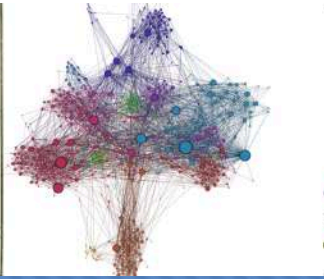
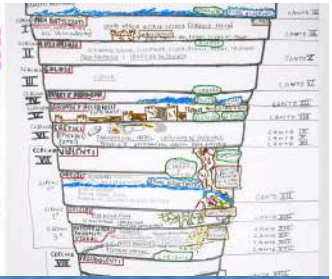
```
# For details, see https://review.docs.microsoft.com/en-us/visualstudio/rtvs/sql-server

# Test code
library(RODBC)
channel <- odbcDriverConnect(dbConnection)
InputDataSet <- sqlQuery(channel, iconv(paste(readLines(
    'c:/proj/rproject1/rproject1/storedprocedure.query.sql',
    encoding = 'UTF-8', warn = FALSE), collapse = '\n'), from = 'UTF-8',
    to = 'ASCII', sub = ''))
odbcClose(channel)

OutputDataSet <- InputDataSet
```





Stylometry

Giovanni Pietro Vitali – University College Cork

giovannipietrovitali@gmail.com

https://github.com/digitalkoine

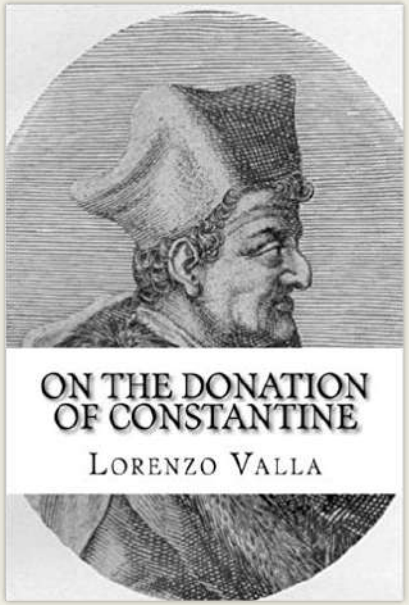
https://ucc-ie.academia.edu/GiovanniPietroVitali

université de TOURS

UCC
Coláiste na hOllscoile Corcaigh, Éire
University College Cork, Ireland

MARIE CURIE ACTIONS

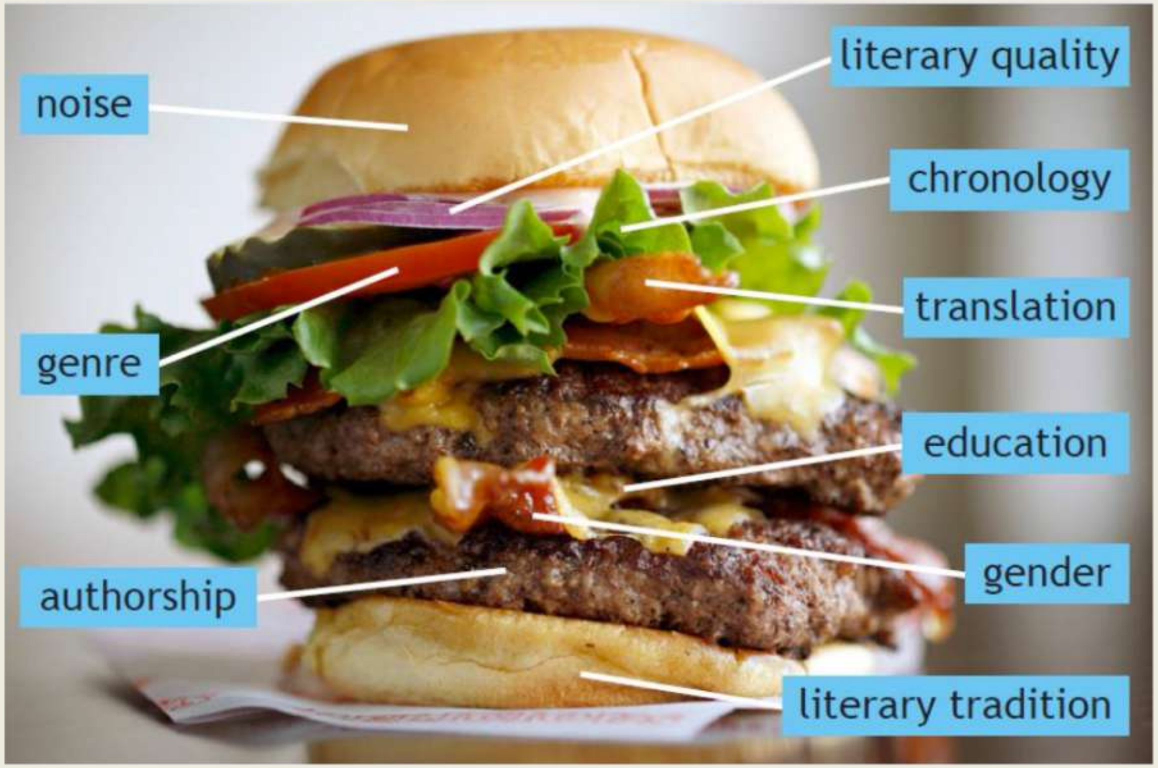
Lorenzo Valla and the *Constantini donatione*



Lorenzo Valla
(c. 1407–1457)

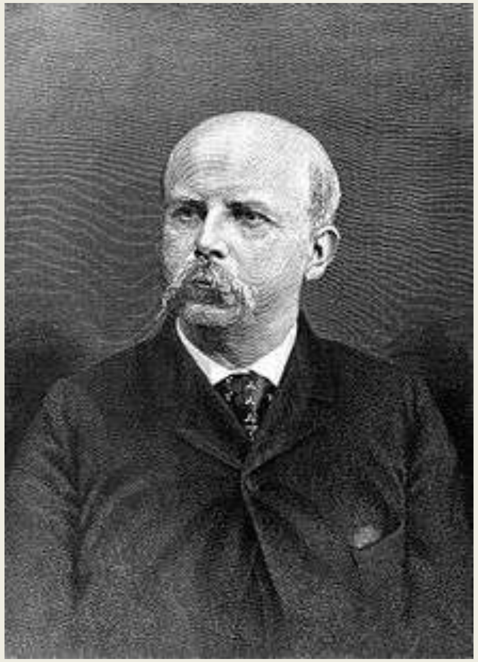
- The ***Constantini donatione*** was a forged decree where the emperor Constantine I transfers authority of the Roman Empire to the Pope.
- In ***De falso credita et ementita Constantini donatione declamatio*** (1517), Loreno Valla shows that the act was done in the eighth century by the same papal chancellery. Some grammatical forms could not have been used in the 4th century
- Act of immense value for the history of philology. The first instance of scholarly-based investigation of style

STYLOmetry



- Author attribution identification of unknown authors
- Genre classification
- Historical study of language change
- Other applications
- Anonymity
- Plagiarism

«Measuring» authorial style



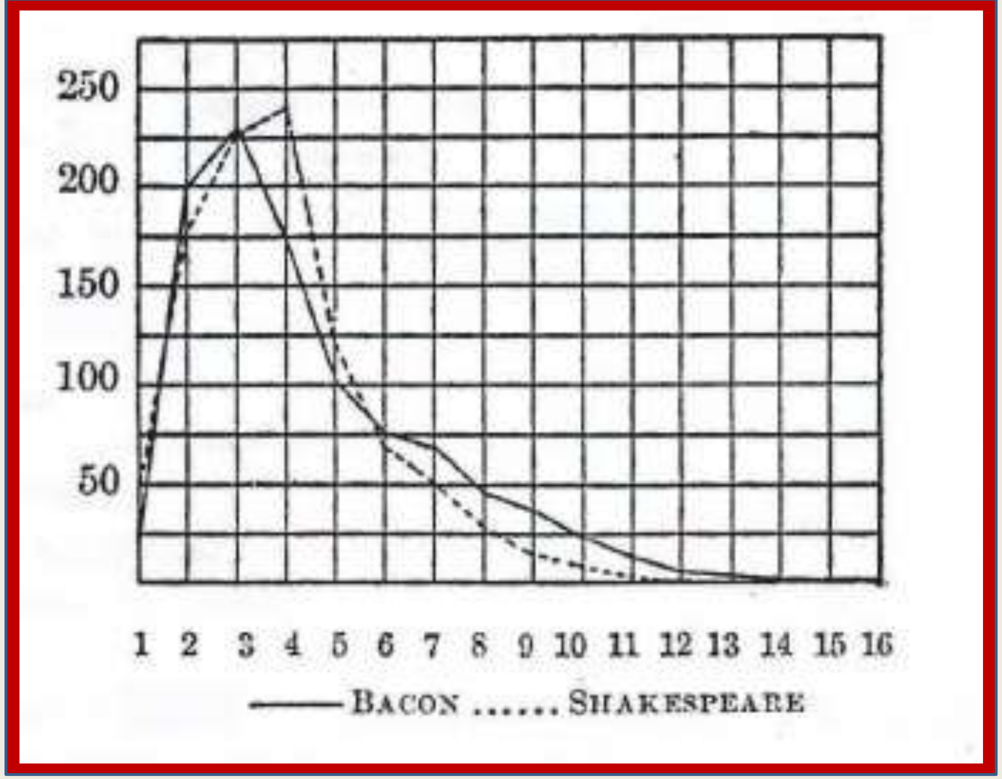
Thomas Corwin Mendenhall (October 4, 1841 – March 23, 1924) was an American autodidact physicist and meteorologist. He was the first professor hired at The Ohio State University in 1873 and the superintendent of the U.S. Coast and Geodetic Survey from 1889 to 1894. Alongside his work, he was also an advocate for the adoption of the metric system by the United States.

He provided the first empirical evidence in favor of de Morgan's assumptions. In two subsequent studies, Mendenhall (1887, 1901) elaborated on de Morgan's ideas, suggesting that in addition to analy- ses "based simply on mean word-length" (1887: 239), one should attempt to graphically exhibit the peculiarities of style in composition: in order to arrive at such graphics, Mendenhall counted the frequency with which words of a given length occur in 1000-word samples from different authors, among them Francis Bacon, Charles Dickens, William M. Thackerey, and John Stuart Mill.

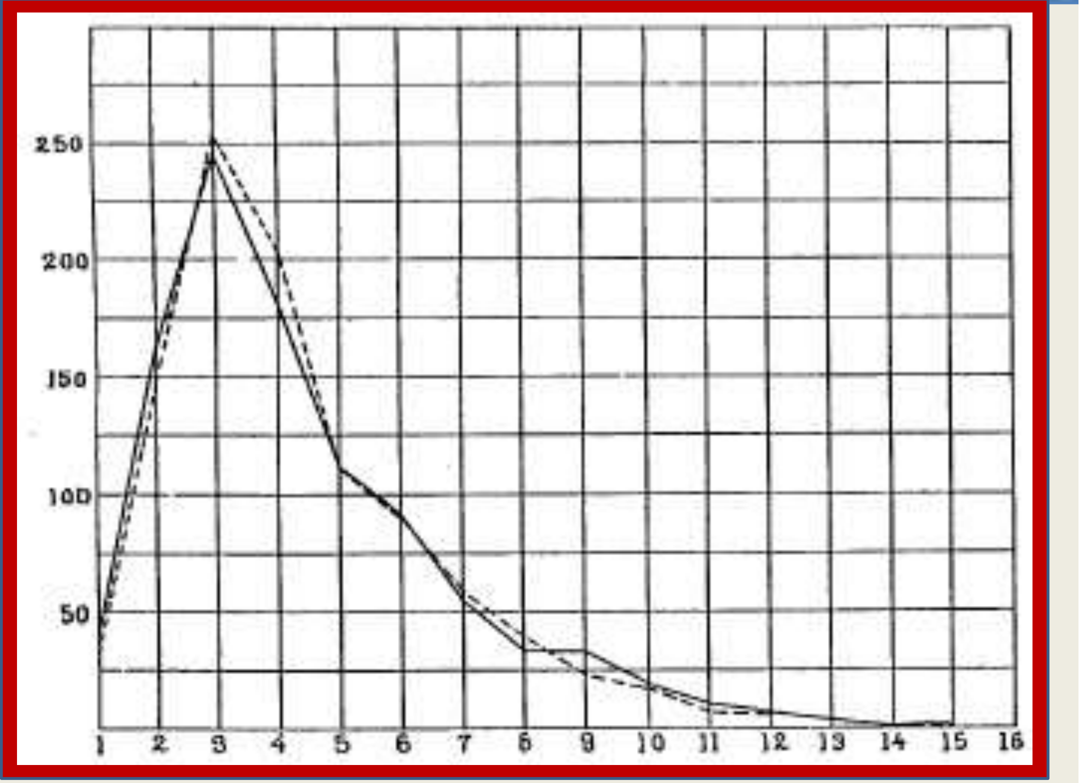
Robert E. Moritz, On The Significance Of Characteristic Curves Of Composition, Popular Science Monthly, volume 65, June 1904.

https://en.wikisource.org/wiki/Popular_Science_Monthly/Volume_65/June_1904/On_the_Significance_of_Characteristic_Curves_of_Composition

Bacon, Shakespeare, Dickens (Mendenhall)



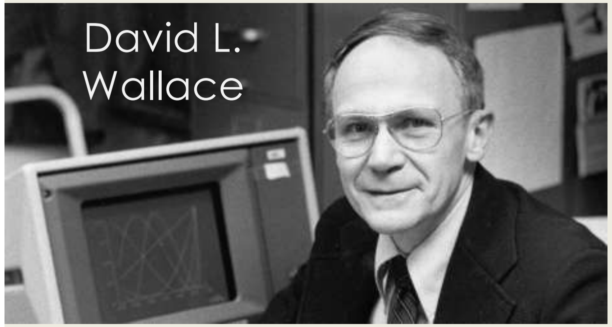
Word Length Frequencies in Bacon's and Shakespeare's Texts (Mendenhall 1901)



Word Length Frequencies in Dickens' Oliver Twist (Mendenhall 1887)

Mosteller and Wallace (1964)

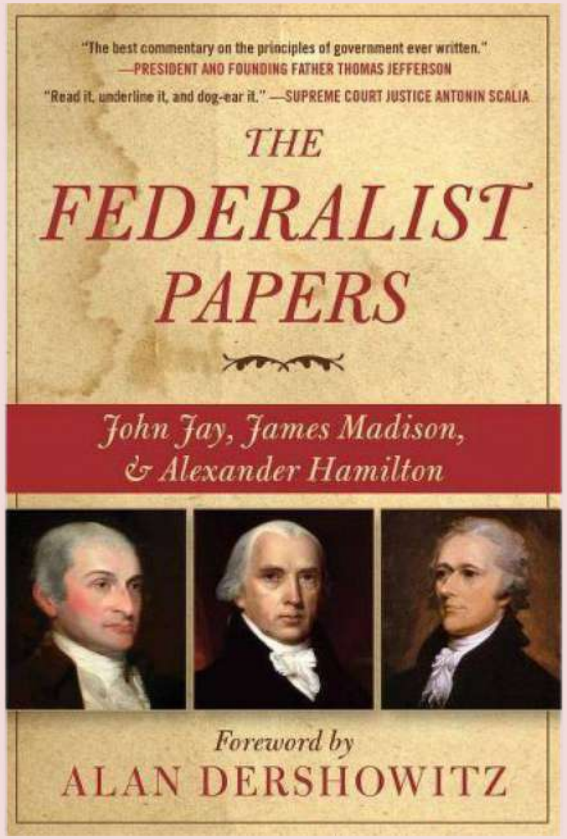
Frederick Mosteller



David L. Wallace

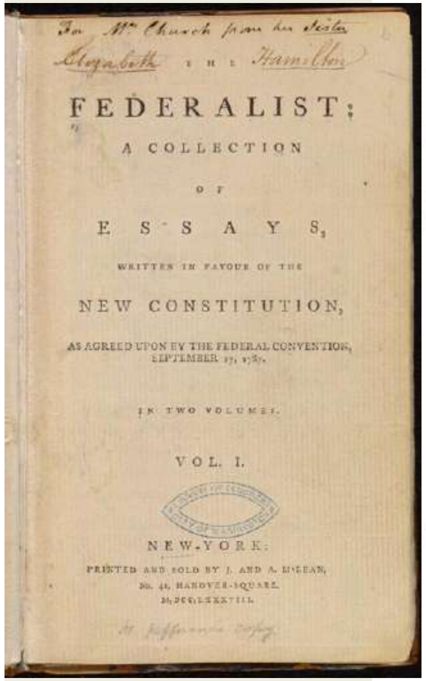
Case study: *The Federalist Papers* (1787-1788)

The Federalist Papers is a collection of 85 articles and essays written by Alexander Hamilton, James Madison, and John Jay under the pseudonym "Publius" to promote the ratification of the United States Constitution. The collection was commonly known as The Federalist until the name The Federalist Papers emerged in the 20th century.
The authors of The Federalist intended to influence the voters to ratify the Constitution.


"The best commentary on the principles of government ever written."
—PRESIDENT AND FOUNDING FATHER THOMAS JEFFERSON
"Read it, underline it, and dog-ear it." —SUPREME COURT JUSTICE ANTONIN SCALIA
THE
FEDERALIST
PAPERS
John Jay, James Madison,
& Alexander Hamilton
Foreword by
ALAN DERSHOWITZ

Statistical Approach

	enough	while	whilst	upon
Hamilton	0.59	0.26	0	2.93
Madison	0	0	0.47	0.16
Disputed texts	0	0	0.34	0.08
Co-authored texts	0.18	0	0.36	0.36



For Mrs Church from her Sister
Elizabeth THE Hamilton

FEDERALIST;

A COLLECTION

OF

ESSAYS,

WRITTEN IN FAVOUR OF THE

NEW CONSTITUTION,

AS AGREED UPON BY THE FEDERAL CONVENTION,
SEPTEMBER 17, 1787.

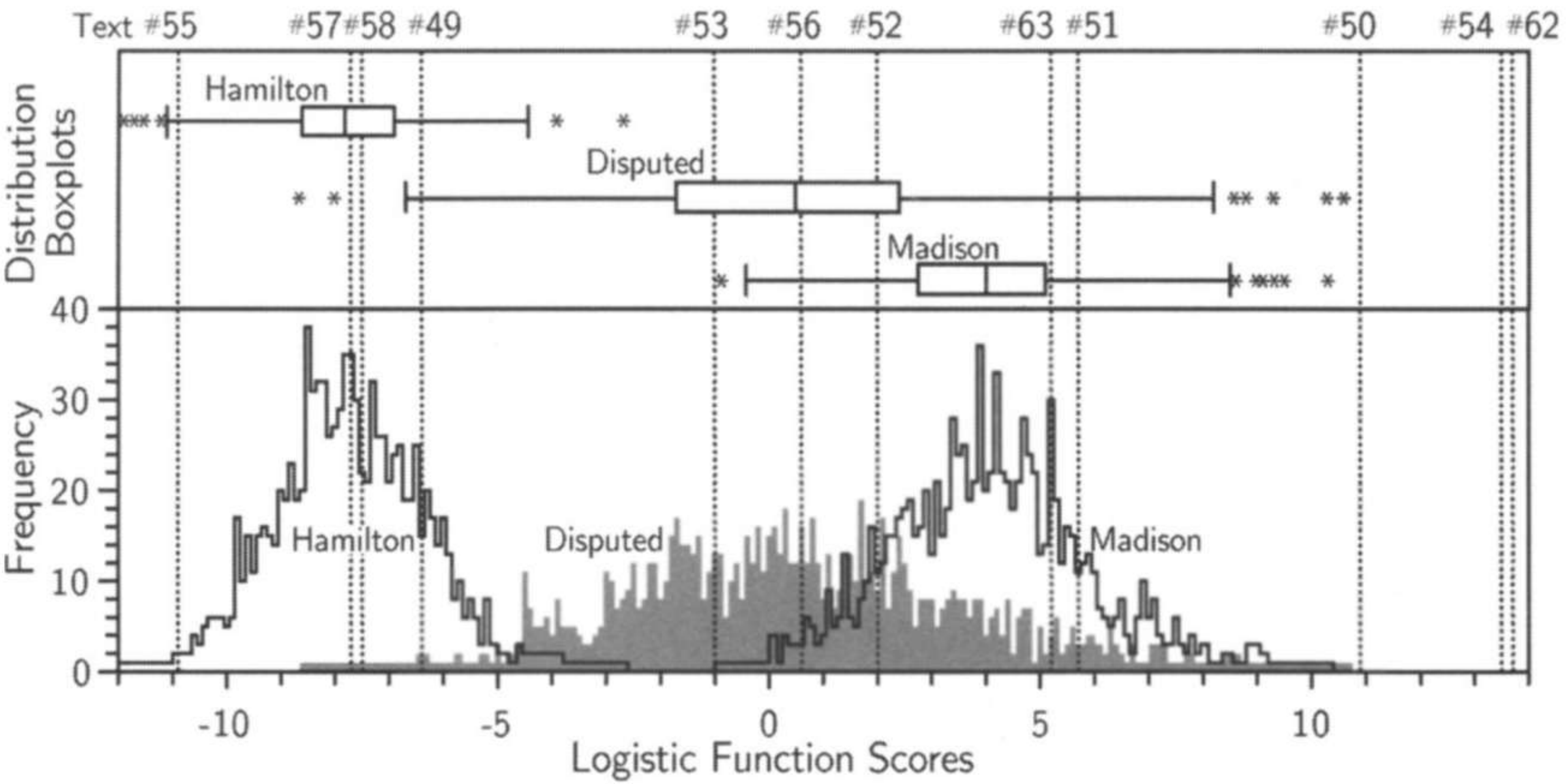
IN TWO VOLUMES.

