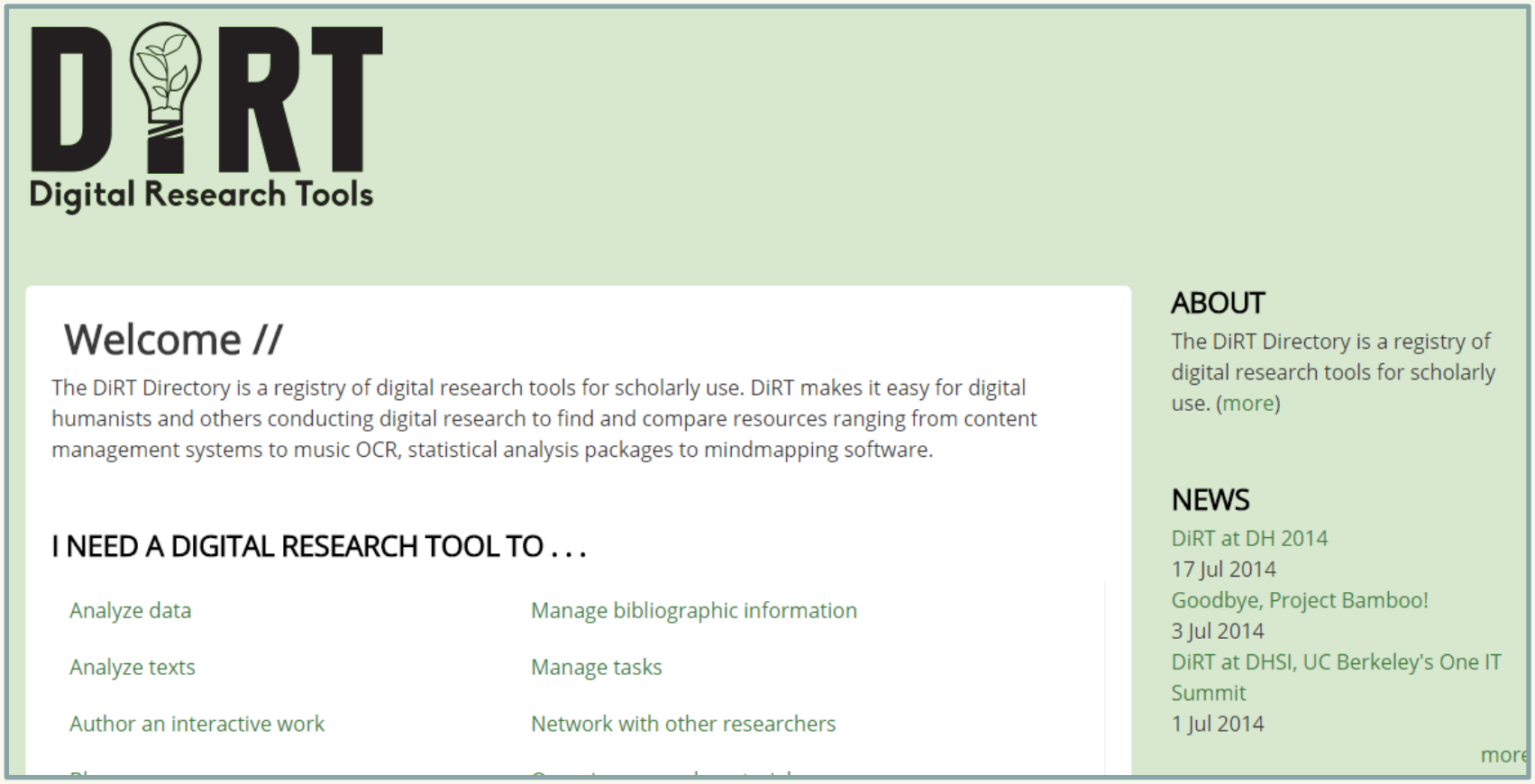
Text Analysis and Visualisation:
An Overview of Tools

