Computational text analysis: Survival guide

ESU 24 @ Cluj-Napoca

Jeremi Ochab, Artjoms Šeļa

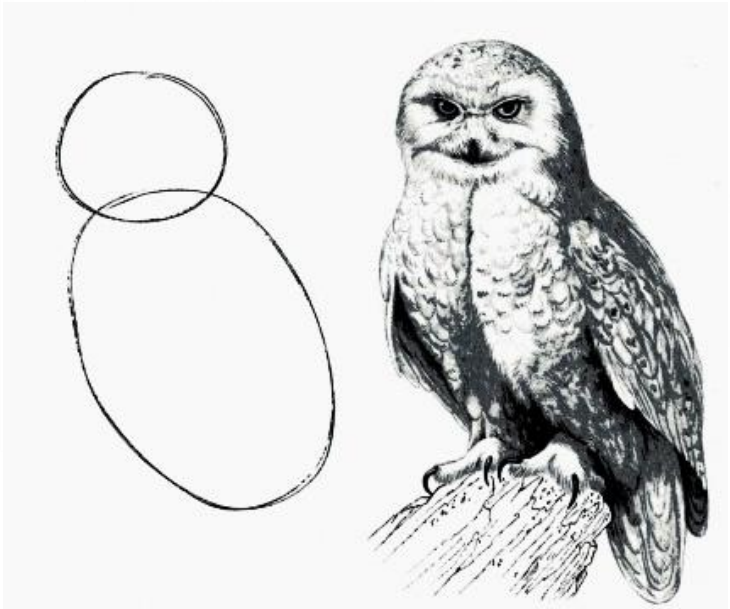


Fig 1. Draw two circles

Fig 2. Draw the rest of the damn Owl

Is there a path between C. Bronte and H. Melville in Manhattan?

TL;DR Multivariate text analysis

1. **Feature space:** count things in multidimensional Manhattan of your design (MFWs, POS, etc…)
2. **Distance measure:** estimate differences between texts (each text == counted things)
3. **Mapping relationships:** trees, projections, networks…

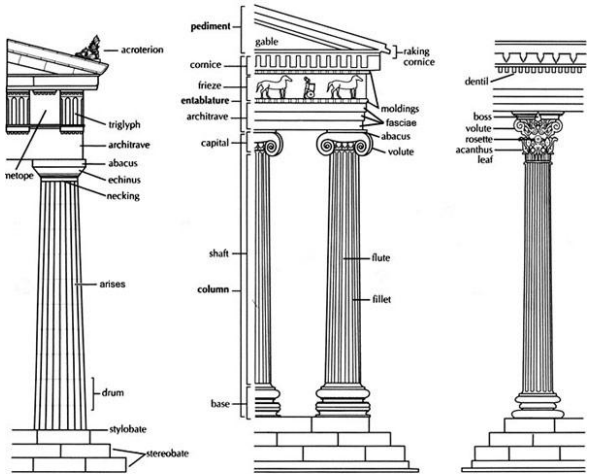
`stylo` package does that in R!

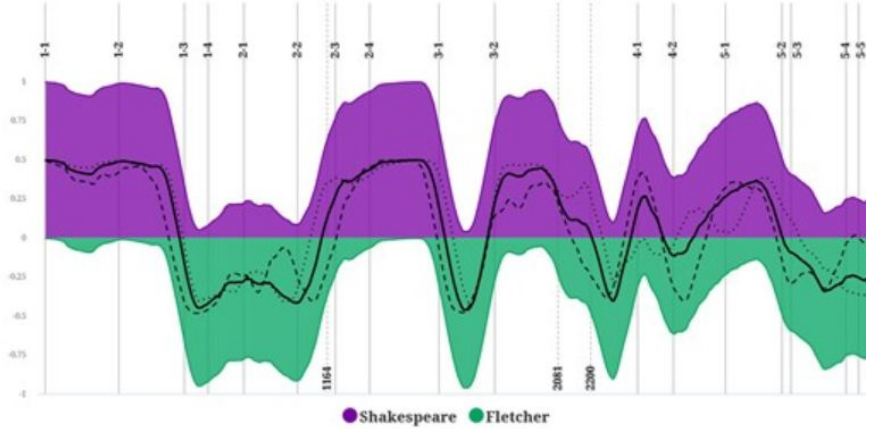
“The Art of Measuring Columns”

- **Stylometry:** A sub-field of computational text analysis that studies **differences** between texts
- Lutosławski 1897: method of “measuring stylistic affinities”

Don’t mix up with another “stylometry”: “the art of measuring columns”!

Stylometrie, *f.*, stylometry, the art of measuring columns (Säulenmeßkunst).





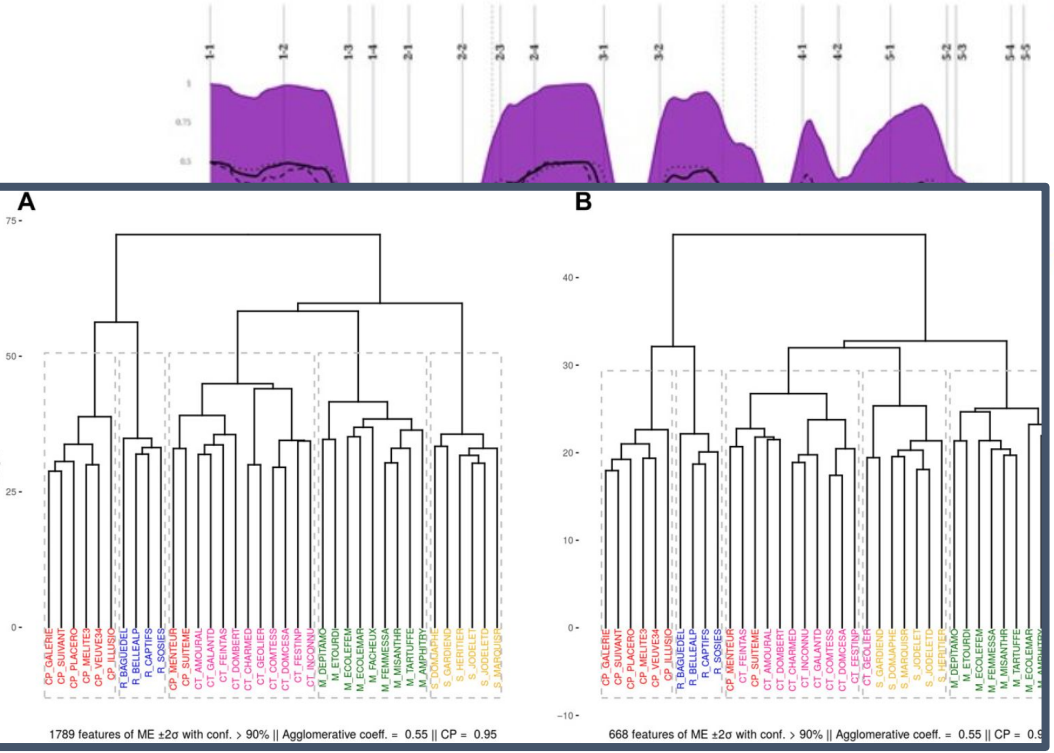
JOURNAL ARTICLE

Relative contributions of Shakespeare and Fletcher in *Henry VIII*: An analysis based on most frequent words and most frequent rhythmic patterns