VOL. I.

NEW-YORK:

PRINTED AND SOLD BY J. AND A. M'LEAN,
No. 41, HANOVER-SQUARE.
M,DCC,LXXXVIII.

Th. Jefferson's copy

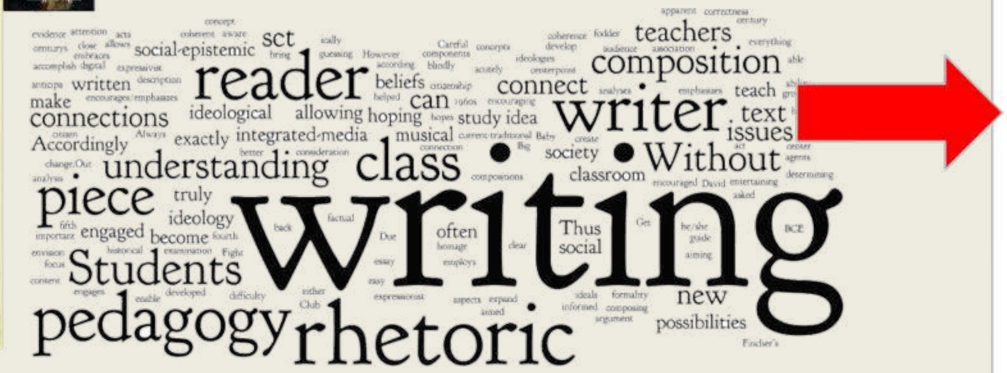


The Methods For Stylometry And Authorship Attribution

- Character-level analysis
- Syntax-level analysis
- Multi-method analysis (e.g. JGAAP, PAN competition software...)
- ...and many others
- In this lesson, just two methods:
  - Zeta method (for the quantitative analysis of style)
  - Delta method (for authorship attribution)

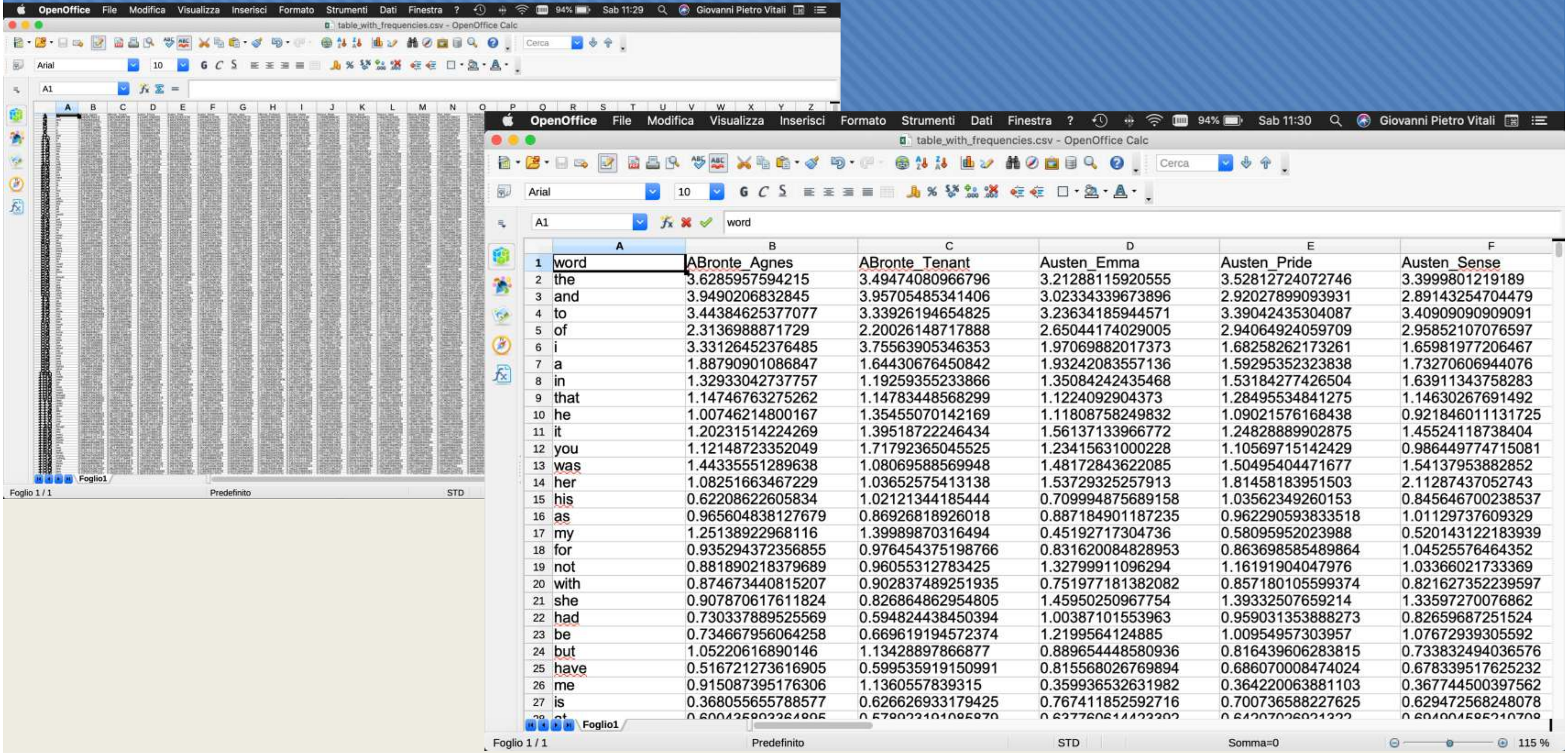
Let's take an example: English Novels

DELTA DISTANCE



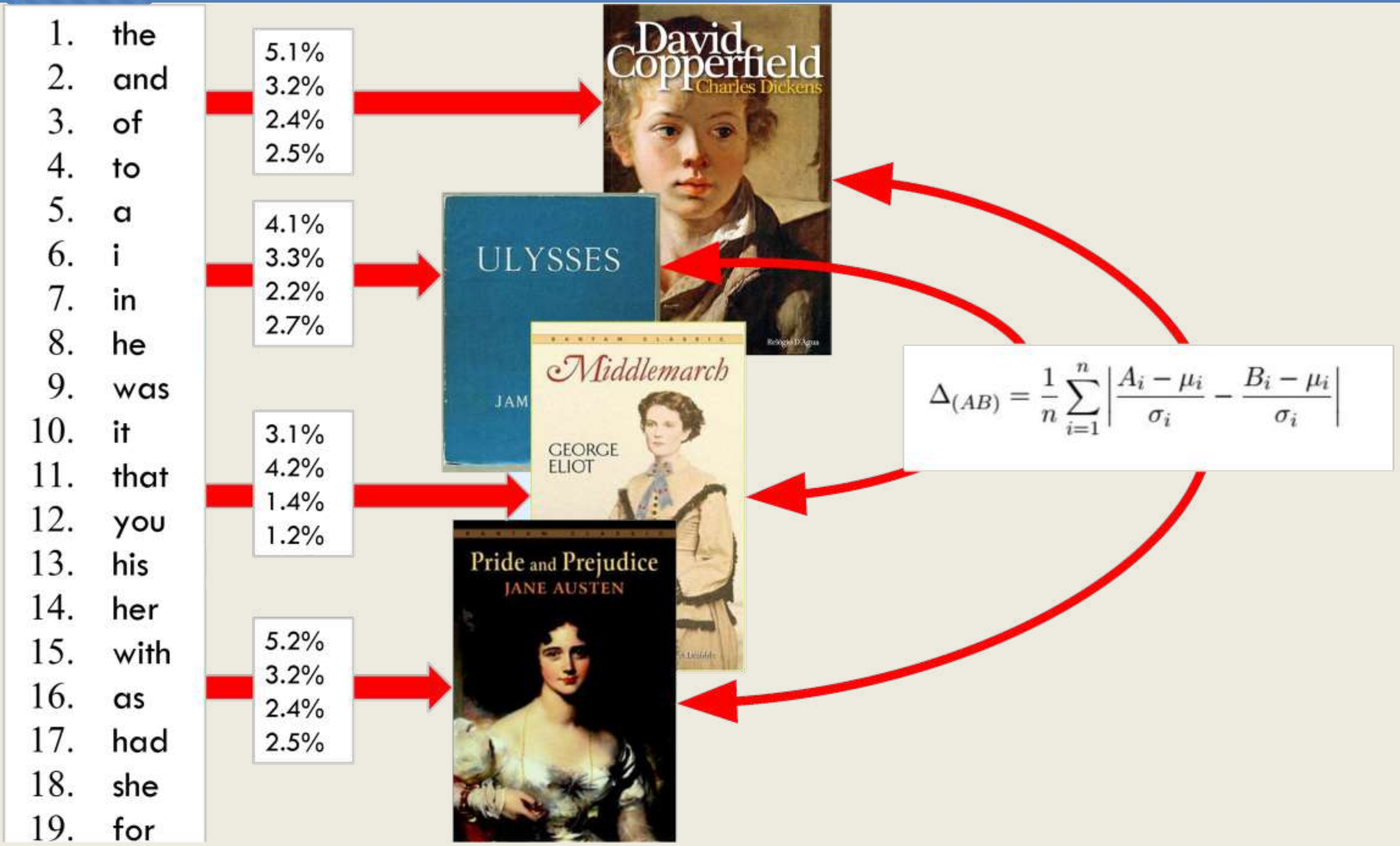
1. the
2. and
3. of
4. to
5. a
6. i
7. in
8. he
9. was
10. it
11. that
12. you
13. his
14. her
15. with
16. as
17. had
18. she
19. for

WORD-FREQUENCY BASED STYLOMETRY


OpenOffice File Modifica Visualizza Inserisci Formato Strumenti Dati Finestra ?
94%
Sab 11:29
Sab 11:30
Giovanni Pietro Vitali
table_with_frequencies.csv - OpenOffice Calc
Cerca
Arial
10
A1
word
Foglio1
Foglio 1 / 1
Predefinito
STD
Somma=0
115 %

	A	B	C	D	E	F
1	word	ABronte_Agnes	ABronte_Tenant	Austen_Emma	Austen_Pride	Austen_Sense
2	the	3.6285957594215	3.49474080966796	3.21288115920555	3.52812724072746	3.3999801219189
3	and	3.9490206832845	3.95705485341406	3.02334339673896	2.92027899093931	2.89143254704479
4	to	3.44384625377077	3.33926194654825	3.23634185944571	3.39042435304087	3.40909090909091
5	of	2.3136988871729	2.20026148717888	2.65044174029005	2.94064924059709	2.95852107076597
6	i	3.33126452376485	3.75563905346353	1.97069882017373	1.68258262173261	1.65981977206467
7	a	1.88790901086847	1.64430676450842	1.93242083557136	1.59295352323838	1.73270606944076
8	in	1.32933042737757	1.19259355233866	1.35084242435468	1.53184277426504	1.63911343758283
9	that	1.14746763275262	1.14783448568299	1.1224092904373	1.28495534841275	1.14630267691492
10	he	1.00746214800167	1.35455070142169	1.11808758249832	1.09021576168438	0.921846011131725
11	it	1.20231514224269	1.39518722246434	1.56137133966772	1.24828889902875	1.45524118738404
12	you	1.12148723352049	1.71792365045525	1.23415631000228	1.10569715142429	0.986449774715081
13	was	1.44335551289638	1.08069588569948	1.48172843622085	1.50495404471677	1.54137953882852
14	her	1.08251663467229	1.03652575413138	1.53729325257913	1.81458183951503	2.11287437052743
15	his	0.62208622605834	1.02121344185444	0.709994875689158	1.03562349260153	0.845646700238537
16	as	0.965604838127679	0.86926818926018	0.887184901187235	0.962290593833518	1.01129737609329
17	my	1.25138922968116	1.39989870316494	0.45192717304736	0.58095952023988	0.520143122183939
18	for	0.935294372356855	0.976454375198766	0.831620084828953	0.863698585489864	1.04525576464352
19	not	0.881890218379689	0.96055312783425	1.32799911096294	1.16191904047976	1.03366021733369
20	with	0.874673440815207	0.902837489251935	0.751977181382082	0.857180105599374	0.821627352239597
21	she	0.907870617611824	0.826864862954805	1.45950250967754	1.39332507659214	1.33597270076862
22	had	0.730337889525569	0.594824438450394	1.00387101553963	0.959031353888273	0.82659687251524
23	be	0.734667956064258	0.669619194572374	1.2199564124885	1.00954957303957	1.07672939305592
24	but	1.05220616890146	1.13428897866877	0.889654448580936	0.816439606283815	0.733832494036576
25	have	0.516721273616905	0.599535919150991	0.815568026769894	0.686070008474024	0.678339517625232
26	me	0.915087395176306	1.1360557839315	0.359936532631982	0.364220063881103	0.367744500397562
27	is	0.368055655788577	0.626626933179425	0.767411852592716	0.700736588227625	0.629472568248078
28	at	0.600435803364895	0.578023191085879	0.637760614423392	0.642070269021322	0.694904585210708

WORD-FREQUENCY BASED STYLOMETRY



1. the
2. and
3. of
4. to
5. a
6. i
7. in
8. he
9. was
10. it
11. that
12. you
13. his
14. her
15. with
16. as
17. had
18. she
19. for

5.1%
3.2%
2.4%
2.5%

4.1%
3.3%
2.2%
2.7%

3.1%
4.2%
1.4%
1.2%

5.2%
3.2%
2.4%
2.5%

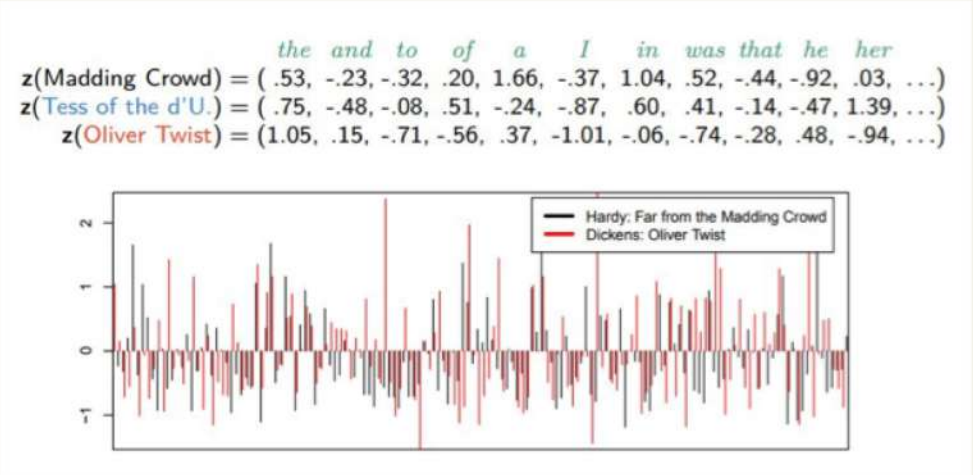
$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i-\mu_i}{\sigma_i} - \frac{B_i-\mu_i}{\sigma_i}\right|$$

Burrows DELTA

John Burrows

Frequencies of 100 – 5,000 most frequent words (MFW) form a “fingerprint” of an author’s style

Standardized to z-scores to give each word equal weight



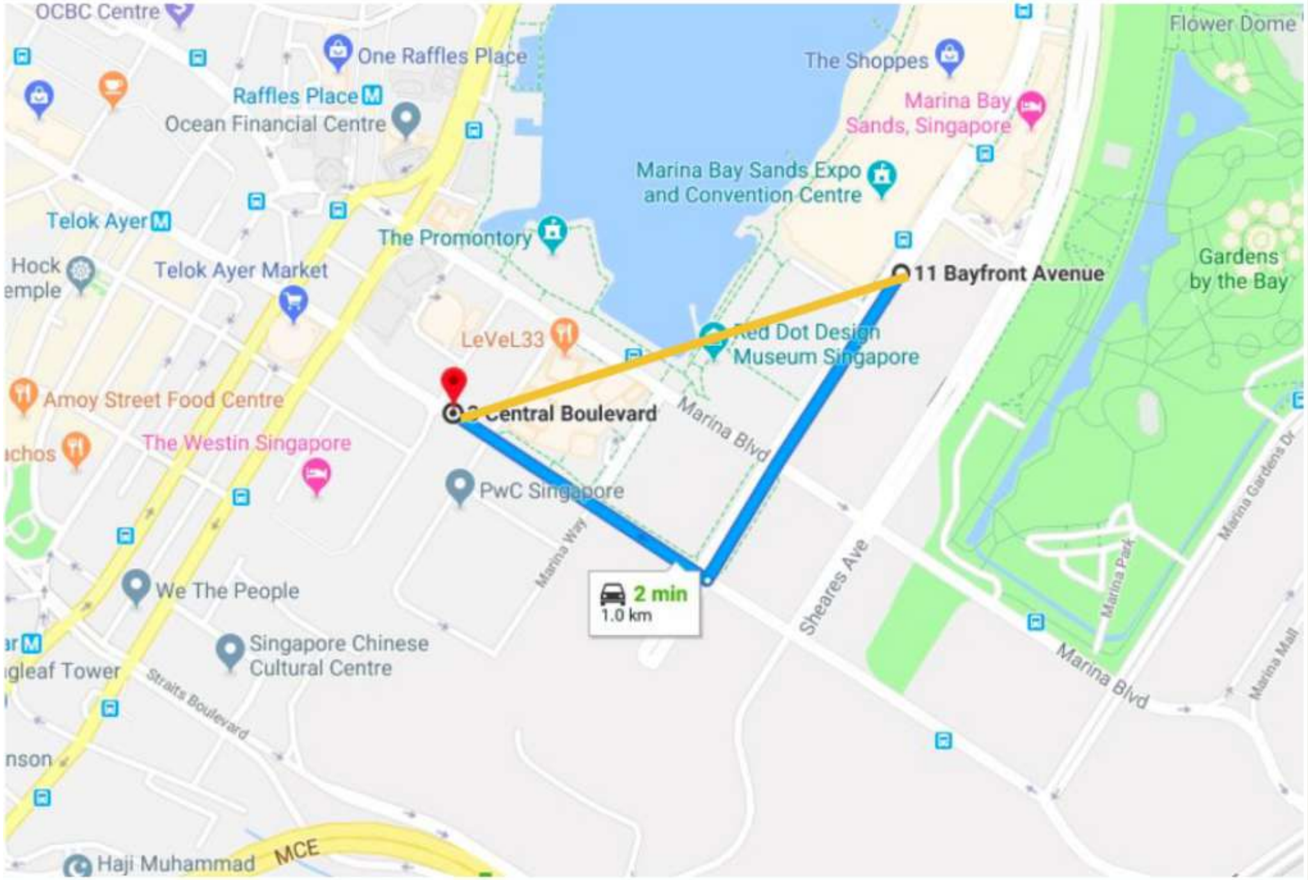
	the	and	to	of	a	I	in	was	that	he	her	
z(Madding Crowd) =	(.53,	-.23,	-.32,	.20,	1.66,	-.37,	1.04,	.52,	-.44,	-.92,	.03,	...)
z(Tess of the d'U.) =	(.75,	-.48,	-.08,	.51,	-.24,	-.87,	.60,	.41,	-.14,	-.47,	1.39,	...)
z(Oliver Twist) =	(1.05,	.15,	-.71,	-.56,	.37,	-1.01,	-.06,	-.74,	-.28,	.48,	-.94,	...)