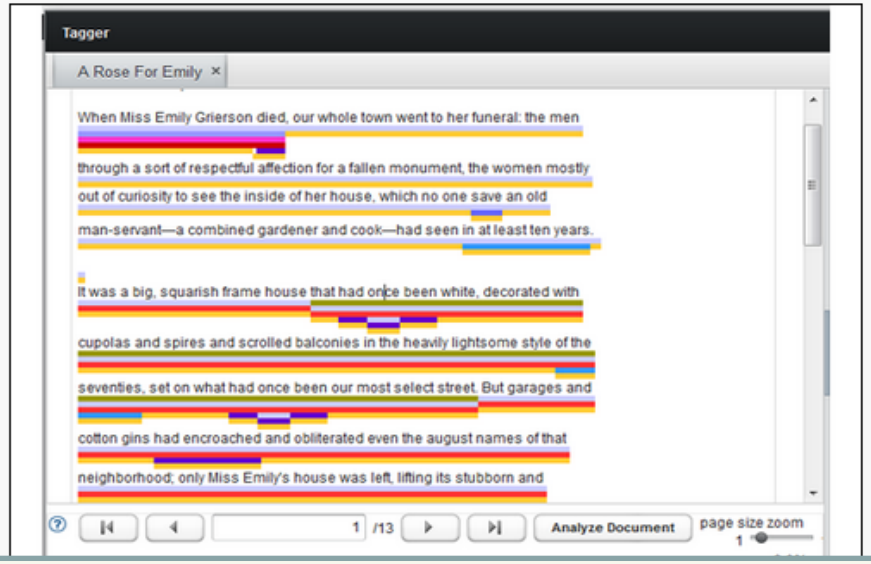

DiRT
Digital Research Tools
Welcome //
The DiRT Directory is a registry of digital research tools for scholarly use. DiRT makes it easy for digital humanists and others conducting digital research to find and compare resources ranging from content management systems to music OCR, statistical analysis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

About TAPoR Tour Useful links Contact Login/Sign up

Search

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



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


dh Commons
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.

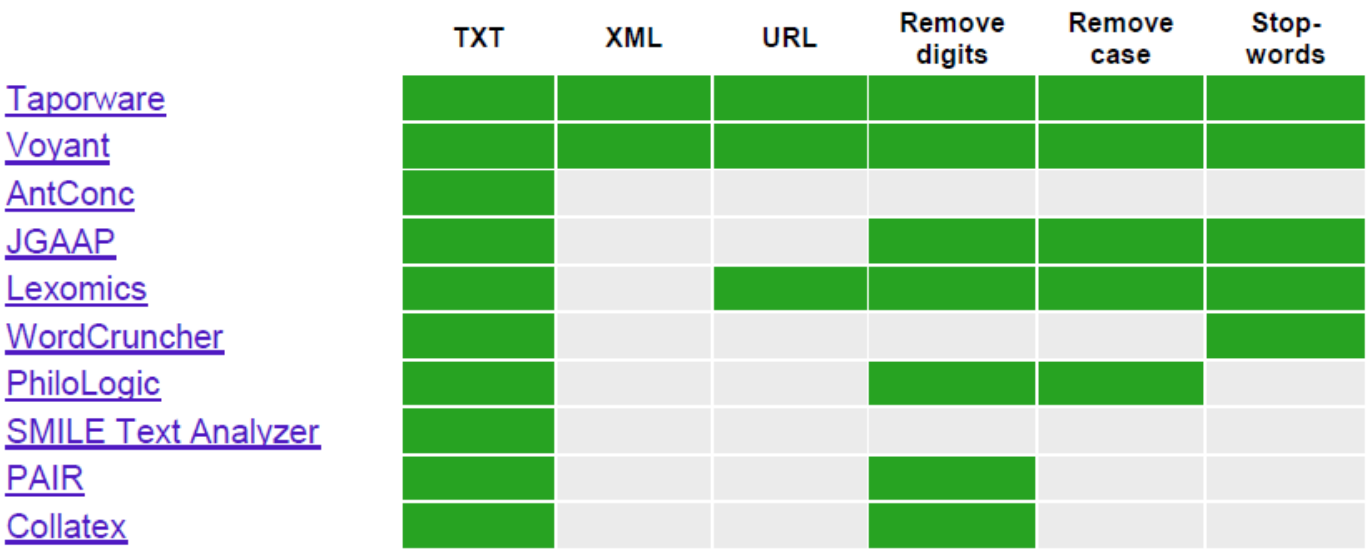
Webpage: http://voyant-tools.org/
Platform: Web-based
Developer: hermeneuti.ca [Stéfan Sinclair and Geoffrey Rockwell]
Cost: Free
Code license: Open source
Last modified: 7/14/14
TaDiRAH goals & methods: Analysis Content Analysis
TaDiRAH research objects: Text
Categories: Text collections Text mining Visualization
Tags: text analysis corpus word cloud