Petr Plecháč ✉

Digital Scholarship in the Humanities, Volume 36, Issue 2, June 2021, Pages 430–438, https://doi.org/10.1093/llc/fqaa032

Published: 26 June 2020 **Article history** ▾



JOURNAL ARTICLE

Relative contributions of Shakespeare and Fletcher in *Henry VIII*: An analysis based on most frequent words and most frequent rhythmic patterns

Petr Plecháč

Digital Scholarship in the Humanities, Volume 36, Issue 2, June 2021, Pages 430–438, https://doi.org/10.1093/llc/fqaa032

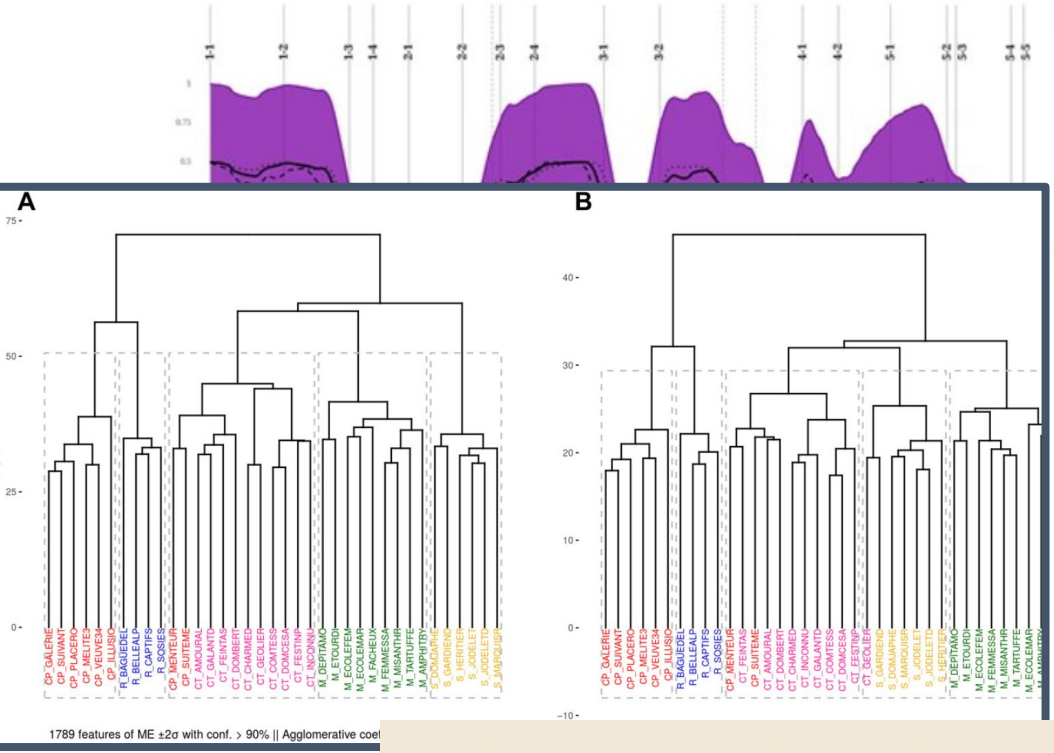
Published: 26 June 2020 **Article history** ▾

RESEARCH ARTICLE | SOCIAL SCIENCES

Why Molière most likely did write his plays

FLORIAN CAFIERO AND JEAN-BAPTISTE CAMPS Authors Info & Affiliations

SCIENCE ADVANCES · 27 Nov 2019 · Vol 5, Issue 11 · DOI: 10.1126/sciadv.aax5489



JOURNAL ARTICLE

Relative contributions of Shakespeare and Fletcher in *Henry VIII*: An analysis based on most frequent words and most frequent rhythmic patterns

Petr Plecháč

Digital Scholarship in the Humanities, Volume 36, Issue 2, June 2021, Pages 430–438, https://doi.org/10.1093/llc/fqaa032

Published: 26 June 2020 **Article history ▾**

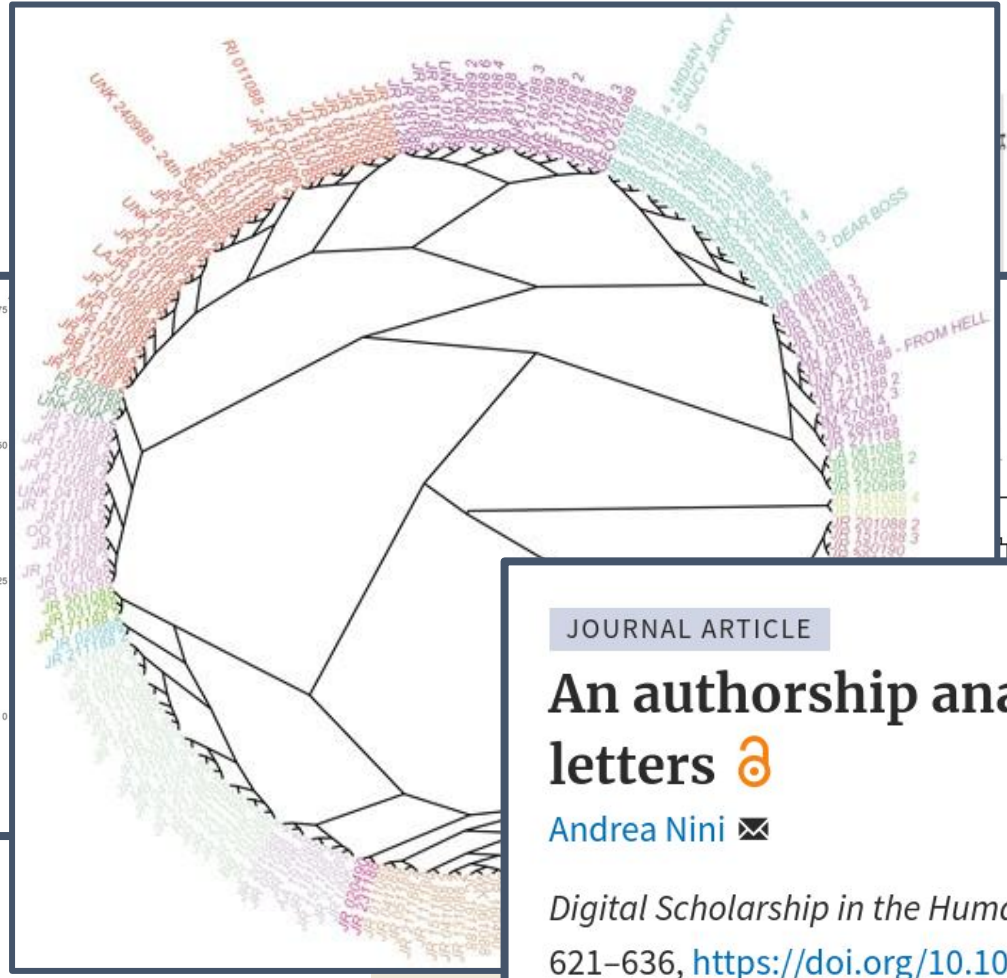
RESEARCH ARTICLE | SOCIAL SCIENCES

Why Molière most likely did write his plays

FLORIAN CAFIERO AND JEAN-BAPTISTE CAMPS Authors Info & Affiliations

SCIENCE ADVANCES · 27 Nov 2019 · Vol 5, Issue 11 · DOI: 10.1126/sciadv.aax5489

Discovering a lost play by Lope de Vega: Álvaro Cuéllar



JOURNAL ARTICLE

Relative contributions of Shakespeare and Fletcher in *Henry VIII*: An analysis based on most frequent words and most frequent rhythmic patterns

