STYLOmetry

The art of measuring style

Agenda

- Software installation
- Why to analyze text?
- History of stylometry?
- What is STYLO?
- STYLO()
- CLASSIFY()
- ROLLING.CLASSIFY()
- Way Forward

About me

- Data science with math background
- Hiking
- Psychology
- People
- yoga

LinkedIn: https://www.linkedin.com/in/janidziak/

Count number of F’s

FINISHED FILES ARE THE RESULT OF YEARS OF SCIENTIFIC STUDY COMBINED WITH THE EXPERIENCE OF YEARS.

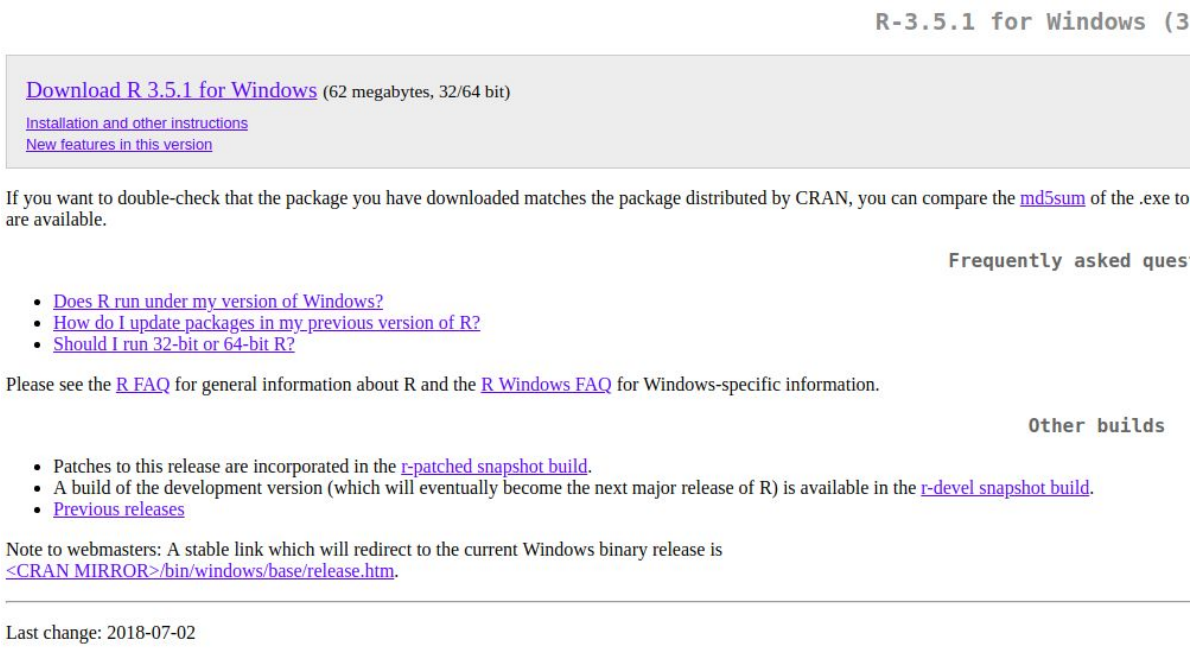
Count number of F’s

FINISHED FILES ARE THE RESULT OF YEARS OF SCIENTIFIC STUDY COMBINED WITH THE EXPERIENCE OF YEARS.

Environment preparation

Install R

https://cran.r-project.org/bin/windows/base



R-3.5.1 for Windows (3

Download R 3.5.1 for Windows (62 megabytes, 32/64 bit)

Installation and other instructions
New features in this version

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .exe to
are available.

Frequently asked ques

- Does R run under my version of Windows?
- How do I update packages in my previous version of R?
- Should I run 32-bit or 64-bit R?

Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.

Other builds

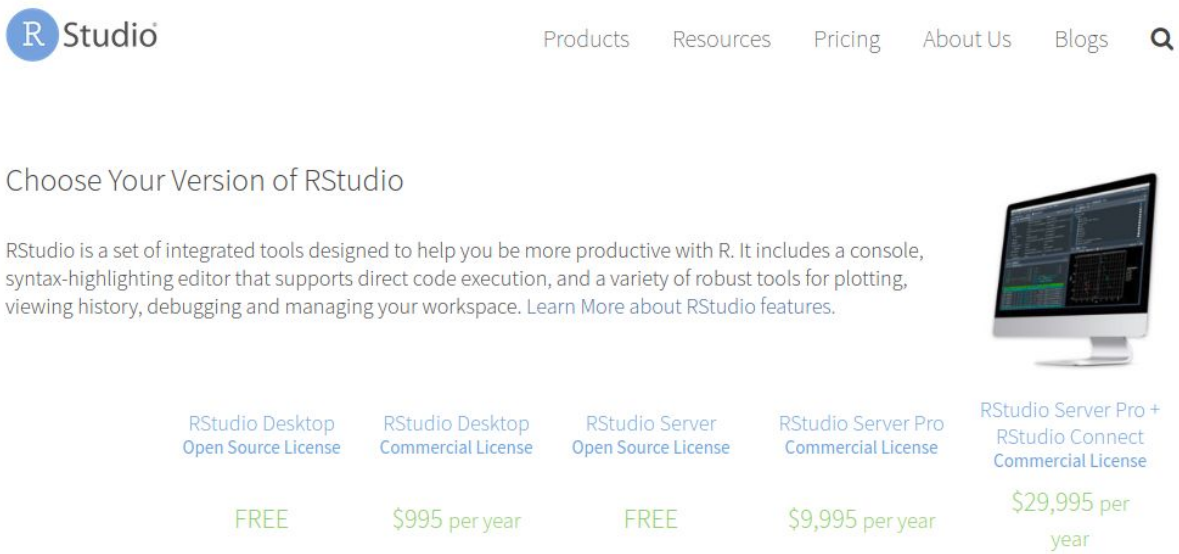
- Patches to this release are incorporated in the r-patched snapshot build.
- A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
- Previous releases

Note to webmasters: A stable link which will redirect to the current Windows binary release is
<CRAN MIRROR>/bin/windows/base/release.htm.