Zeta score

$$z = \frac{X - \bar{X}}{s}$$

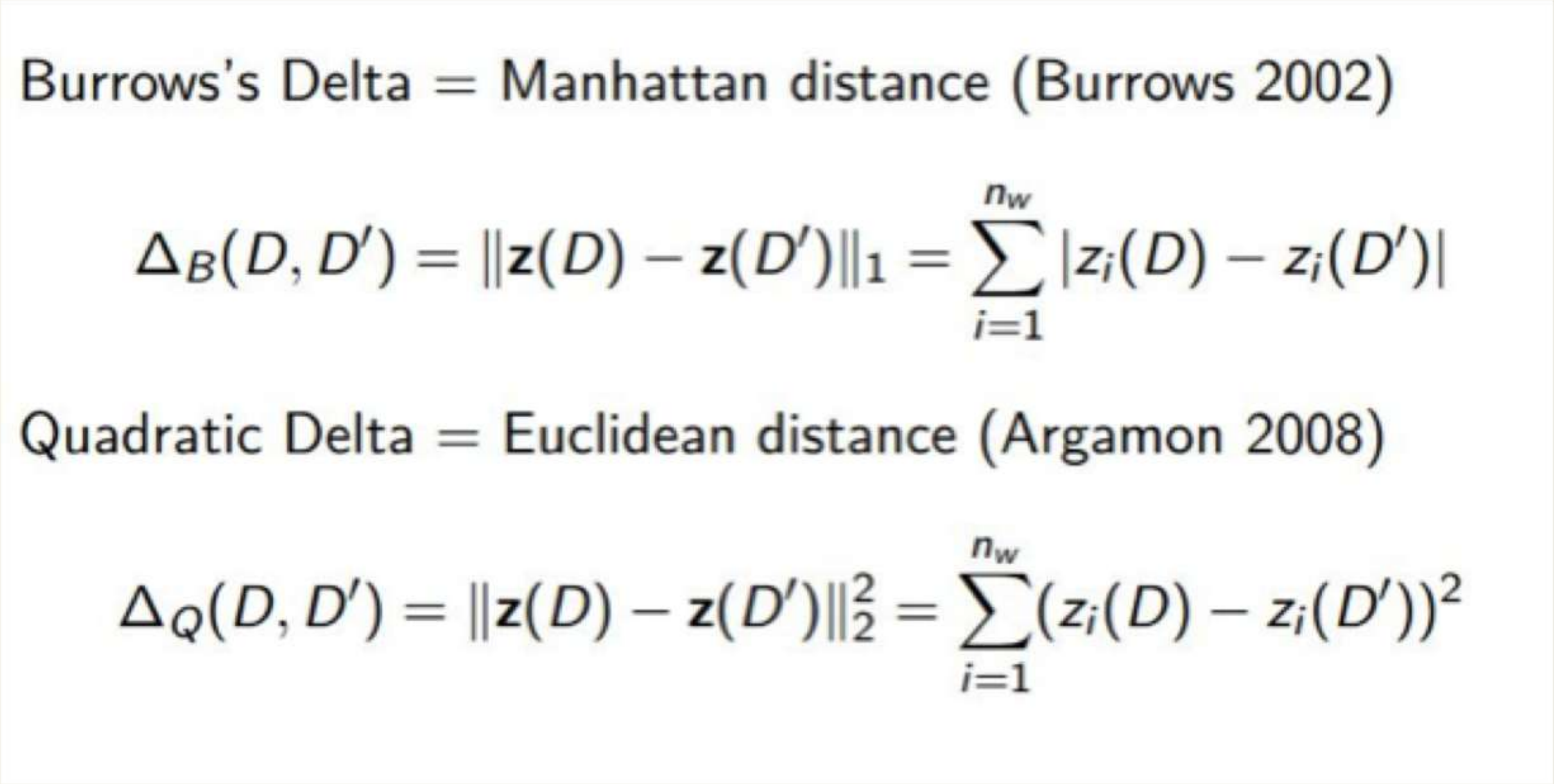
- X - frequency of term
- Mean(X) - mean frequency of term
- S - standard deviation

Distances



- Time
- Manhattan
- Euclidean

DELTA measures



Burrows's Delta = Manhattan distance (Burrows 2002)

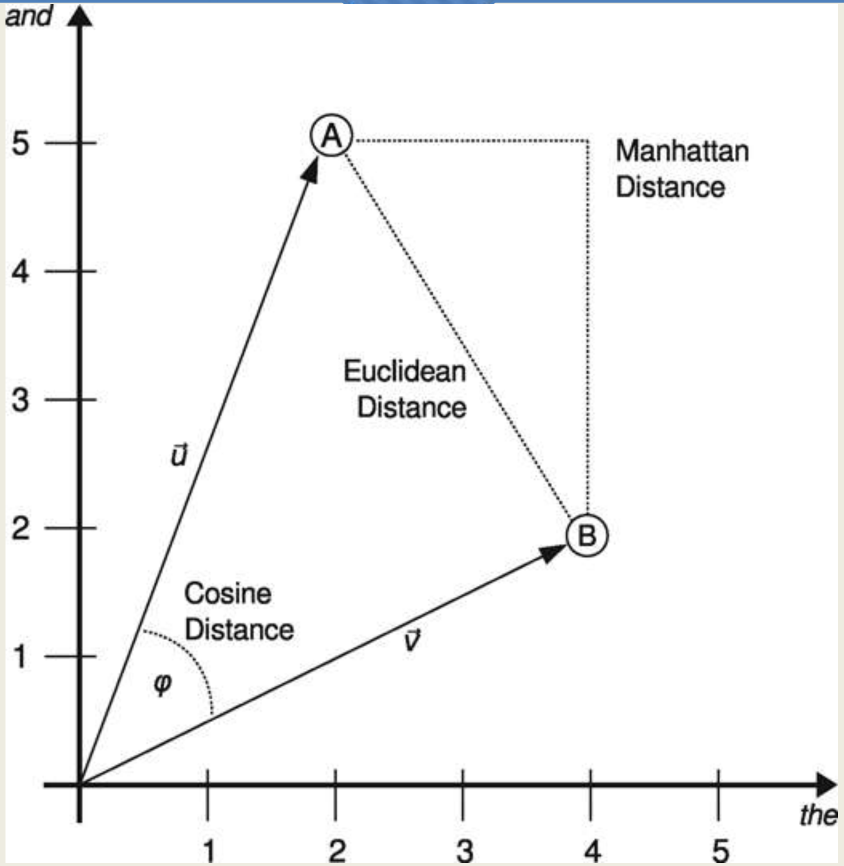
$$\Delta_B(D, D') = \|\mathbf{z}(D) - \mathbf{z}(D')\|_1 = \sum_{i=1}^{n_w} |z_i(D) - z_i(D')|$$

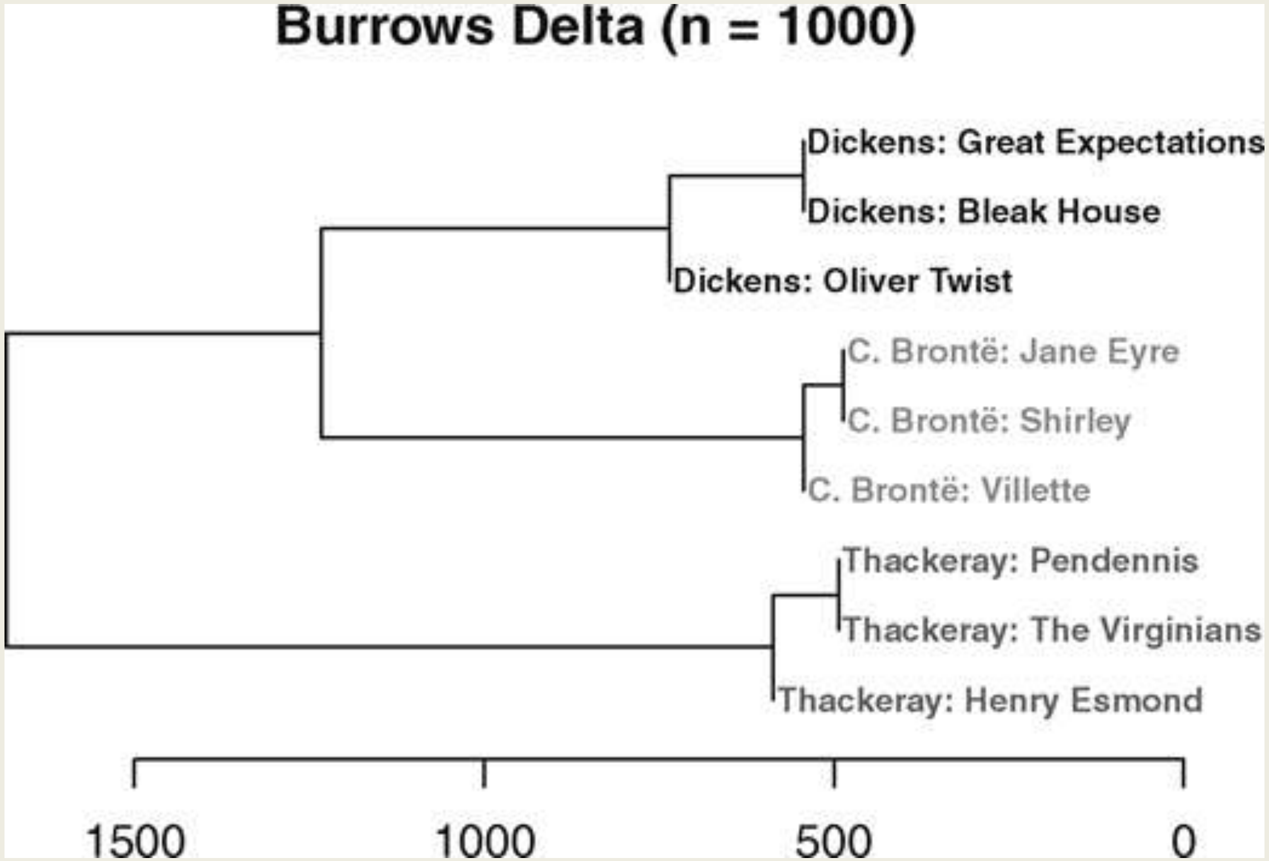
Quadratic Delta = Euclidean distance (Argamon 2008)

$$\Delta_Q(D, D') = \|\mathbf{z}(D) - \mathbf{z}(D')\|_2^2 = \sum_{i=1}^{n_w} (z_i(D) - z_i(D'))^2$$

Deltas & Distance

Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, Thorsten Vitt, "Understanding and explaining Delta measures for authorship attribution", *Digital Scholarship in the Humanities*, Volume 32, Issue suppl_2, December 2017, Pages ii4–ii16, https://doi.org/10.1093/llc/fqx023





Stylo package

https://cran.r-project.org/web/packages/stylo/index.html

https://sites.google.com/site/computationalstylistics/stylo

https://cran.r-project.org/web/packages/stylo/stylo.pdf

https://computationalstylistics.github.io/

10Computational01
01Stylistics0101000
11Group011010110

Stylo needs XQuartz to work on Mac

Command to Start

```
> install.packages("stylo")

> library(stylo)

> stylo()
```

Installing stylo

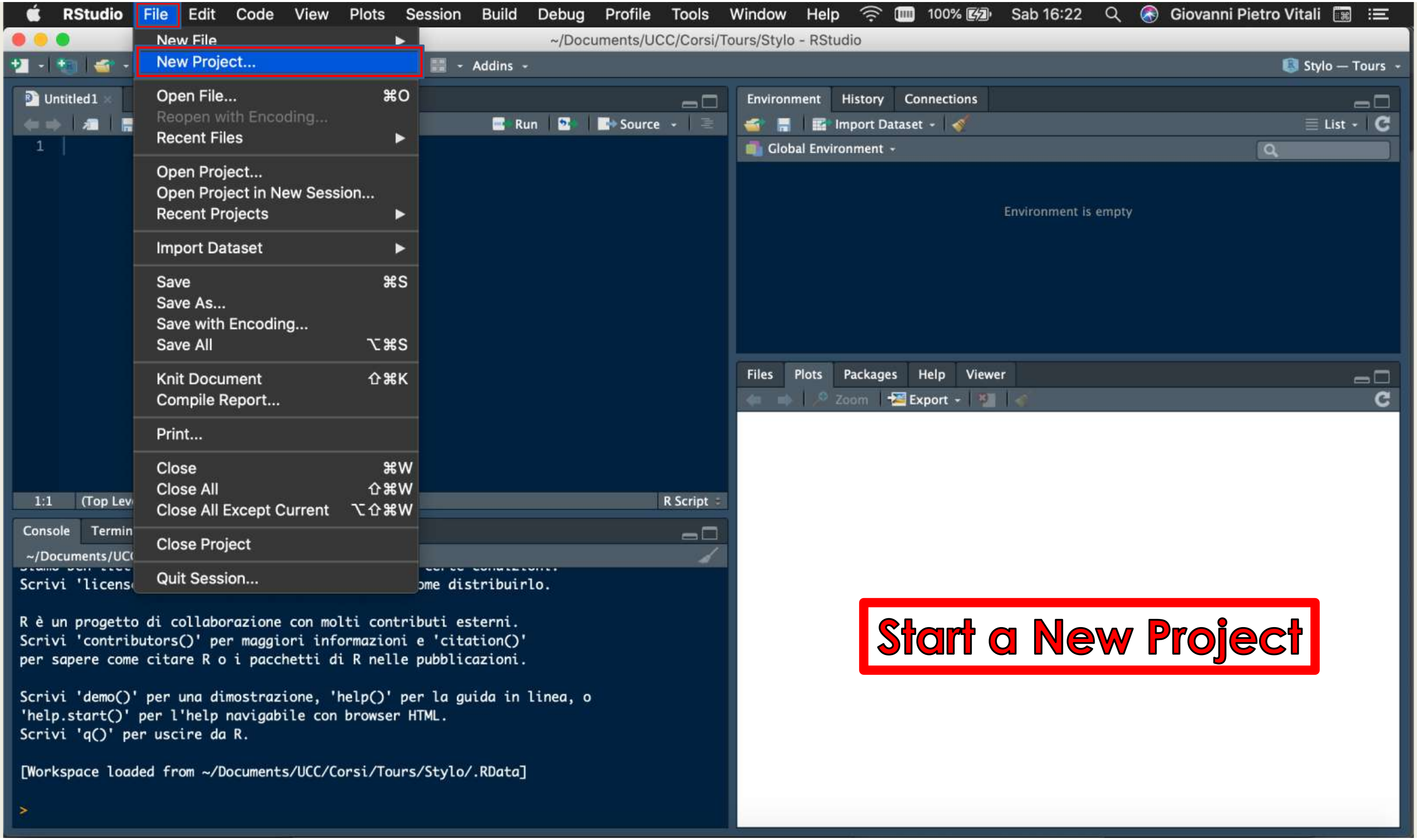
- run R
- type install.packages("stylo")
- pick your R server
- click OK
- done!

Some basic R functions

- to activate a package: library(stylo)
- to set working directory: setwd("path/to/my/stuff")
- to find your current location: getwd()
- to list files in your current location: list.files()
- to get help: help(<function>), e.g. help(stylo)
- to quit R: q()

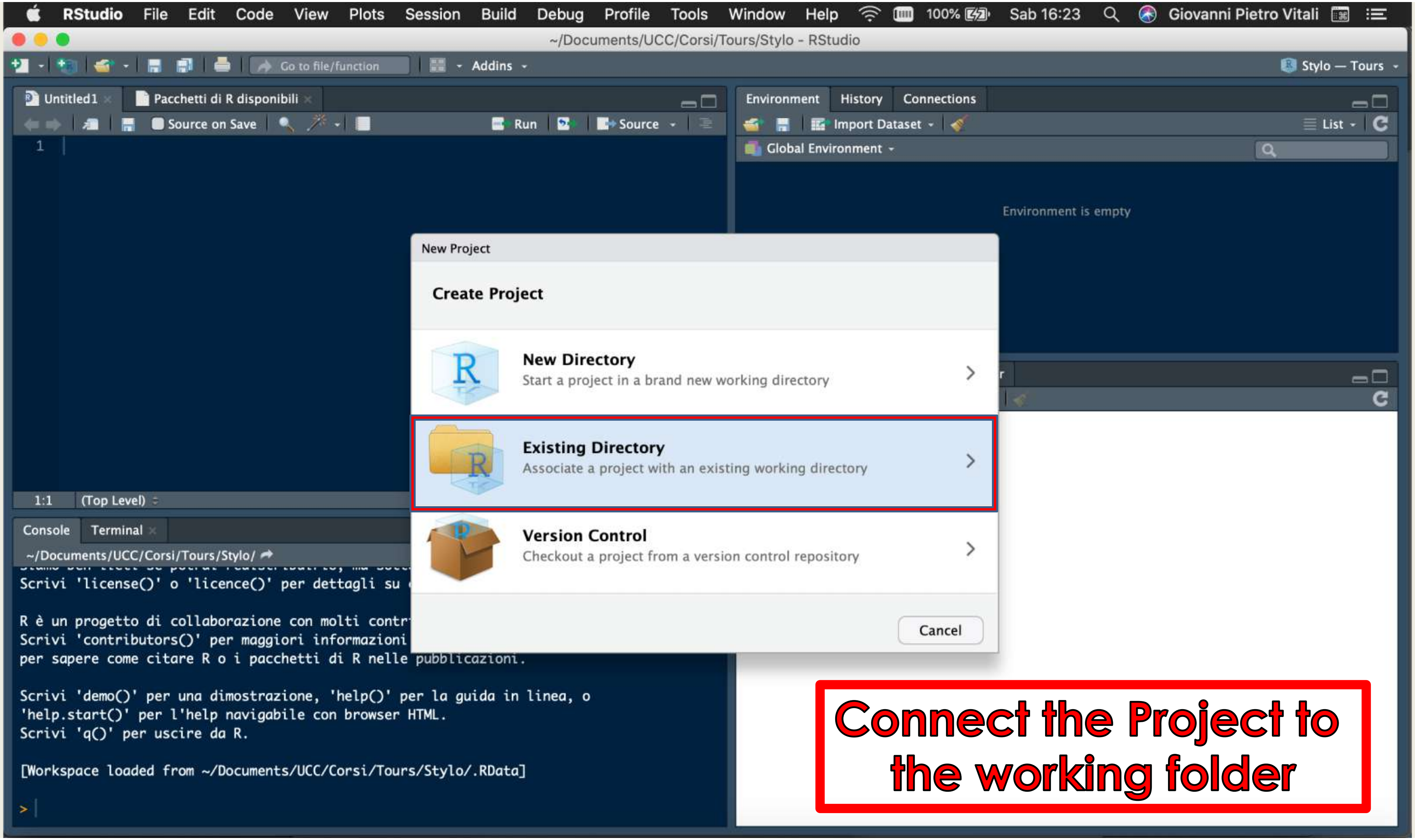
Main functions: stylo()