Petr Plecháč ✉

Digital Scholarship in the Humanities, Volume 36, Issue 2, June 2021, Pages 430–438, https://doi.org/10.1093/llc/fqaa032

Published: 26 June 2020 **Article history** ▾

RESEARCH ARTICLE | SOCIAL SCIENCES

JOURNAL ARTICLE

An authorship analysis of the Jack the Ripper letters

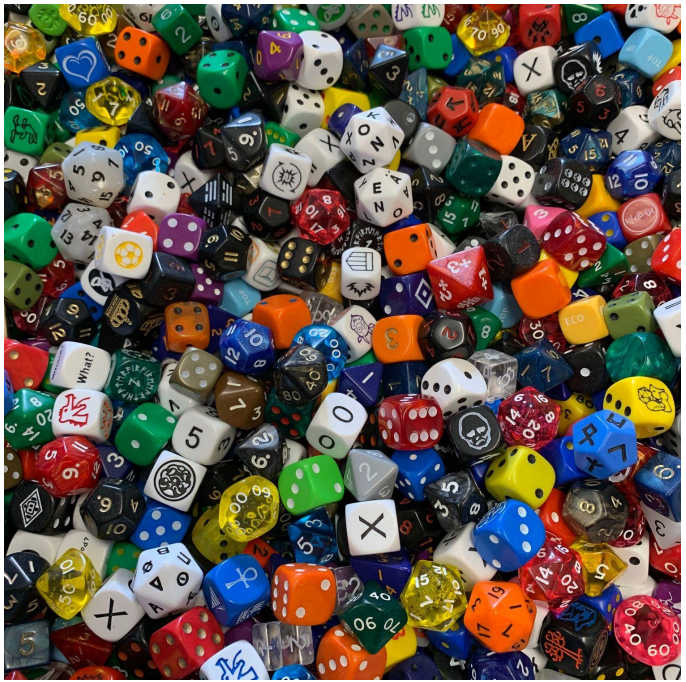
Andrea Nini ✉

Digital Scholarship in the Humanities, Volume 33, Issue 3, September 2018, Pages 621–636, https://doi.org/10.1093/llc/fqx065

Published: 25 January 2018

Proxies

- Word frequencies may serve as a proxy to **things we care about** in texts
- Word frequencies are the result of word choice -> word choice is a result of **forces that organize texts** (intention, cultural and social conditions)

