

Text Analysis and Visualisation: An Overview of Tools



Welcome //

The DiRT Directory is a registry of digital research tools for scholarly use. DiRT makes it easy for digital humanists and others conducting digital research to find and compare resources ranging from content management systems to music OCR, statistical analysis packages to mindmapping software.

I NEED A DIGITAL RESEARCH TOOL TO . . .

Analyze data

Analyze texts

Author an interactive work

Manage bibliographic information

Manage tasks

Network with other researchers

ABOUT

The DiRT Directory is a registry of digital research tools for scholarly use. [\(more\)](#)

NEWS

DiRT at DH 2014

17 Jul 2014

Goodbye, Project Bamboo!

3 Jul 2014

DiRT at DHSI, UC Berkeley's One IT Summit

1 Jul 2014

[more](#)

TAPoR

Discover Research Tools for Textual Study

- Browse Tools by Type or Tag
- Search and Use Tools
- Read and Create Tool Reviews
- Contribute and Advertise Tools

The screenshot shows the CATMA Tagger interface. At the top, there's a tab labeled "A Rose For Emily x". Below it, a text document is displayed with several lines of text. Each line has horizontal bars of various colors (red, yellow, blue, green) underneath it, representing annotations. The text includes phrases like "When Miss Emily Grierson died, our whole town went to her funeral: the men", "through a sort of respectful affection for a fallen monument, the women mostly", "out of curiosity to see the inside of her house, which no one save an old", "man-servant—a combined gardener and cook—had seen in at least ten years.", "It was a big, squarish frame house that had onjce been white, decorated with", "cupolas and spires and scrolled balconies in the heavily lightsome style of the", "seventies, set on what had once been our most select street. But garages and", "cotton gins had encroached and obliterated even the august names of that", "neighborhood; only Miss Emily's house was left, lifting its stubborn and". At the bottom of the interface, there are navigation buttons (back, forward, search, etc.) and a "page size zoom" control.

Featured tool:

CATMA (Computer Aided Textual Markup and Analysis)

Site: <http://www.catma.de/>
Author(s): [Department of Languages, Literature and Media, University of Hamburg](#)

CATMA (Computer Aided Textual Markup and Analysis) is a free, open source markup and analysis tool from the University of Hamburg's Department of Languages, Literature and Media. It incorporates three interactive modules, a tagger enabling textual markup and markup editing, an analyzer incorporating a query language and predefined functions, and a query builder that allows users to construct queries from combinations of pre-defined questions while allowing for manual modification for

View tools by:

All (464)	Reviewed (244)
New (10)	Popular (380)
Concording (63)	Editing (21)
Miscellaneous (247)	Programming language (17)
Search (123)	Statistical (137)
Text cleaning (16)	Text gathering
Visualization (148)	

View tools by tag:

1960s 1970s 1980s 1990s 2000s
2010s American Canadian
Comparator Dutch English English



an initiative of
centernet

[About](#)

[Journal](#)

[Projects](#)

[Collaborators](#)

[Contribute](#)

[Blog](#)

[Help](#)

[Log in](#)

Welcome

DHCommons is a hub for people and organizations to find projects to work with, and for projects to find collaborators. [Log in](#) or [sign up](#) to get started.

Tweet us! @DHCommons.

Voyant Tools

Description

Screenshots

Contributors

Status: [Active](#)

Voyant Tools is a web-based reading and analysis environment for digital texts.

Webpage: <http://voyant-tools.org/>

Platform: Web-based

Developer: hermeneuti.ca [Stéfan Sinclair and Geoffrey Rockwell]

Cost: Free

Code license: [Open source](#)

Last modified: 7/14/14

TaDiRAH goals & methods: [Analysis](#) [Content Analysis](#)

TaDiRAH research objects: [Text](#)

Categories: [Text collections](#) [Text mining](#) [Visualization](#)

Tags: [text analysis](#) [corpus](#) [word cloud](#)

Text Analysis

	Distribution	Types and Tokens	Collocation	Co-occurrence	POS
Taporware	Green	Green	Green	Green	Green
Voyant	Green	Green	Green	Green	Grey
AntConc	Grey	Grey	Grey	Grey	Grey
JGAAP	Green	Green	Green	Green	Grey
Lexomics	Grey	Green	Grey	Green	Grey
WordCruncher	Grey	Green	Grey	Green	Grey
PhiloLogic	Grey	Green	Green	Green	Grey
SMILE Text Analyzer	Grey	Grey	Grey	Grey	Green
PAIR	Grey	Grey	Grey	Grey	Grey
Collatex	Grey	Grey	Grey	Grey	Grey