Text Analysis

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

	Concordance	Data extraction	Comparison	Word cloud
Taporware	■	■		■

Import and Data cleaning



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

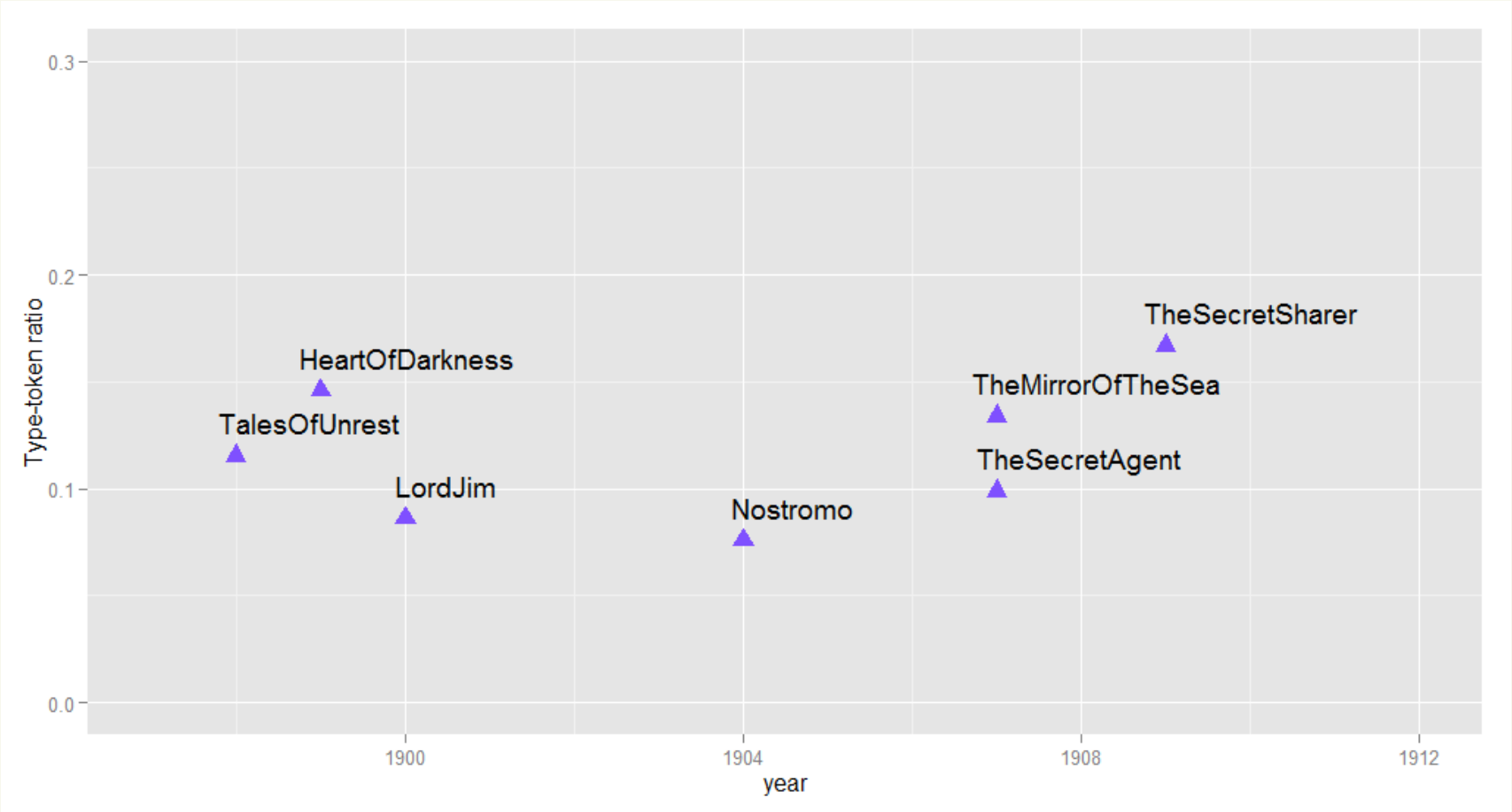
Stopword filtering

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

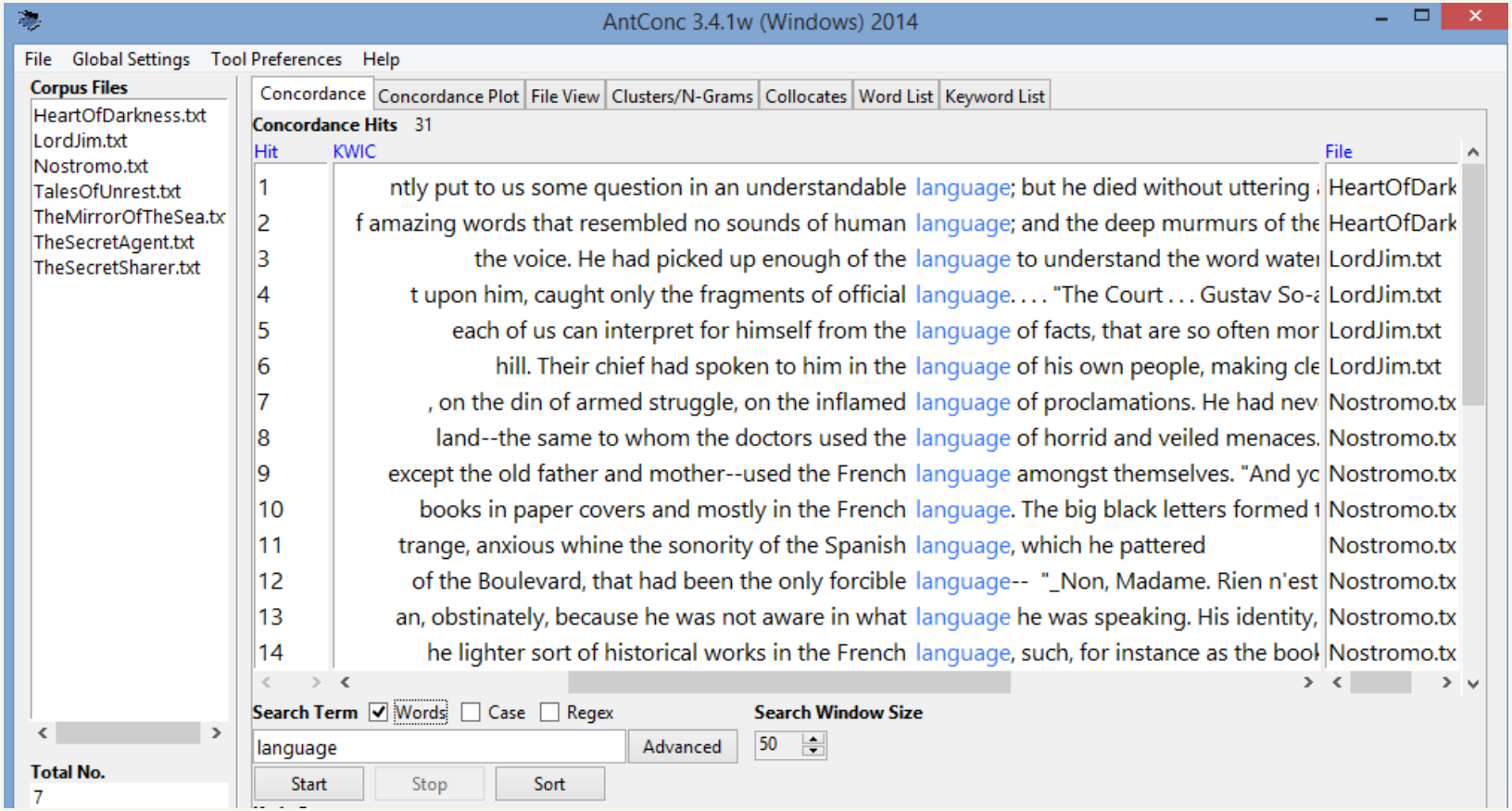
Frequency list produced using TaporWare