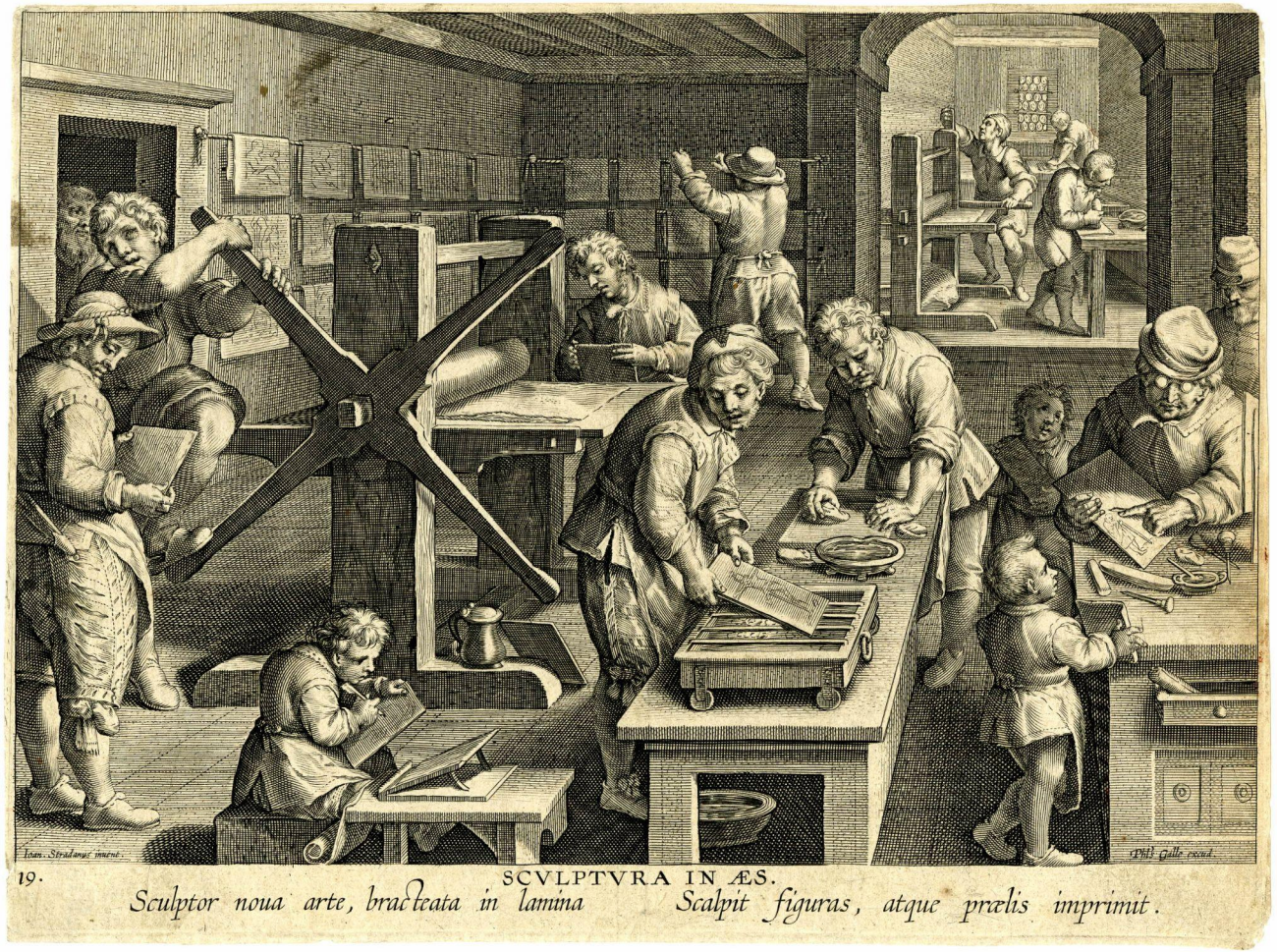


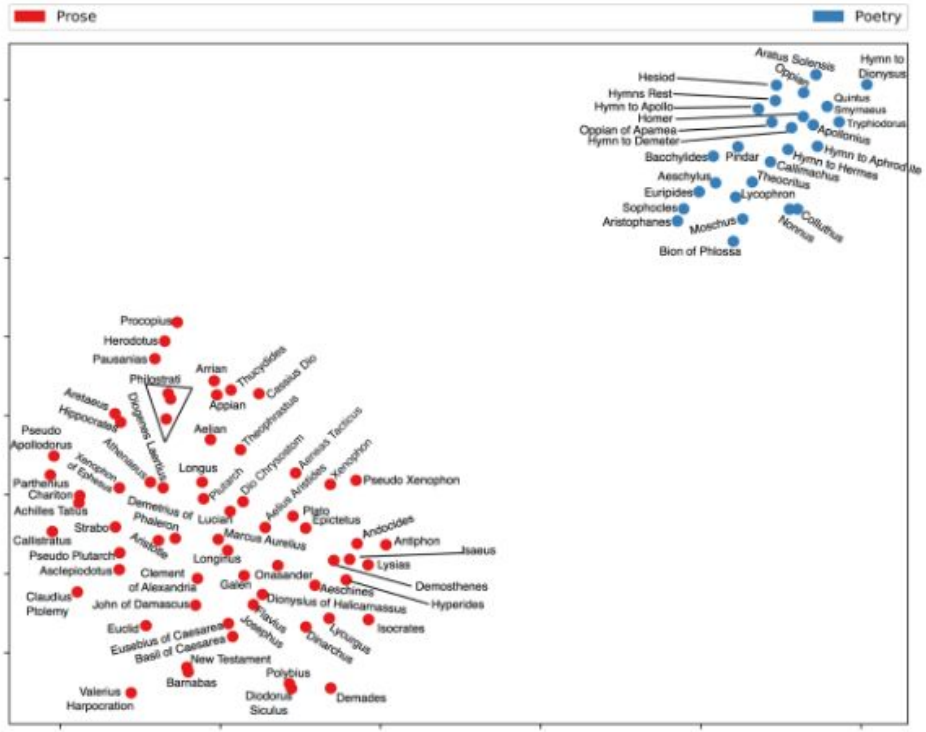
A model of text

- Differences between texts can be expressed in a multitude of ways
- Central question is **how to represent** a text so it could be placed on a quantitative scale? i.e. how to **model** it?
- Short answer: all representations are ‘wrong’, but some are useful (or more useful than others)

Can we tell apart...

prose from poetry?

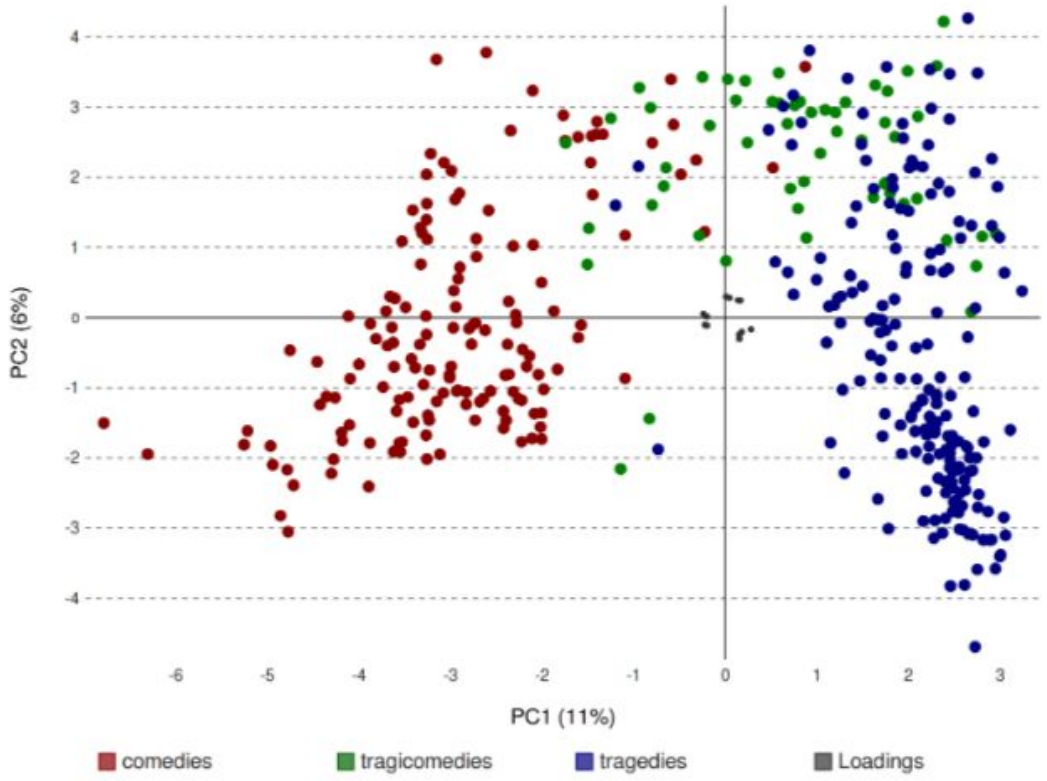




Storey & Mimno 2020: Like Two Pis in a Pod: Author Similarity Across Time in the Ancient Greek Corpus

Can we tell apart...

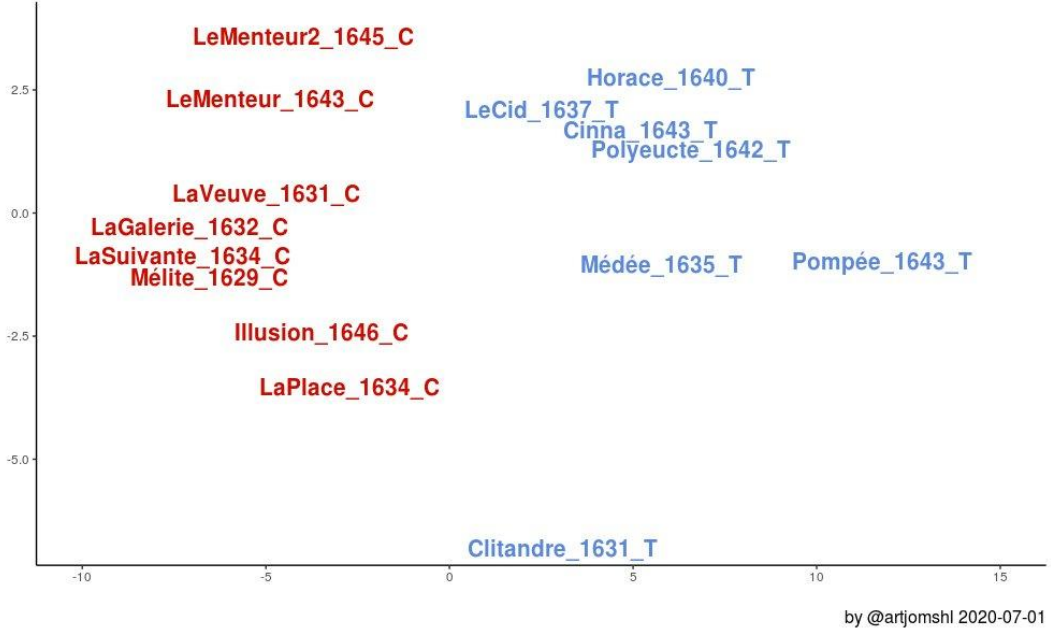
comedy from tragedy?