- It computes distances (differences) between texts, …
- … represented as rows of frequencies of most frequent words.
- Then it plots graphs of those distances:
  - **Cluster Analysis** plots (dendrograms)
  - **Multidimensional Scaling** scatterplots
  - **Principal Components Analysis** scatterplots
  - **Bootstrap Consensus Trees** plots (for multiple parameter settings)
  - **Bootstrap Consensus Networks** (other software will be needed to take over)
- The plots can be both displayed on screen and saved to a file (e.g. PNG).



```
Scrivi 'licens                           ome distribuirlo.

R è un progetto di collaborazione con molti contributi esterni.
Scrivi 'contributors()' per maggiori informazioni e 'citation()'
per sapere come citare R o i pacchetti di R nelle pubblicazioni.

Scrivi 'demo()' per una dimostrazione, 'help()' per la guida in linea, o
'help.start()' per l'help navigabile con browser HTML.
Scrivi 'q()' per uscire da R.

[Workspace loaded from ~/Documents/UCC/Corsi/Tours/Stylo/.RData]

>
```

Start a New Project



New Project

Create Project

New Directory
Start a project in a brand new working directory

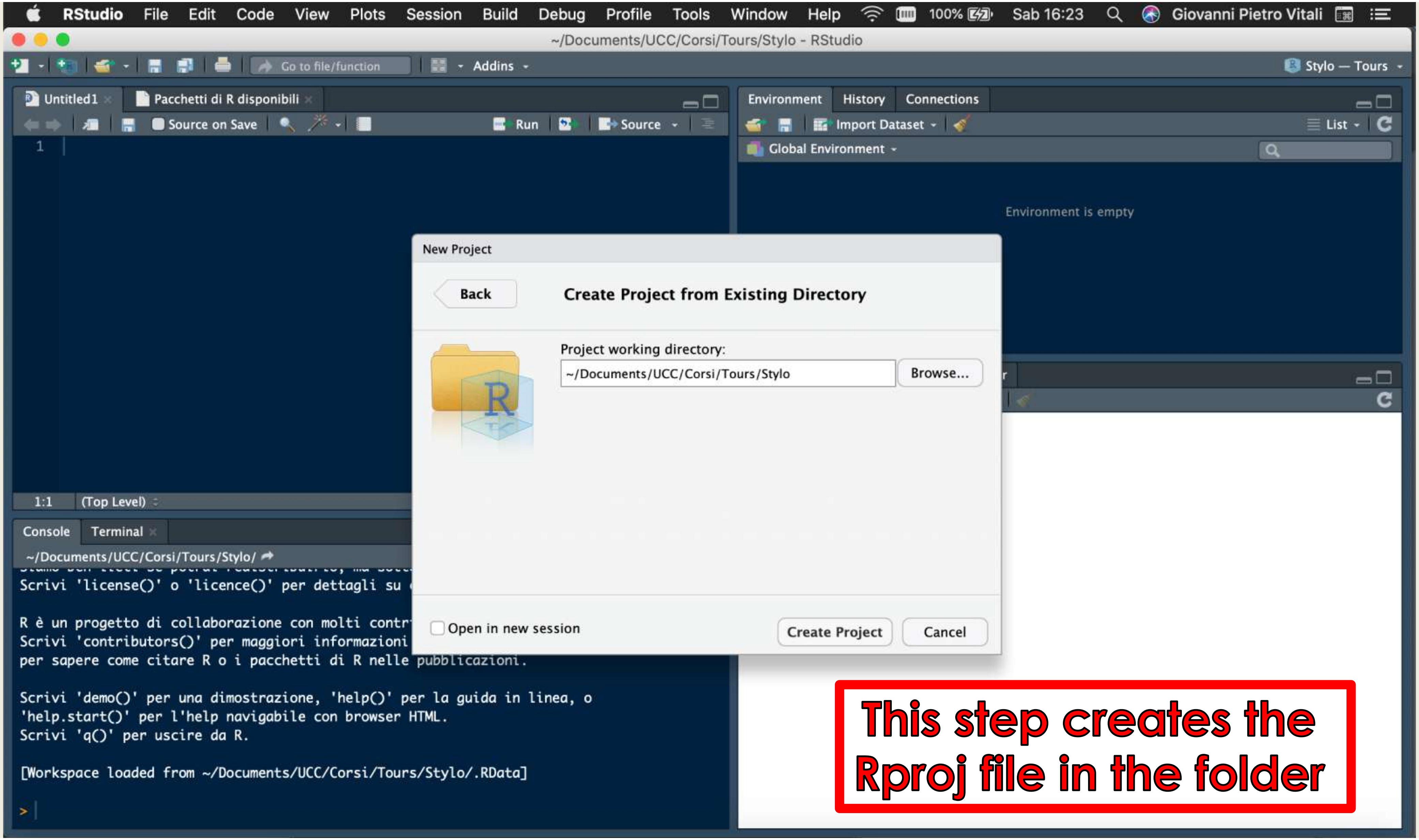
Existing Directory
Associate a project with an existing working directory

Version Control
Checkout a project from a version control repository

Cancel

```
Scrivi 'license()' o 'licence()' per dettagli su c

R è un progetto di collaborazione con molti contr
Scrivi 'contributors()' per maggiori informazioni
per sapere come citare R o i pacchetti di R nelle pubblicazioni.

Scrivi 'demo()' per una dimostrazione, 'help()' per la guida in linea, o
'help.start()' per l'help navigabile con browser HTML.
Scrivi 'q()' per uscire da R.

[Workspace loaded from ~/Documents/UCC/Corsi/Tours/Stylo/.RData]

>
```

Connect the Project to the working folder



Project working directory: ~/Documents/UCC/Corsi/Tours/Stylo

```
~/Documents/UCC/Corsi/Tours/Stylo/
Scrivi 'license()' o 'licence()' per dettagli su

R è un progetto di collaborazione con molti contr
Scrivi 'contributors()' per maggiori informazioni
per sapere come citare R o i pacchetti di R nelle pubblicazioni.

Scrivi 'demo()' per una dimostrazione, 'help()' per la guida in linea, o
'help.start()' per l'help navigabile con browser HTML.
Scrivi 'q()' per uscire da R.

[Workspace loaded from ~/Documents/UCC/Corsi/Tours/Stylo/.RData]

>
```

This step creates the Rproj file in the folder

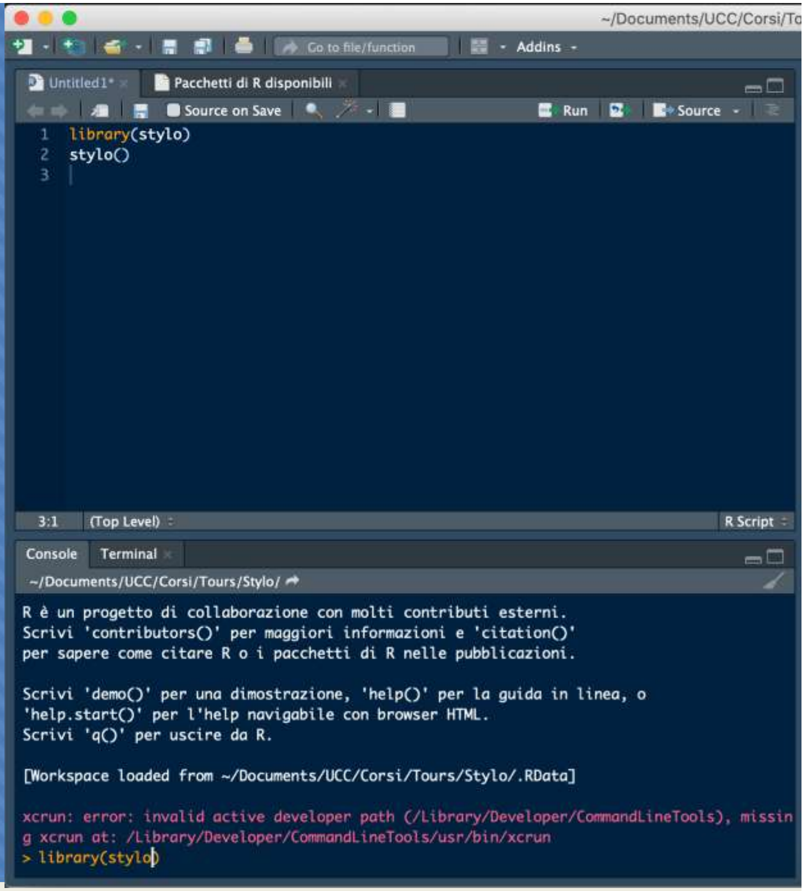
Main functions: stylo()

- It computes distances (differences) between texts, …
- … represented as rows of frequencies of most frequent words.
- Then it plots graphs of those distances:
  - **Cluster Analysis** plots (dendrograms)
  - **Multidimensional Scaling** scatterplots
  - **Principal Components Analysis** scatterplots
  - **Bootstrap Consensus Trees** plots (for multiple parameter settings)
  - **Bootstrap Consensus Networks** (other software will be needed to take over)
- The plots can be both displayed on screen and saved to a file (e.g. PNG).

> library(stylo)

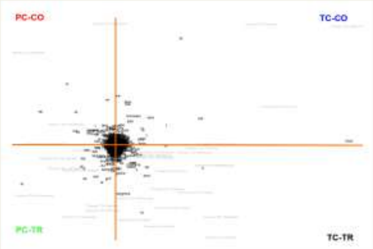
[Then]

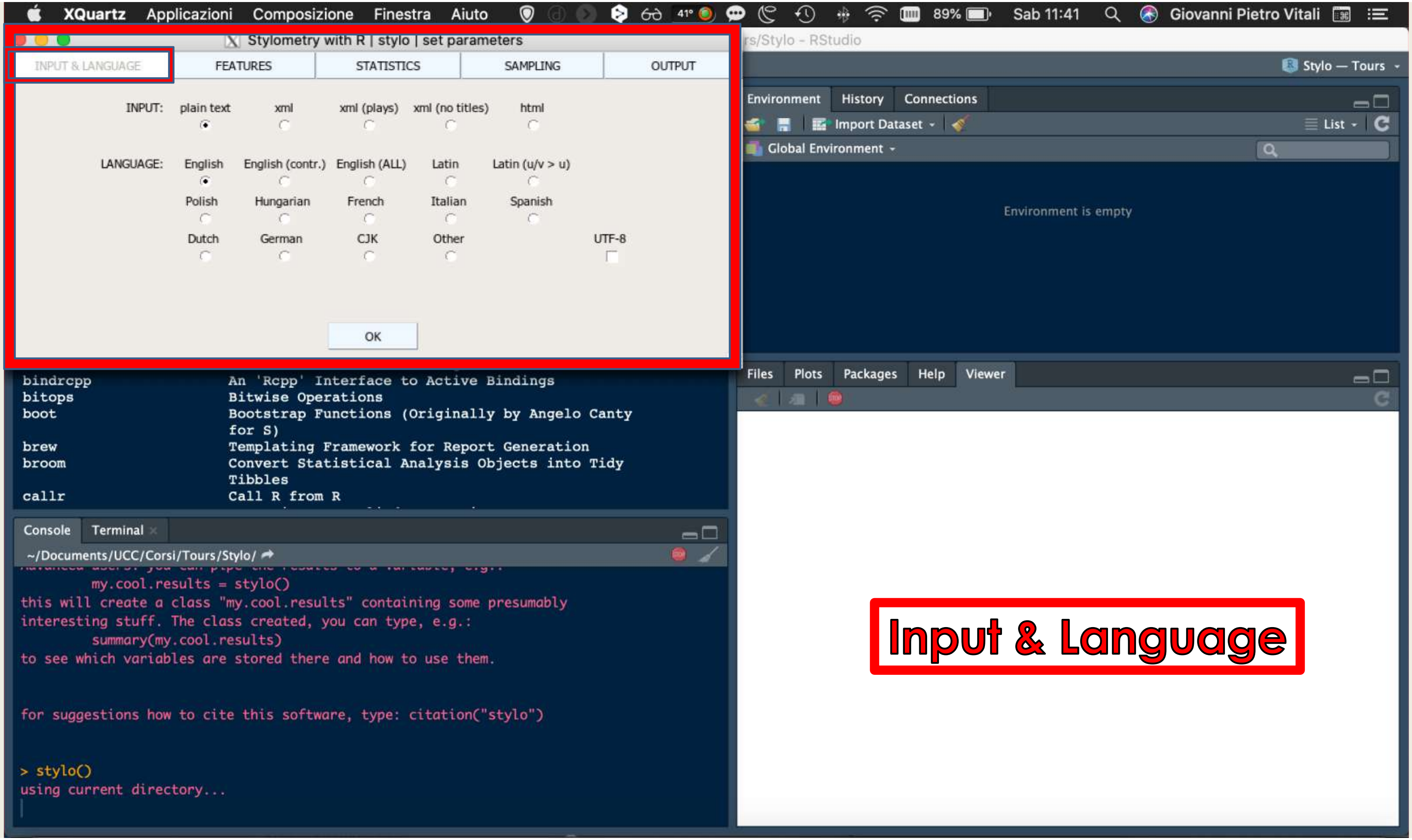
> stylo()



```
library(stylo)
stylo()
```

```
R è un progetto di collaborazione con molti contributi esterni.
Scrivi 'contributors()' per maggiori informazioni e 'citation()'
per sapere come citare R o i pacchetti di R nelle pubblicazioni.

Scrivi 'demo()' per una dimostrazione, 'help()' per la guida in linea, o
'help.start()' per l'help navigabile con browser HTML.
Scrivi 'q()' per uscire da R.

[Workspace loaded from ~/Documents/UCC/Corsi/Tours/Stylo/.RData]

xcrun: error: invalid active developer path (/Library/Developer/CommandLineTools), missin
g xcrun at: /Library/Developer/CommandLineTools/usr/bin/xcrun
> library(stylo)
```





XQuartz Applicazioni Composizione Finestra Aiuto
89% Sab 11:41 Giovanni Pietro Vitali
rs/Stylo - RStudio
Stylo — Tours
Environment History Connections
Import Dataset List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Console Terminal
~/Documents/UCC/Corsi/Tours/Stylo/

Stylometry with R | stylo | set parameters

INPUT & LANGUAGE | FEATURES | STATISTICS | SAMPLING | OUTPUT

INPUT: ● plain text ○ xml ○ xml (plays) ○ xml (no titles) ○ html

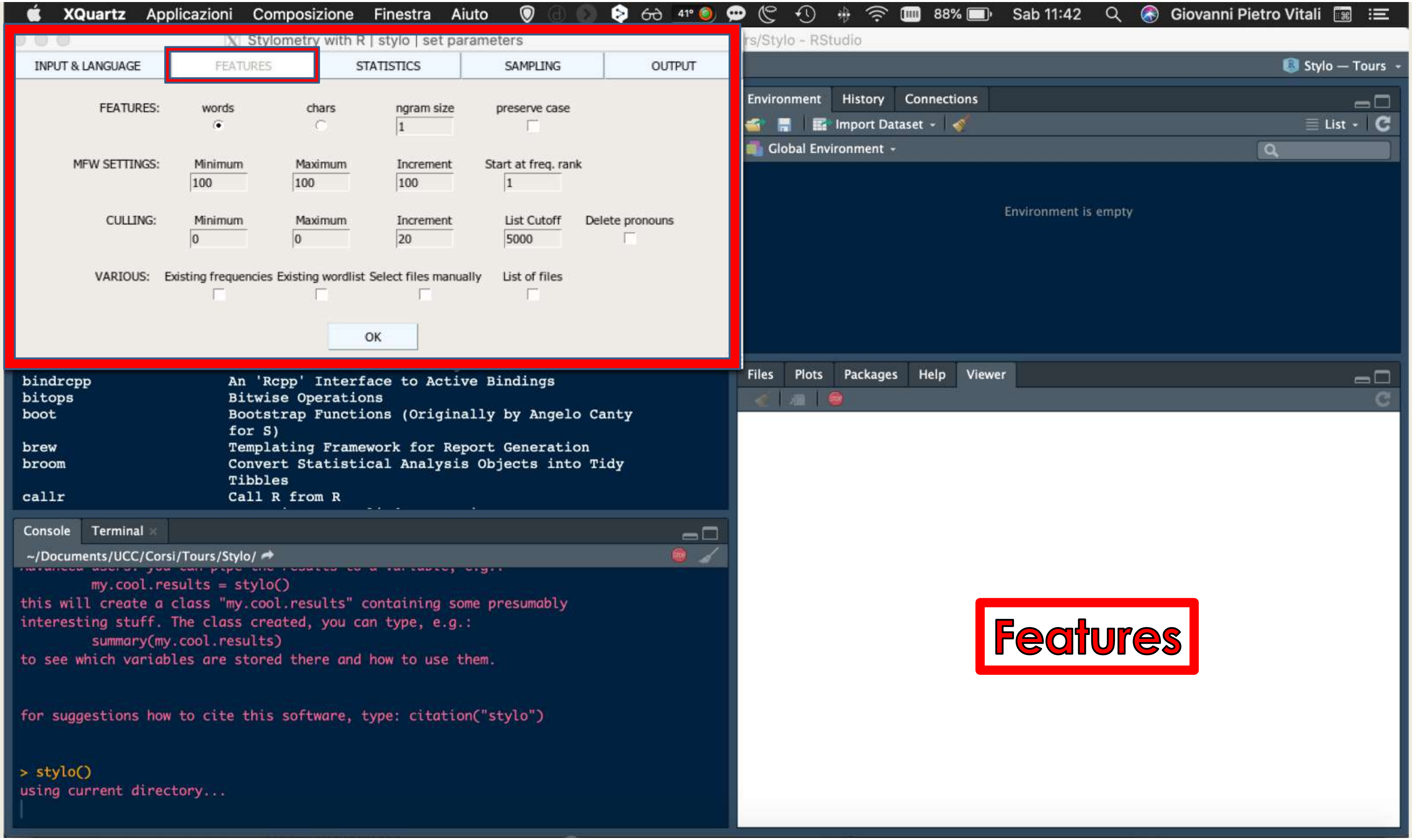
LANGUAGE: ● English ○ English (contr.) ○ English (ALL) ○ Latin ○ Latin (u/v > u)
○ Polish ○ Hungarian ○ French ○ Italian ○ Spanish
○ Dutch ○ German ○ CJK ○ Other □ UTF-8

OK

```
bindrcpp                 An 'Rcpp' Interface to Active Bindings
bitops                   Bitwise Operations
boot                     Bootstrap Functions (Originally by Angelo Canty
                         for S)
brew                     Templating Framework for Report Generation
broom                    Convert Statistical Analysis Objects into Tidy
                         Tibbles
callr                    Call R from R
```

```
        my.cool.results = stylo()
this will create a class "my.cool.results" containing some presumably
interesting stuff. The class created, you can type, e.g.:
        summary(my.cool.results)
to see which variables are stored there and how to use them.


for suggestions how to cite this software, type: citation("stylo")


> stylo()
using current directory...
```

Input & Language



Stylometry with R | stylo | set parameters

INPUT & LANGUAGE | FEATURES | STATISTICS | SAMPLING | OUTPUT

FEATURES: words ● chars ○ ngram size 1 preserve case □

MFW SETTINGS: Minimum 100 Maximum 100 Increment 100 Start at freq. rank 1

CULLING: Minimum 0 Maximum 0 Increment 20 List Cutoff 5000 Delete pronouns □

VARIOUS: Existing frequencies □ Existing wordlist □ Select files manually □ List of files □

