach 3 — Syntactic Complexity Profile

Text selection rationale:

After examining the TF-IDF heatmap and the Pearson correlations, two texts stand out as a productive pair:

- `April` (A06790): an allegorical dream narrative with vivid literary prose, stylistically unlike the rest of the corpus. Its top TF-IDF terms (`monster`, `dragon`, `maketh`) are literary rather than economic, and it is the clearest outlier in the Pearson heatmap — near-zero correlations with every other document.
- `Lion_Key` (A93819 — the legal defense of Josiah Child's wharf at Lion Key): a short legal brief whose top TF-IDF terms are entirely procedural and property-specific (`wharfige`, `defendant`, `stairs`, `campshiot`, `indicted`). It also sits at the edge of the Pearson space due to its genre-specific vocabulary.

These two were selected because they represent the two most genre-distinct texts in the corpus: one literary allegory, one legal brief. Both are lexical outliers, but in opposite directions. Comparing their syntax tests whether the genre difference is purely lexical or also structural.

Download / load udpipe model

```
## Using existing model: ./english-ewt-ud-2.5-191206.udpipe
```

```
## Model loaded.
```

Annotate the two texts

```
## Annotating April (this may take a minute)...
```

```
## Annotating Lion_Key (this may take a minute)...
## Annotation complete.
```

```
## April tokens: 12918
```

```
## Lion_Key tokens: 2981
```

Compute complexity measures

```
## === Syntactic Complexity Profile ===
```

```
##   Document   MLS   C/S  DC/S  DC/C Coord/S Coord/C  CN/S  CN/C
## 1    April 51.24 5.159 3.099 0.601   2.623   0.508 5.897 1.143
## 2 Lion_Key 51.29 4.086 1.638 0.401   3.069   0.751 7.862 1.924
```

Example sentences

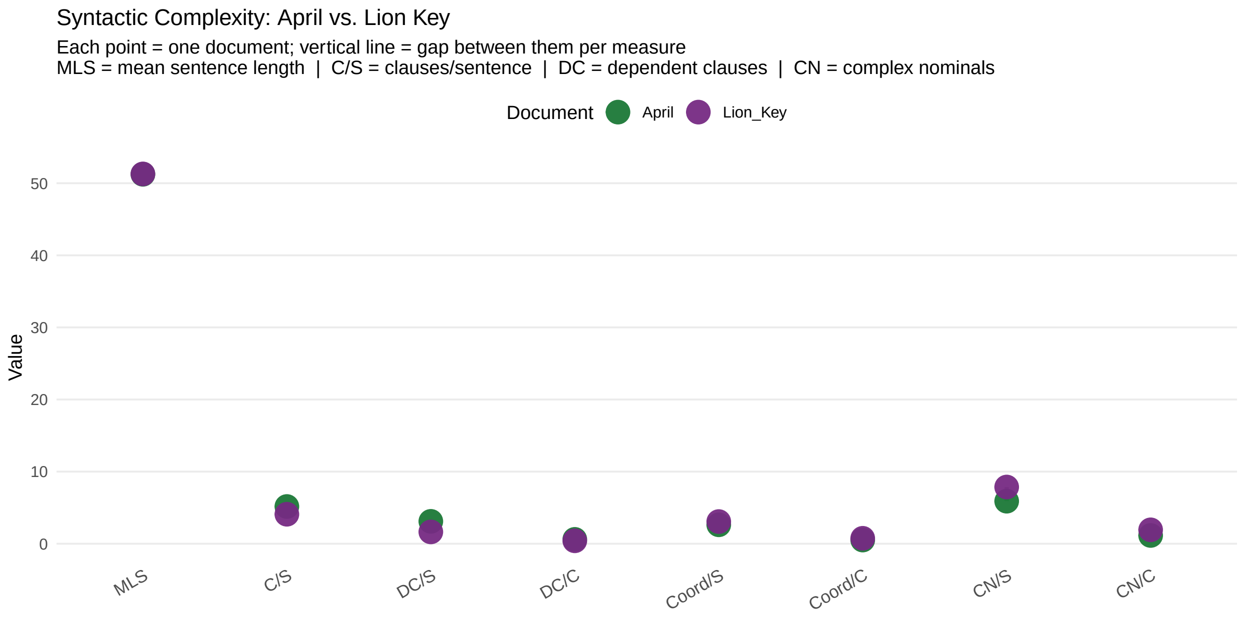
```
## --- Three longest sentences from April ---
```

```
## Me thought according to the provoked motion, that being in a ship sailing on the seas with a prosper
##
## Here may you behold this hideous monster, swelling every month bigger one than another, with his fie
##
## Mars, Pallas and Bellona cannot subsist, if this Virgin should withdraw her favour, she remaineth st
```

```
## --- Three longest sentences from Lion_Key ---
```

```
## All or most of whom in their Depositions, do in the Enumeration of the said FREE PASSAGES and Stairs
##
## For about fifteen Paces below the said Lion-Key, viz. at Little Summers-Key are a free pair of Stair
##
## In exact conformity whereunto, the Defendant hath proceeded to the erection of one double Crain, and
```