Schöch, C. 2017. Topic Modeling Genre…

Comedies and **Tragedies** of Pierre Corneille

Data come from large-scale quantiative study on distinistive features of classic dramatic genres in Corneille **done by Boris I. Yarkho in 1920s**. Each text was represented across 15 features that Yarkho tried to synthesise into clear 'comedy' vs. 'tragedy' cut. This study served as a general demonstration of Yarkho's grand project of quantiative methodology for literary studies. 120 pages long work was first published only in 2006.

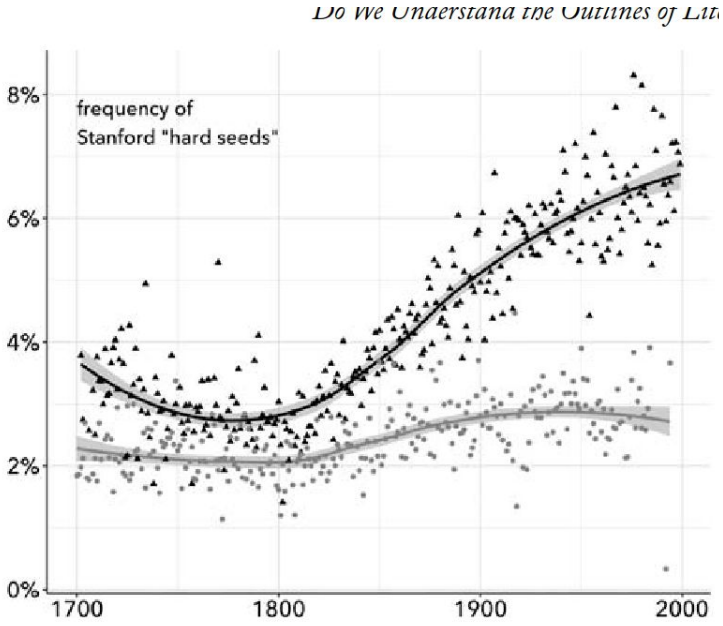


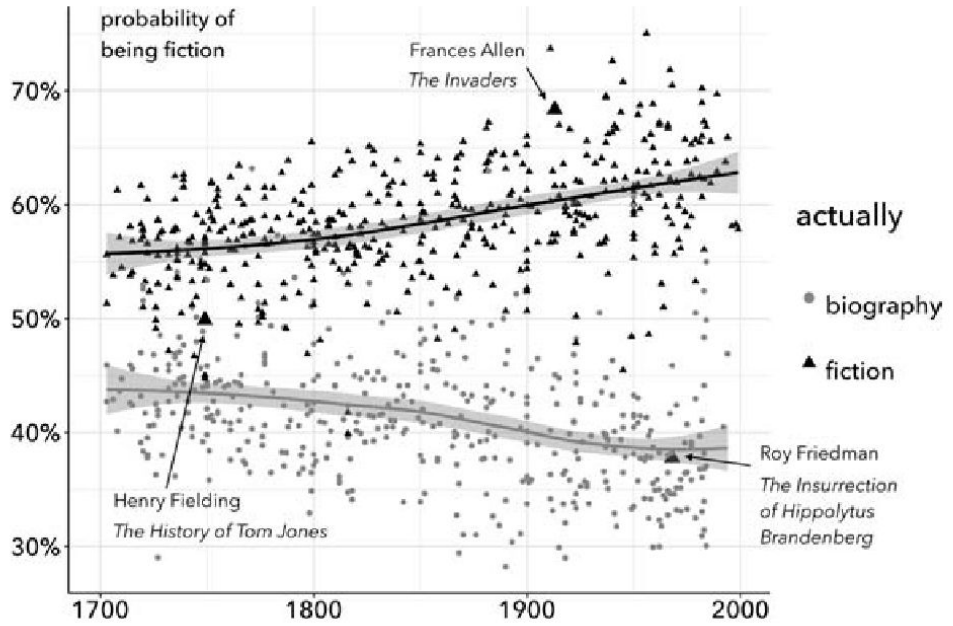
by @artjomshl 2020-07-01

Boris Yarkho (1889-1942)

Can we tell apart...

Fiction from non-fiction?