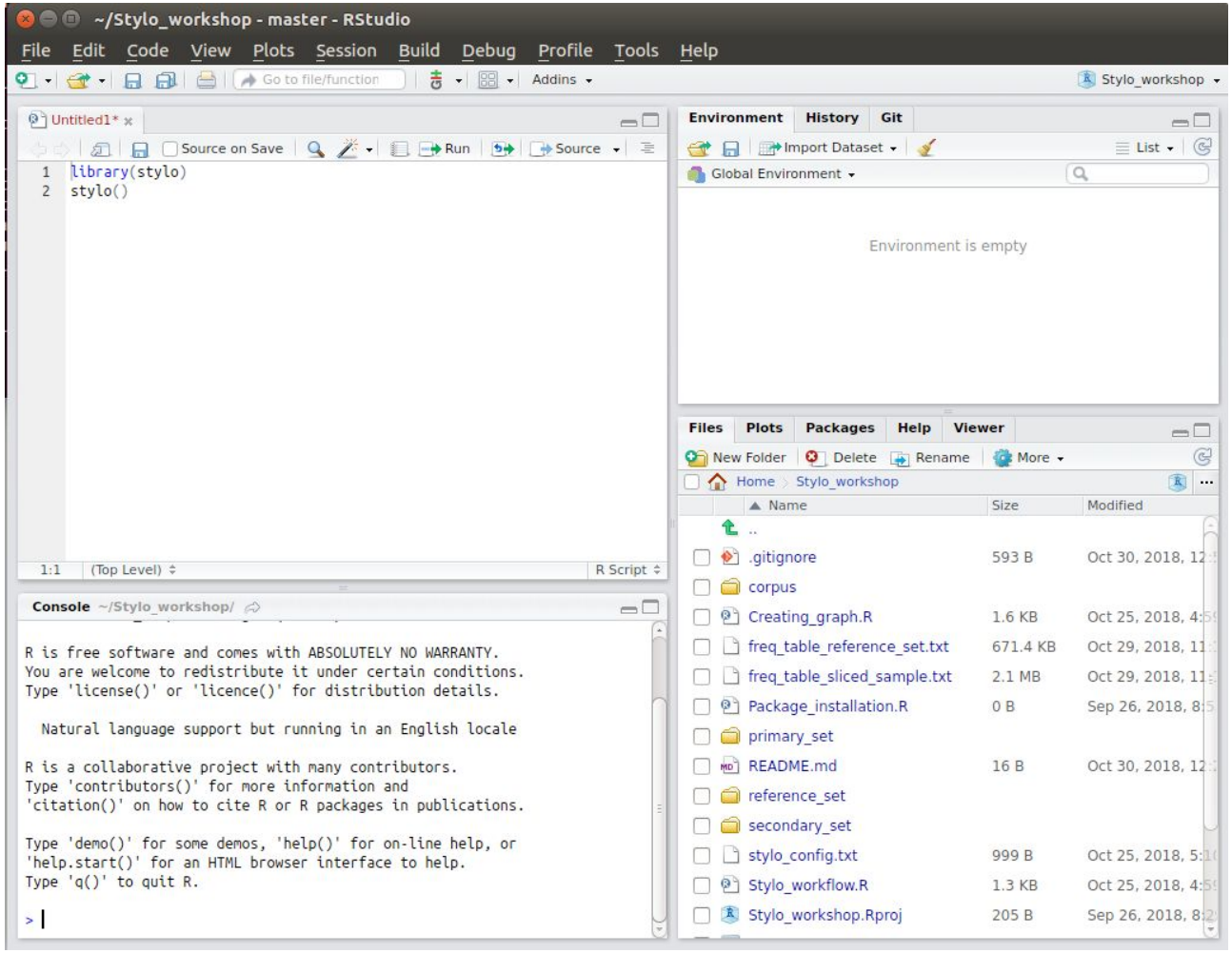
Last change: 2018-07-02

Install RStudio

www.rstudio.com/products/rstudio/download

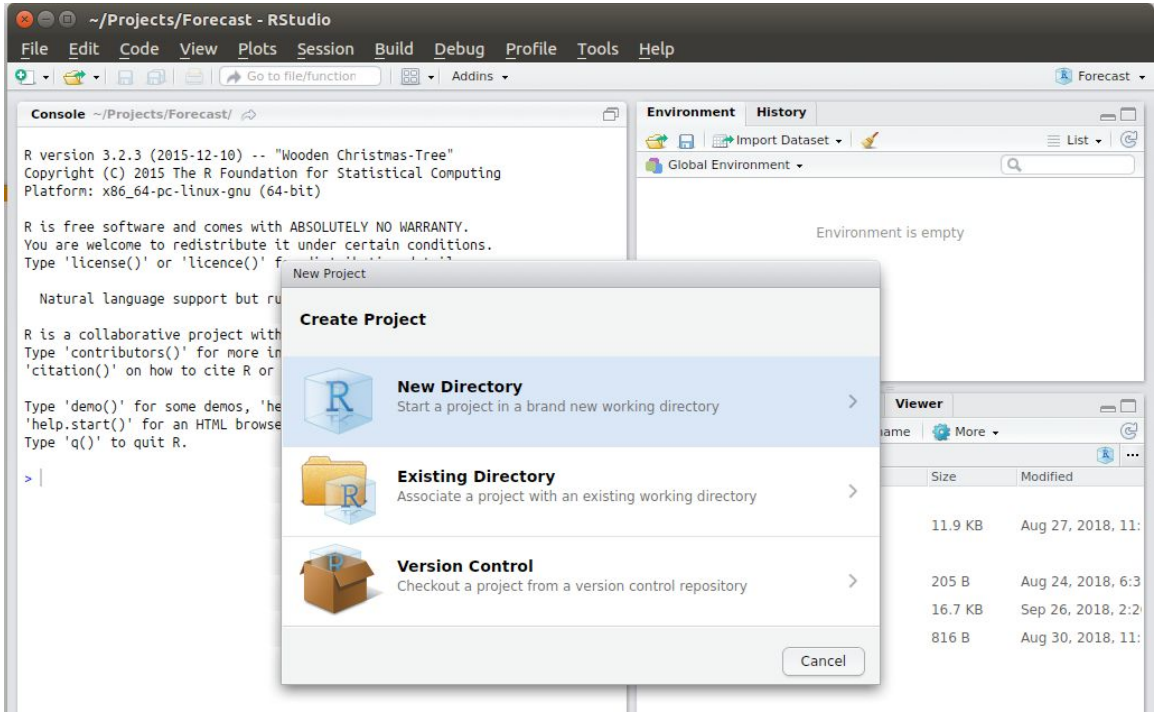


RStudio


~/Stylo_workshop - master - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Stylo_workshop
Untitled1*
Source on Save
Run
Source
1 library(stylo)
2 stylo()
1:1
(Top Level)
R Script
Console ~/Stylo_workshop/
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Environment History Git
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder
Delete
Rename
More
Home
Stylo_workshop
Name
Size
Modified
..
.gitignore 593 B Oct 30, 2018, 12:
corpus
Creating_graph.R 1.6 KB Oct 25, 2018, 4:5
freq_table_reference_set.txt 671.4 KB Oct 29, 2018, 11:
freq_table_sliced_sample.txt 2.1 MB Oct 29, 2018, 11:
Package_installation.R 0 B Sep 26, 2018, 8:5
primary_set
README.md 16 B Oct 30, 2018, 12:
reference_set
secondary_set
stylo_config.txt 999 B Oct 25, 2018, 5:1
Stylo_workflow.R 1.3 KB Oct 25, 2018, 4:5
Stylo_workshop.Rproj 205 B Sep 26, 2018, 8:2