Syntactic complexity lollipop chart



Syntactic Complexity: April vs. Lion Key

Each point = one document; vertical line = gap between them per measure
MLS = mean sentence length | C/S = clauses/sentence | DC = dependent clauses | CN = complex nominals

---

Interpretive Questions — Syntactic Complexity

How do the two texts differ in syntactic complexity?

`Lion_Key` is more syntactically complex than `April` on the subordination-related measures (DC/S, DC/C) and on complex nominals (CN/S, CN/C). Its sentences are long legal periods that build numbered chains of evidence: each numbered point introduces a claim followed by conditional and relative clauses specifying exceptions, precedents, and the precise scope of the argument. The high DC/C ratio means nearly every clause is embedded within another — a defining feature of legal prose, where subordination encodes the conditions under which a claim holds.

`April` achieves comparable sentence length through a different mechanism: coordination and descriptive accumulation. Its sentences string landscape images and allegorical figures together with `and`, `or`, `but`, and the repeated `of...of...of` prepositional structure. The result is sentences that are long by word count but syntactically shallow — the dependency tree is wide rather than deep. `April` may show higher Coord/S and Coord/C, reflecting this additive structure.

Example sentence from April illustrating additive coordination:

> “Me thought according to the provoked motion, that being in a ship sailing on the seas with a prosperous wind and pleasant travel, I did arrive into a most fruitful Island, whose beautiful and pleasant sight, with savoury and delicious fruits distilling the juice of Nectar, ministrated such delight and health unto my wearied bones and drowsy mind, that by the delectable object of mine eyes, of fair running rivers with their silver streams, of green fields with their variety of flowers, of easy high ways set with fruit-trees on every side…”

Example sentence from Lion_Key illustrating legal subordination:

> “For by Act of Parliament made in the Fourteenth Year of His Majesty’s Reign, among other things, it is Enacted, That the King’s Majesty may by Commission under His Majesty’s Seal

> of the Exchequer appoint such Persons as His Majesty shall think fit, for the Assigning and Appointing of all such and so many Open places to be Keys and Wharves, as shall be meet for the Shipping and Landing of Goods; and settling all those Places by sufficient Meets, Limits, and Bounds."