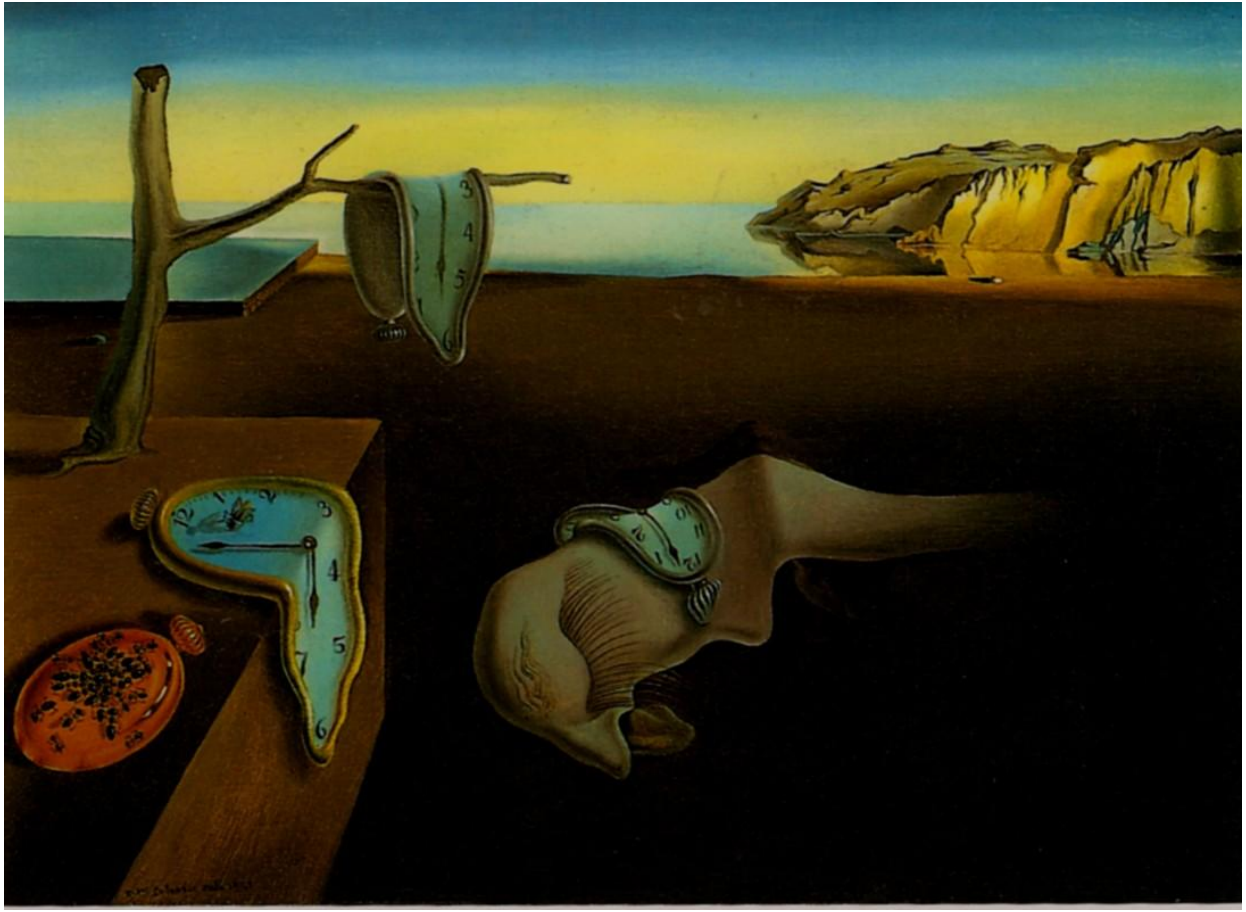


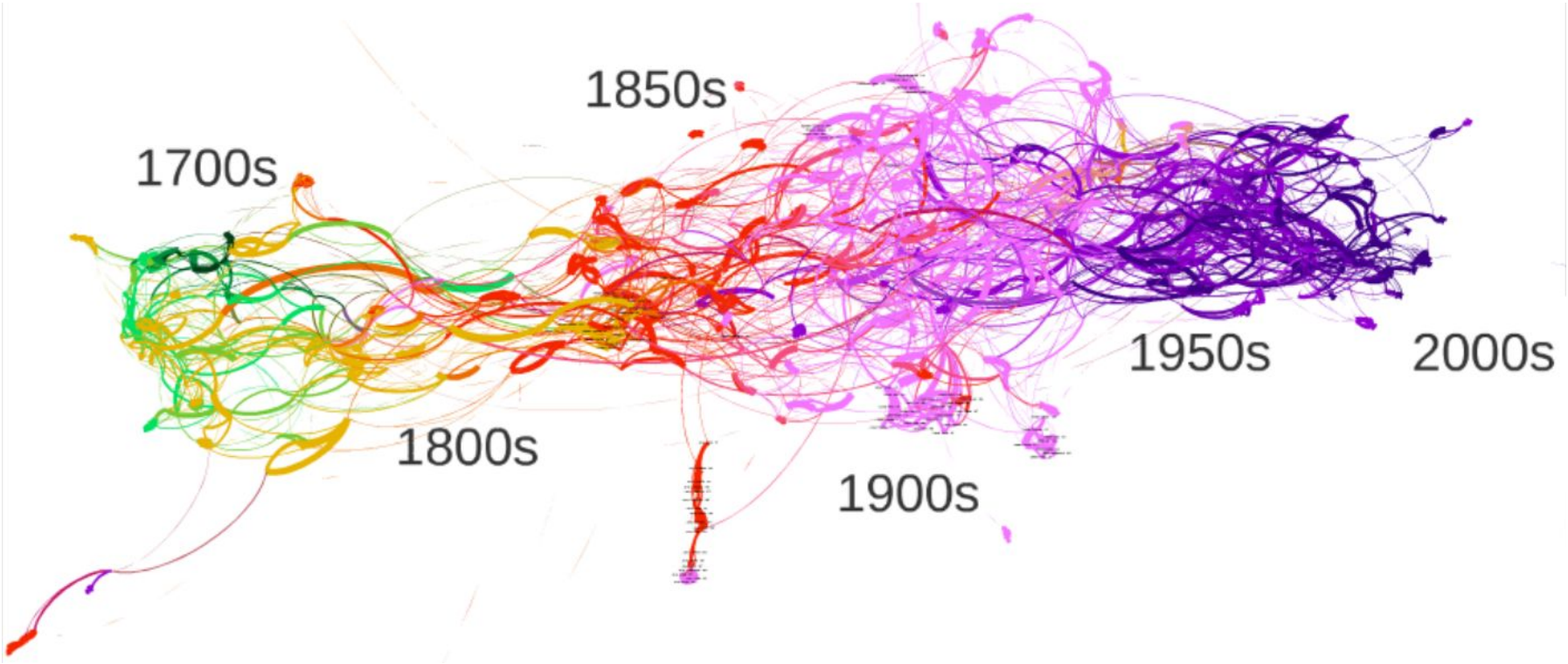


Underwood T. 2019 *Distant Horizons*

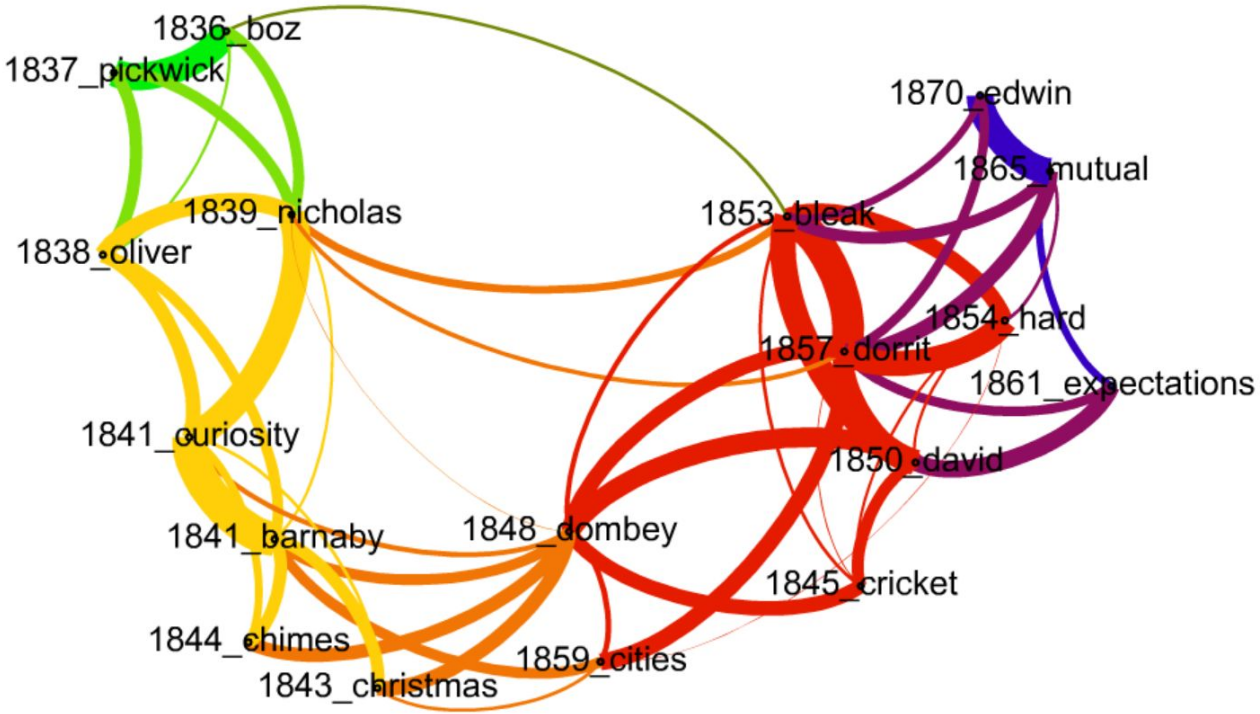
Can we tell apart...

a 19th c. text from 18th c. text?