	Concordance	Data extraction	Comparison	Word cloud
Taporware	Green	Green	Grey	Green

Import and Data cleaning

	TXT	XML	URL	Remove digits	Remove case	Stop-words
Taporware	Green	Green	Green	Green	Green	Green
Voyant	Green	Green	Green	Green	Green	Green
AntConc	Green	Grey	Grey	Grey	Grey	Grey
JGAAP	Green	Grey	Grey	Green	Green	Green
Lexomics	Green	Grey	Green	Green	Green	Green
WordCruncher	Green	Grey	Grey	Grey	Grey	Green
PhiloLogic	Green	Grey	Grey	Green	Green	Grey
SMILE Text Analyzer	Green	Grey	Grey	Grey	Grey	Grey
PAIR	Green	Grey	Grey	Green	Grey	Grey
Collatex	Green	Grey	Grey	Green	Grey	Grey

Studies based on vocabulary






- Segmentation or tokenisation
- Often based on the fact that there are generally spaces in between words
- Types are the unique words in a document; tokens are the total number of words

He cried in a whisper at some image, at some vision,--he cried out twice, a cry that was no more than a breath-- 'The horror! The horror!'

28 tokens and 21 types

Frequency lists





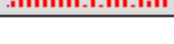
Summary: There are 6053 unique words , and there are 37984 words in total. 3524 words occurred once and 920 words occurred twice.

Words	Distribution	Counts
the		2254
of		1365
a		1123
i		1094
and		930
to		870
was		668
in		602
he		580
had		501
it		463
that		407
with		367
his		334
on		309
you		286
as		285
for		262
at		261

Frequency list produced using [TaporWare](#)