OK

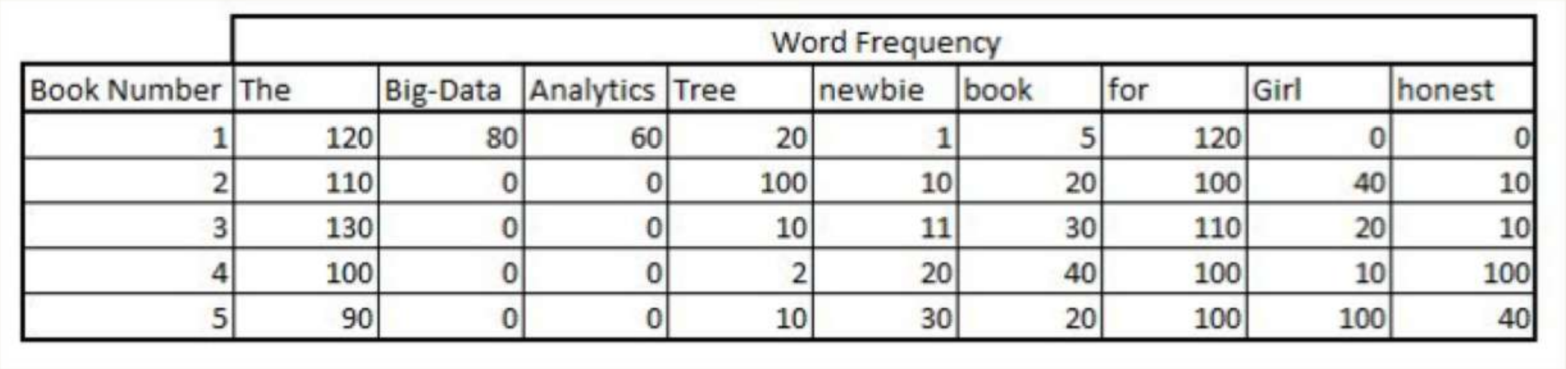
```
bindrcpp                An 'Rcpp' Interface to Active Bindings
bitops                  Bitwise Operations
boot                    Bootstrap Functions (Originally by Angelo Canty
                        for S)
brew                    Templating Framework for Report Generation
broom                   Convert Statistical Analysis Objects into Tidy
                        Tibbles
callr                   Call R from R
```

```
Advanced users: you can pipe the results to a variable, e.g.:
        my.cool.results = stylo()
this will create a class "my.cool.results" containing some presumably
interesting stuff. The class created, you can type, e.g.:
        summary(my.cool.results)
to see which variables are stored there and how to use them.


for suggestions how to cite this software, type: citation("stylo")


> stylo()
using current directory...
```

Features

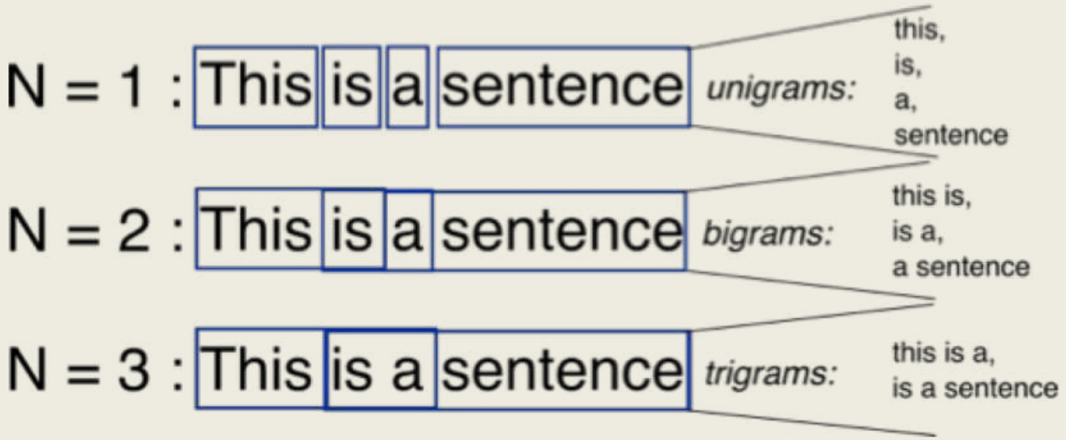
Features

- words: words are used as the unit.
  - characters: characters are used as the unit.
  - *n*-gram size: this is where you can specify the
- value of *n* for your *n*-grams
  - preserve case: normally, all the words from
- the input texts are turned into lowercase



	Word Frequency								
Book Number	The	Big-Data	Analytics	Tree	newbie	book	for	Girl	honest
1	120	80	60	20	1	5	120	0	0
2	110	0	0	100	10	20	100	40	10
3	130	0	0	10	11	30	110	20	10
4	100	0	0	2	20	40	100	10	100
5	90	0	0	10	30	20	100	100	40

WHAT IS N-GRAM

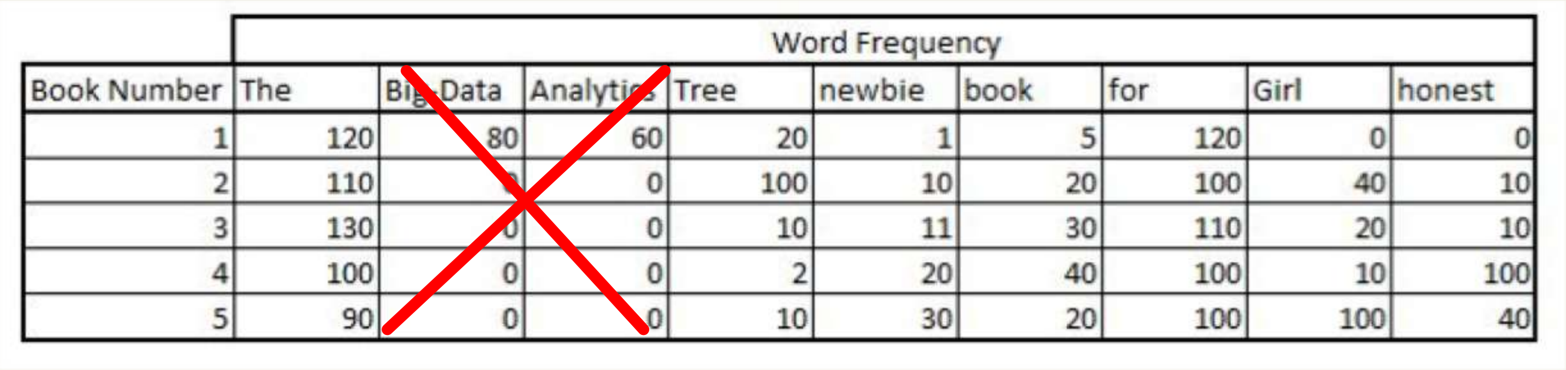


MFV (most-frequent-word) settings

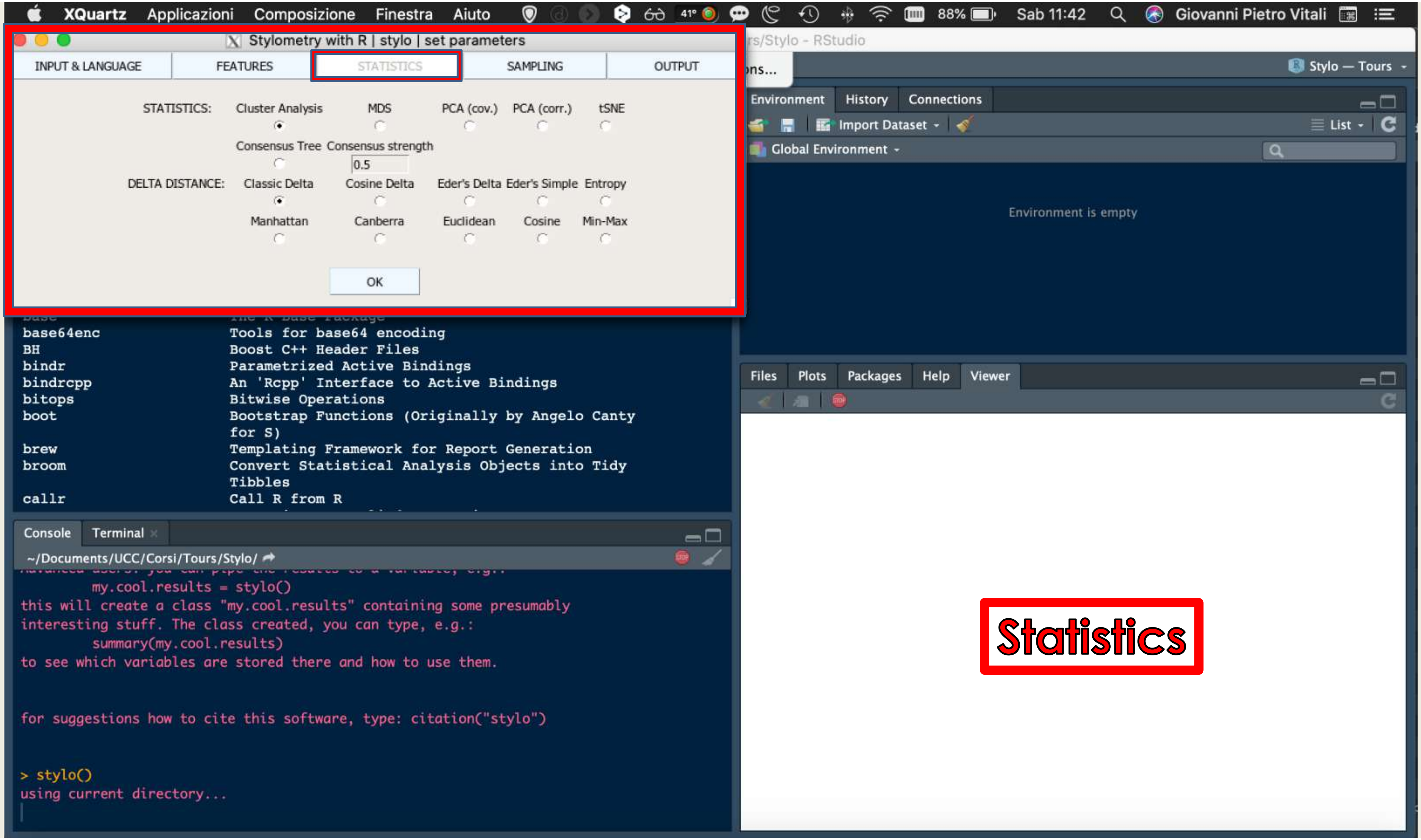
- **Minimum**: this setting determines how many words (or features) from the top of the frequency list will be used
- **Maximum**: this setting determines how many words from the top of the word frequency list for the entire corpus will be used
- **Increment**: defines the value by which the value of Minimum will be increased at each subsequent run of your analysis
- **Start at freq. Rank**: how many words from the top overall frequency rank list to be skipped

Culling

The culling values specify the degree to which words that do not appear in all the texts of your corpus will be removed. Thus, a culling value of 20 indicates that words that appear in at least 20% of the texts in the corpus will be considered in the analysis. A culling setting of 0 means that no words will be removed.



	Word Frequency								
Book Number	The	Big Data	Analytics	Tree	newbie	book	for	Girl	honest
1	120	80	60	20	1	5	120	0	0
2	110	0	0	100	10	20	100	40	10
3	130	0	0	10	11	30	110	20	10
4	100	0	0	2	20	40	100	10	100
5	90	0	0	10	30	20	100	100	40



INPUT & LANGUAGE | FEATURES | STATISTICS | SAMPLING | OUTPUT

STATISTICS: Cluster Analysis ● MDS ○ PCA (cov.) ○ PCA (corr.) ○ tSNE ○

Consensus Tree ○ Consensus strength 0.5

DELTA DISTANCE: Classic Delta ● Cosine Delta ○ Eder's Delta ○ Eder's Simple ○ Entropy ○

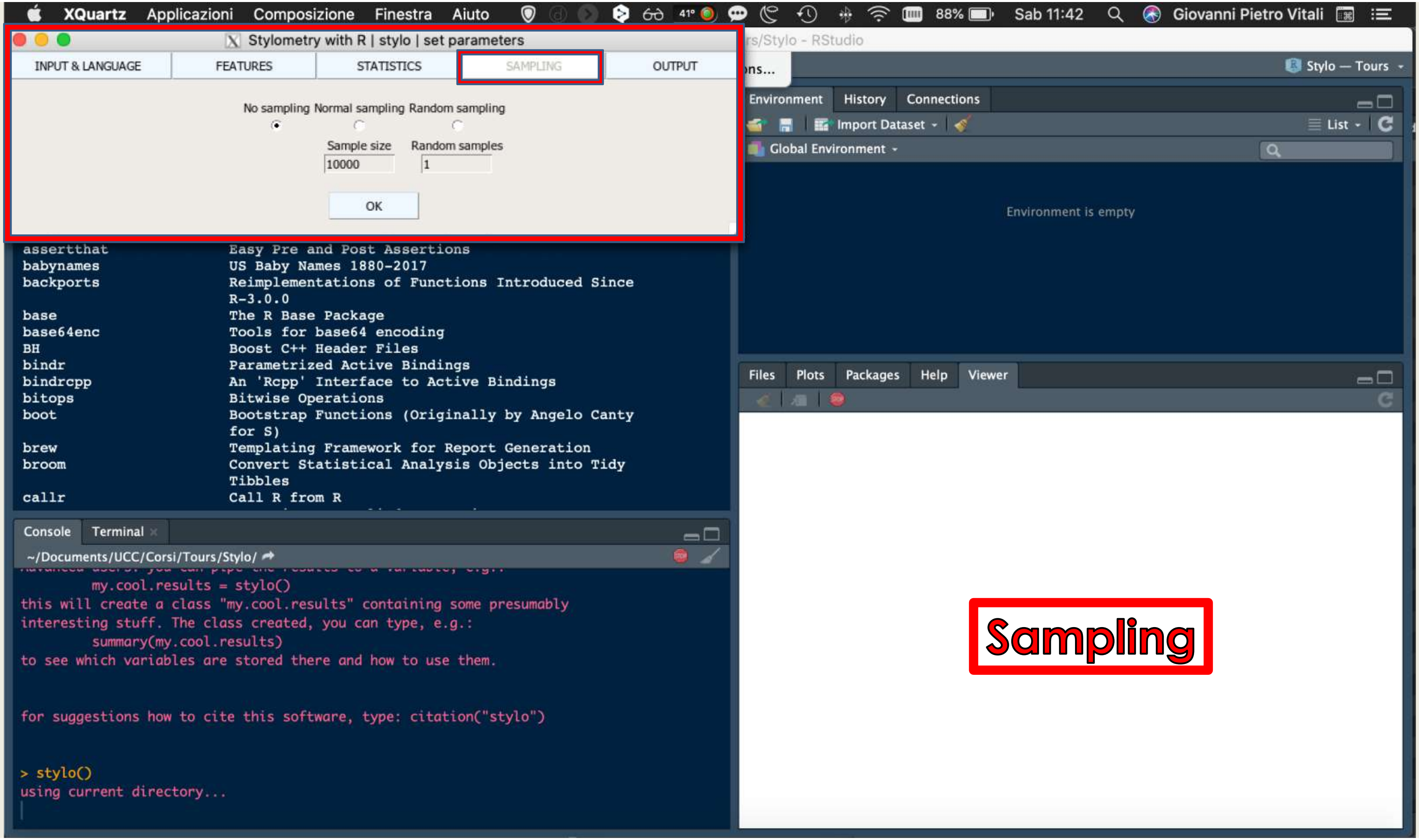
Manhattan ○ Canberra ○ Euclidean ○ Cosine ○ Min-Max ○

OK

```
base64enc                Tools for base64 encoding
BH                       Boost C++ Header Files
bindr                    Parametrized Active Bindings
bindrcpp                 An 'Rcpp' Interface to Active Bindings
bitops                   Bitwise Operations
boot                     Bootstrap Functions (Originally by Angelo Canty
                         for S)
brew                     Templating Framework for Report Generation
broom                    Convert Statistical Analysis Objects into Tidy
                         Tibbles
callr                    Call R from R
```

```
        my.cool.results = stylo()
this will create a class "my.cool.results" containing some presumably
interesting stuff. The class created, you can type, e.g.:
        summary(my.cool.results)
to see which variables are stored there and how to use them.


for suggestions how to cite this software, type: citation("stylo")


> stylo()
using current directory...
```

Statistics



No sampling ● Normal sampling ○ Random sampling ○

Sample size: 10000

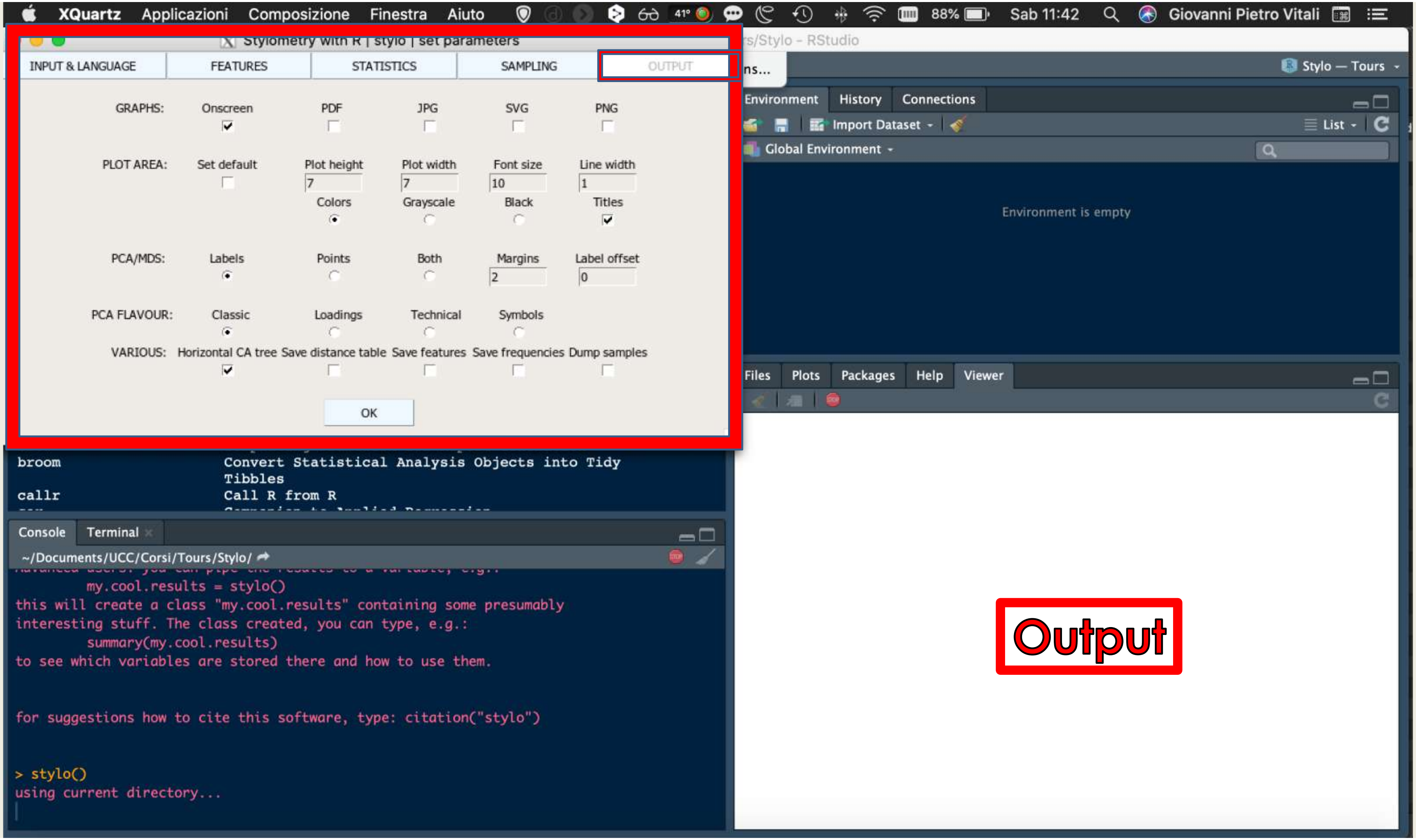
Random samples: 1

OK

```
assertthat                 Easy Pre and Post Assertions
babynames                  US Baby Names 1880-2017
backports                  Reimplementations of Functions Introduced Since
                           R-3.0.0
base                       The R Base Package
base64enc                  Tools for base64 encoding
BH                         Boost C++ Header Files
bindr                      Parametrized Active Bindings
bindrcpp                   An 'Rcpp' Interface to Active Bindings
bitops                     Bitwise Operations
boot                       Bootstrap Functions (Originally by Angelo Canty
                           for S)
brew                       Templating Framework for Report Generation
broom                      Convert Statistical Analysis Objects into Tidy
                           Tibbles
callr                      Call R from R
```

```
        my.cool.results = stylo()
this will create a class "my.cool.results" containing some presumably
interesting stuff. The class created, you can type, e.g.:
        summary(my.cool.results)
to see which variables are stored there and how to use them.


for suggestions how to cite this software, type: citation("stylo")


> stylo()
using current directory...
```