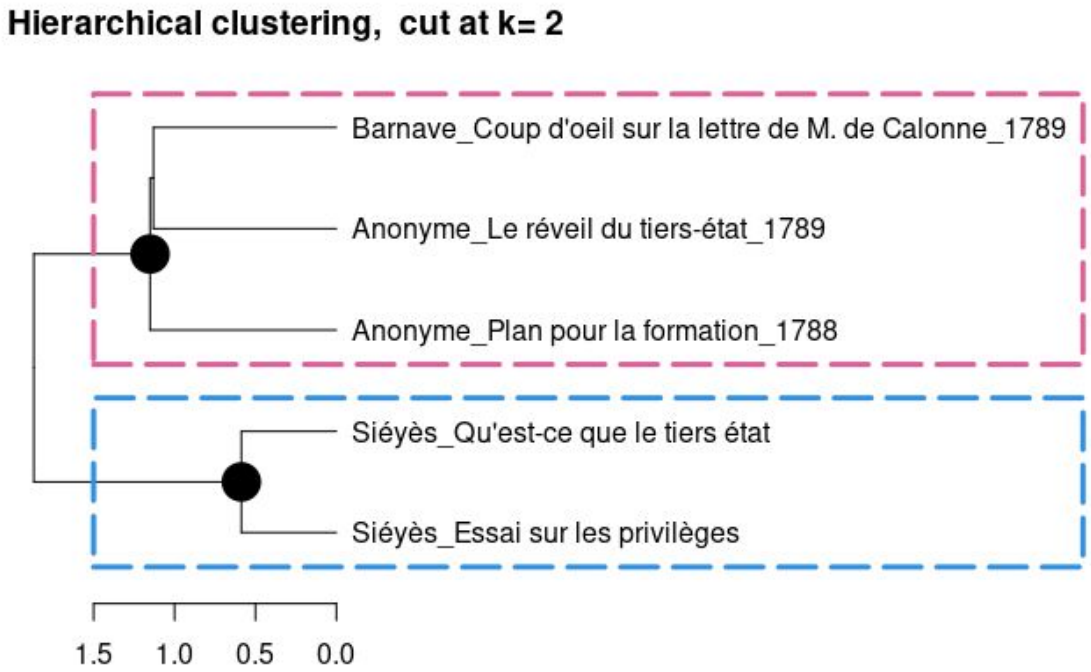


Rybicky 2016



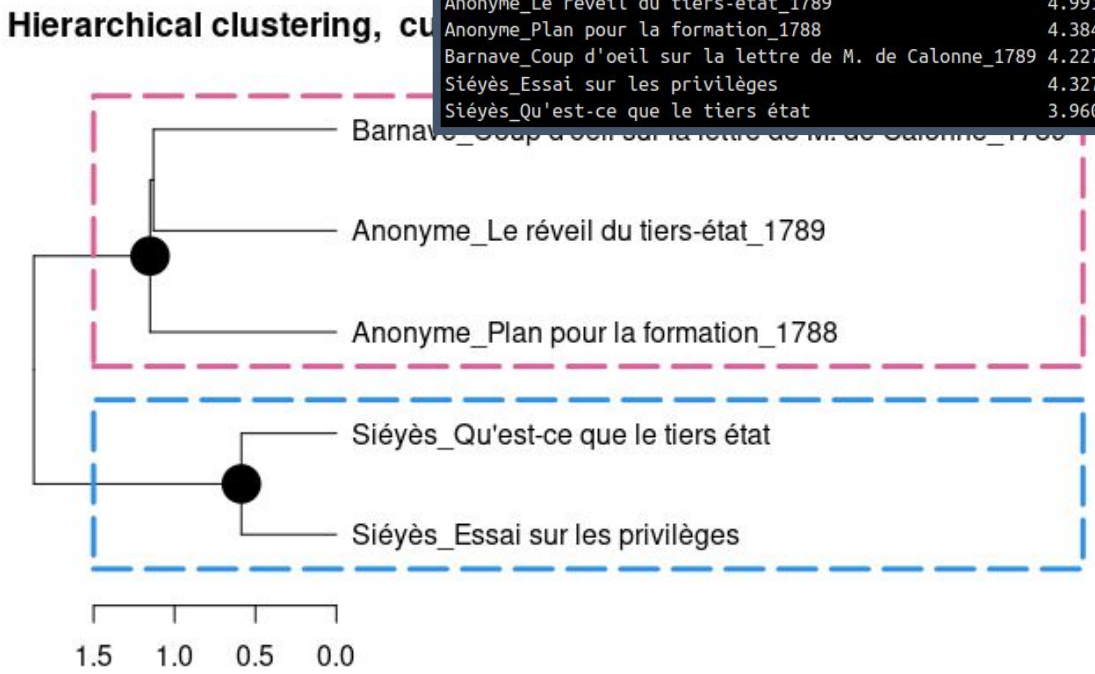
Rybicky 2016

Cutting trees for explainable results!



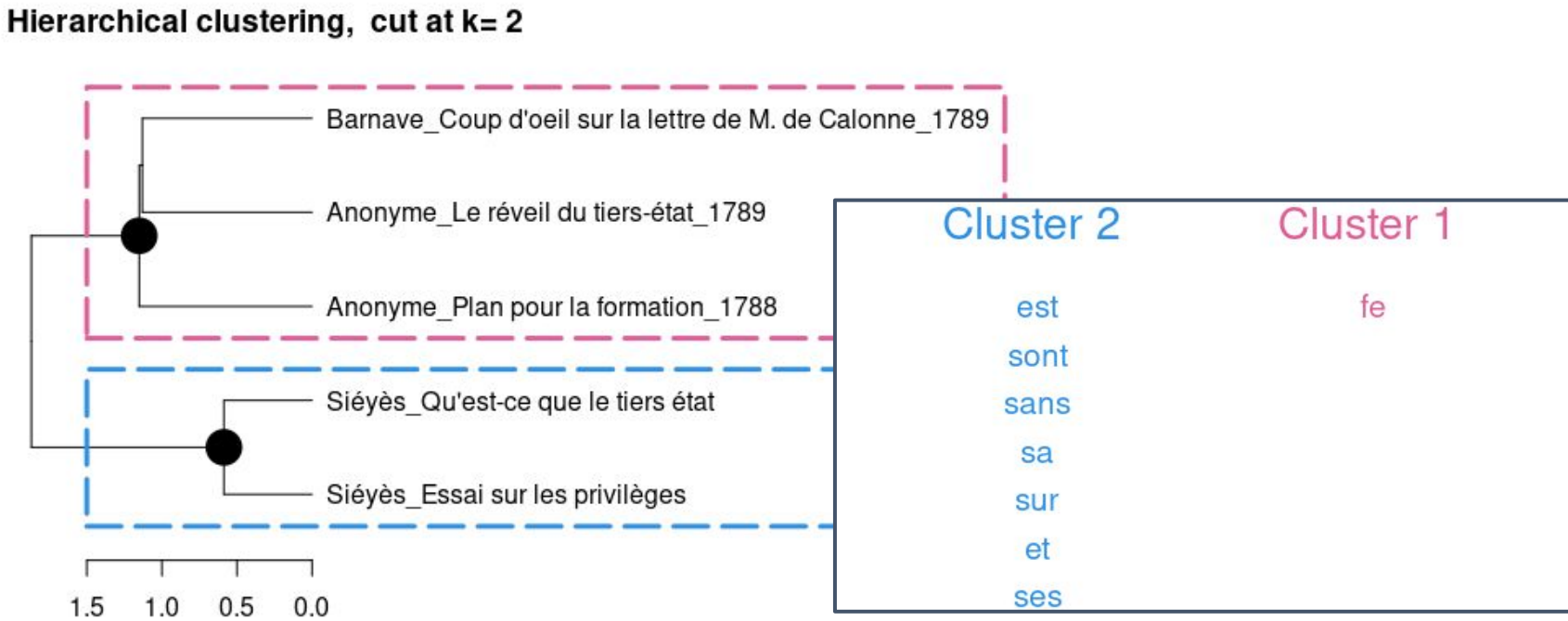
Cutting trees for explainable results!