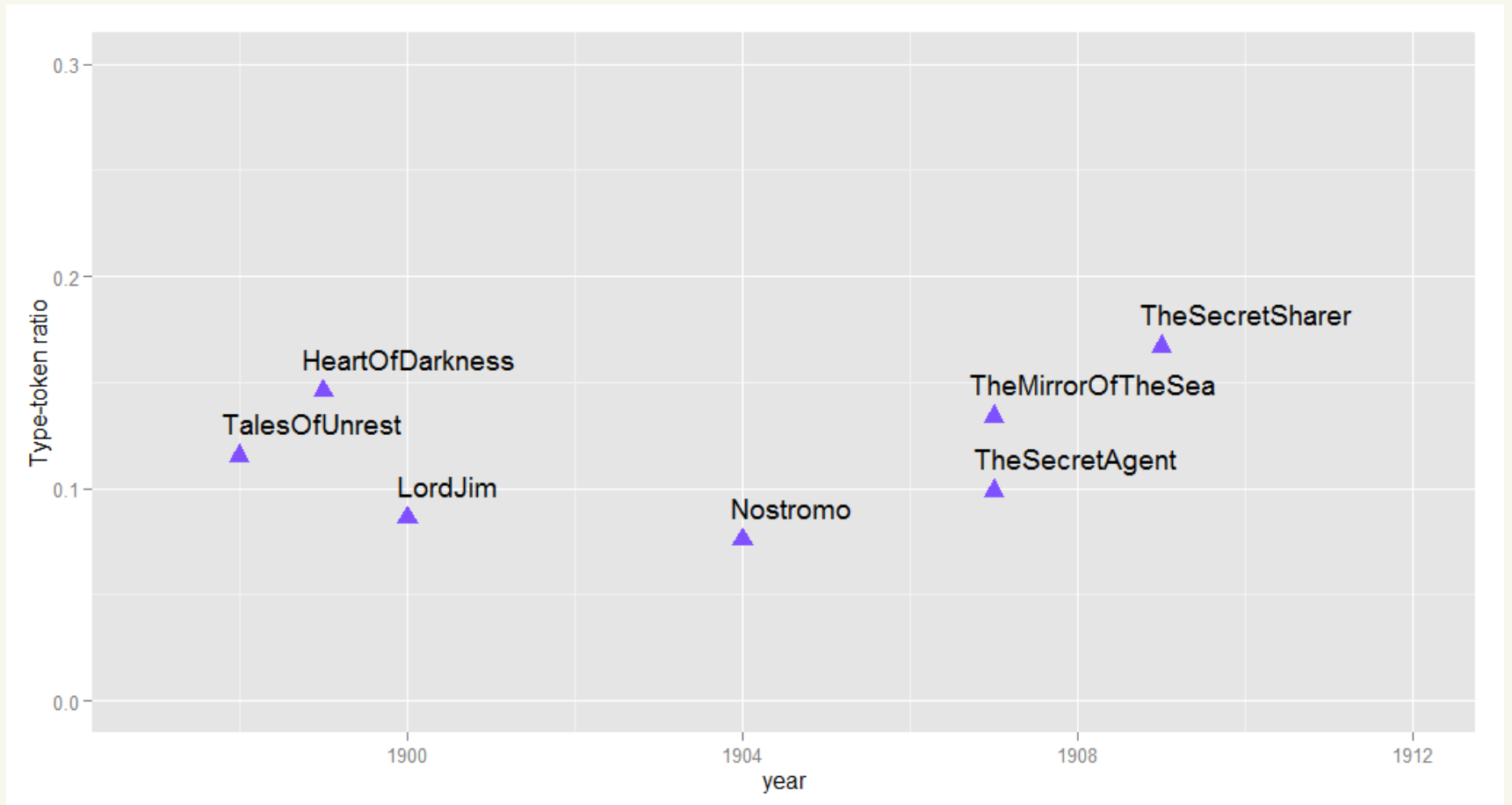
Stopword filtering

Summary: There are 5804 unique words other than those in the stop list, there are 15892 words other than those in the stop list. There are 37984 words in total including the stop words.

Words	Distribution	Counts
said		130
like		115
man		103
kurtz		93
know		81
time		76
little		62
came		62
looked		56
river		54
mr		51
long		50
men		48
say		47

Frequency list produced using [TaporWare](#)

Type-token ratio



Graph produced using [R](#)

Concordances

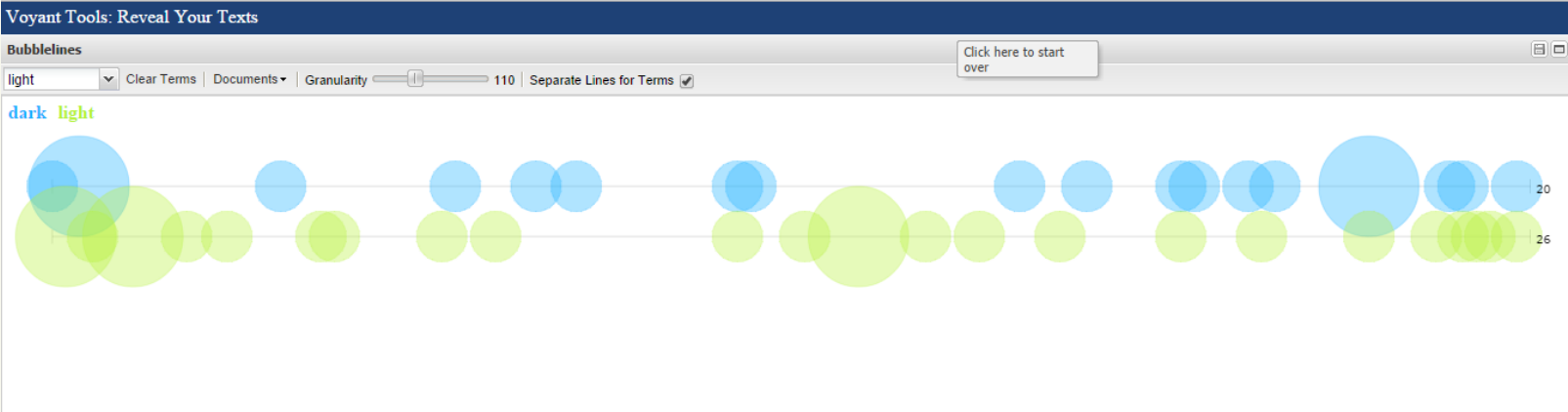
The screenshot shows the AntConc 3.4.1w (Windows) 2014 interface. The main window displays a concordance search for the term "language". The search results are shown in a table with columns for Hit, KWIC, and File. The search term is "language" and the search window size is set to 50. The results show 14 hits across 14 lines of text, with the word "language" highlighted in blue in the original image. The files listed in the File column are HeartOfDarkness.txt, LordJim.txt, Nostromo.txt, TalesOfUnrest.txt, TheMirrorOfTheSea.txt, TheSecretAgent.txt, and TheSecretSharer.txt.