Type-token ratio



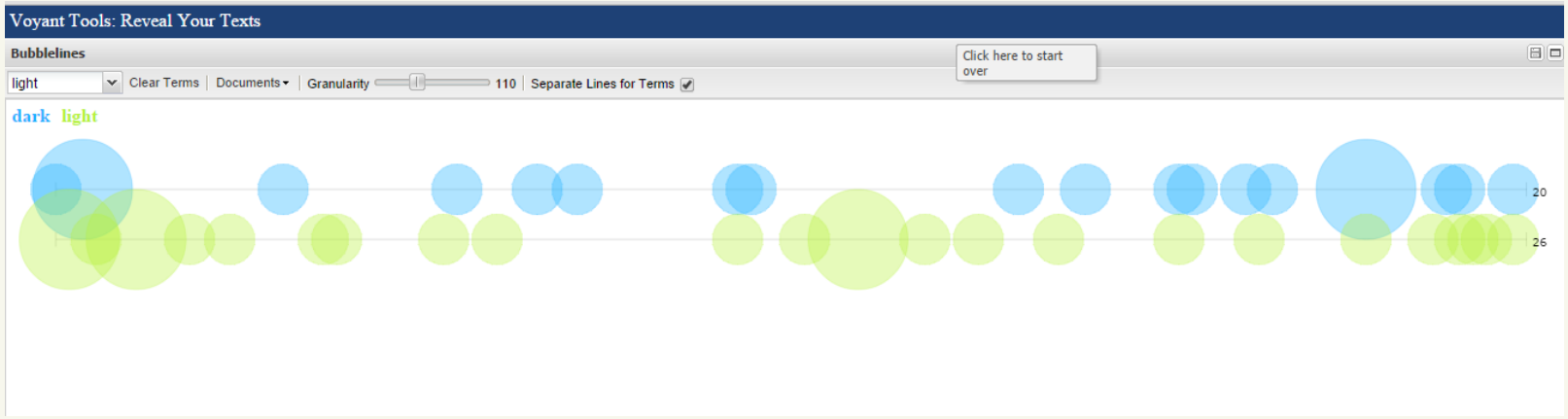
Graph produced using R

Concordances

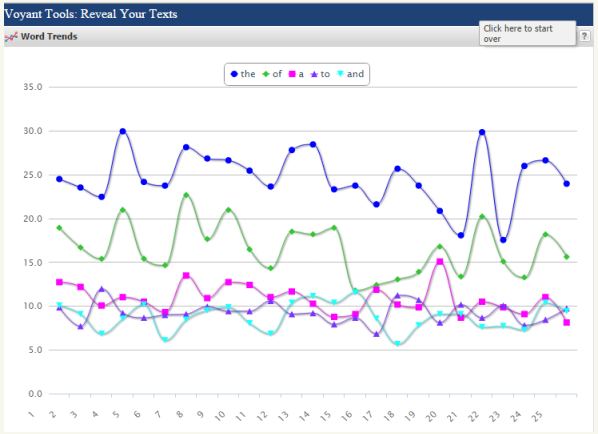


Concordance produced using [AntConc]