Sampling



INPUT & LANGUAGE | FEATURES | STATISTICS | SAMPLING | OUTPUT

GRAPHS: Onscreen ☑ PDF □ JPG □ SVG □ PNG □

PLOT AREA: Set default □ Plot height 7 Plot width 7 Font size 10 Line width 1

Colors ● Grayscale ○ Black ○ Titles ☑

PCA/MDS: Labels ● Points ○ Both ○ Margins 2 Label offset 0

PCA FLAVOUR: Classic ● Loadings ○ Technical ○ Symbols ○

VARIOUS: Horizontal CA tree ☑ Save distance table □ Save features □ Save frequencies □ Dump samples □

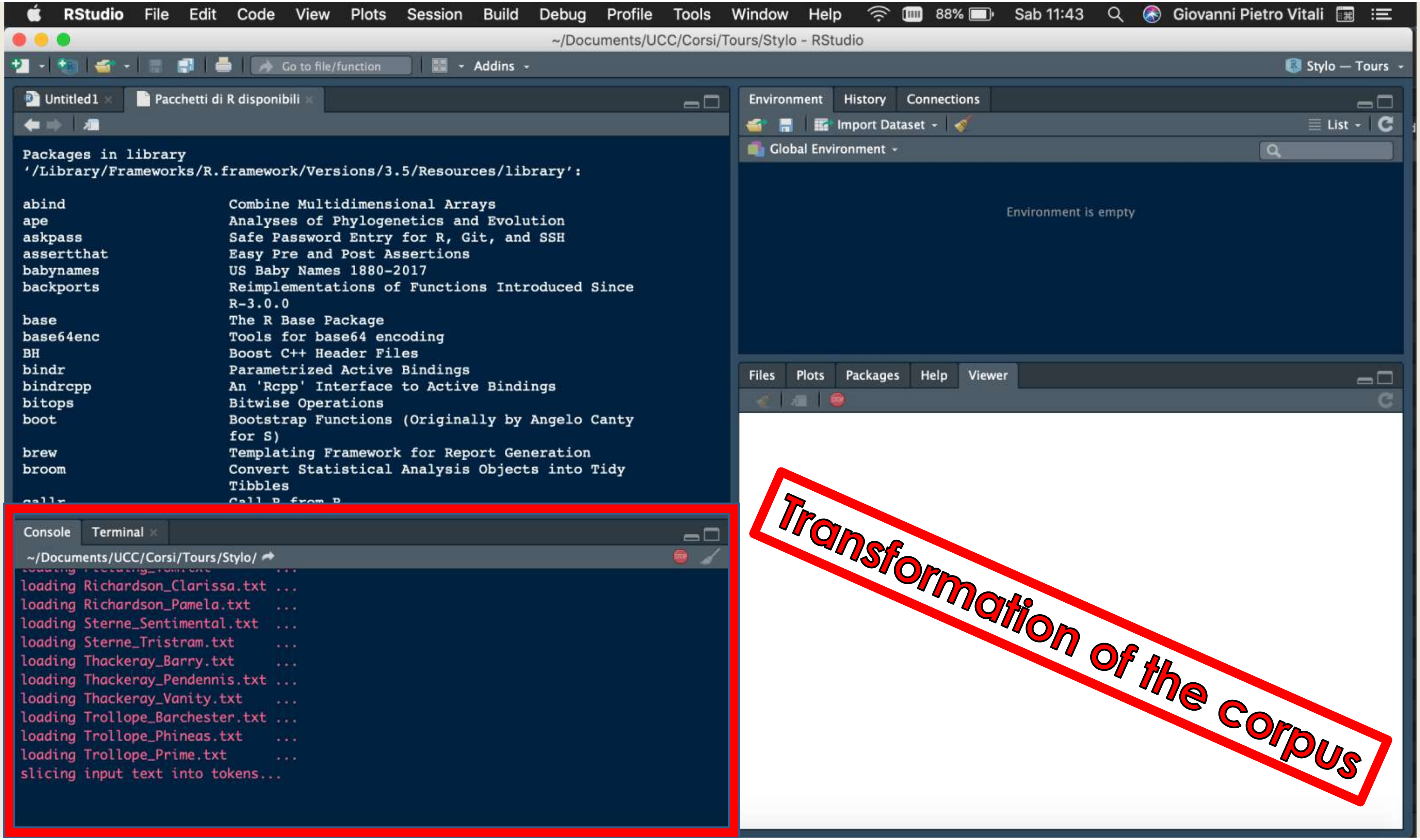
OK

```
broom                     Convert Statistical Analysis Objects into Tidy
                          Tibbles
callr                     Call R from R
```

~/Documents/UCC/Corsi/Tours/Stylo/

```
         my.cool.results = stylo()
this will create a class "my.cool.results" containing some presumably
interesting stuff. The class created, you can type, e.g.:
         summary(my.cool.results)
to see which variables are stored there and how to use them.


for suggestions how to cite this software, type: citation("stylo")


> stylo()
using current directory...
```

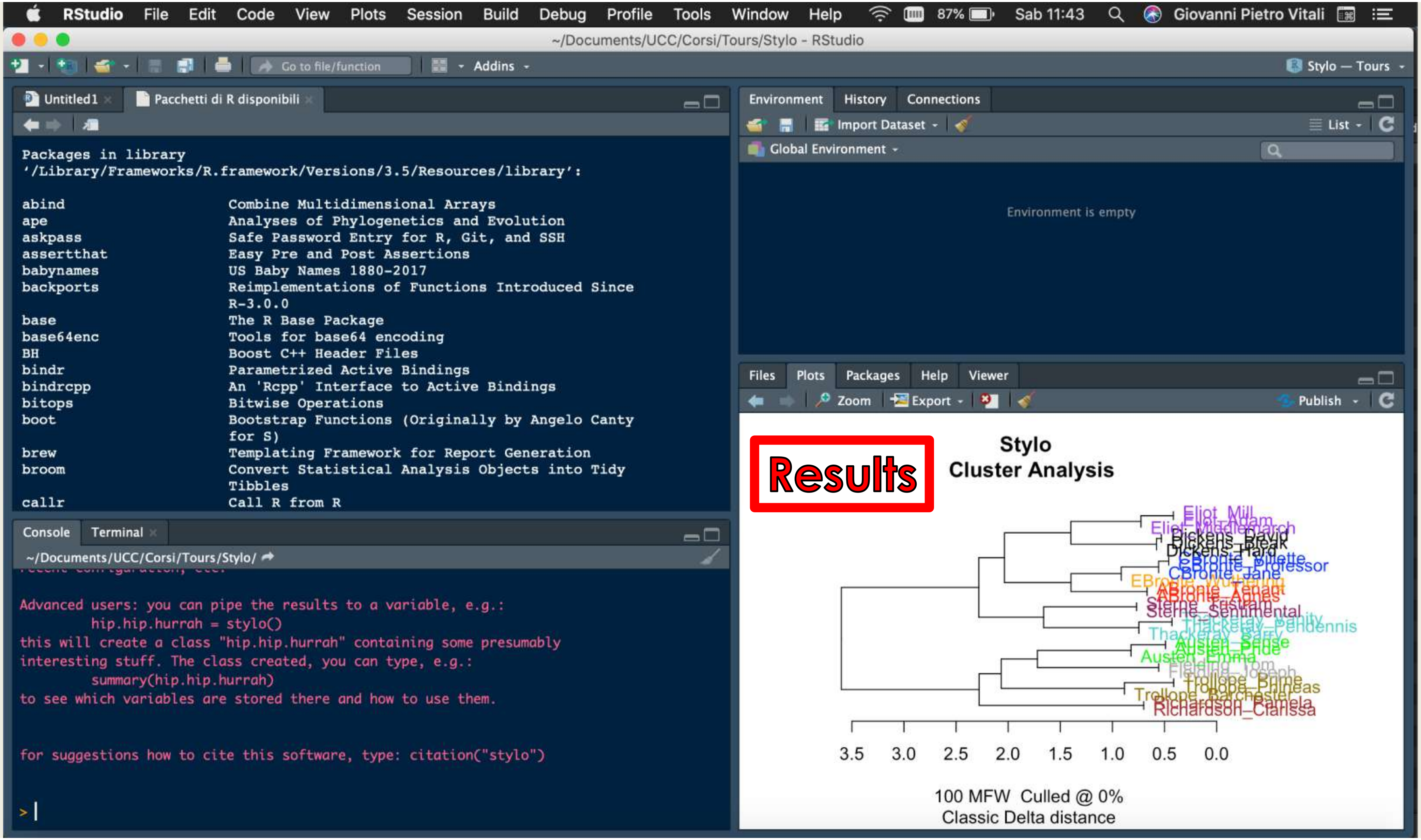
Output



```
Packages in library
'/Library/Frameworks/R.framework/Versions/3.5/Resources/library':

abind                   Combine Multidimensional Arrays
ape                     Analyses of Phylogenetics and Evolution
askpass                 Safe Password Entry for R, Git, and SSH
assertthat              Easy Pre and Post Assertions
babynames               US Baby Names 1880-2017
backports               Reimplementations of Functions Introduced Since
                        R-3.0.0
base                    The R Base Package
base64enc               Tools for base64 encoding
BH                      Boost C++ Header Files
bindr                   Parametrized Active Bindings
bindrcpp                An 'Rcpp' Interface to Active Bindings
bitops                  Bitwise Operations
boot                    Bootstrap Functions (Originally by Angelo Canty
                        for S)
brew                    Templating Framework for Report Generation
broom                   Convert Statistical Analysis Objects into Tidy
                        Tibbles
callr                   Call R from R
```

```
loading Richardson_Clarissa.txt ...
loading Richardson_Pamela.txt   ...
loading Sterne_Sentimental.txt  ...
loading Sterne_Tristram.txt     ...
loading Thackeray_Barry.txt     ...
loading Thackeray_Pendennis.txt ...
loading Thackeray_Vanity.txt    ...
loading Trollope_Barchester.txt ...
loading Trollope_Phineas.txt    ...
loading Trollope_Prime.txt      ...
slicing input text into tokens...
```

Transformation of the corpus



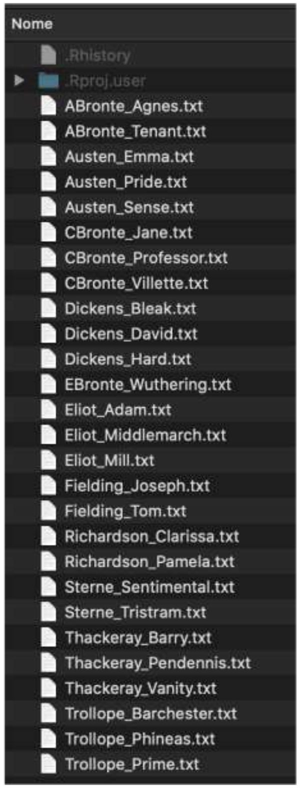
```
Packages in library
'/Library/Frameworks/R.framework/Versions/3.5/Resources/library':

abind                   Combine Multidimensional Arrays
ape                     Analyses of Phylogenetics and Evolution
askpass                 Safe Password Entry for R, Git, and SSH
assertthat              Easy Pre and Post Assertions
babynames               US Baby Names 1880-2017
backports               Reimplementations of Functions Introduced Since
                        R-3.0.0
base                    The R Base Package
base64enc               Tools for base64 encoding
BH                      Boost C++ Header Files
bindr                   Parametrized Active Bindings
bindrcpp                An 'Rcpp' Interface to Active Bindings
bitops                  Bitwise Operations
boot                    Bootstrap Functions (Originally by Angelo Canty
                        for S)
brew                    Templating Framework for Report Generation
broom                   Convert Statistical Analysis Objects into Tidy
                        Tibbles
callr                   Call R from R
```

```
recent configuration, etc.

Advanced users: you can pipe the results to a variable, e.g.:
        hip.hip.hurrah = stylo()
this will create a class "hip.hip.hurrah" containing some presumably
interesting stuff. The class created, you can type, e.g.:
        summary(hip.hip.hurrah)
to see which variables are stored there and how to use them.


for suggestions how to cite this software, type: citation("stylo")

> |
```

Results

Cluster analysis

Builds "tree" based on the most similar texts based on the MFV. It is not robust on the changes of the parameters.



Nome

.Rhistory

.Rproj.user

ABronte_Agnes.txt

ABronte_Tenant.txt

Austen_Emma.txt

Austen_Pride.txt

Austen_Sense.txt

CBronte_Jane.txt

CBronte_Professor.txt

CBronte_Villette.txt

Dickens_Bleak.txt

Dickens_David.txt

Dickens_Hard.txt

EBronte_Wuthering.txt

Eliot_Adam.txt

Eliot_Middlemarch.txt

Eliot_Mill.txt

Fielding_Joseph.txt

Fielding_Tom.txt

Richardson_Clarissa.txt

Richardson_Pamela.txt

Sterne_Sentimental.txt

Sterne_Tristram.txt

Thackeray_Barry.txt

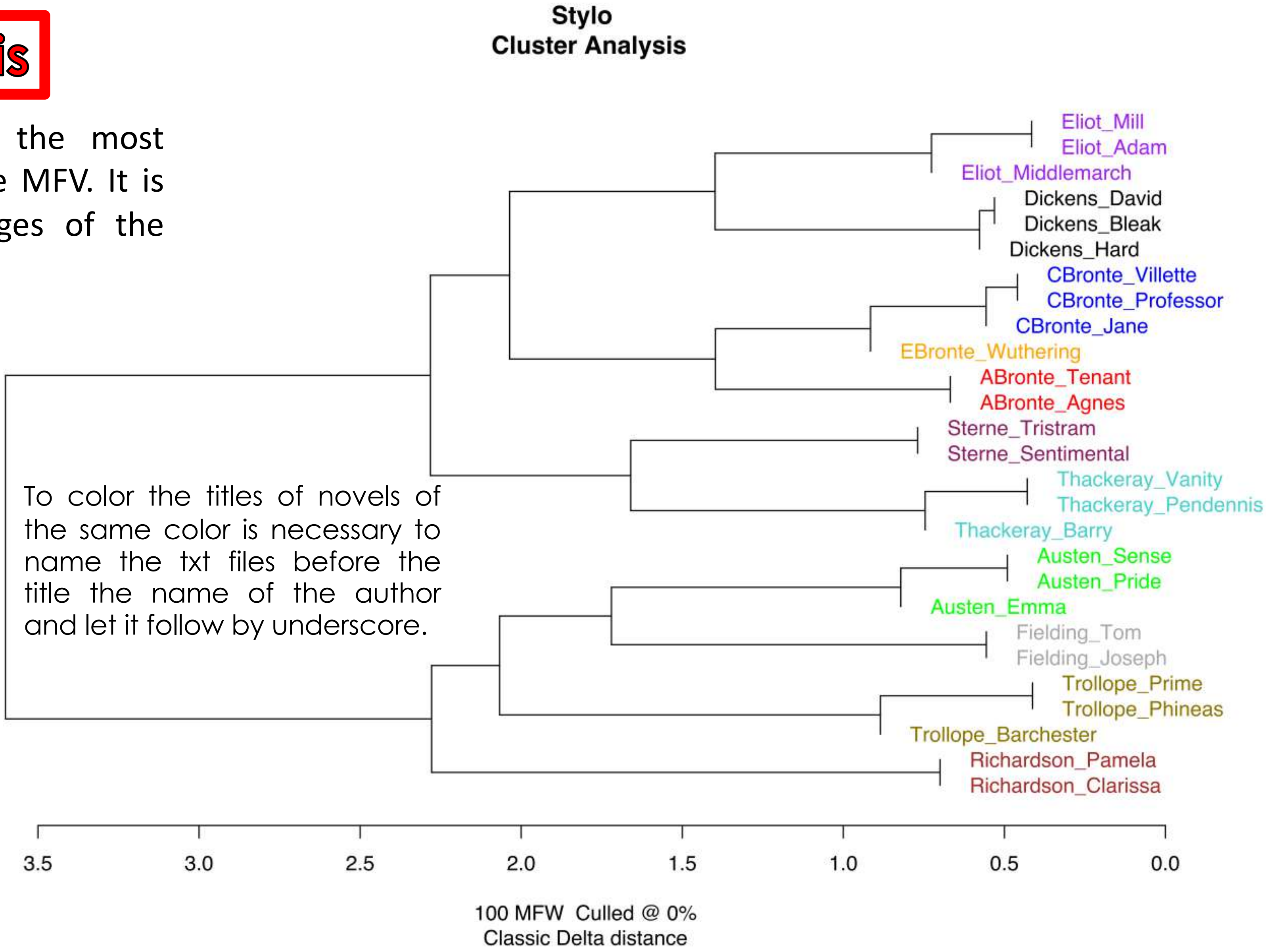
Thackeray_Pendennis.txt

Thackeray_Vanity.txt

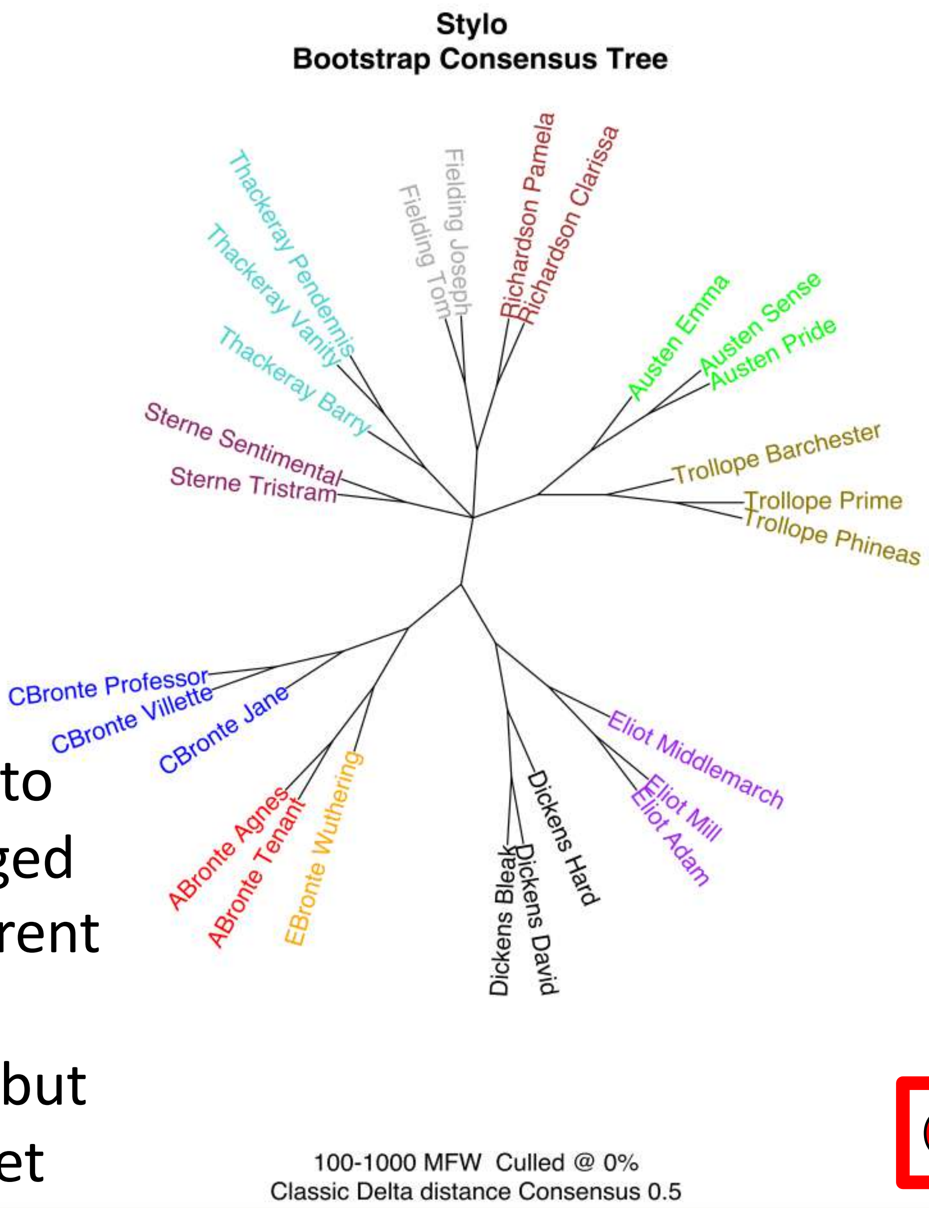
Trollope_Barchester.txt

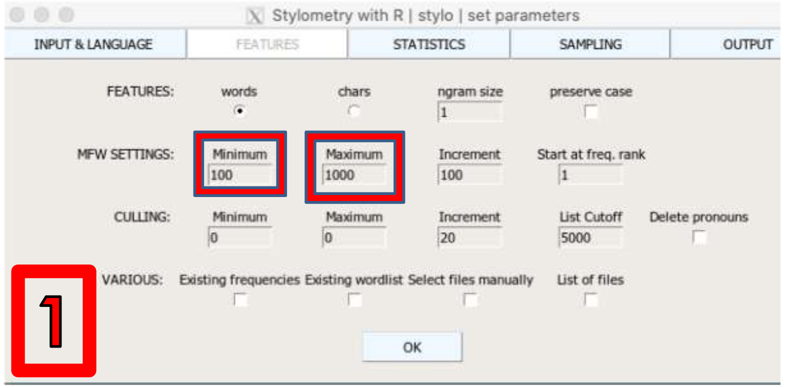
Trollope_Phineas.txt

Trollope_Prime.txt



To color the titles of novels of the same color is necessary to name the txt files before the title the name of the author and let it follow by underscore.