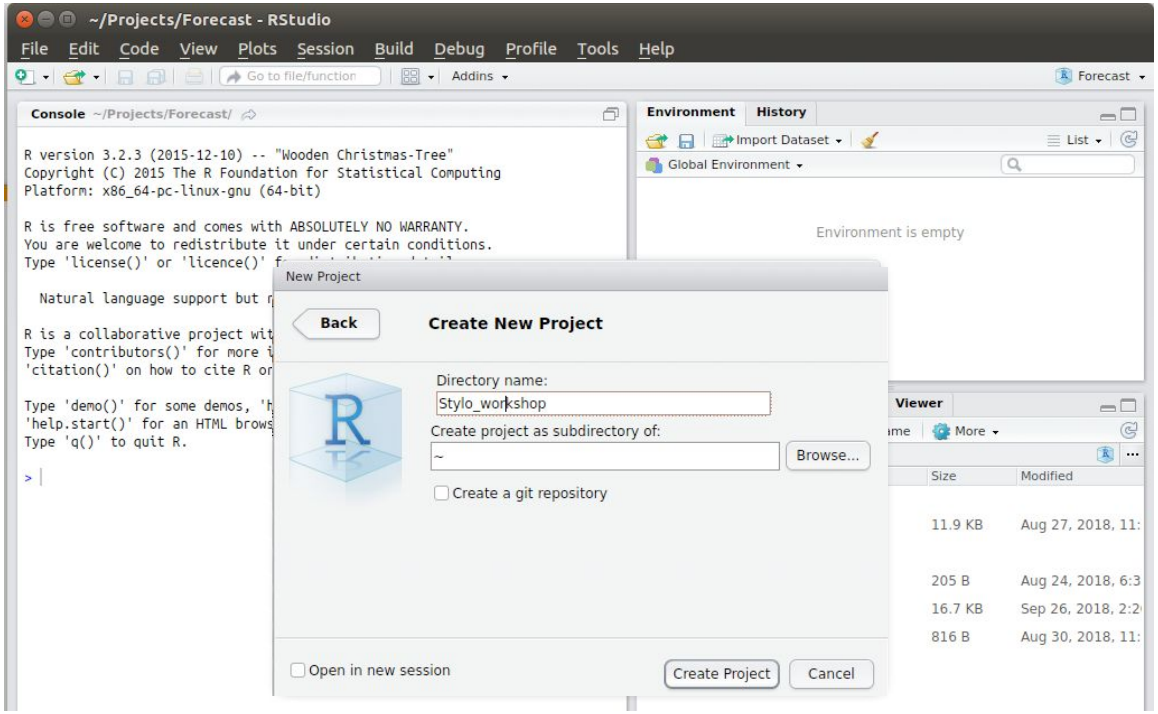
Create project directory

File >> New Project… >> New Directory >> Empty Project



Project Name

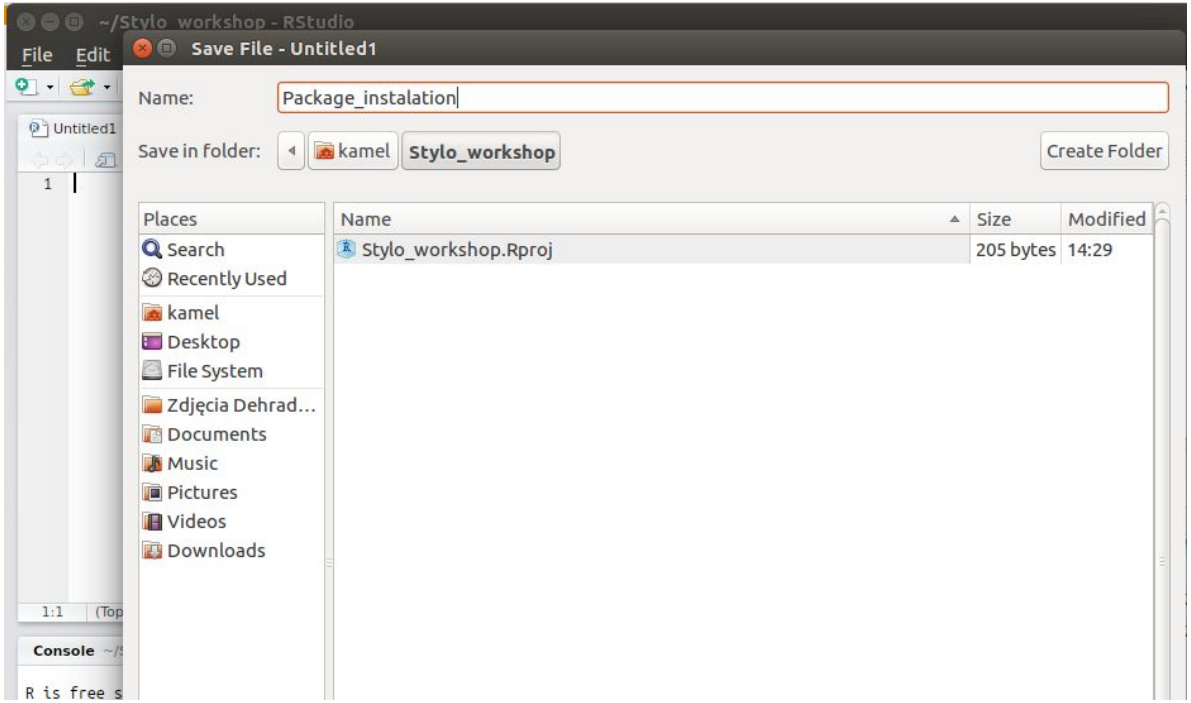
Stylo_workshop



Create script

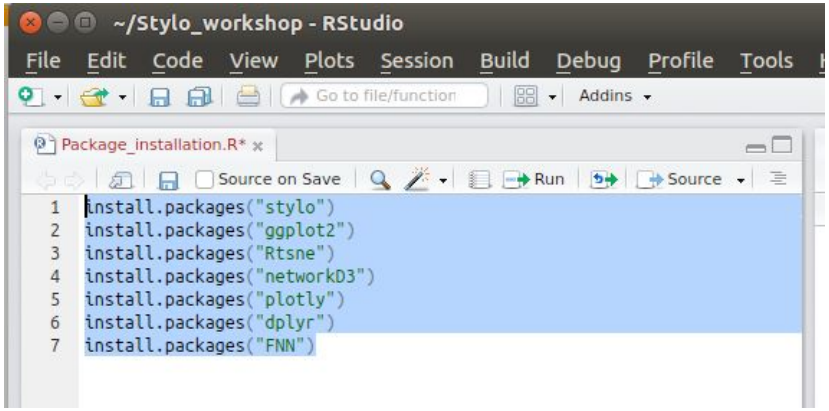
File >> New File >> R Script

File >> Save >> Package_installation

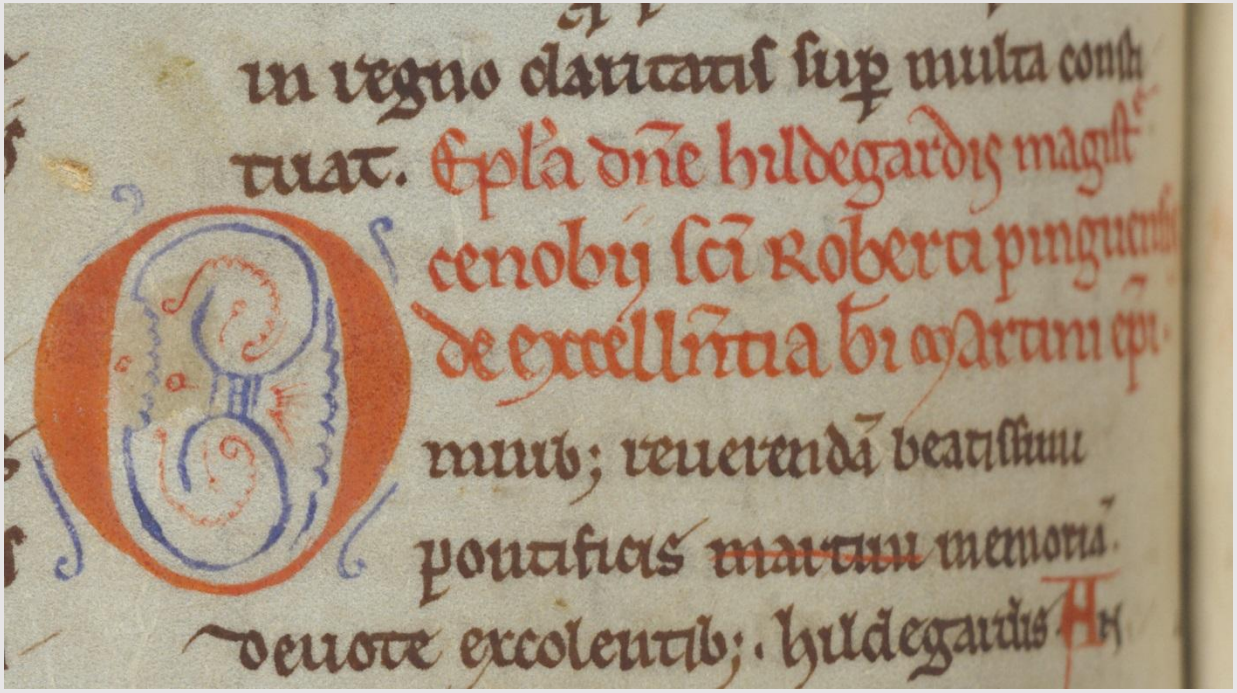


Install packages

```
install.packages("stylo")
install.packages("ggplot2")
install.packages("Rtsne")
install.packages("networkD3")
install.packages("plotly")
install.packages("dplyr")
install.packages("FNN")
```