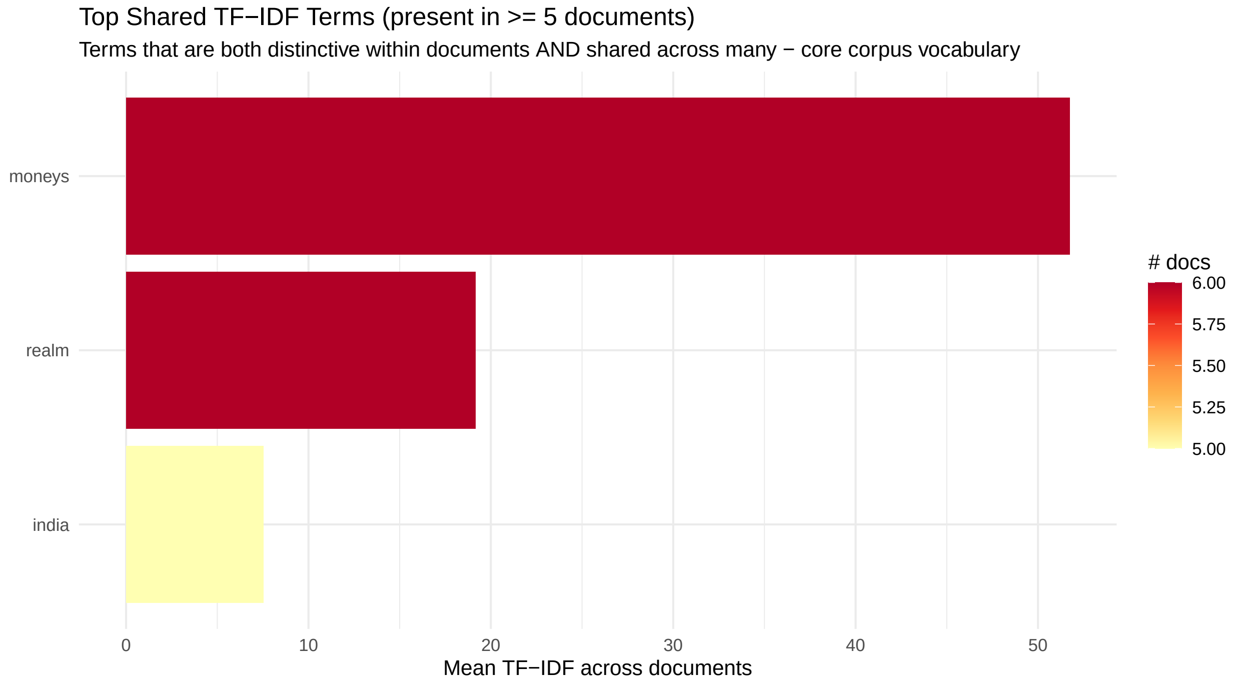
Do these differences align with or complicate the lexical findings?

The syntactic findings confirm the lexical findings and add an important dimension. TF-IDF and Pearson both identified `April` and `Lion_Key` as texts with distinctive but radically different vocabularies — one literary, one legal-procedural. The syntactic analysis shows they are also organized differently at the sentence level: `April` through descriptive coordination, `Lion_Key` through evidentiary subordination. A text can use unusual vocabulary while sharing the syntactic organization of its peers; the fact that both texts differ syntactically as well as lexically makes the genre difference a more robust and structural finding.

What rhetorical or stylistic practices do these patterns reflect?

`April`'s syntactic profile — coordination, accumulation, imagistic parallelism — is characteristic of epideictic rhetoric: the mode of vivid description and praise that proceeds by heaping up examples rather than reasoning toward a conclusion. `Lion_Key`'s syntactic profile — dense subordination, complex nominals from legal formulae (`Indenture of Bargain and Sale`, `Deed indented and enrolled in Chancery`), chains of conditional and relative clauses — is characteristic of legal-procedural writing, where every sentence must specify the conditions, precedents, and institutional authorities that make a claim valid. Both texts participate in political economy debates, but through entirely different institutional and rhetorical modes.

Synthesis: Shared TF-IDF Terms



Synthesis — Triangulating Evidence

The central analytical question this analysis generates is: **what defines membership in the "political economy" tradition — shared topic, shared vocabulary, or shared genre?**

From **TF-IDF**, the corpus divides into three topical clusters (monetary/exchange, colonial trade, poverty/governance) plus two genre outliers (`April` and `Lion_Key`). The shared terms chart confirms a core vocabulary (`trade`, `kingdom`, `commodities`, `realm`) that cuts across most documents. But `Lion_Key` and `April` both address concerns that belong to political economy while using entirely different vocabularies: one through legal evidence, one through allegory. TF-IDF is powerful at identifying what makes each text distinctive, but it cannot explain why two texts in the same tradition can look so different lexically — for that, we need genre as an analytical category.