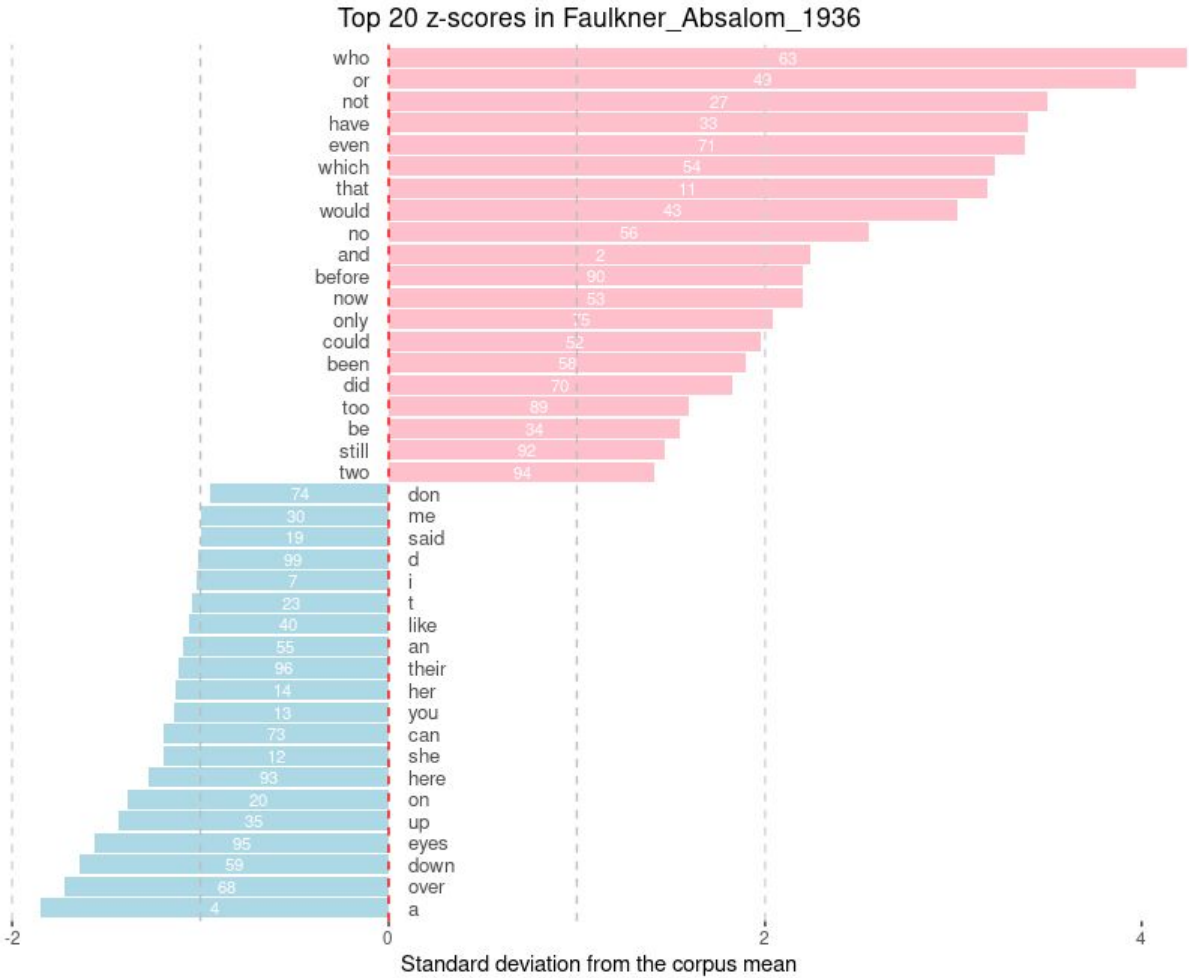
	de	la	les	l	à	le
Anonyme_Le réveil du tiers-état_1789	4.9919255	3.1490453	2.9068111	2.0993635	1.8333808	1.9663722
Anonyme_Plan pour la formation_1788	4.3847242	3.1117397	4.6676096	1.6030174	1.3672796	1.7444602
Barnave_Coup d'oeil sur la lettre de M. de Calonne_1789	4.2275472	3.0196766	2.8443405	2.2209234	1.1494253	2.1235145
Siéyès_Essai sur les privilèges	4.3275072	2.729227	2.4114403	2.0095336	2.3460136	1.8412936
Siéyès_Qu'est-ce que le tiers état	3.9608393	2.9340656	2.3968003	1.874459	2.0416082	1.8983375



	de	la	les	l	à	le
Anonyme_Le réveil du tiers-état_1789	4.9919255	3.1490453	2.9068111	2.0993635	1.8333808	1.9663722
Anonyme_Plan pour la formation_1788	4.3847242	3.1117397	4.6676096	1.6030174	1.3672796	1.7444602
Barnave_Coup d'oeil sur la lettre de M. de Calonne_1789	4.2275472	3.0196766	2.8443405	2.2209234	1.1494253	2.1235145
Siéyès_Essai sur les privilèges	4.3275072	2.729227	2.4114403	2.0095336	2.3460136	1.8412936
Siéyès_Qu'est-ce que le tiers état	3.9608393	2.9340656	2.3968003	1.874459	2.0416082	1.8983375

s!




Top 20 z-scores in Faulkner_Absalom_1936
who 63
or 43
not 27
have 33
even 71
which 54
that 11
would 43
no 56
and 2
before 90
now 53
only 75
could 52
been 58
did 70
too 89
be 34
still 92
two 94
74 don
30 me
19 said
99 d
7 i
23 t
40 like
55 an
96 their
14 her
13 you
73 can
12 she
93 here
20 on
35 up
95 eyes
59 down
68 over
4 a
-2
0
2
4
Standard deviation from the corpus mean

Comparing ‘profiles’