SHORT HISTORY OF STYLOMERTY



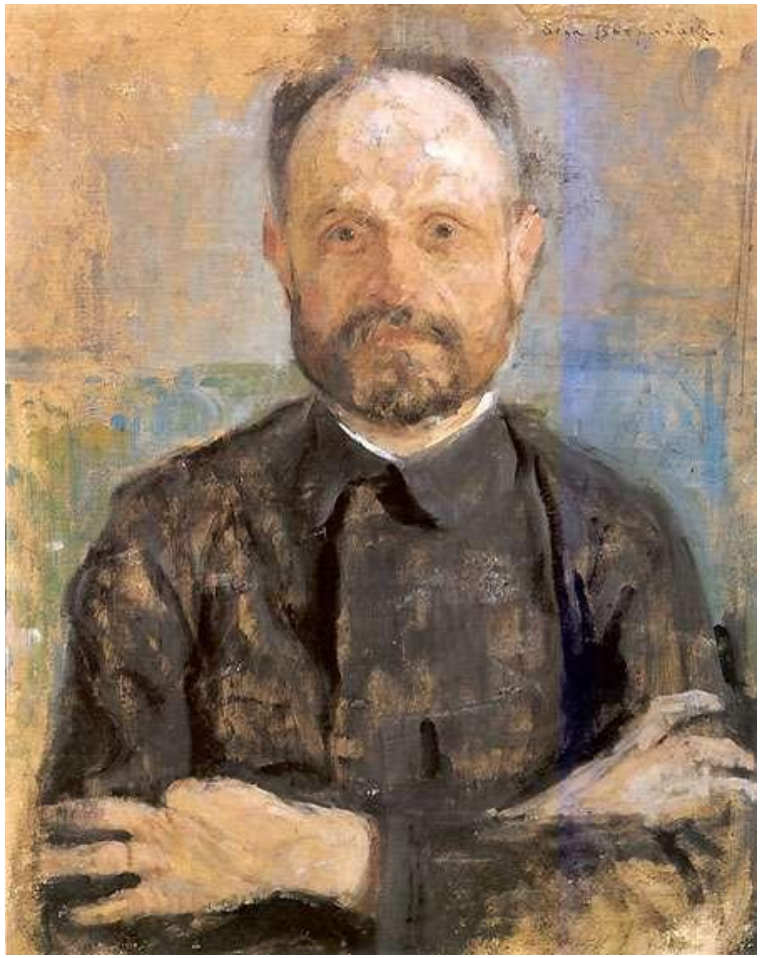
Lorenzo Valla (c. 1407–1457)

Constantini Donatione declamatio

Constantini Donatione declamatio

- A forged decree where the emperor Constantine I transfers authority of the Roman Empire to the Pope
- Written probably in the 8th century, claimed to be authentic for many centuries
- The first instance of scholarly-based investigation of style:
- Some grammatical forms could not have been used in the 4th century

Wincenty Lutosławski (1863–1954)



Wincenty Lutosławski

- 1890: chronology of Plato's dialogues
- The basics of stylometry were set out by Polish philosopher Wincenty Lutosławski in *Principes de stylométrie* (1890)

- Pionier of yoga in Poland

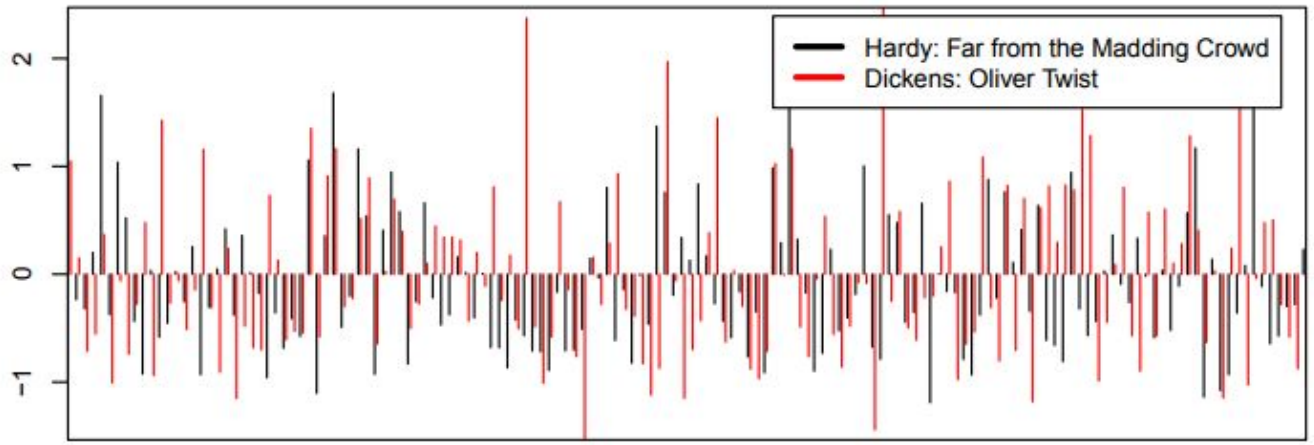
John Burrows

Burrows DELTA

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,	...)



Z Score

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

$$Z = \frac{X - \bar{X}}{S}$$

DELTA measures

Burrows's Delta = Manhattan distance (Burrows 2002)

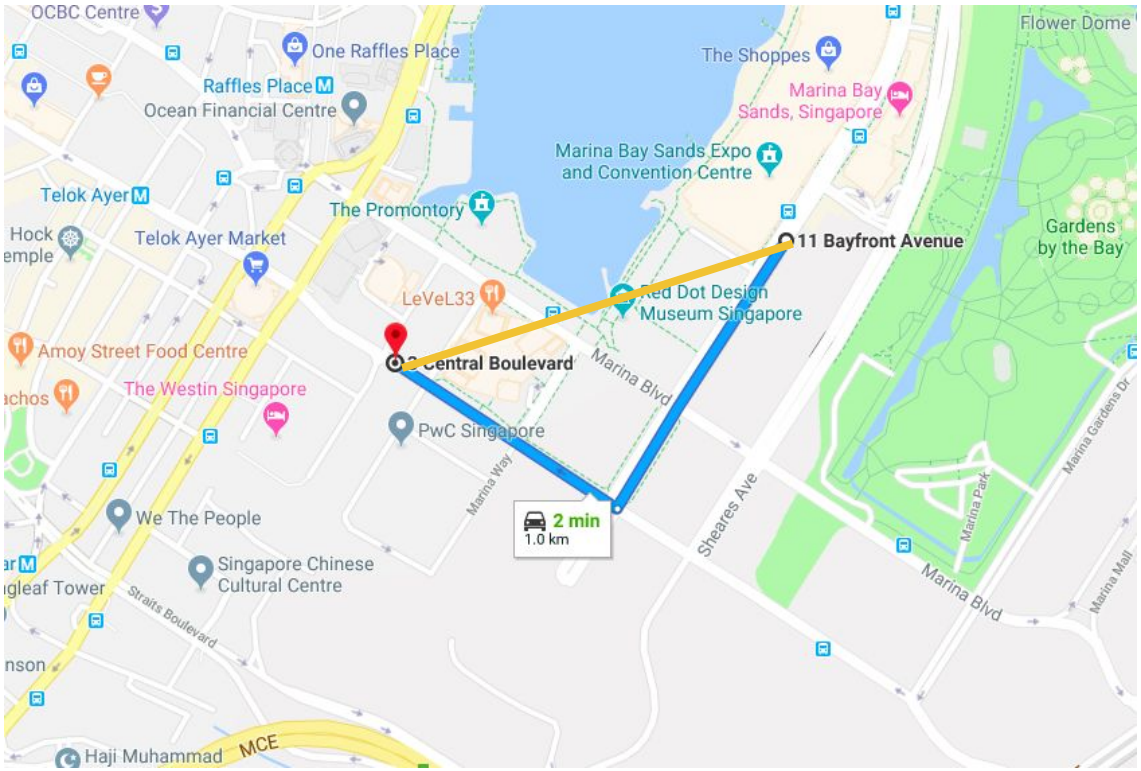
$$\Delta_B(D, D') = \|\mathbf{z}(D) - \mathbf{z}(D')\|_1 = \sum_{i=1}^{n_w} |z_i(D) - z_i(D')|$$