Hit	KWIC	File
1	ntly put to us some question in an understandable language; but he died without uttering :	HeartOfDark
2	f amazing words that resembled no sounds of human language; and the deep murmurs of the	HeartOfDark
3	the voice. He had picked up enough of the language to understand the word water	LordJim.txt
4	t upon him, caught only the fragments of official language. . . "The Court . . . Gustav So-a	LordJim.txt
5	each of us can interpret for himself from the language of facts, that are so often mor	LordJim.txt
6	hill. Their chief had spoken to him in the language of his own people, making cle	LordJim.txt
7	, on the din of armed struggle, on the inflamed language of proclamations. He had nev	Nostromo.tx
8	land--the same to whom the doctors used the language of horrid and veiled menaces.	Nostromo.tx
9	except the old father and mother--used the French language amongst themselves. "And yc	Nostromo.tx
10	books in paper covers and mostly in the French language. The big black letters formed t	Nostromo.tx
11	trange, anxious whine the sonority of the Spanish language, which he pattered	Nostromo.tx
12	of the Boulevard, that had been the only forcible language-- "_Non, Madame. Rien n'est	Nostromo.tx
13	an, obstinately, because he was not aware in what language he was speaking. His identity,	Nostromo.tx
14	he lighter sort of historical works in the French language, such, for instance as the boo	Nostromo.tx

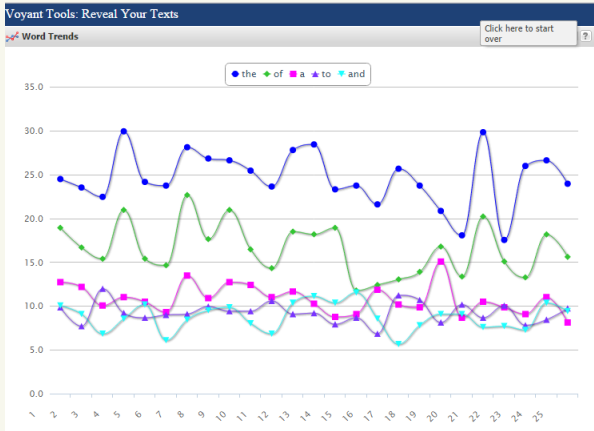
Search Term: Words Case Regex Search Window Size: 50
Start Stop Sort

Concordance produced using [AntConc](#)