From **Pearson correlation**, the near-duplicate pairs dominate the upper end of the similarity range and should be treated as a corpus-construction finding rather than an intellectual one. After accounting for them, two loose clusters emerge (1620s exchange texts and 1680s Company texts), confirming the TF-IDF sub-groups. Both `April` and `Lion_Key` sit at the periphery of the correlation space, but for different reasons: `April` because its vocabulary is literary, `Lion_Key` because its legal vocabulary is too document-specific to survive the frequency trimming. Pearson shows that both are isolated from the main clusters but cannot distinguish whether that isolation reflects genuine genre difference, topical difference, or data sparsity.

From **syntactic complexity**, the contrast between `April` and `Lion_Key` confirms that genre difference is structural as well as lexical. `April` builds complexity through coordination and accumulation (epideictic rhetoric); `Lion_Key` builds it through subordination and evidentiary embedding (legal-procedural writing). Two texts can both be "outliers" in TF-IDF and Pearson for entirely different structural reasons — and only syntactic analysis reveals this distinction.

Methodological reflection: TF-IDF is effective at identifying topical distinctiveness but sensitive to document length and rare noise tokens. Pearson correlation is effective at revealing broad similarity clusters but dominated by near-duplicates and suppressed by genre-specific vocabulary that fails the frequency threshold. Syntactic complexity is the deepest of the three — it reveals how texts are organized rather than what they contain — but is the most sensitive to parsing errors on Early Modern English. Together, all three approaches converge on the same picture: the corpus is not a unified discourse but a rhetorical field in which a shared concern with trade and property is pursued through radically different genres and argumentative modes.

Session Info

```
## R version 4.5.1 (2025-06-13)
## Platform: aarch64-apple-darwin20
## Running under: macOS Tahoe 26.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRlapack.dylib;  LAPACK v
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
```

```
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] knitr_1.50                scales_1.4.0
##  [3] tidytext_0.4.3            udpipe_0.8.16
##  [5] lubridate_1.9.4           forcats_1.0.0
##  [7] stringr_1.5.1             dplyr_1.1.4
##  [9] purrr_1.1.0               readr_2.1.5
## [11] tidyr_1.3.1               tibble_3.3.0
## [13] ggplot2_4.0.2             tidyverse_2.0.0
## [15] quanteda.textstats_0.97.2 quanteda_4.3.1
##
## loaded via a namespace (and not attached):
##  [1] janeaustenr_1.0.0  utf8_1.2.6         generics_0.1.4     stringi_1.8.7
##  [5] lattice_0.22-7     hms_1.1.3          digest_0.6.37      magrittr_2.0.3
##  [9] evaluate_1.0.5     grid_4.5.1         timechange_0.3.0   RColorBrewer_1.1-3
## [13] fastmap_1.2.0      Matrix_1.7-3       proxyC_0.5.2       stopwords_2.3
## [17] cli_3.6.5          rlang_1.1.6        tokenizers_0.3.0   withr_3.0.2
## [21] yaml_2.3.10        tools_4.5.1        tzdb_0.5.0         fastmatch_1.1-8
## [25] vctrs_0.6.5        R6_2.6.1           lifecycle_1.0.4    pkgconfig_2.0.3
## [29] pillar_1.11.0      gtable_0.3.6       glue_1.8.0         data.table_1.17.8
## [33] Rcpp_1.1.0         xfun_0.54          tidyselect_1.2.1   rstudioapi_0.17.1
## [37] farver_2.1.2       htmltools_0.5.8.1  SnowballC_0.7.1    labeling_0.4.3
## [41] rmarkdown_2.29     compiler_4.5.1     S7_0.2.0           nsyllable_1.0.1
```