Quadratic Delta = Euclidean distance (Argamon 2008)

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

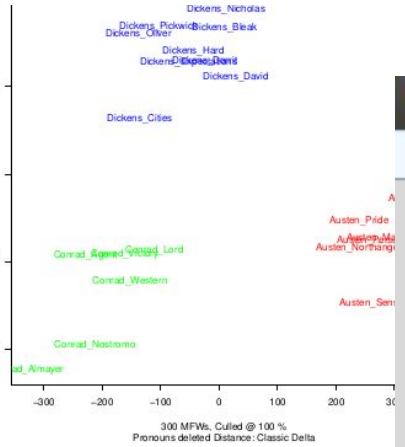
Distances

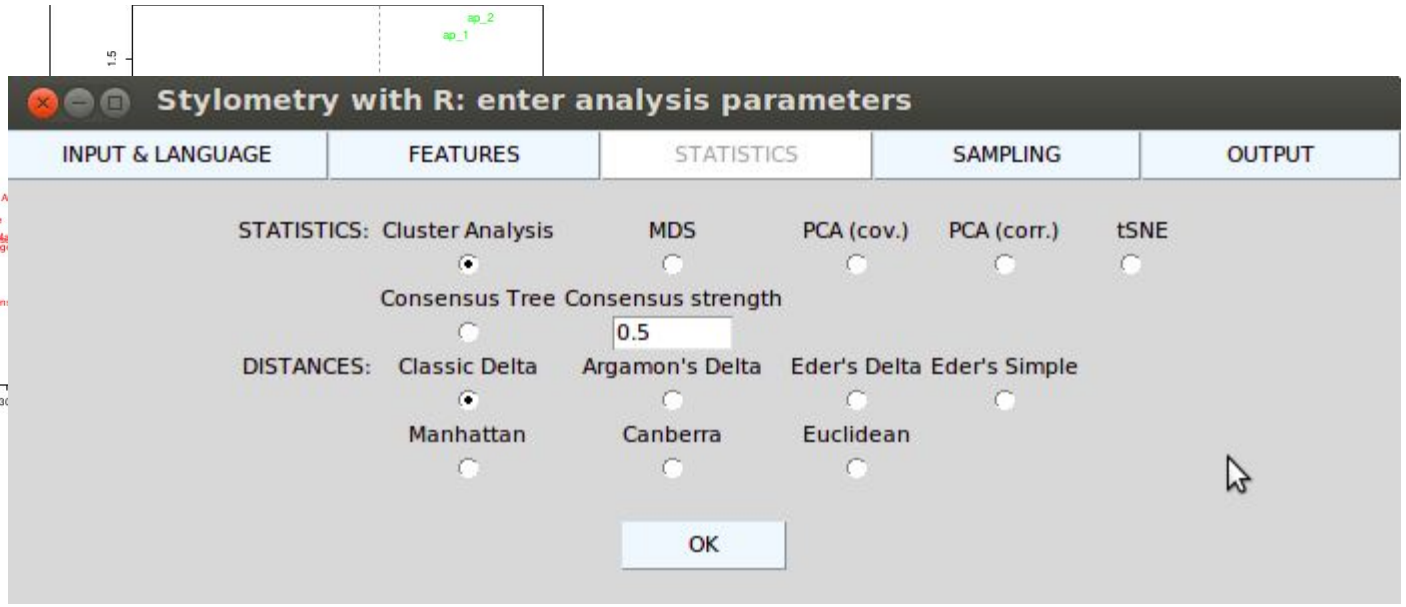
- Time
- Manhattan
- Euclidean

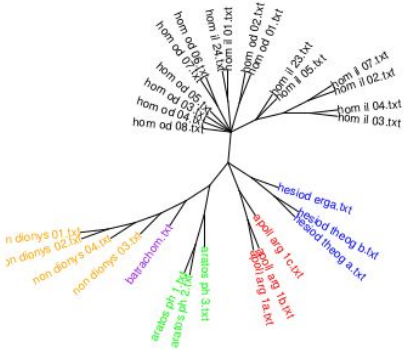


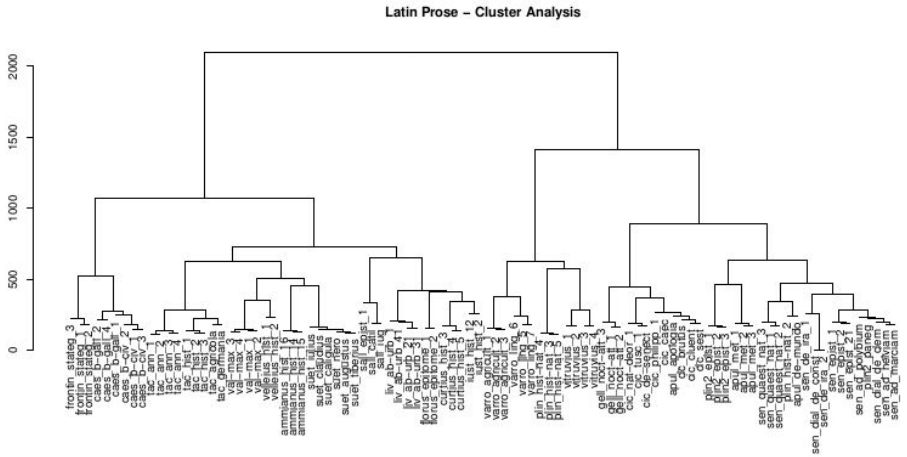
Maciej Eder

Stylo R software



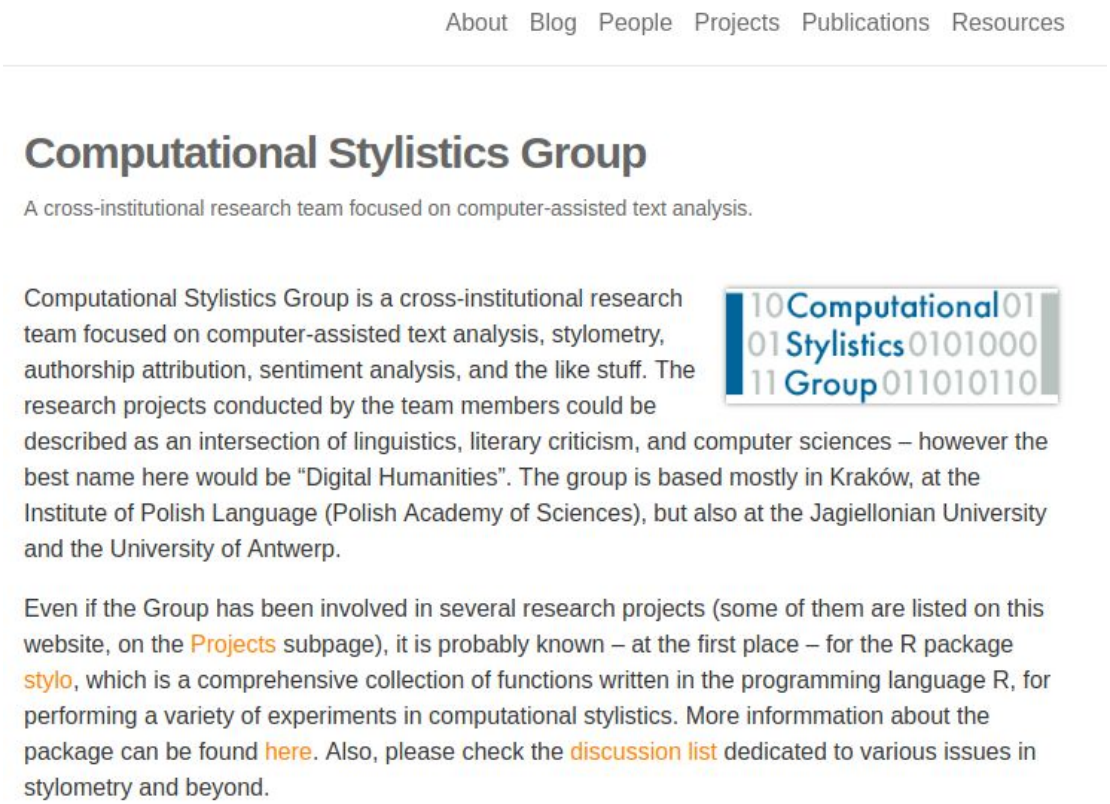




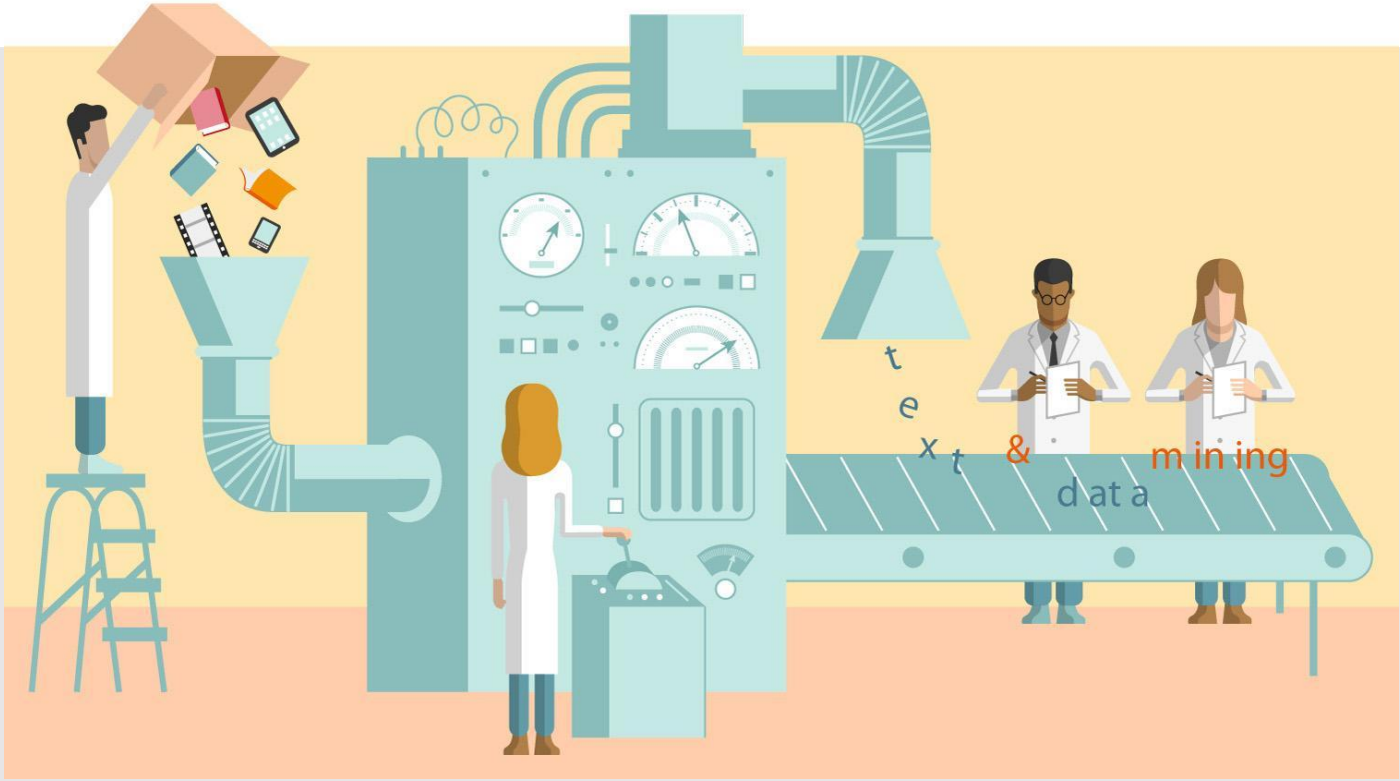


COMPUTATIONAL STYLISTICS GROUP

https://computationalstylistics.github.io/