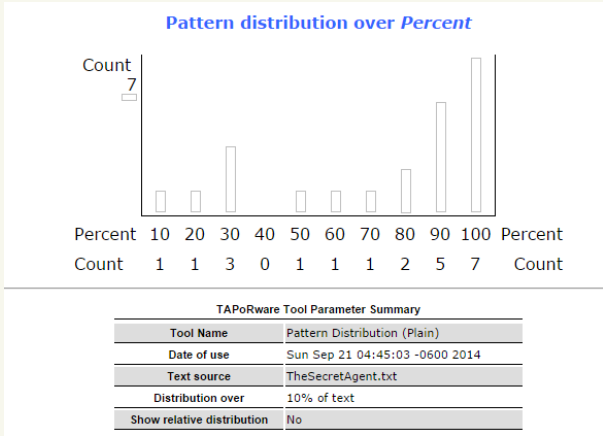
Distribution



Voyant BubbleLines



Voyant Type Frequencies Chart

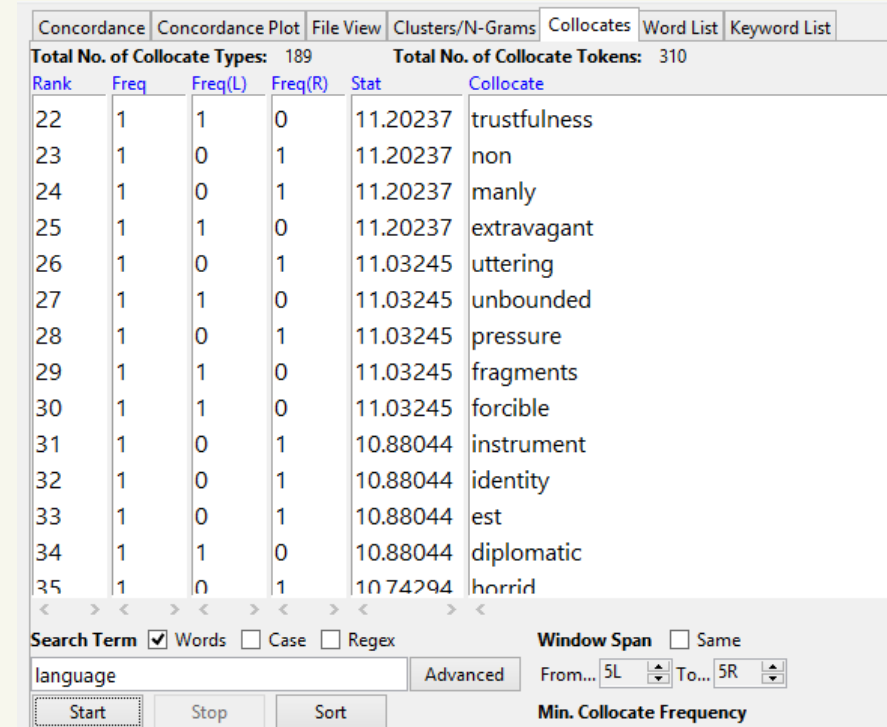


Tapor Distribution

Collocation

Words	Counts
	88
Told	6
Tell	5
Short	4
Love	3
Th	3
Knew	3
Remember	3
Seen	3
Know	3
Make	3
Course	3
Heard	3
Like	3
Lives	2
Long	2
Telling	2
Goes	2
True	2

List produced
using [TaporWare](#)



The screenshot shows the AntConc software interface. At the top, there are menu options: Concordance, Concordance Plot, File View, Clusters/N-Grams, Collocates (selected), Word List, and Keyword List. Below the menu, it displays 'Total No. of Collocate Types: 189' and 'Total No. of Collocate Tokens: 310'. The main table has columns: Rank, Freq, Freq(L), Freq(R), Stat, and Collocate. The table lists 35 items, each with a rank, frequency, left and right frequencies, a statistic, and a collocate word. At the bottom, there are search options: Search Term (checked), Words (checked), Case (unchecked), and Regex (unchecked). The search term is 'language'. There are also buttons for Start, Stop, and Sort, and a Window Span section with 'From... 5L' and 'To... 5R'. A 'Min. Collocate Frequency' section is also visible.

Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate
22	1	1	0	11.20237	trustfulness
23	1	0	1	11.20237	non
24	1	0	1	11.20237	manly
25	1	1	0	11.20237	extravagant
26	1	0	1	11.03245	uttering
27	1	1	0	11.03245	unbounded
28	1	0	1	11.03245	pressure
29	1	1	0	11.03245	fragments
30	1	1	0	11.03245	forcible
31	1	0	1	10.88044	instrument
32	1	0	1	10.88044	identity
33	1	0	1	10.88044	est
34	1	1	0	10.88044	diplomatic
35	1	0	1	10.74294	horrid

List produced
using [AntConc](#)

Co-occurrence

3 co-occurrences found

his passive heroism , to feel the sting of his abstention . The boat was heavy ; they pushed at the bow with no breath to spare for an encouraging **word** : but the turmoil of terror that had scattered their **self** -command like chaff before the wind , converted their desperate exertions into a bit of fooling , upon my word

word : but the turmoil of terror that had scattered their **self** -command like chaff before the wind , converted their desperate exertions into a bit of fooling , upon my **word** , fit for knockabout clowns in a farce . They pushed with their hands , with their heads , they pushed for dear life with all the weight of their

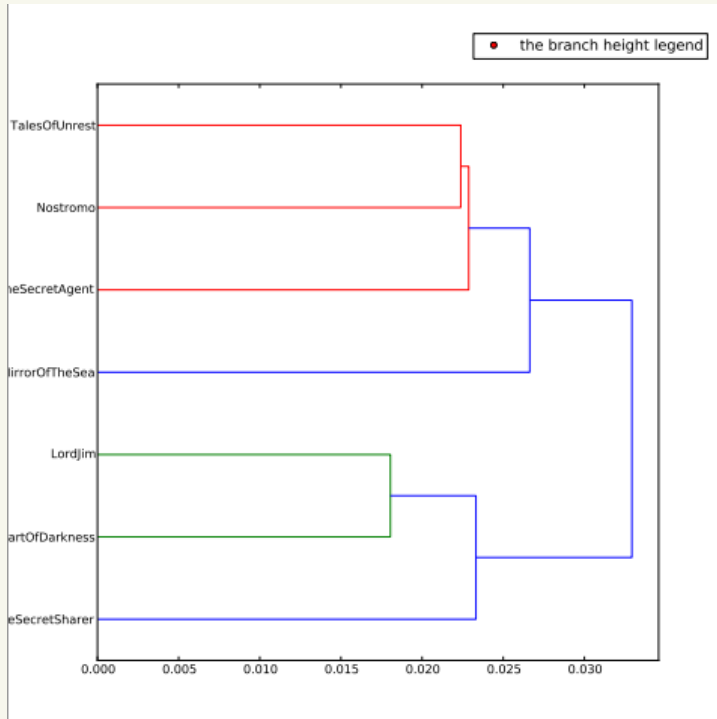
in his careless yet feeling voice , with his offhand manner , a little puzzled , a little bothered , a little hurt , but now and then by a **word** or a phrase giving one of these glimpses of his very own **self** that were never any good for purposes of orientation . It's difficult to believe he will never