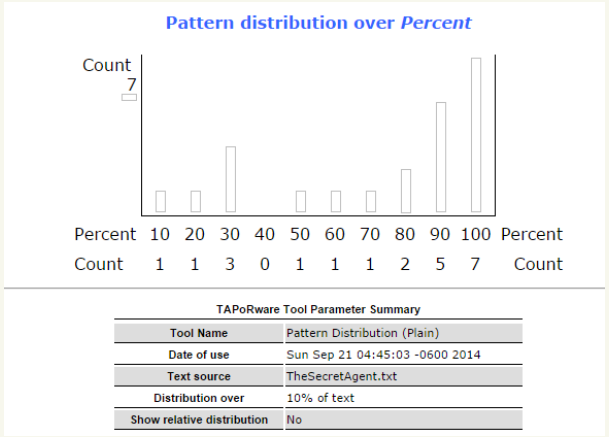
Distribution



Voyant BubbleLines



Voyant Type Frequencies Chart

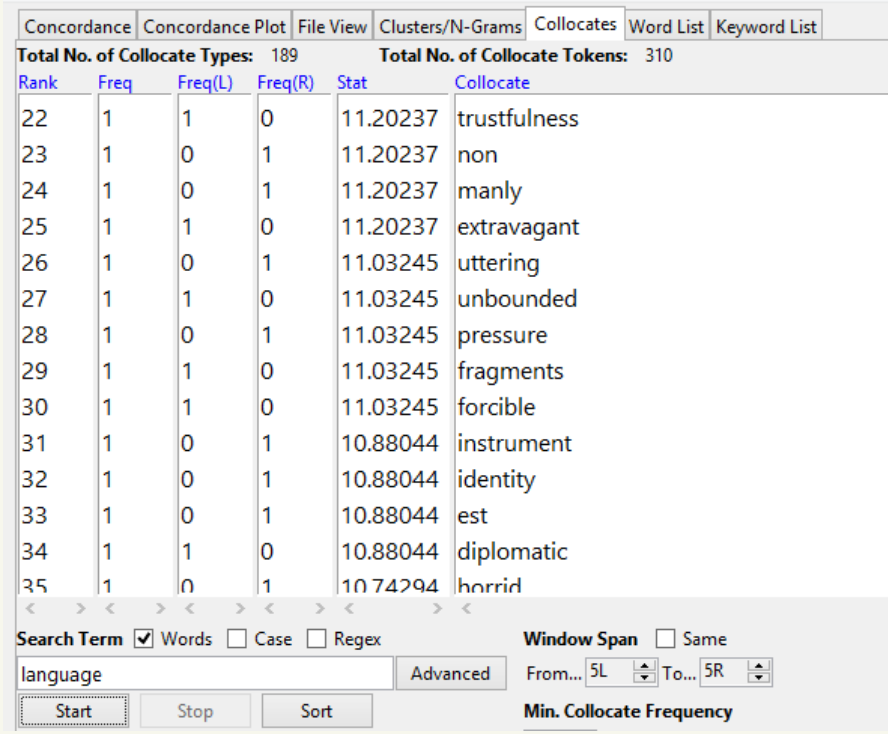


Tapor Distribution

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



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

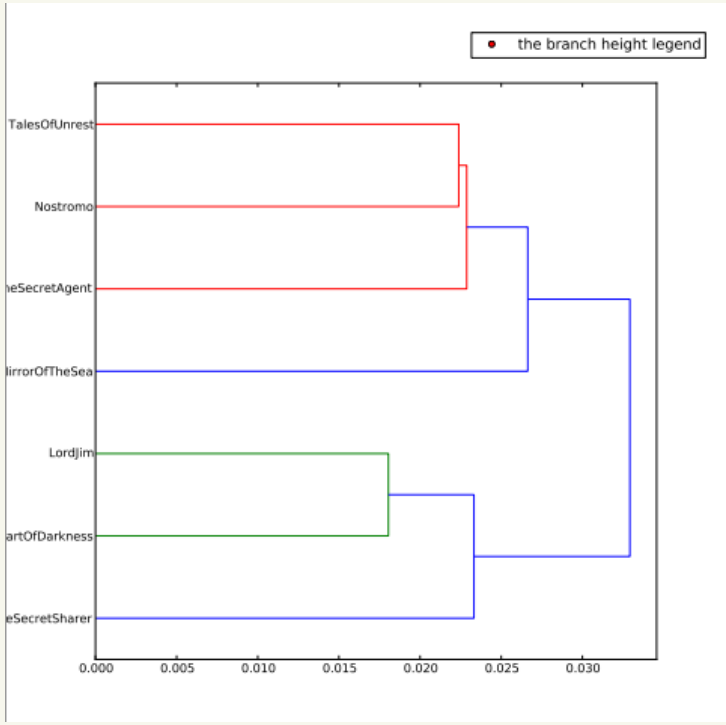
in his careless yet feeling voice , with his offhand manner , a little puzzled , a little bothered , a little hurt , but now and then by a **word** or a phrase giving one of these glimpses of his very own ***self*** that were never any good for purposes of orientation . It's difficult to believe he will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