Why to analyze text?



Applications of stylometry

Digital Humanities

- Author attribution identification of unknown authors
- Genre classification
- Historical study of language change

Other applications

- Anonymity
- Plagiarism
- Criminal civil security- (detection if someone's suicide note is real)

Agatha Christie

5 September 1890 – 12 January 1976

66 detective novels and 14 short story collections

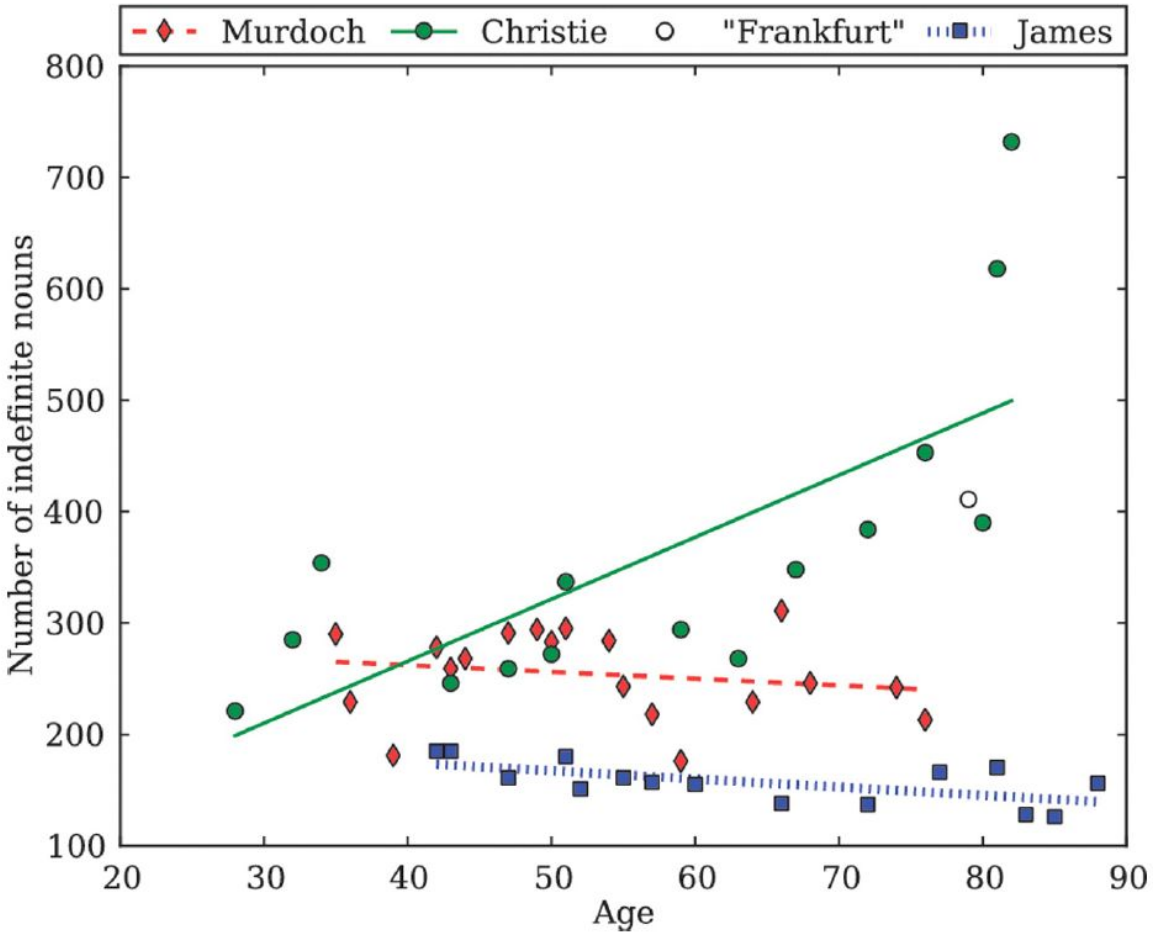
from 1971 to 1974, Christie's health began to fail...