TAPoRware Tool Parameter Summary

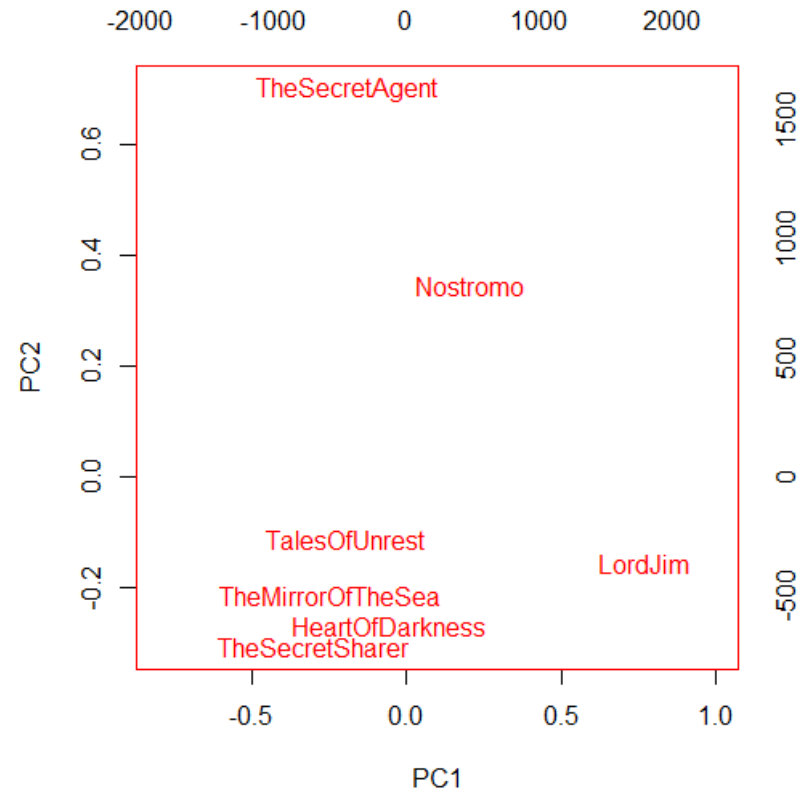
Tool Name	Find Text -- Co-occurrence (Plain)
Date of use	Sun Sep 21 04:58:19 -0600 2014
Text source	LordJim.txt
Primary pattern	word
Co-pattern	self
Context	Word
Context length	30
Display format	HTML

List produced using [TaporWare](#)

Clustering



Dendrogram produced using [Lexomics](#)



PCA created using [Tapor](#) and [R](#)

Information extraction

7 unique date entries found

Year

1879 1907

Month

May Jun

Week

Sun

Season

Spring Summer

1879

- There was a date, 24th **June 1879**, engraved inside.

1907

- **_September_ 1907 _Second Edition_ .**
- **_October_ 1907 TO H.**

May

- And meantime the Chief Inspector went on, peering at the table with a calm face and the slightly anxious attention of an indigent customer bending over what may be called the by-products of a butcher's shop with a view to an inexpensive **Sunday** dinner.
- "â **May** want it soon," snuffled vaguely Mr Verloc, who was coming to the end of his calculated indiscretions.
- "â **May** I ask you where you were going?"

Jun

- There was a date, 24th **June 1879**, engraved inside.

Sun

- And meantime the Chief Inspector went on, peering at the table with a calm face and the slightly anxious attention of an indigent customer bending over what may be called the by-products of a butcher's shop with a view to an inexpensive **Sunday** dinner.
- "âAnd, my dear, I must see that poor boy every **Sunday**."

Conclusions

- Text analysis tools may produce new views on the text
- There are caveats; Tools are based on assumptions on how texts ought to be analysed
- Customisations of existing tools are generally needed for more specific research question
- Identification of appropriate tools via library support