TAPoRware Tool Parameter Summary

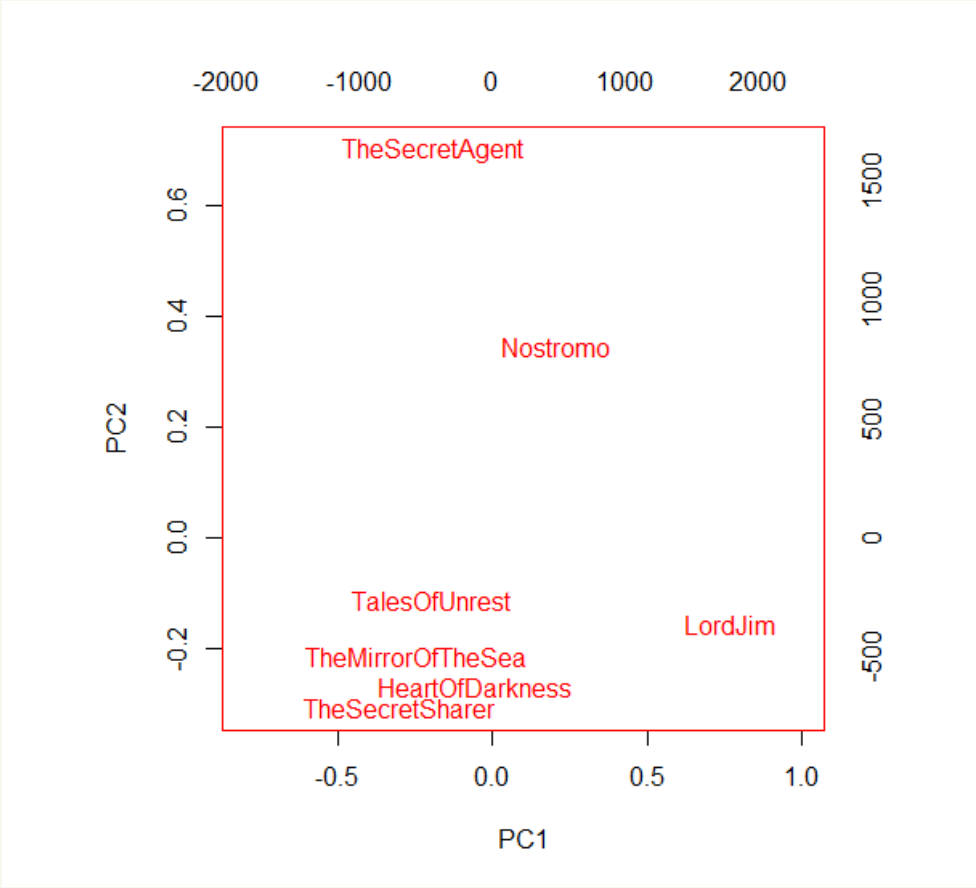
Tool Name	Find Text -- Co-occurrence (Plain)
Date of use	Sun Sep 21 04:58:19 -0600 2014
Text source	LordJim.txt
Primary pattern	word
Co-pattern	self
Context	Word
Context length	30
Display format	HTML

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 butcherâs shop with a view to an inexpensive **Sunday** dinner.
- â â**May** want it soon,â 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 butcherâs 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