...although she continued to write.

Four indefinite nouns

- thing(s),
- something,
- anything,
- nothing

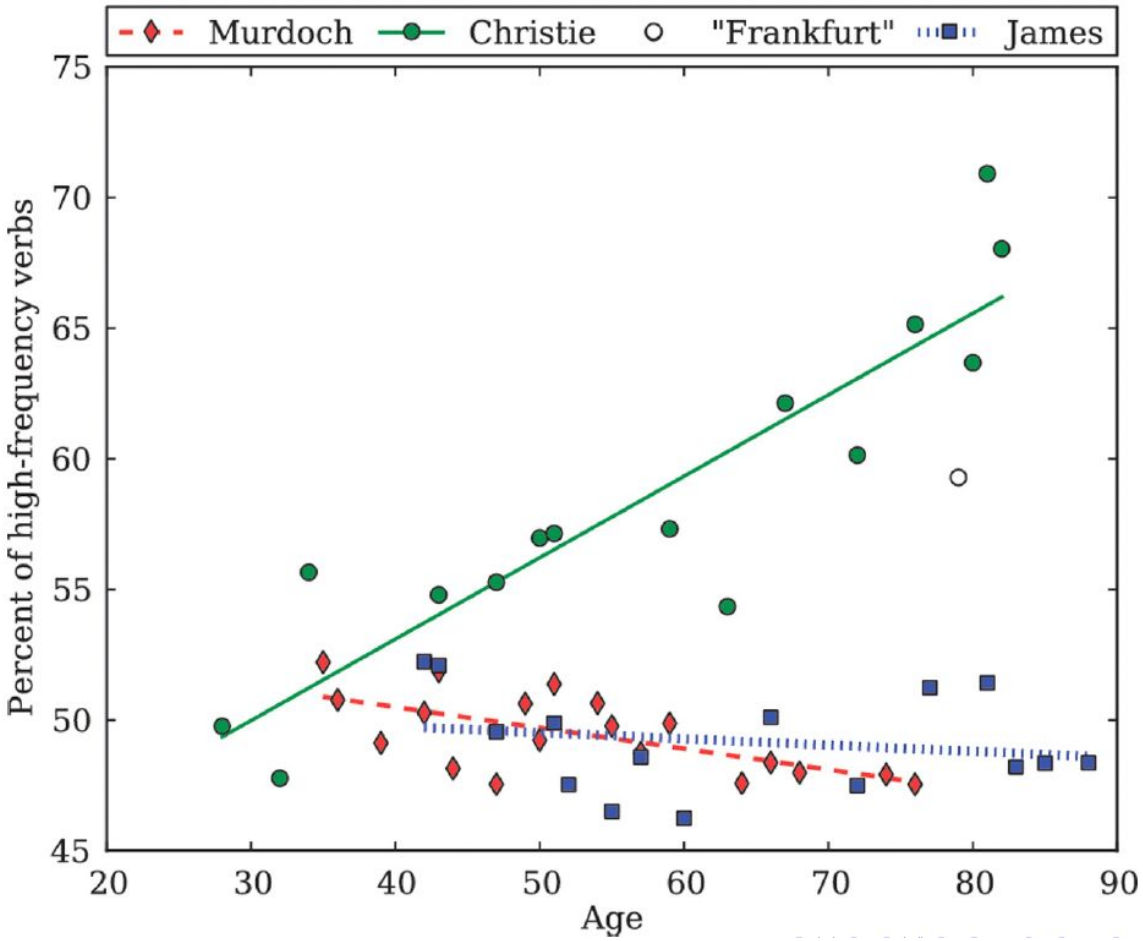
Number of indefinite noun occurrences



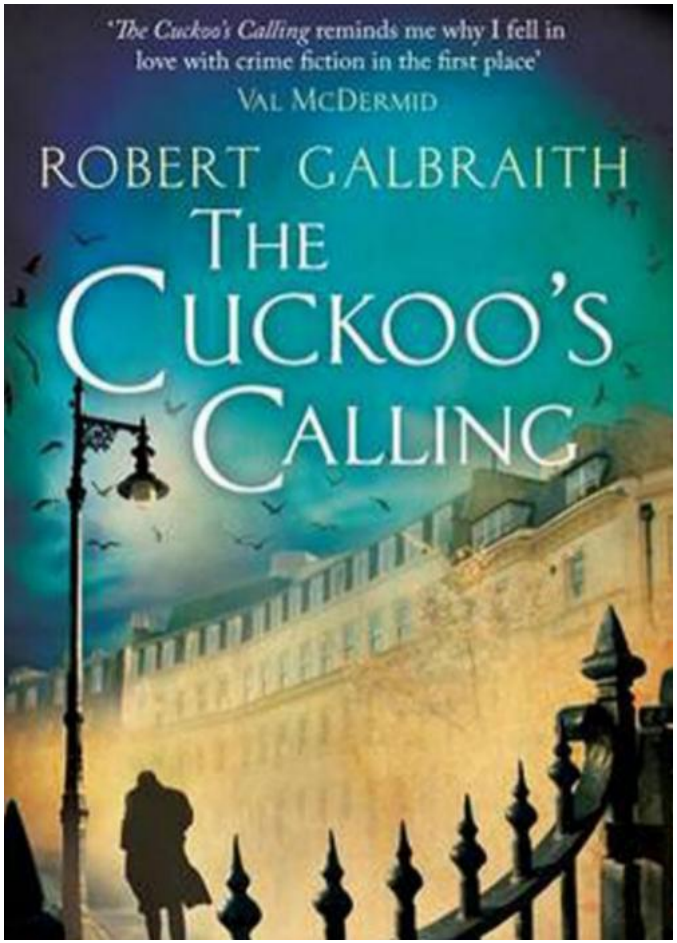
High-frequency verbs of low specificity

be, come, do, get, give, go, have, know, look, make, see, tell, think, want, ask, feel, find, forget, happen, hear, like, live, mean, meet, put, remember, run, say, seem, speak, suppose, take, use, walk, wonder