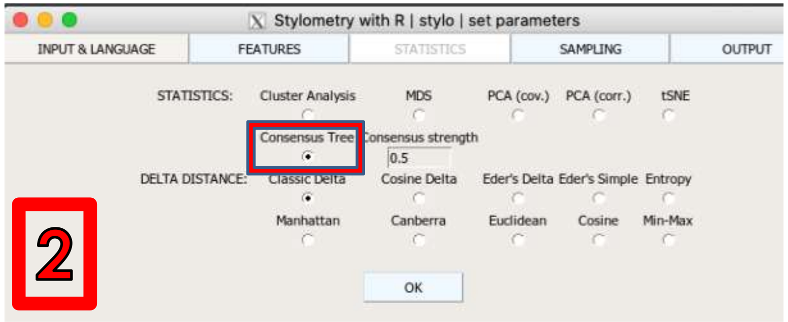


FEATURES: words ● chars ○ ngram size 1 preserve case □

MFW SETTINGS: Minimum 100 Maximum 1000 Increment 100 Start at freq. rank 1

CULLING: Minimum 0 Maximum 0 Increment 20 List Cutoff 5000 Delete pronouns □

VARIOUS: Existing frequencies □ Existing wordlist □ Select files manually □ List of files □



STATISTICS: Cluster Analysis ○ MDS ○ PCA (cov.) ○ PCA (corr.) ○ tSNE ○

Consensus Tree ● Consensus strength 0.5

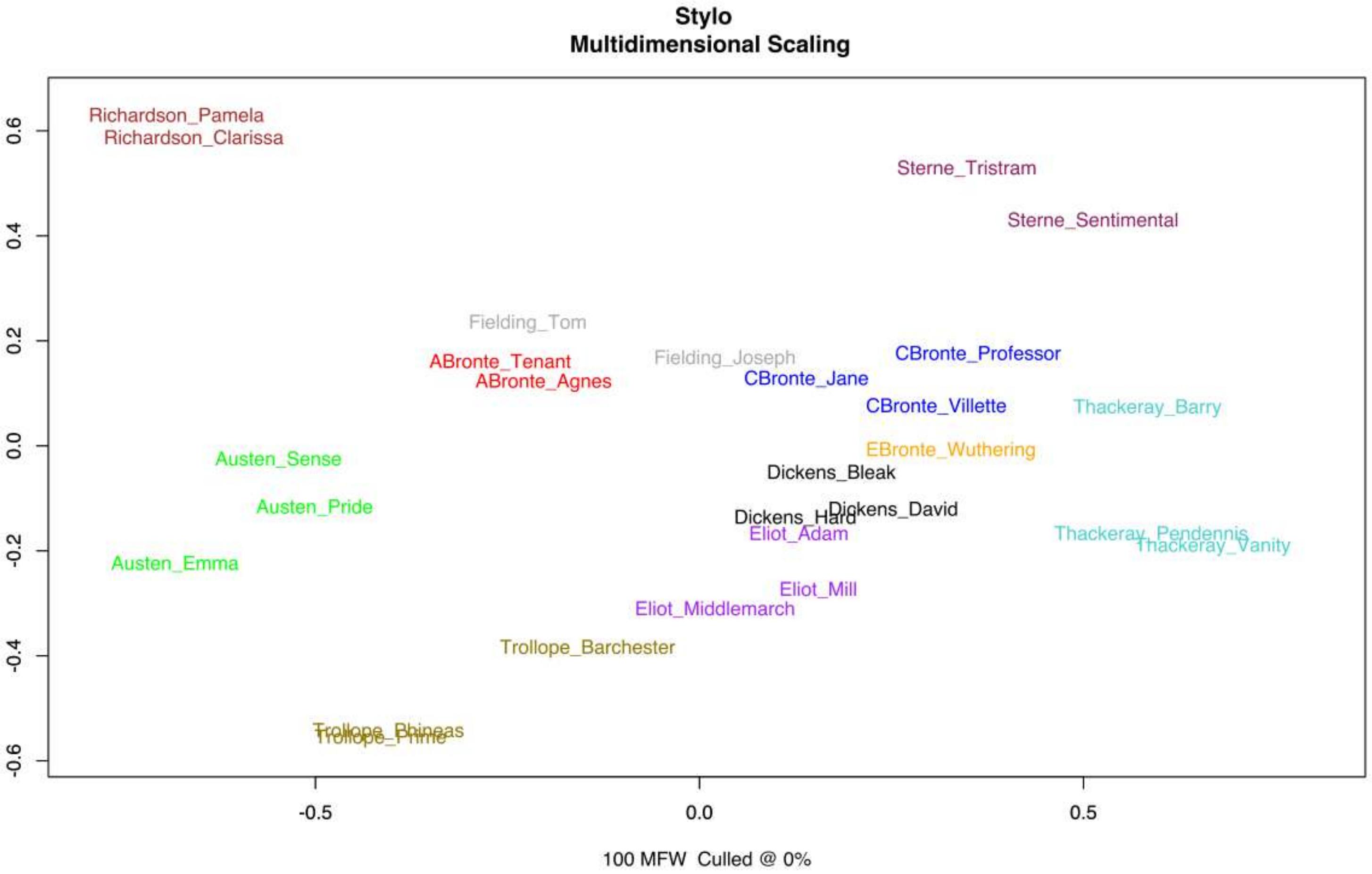
DELTA DISTANCE: Classic Delta ● Cosine Delta ○ Eder's Delta ○ Eder's Simple ○ Entropy ○

Manhattan ○ Canberra ○ Euclidean ○ Cosine ○ Min-Max ○

Uses many trees to discover unchanged patterns for different parameters.

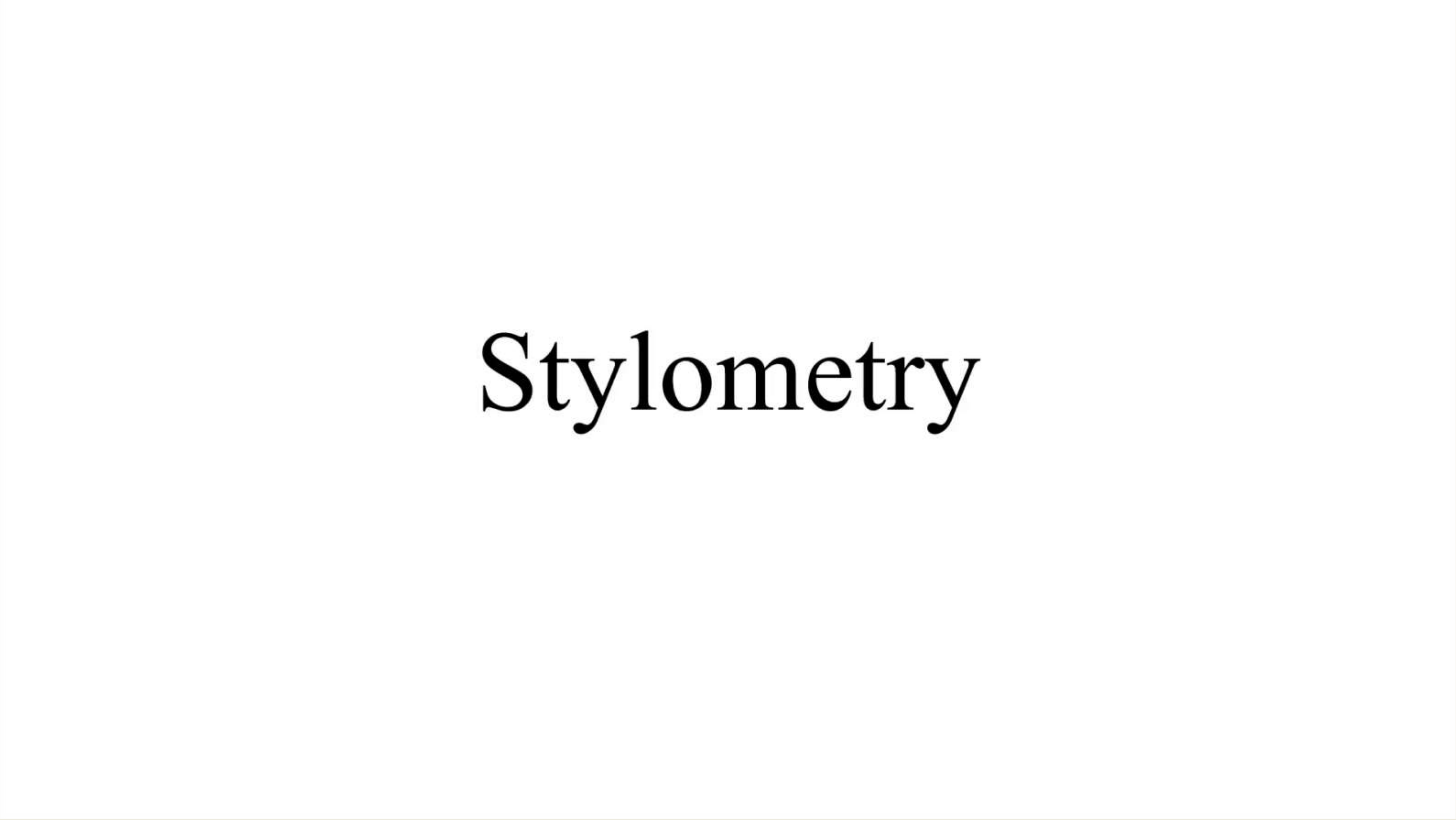
It is more robust but harder to interpret

Consensus Tree



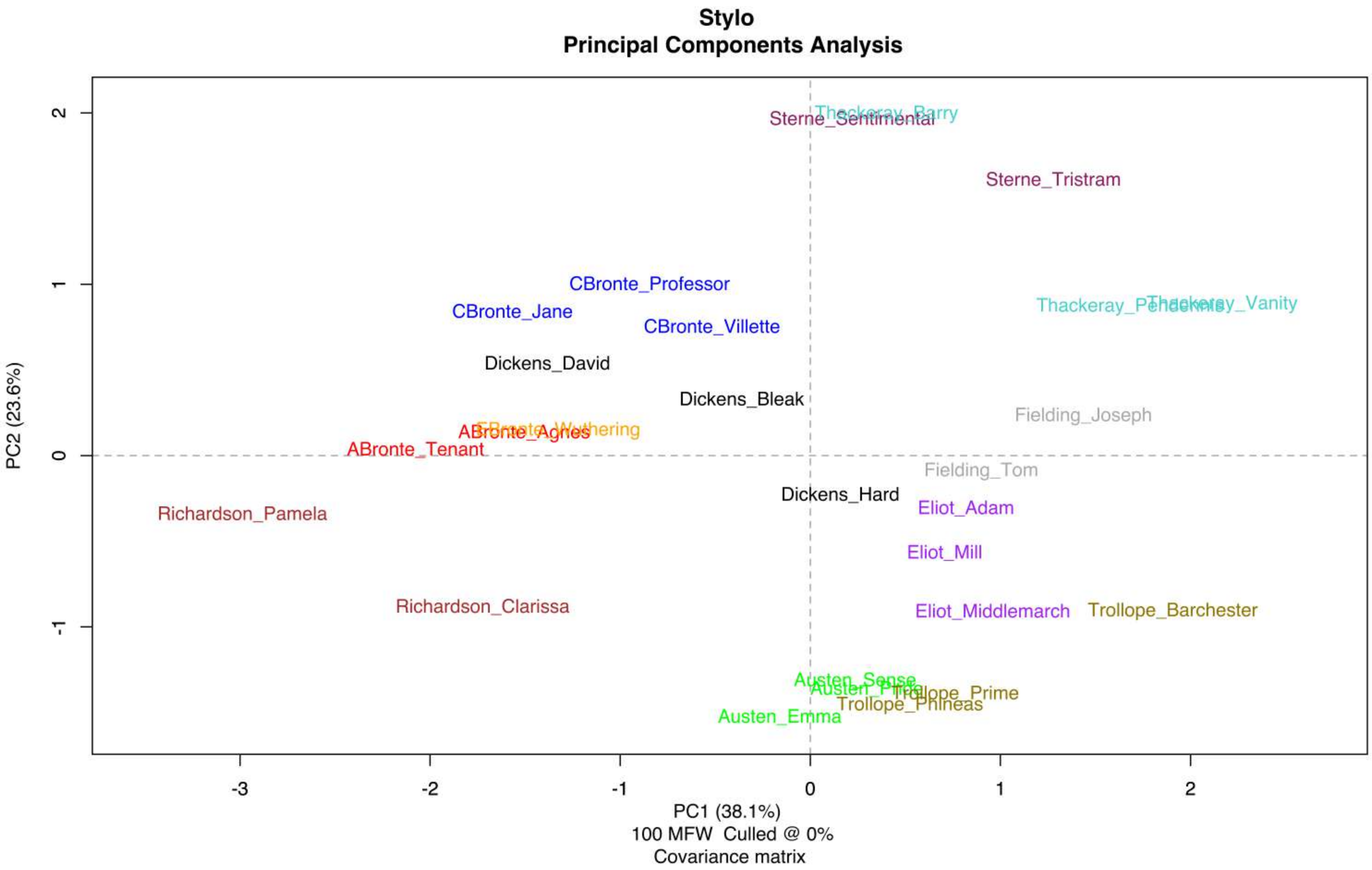
Multidimensional scaling is a statistical analysis technique often used to graphically show differences or similarities between elements of a set. It is a generalization of the concept of sorting: starting from a square matrix, containing the "similarity" of each row element with each column element, the multidimensional scaling algorithm assigns to each element a position in an N-dimensional space, with N established a priori..

If N is small enough, this space can be represented with a 3D graph or display. In practice this technique starts with a system with as many dimensions as the elements of the system, and reduces the dimensions up to a certain number N. In doing this then there is an inevitable loss of information (loss) and there are therefore different algorithms to do multidimensional scaling, which are better suited to different situations of use: in particular we distinguish between metric and non-metric algorithms.



Stylometry

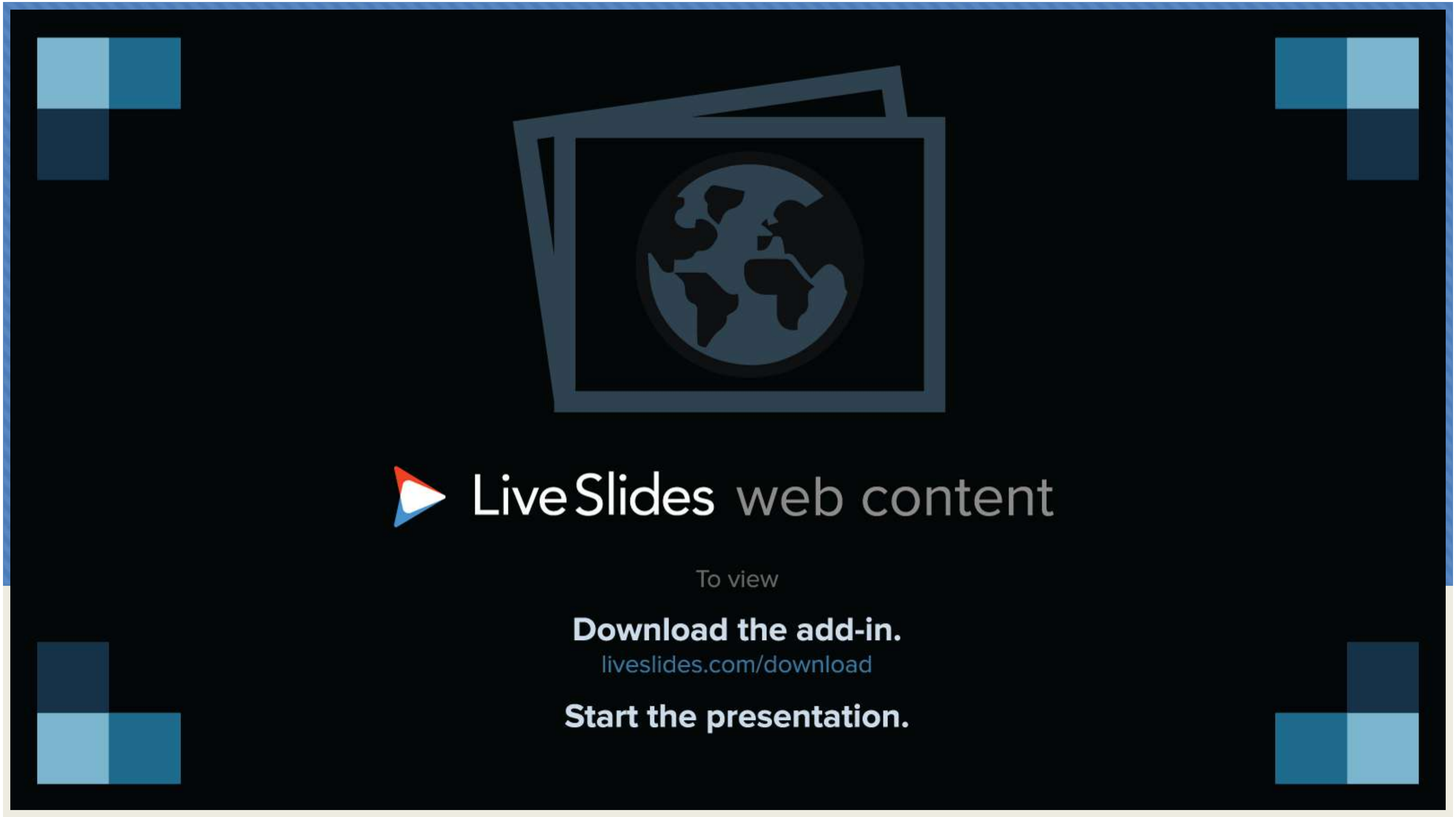
https://www.youtube.com/watch?v=jZ532ucT6Ik


Stylo
Principal Components Analysis
PC2 (23.6%)
2
1
0
-1
Sterne_Sentimental
Thackeray_Barry
Sterne_Tristram
CBronte_Professor
CBronte_Jane
CBronte_Villette
Thackeray_Pendennis
Thackeray_Vanity
Dickens_David
Dickens_Bleak
Fielding_Joseph
ABronte_Agnes
EBronte_Wuthering
ABronte_Tenant
Fielding_Tom
Dickens_Hard
Richardson_Pamela
Eliot_Adam
Eliot_Mill
Richardson_Clarissa
Eliot_Middlemarch
Trollope_Barchester
Austen_Sense
Austen_Pride
Trollope_Prime
Trollope_Phineas
Austen_Emma
-3
-2
-1
0
1
2
PC1 (38.1%)
100 MFW Culled @ 0%
Covariance matrix

Functions: stylo.network()

https://prismatic.phon.ox.ac.uk/index_network.html

- It is an extended version of the function stylo().
- It performs Bootstrap Consensus Networks, or a network-like generalization of the Bootstrap Consensus Trees method.
- It produces interactive visualizations in a web browser: to make it happen, you have to install an additional R package first.
  Type: install.packages("networkD3")



LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

Main functions: classify()

- It trains a model for pre-defined groups of texts, e.g. authors.
- Then it computes distances (differences) between texts, …
- … represented as rows of frequencies of most frequent words.
- Finally, it compares the trained models with test texts, using:
  - **Delta** classifier (lazy learner introduced by Burrows)
  - **k-NN** classifier (lazy learner relying on >1 neighbors)
  - **Suppor Vector Machines**, a high-performance non-probabilistic classifier
  - **Naive Bayes**, a classical yet slightly outdated classifier
  - **Nearest Shrunken Centroids**, a classifier for high-dimensional datasets
- A final report of the classifier's performance is outputted.

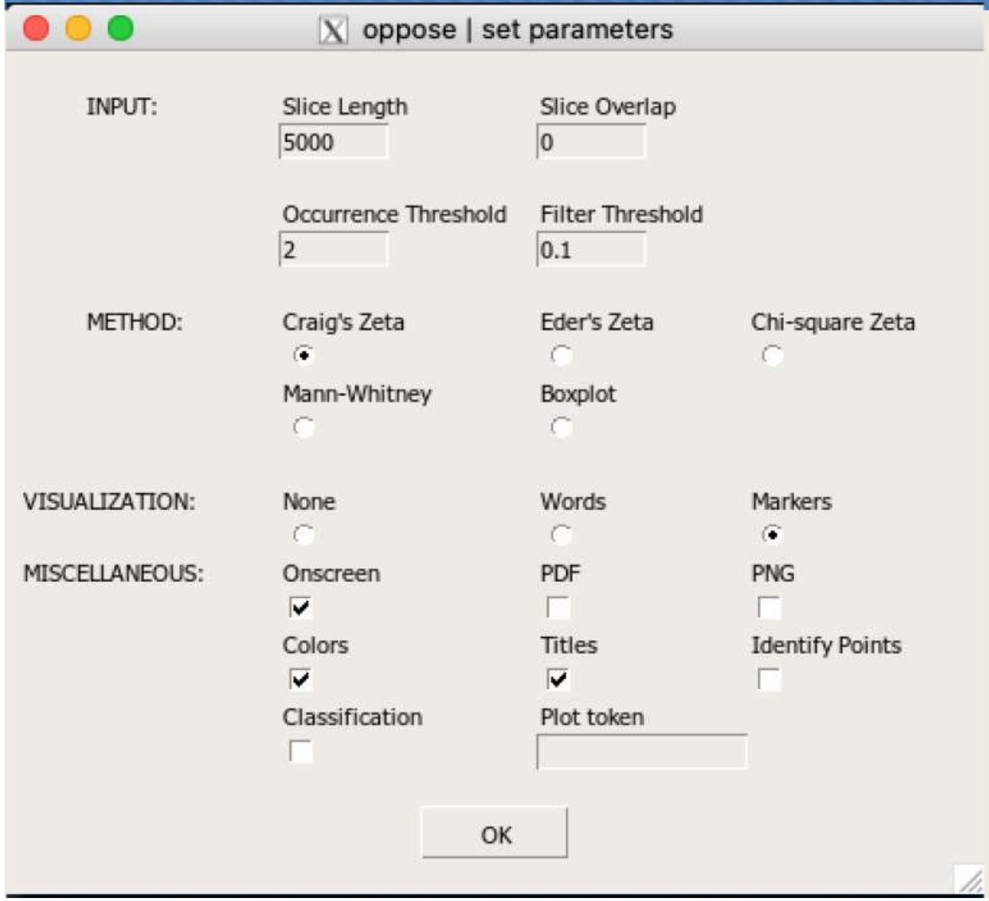
Main functions: oppose()

- Designed to compare two (groups of) texts
- It cuts input texts into equal-sized samples
- Finds words characteristic for two (groups) of texts
  - These can be reused with stylo() or classify()
- Produces a diagram of the use of each group's words

Running oppose()

- Different subfolder structure:
  - primary_set
  - secondary_set
  - test_set (optional)
- Running the function:
  - library(stylo)
  - oppose()
- What we get:
  - words_preferred.txt characteristic for the primary_set texts
  - words_avoided.txt characteristic for the secondary_set texts
- word frequency graph

oppose() parameters



oppose | set parameters

INPUT:

Slice Length: 5000

Slice Overlap: 0

Occurrence Threshold: 2

Filter Threshold: 0.1

METHOD:

Craig's Zeta ●	Eder's Zeta ○	Chi-square Zeta ○
Mann-Whitney ○	Boxplot ○	

VISUALIZATION:

None ○	Words ○	Markers ●

MISCELLANEOUS:

Onscreen ☑	PDF □	PNG □
Colors ☑	Titles ☑	Identify Points □
Classification □	Plot token	

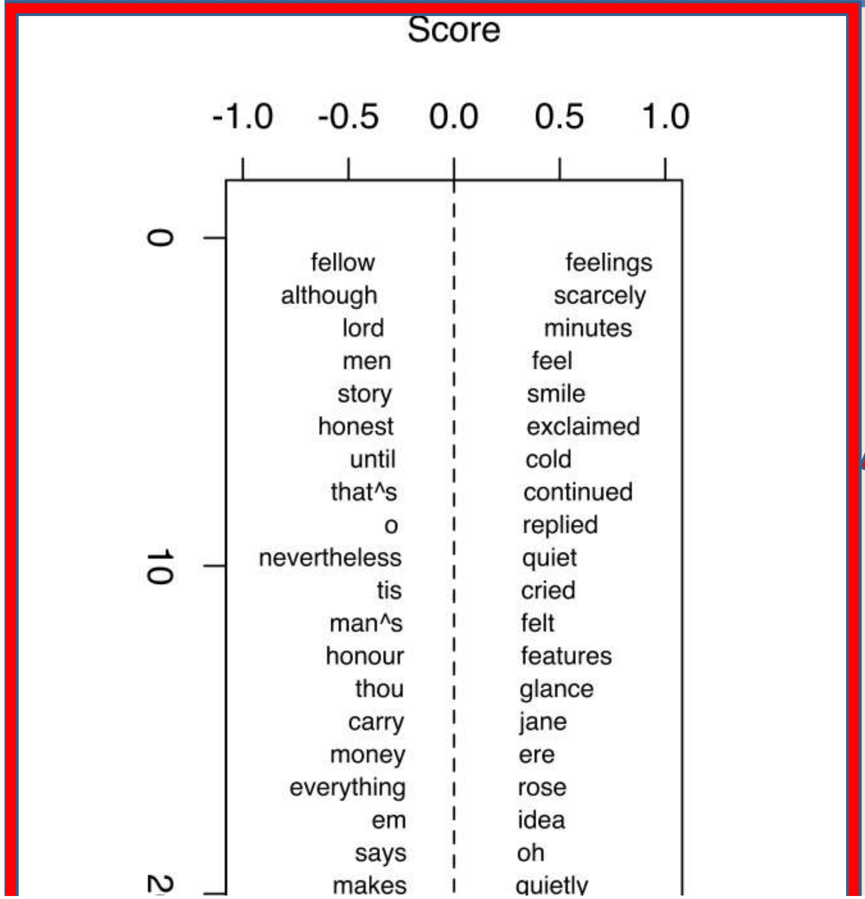
- Slice length: size (in words) of the samples (5000)
- Slice overlap: (0)
- Method: (Craig’s Zeta)
- Visualization: type of graph (Markers)

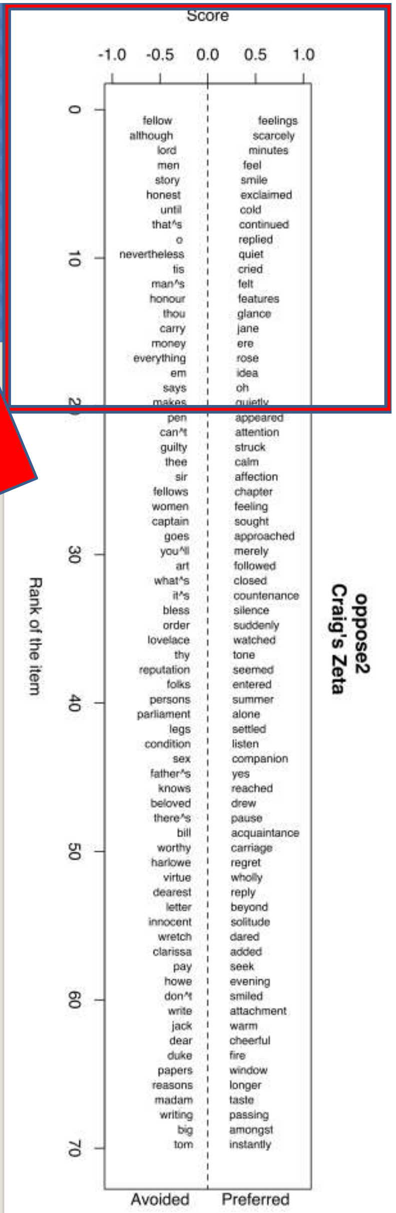
oppose() parameters

Most of the parameters for this somewhat underdeveloped function are not on GUI. You can switch them on as command line parameters

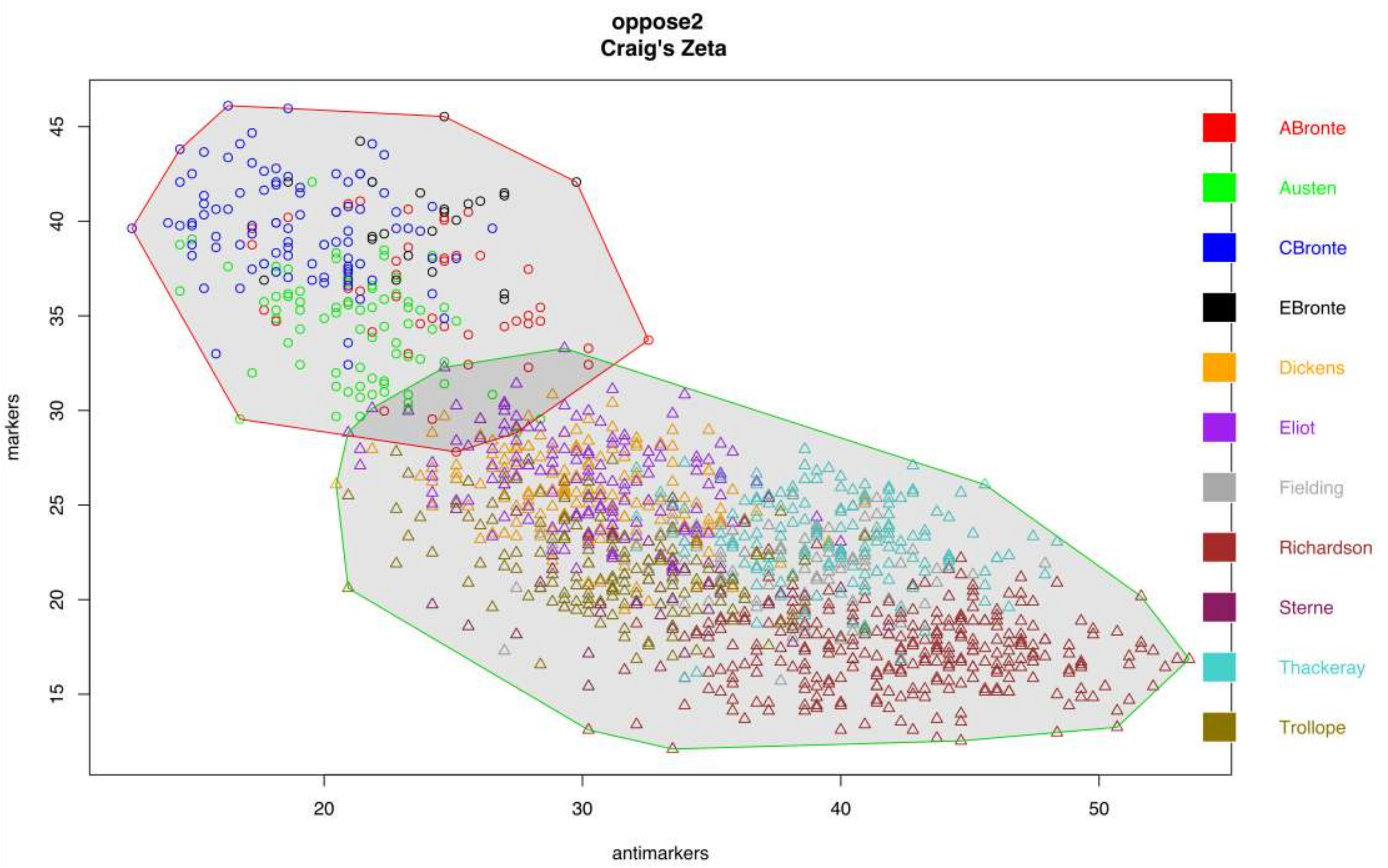
- when your corpus contains non-Latin diacritics:
  - oppose(encoding = "UTF-8", corpus.lang = "Spanish")

Men s Women - Words



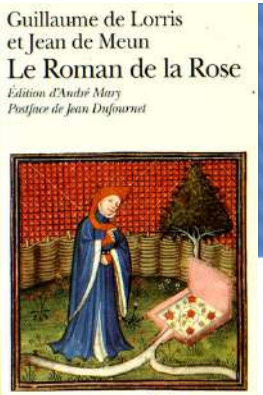


Men Vs Women - Markers

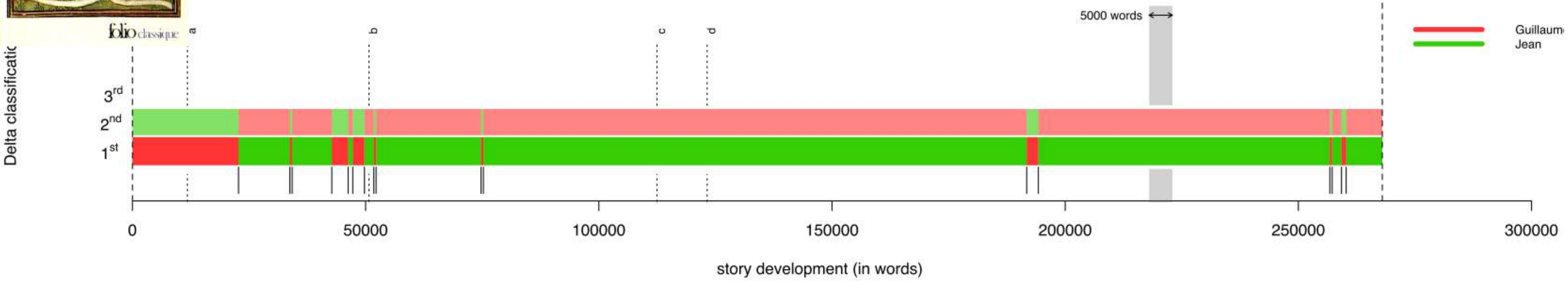


Functions: rolling.classify()

- Looks for traces of authors in a co-authored text…
- … by sliding through this text sequentially in order to detect peculiarities.
- Produces a graph of the respective strengths of these traces.



Le Roman de la Rose



Le Roman de la Rose (English: The Romance of the Rose) is a medieval French poem styled as an allegorical dream vision. It is a notable instance of courtly literature. The work's stated purpose is both to entertain and to teach others about the art of romantic love. Throughout the poem, Rose is used both as the name of the titular lady and as a symbol of female sexuality. The other characters' names also function both as regular names as abstractions illustrating the various factors that are involved in a love affair.

The poem was written in two stages. The first 4,058 lines, written by Guillaume de Lorris circa 1230, describe the attempts of a courtier to woo his beloved. This part of the story is set in a walled garden (a locus amoenus), a traditional literary topos in epic and chivalric literature. Around 1275, Jean de Meun composed an additional 17,724 lines. In this enormous coda, allegorical personages (Reason, Genius, and so on) hold forth on love.

Bibliography

BURROWS J. (2002), "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship," in Literary and Linguistic Computing, XVII: 3, pp. 267-287.

EDER M. (2015), "Does size matter? Authorship attribution, small samples, big problem," in Digital Scholarship in the Humanities, XXX: 2, pp. 167-182.

EDER M. (2017), "Visualization in stylometry: Cluster analysis using networks," in Digital Scholarship in the Humanities, XXXII: 1, pp. 50-64.

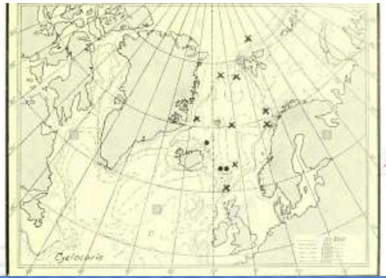
EVERT S., PROISL T., JANNIDIS F., REGER I., PIELSTRÖM S., SCHÖCH C., and VITT T. (2017), "Understanding and explaining Delta measures for authorship attribution" in Digital Scholarship in the Humanities, XXXII: suppl_2, pp. ii4-ii16.

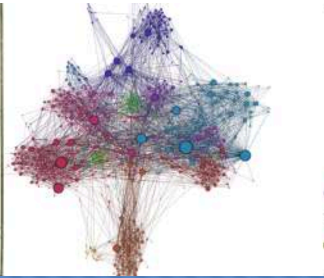
MENDENHALL T. C. (1887), "The characteristic curves of composition," in Science, IX: 214, pp. 237-249.

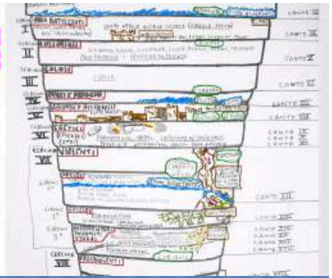
MORTON A. Q. (1978), Literary Detection: How to Prove Authorship and Fraud in Literature and Documents, Scribner, New York.

MOSTELLER F. and WALLACE D. L. (1964), Inference and Disputed Authorship: The Federalist, Addison-Wesley, Reading Mass.

```
# For details, see https://review.docs.microsoft.com/en-us/visualstudio/rtvs/sql-server

# Test code
library(RODBC)
channel <- odbcDriverConnect(dbConnection)
InputDataSet <- sqlQuery(channel, iconv(paste(readLines(
    'c:/proj/rproject1/rproject1/storedprocedure.query.sql',
    encoding = 'UTF-8', warn = FALSE), collapse = '\n'), from = 'UTF-8',
    to = 'ASCII', sub = ''))
odbcClose(channel)

OutputDataSet <- InputDataSet
```







Stylometry _end

Giovanni Pietro Vitali – University College Cork

giovannipietrovitali@gmail.com

https://github.com/digitalkoine

https://ucc-ie.academia.edu/GiovanniPietroVitali

université de TOURS

UCC
Coláiste na hOllscoile Corcaigh, Éire
University College Cork, Ireland

MARIE CURIE ACTIONS