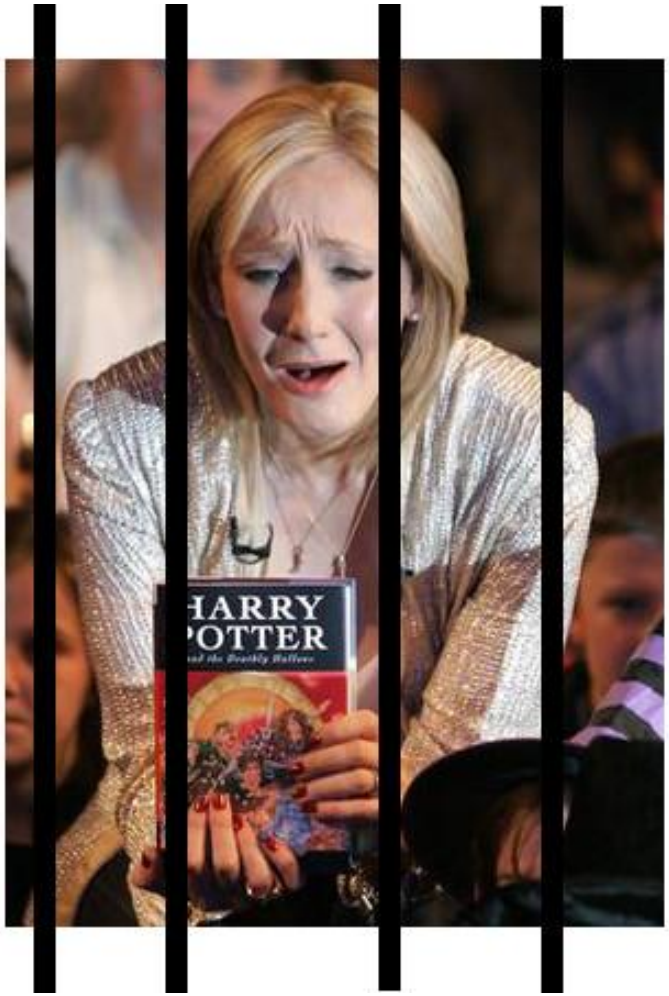
Proportion of 35 high-frequency verbs



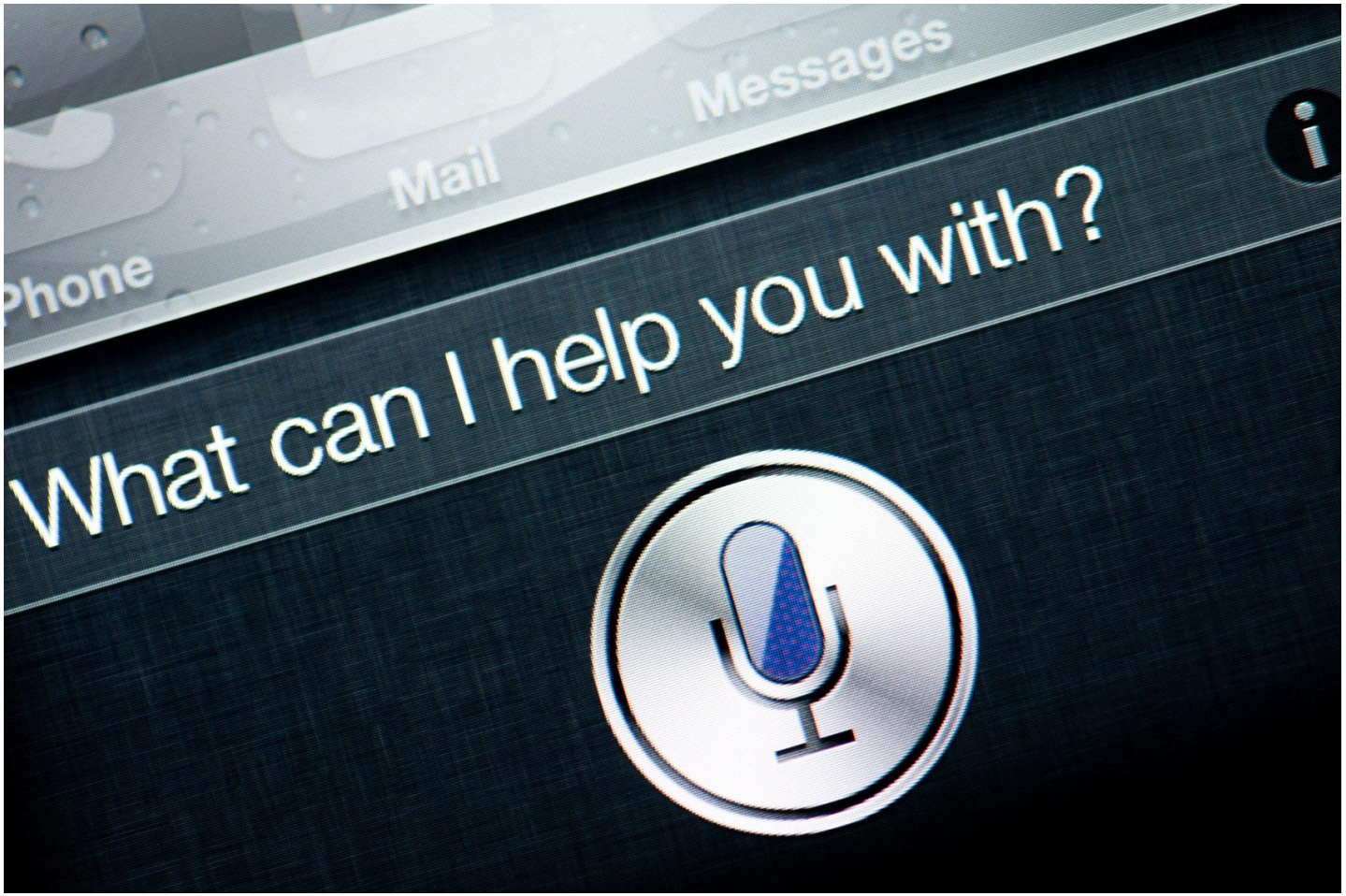
Who is Robert Galbraith?



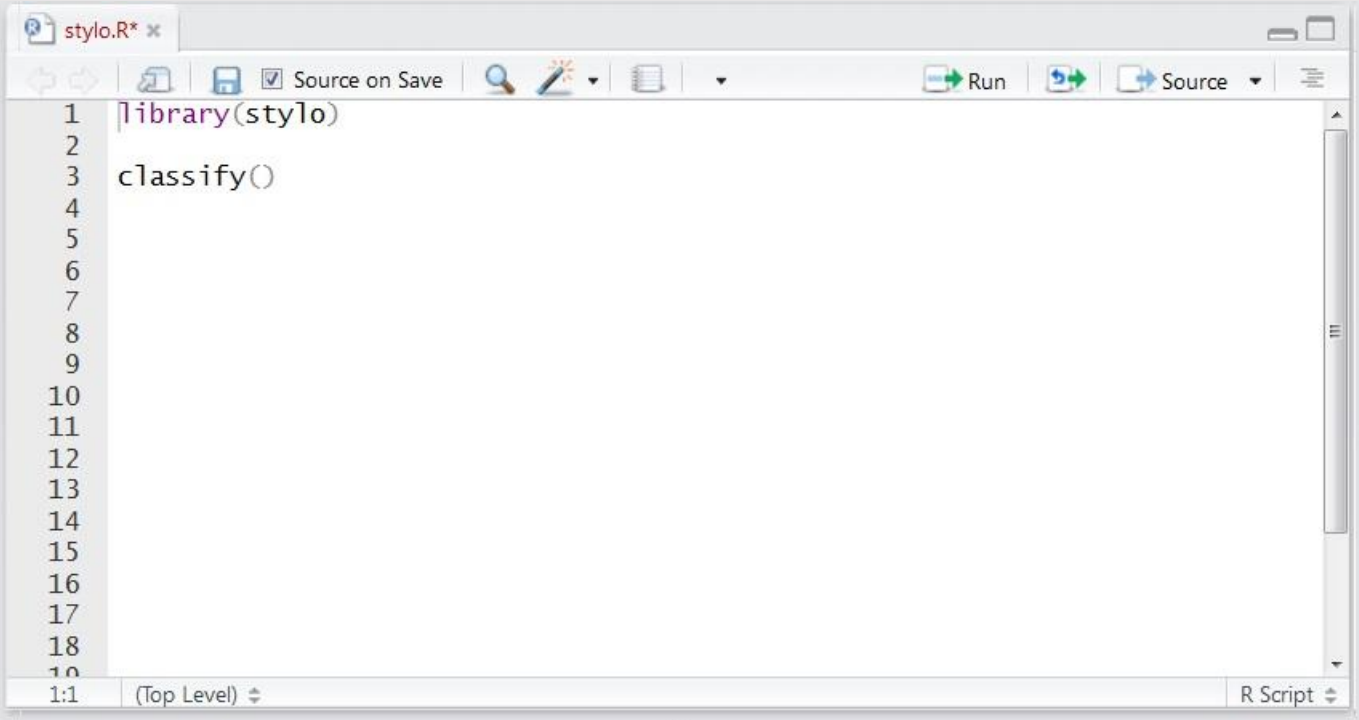
J. K. Rowling . . .



SIRI



How to use stylo?



FUNCTIONS PROVIDED

- stylo() - an all-in-one tool for a variety of experiments in computational stylistics.
- classify() -Function that performs a number of machine-learning methods for classification used in computational stylistics
- rolling.classify() - Function that splits a text into equal-sized consecutive blocks (slices) and performs a supervised classification of these blocks against a training

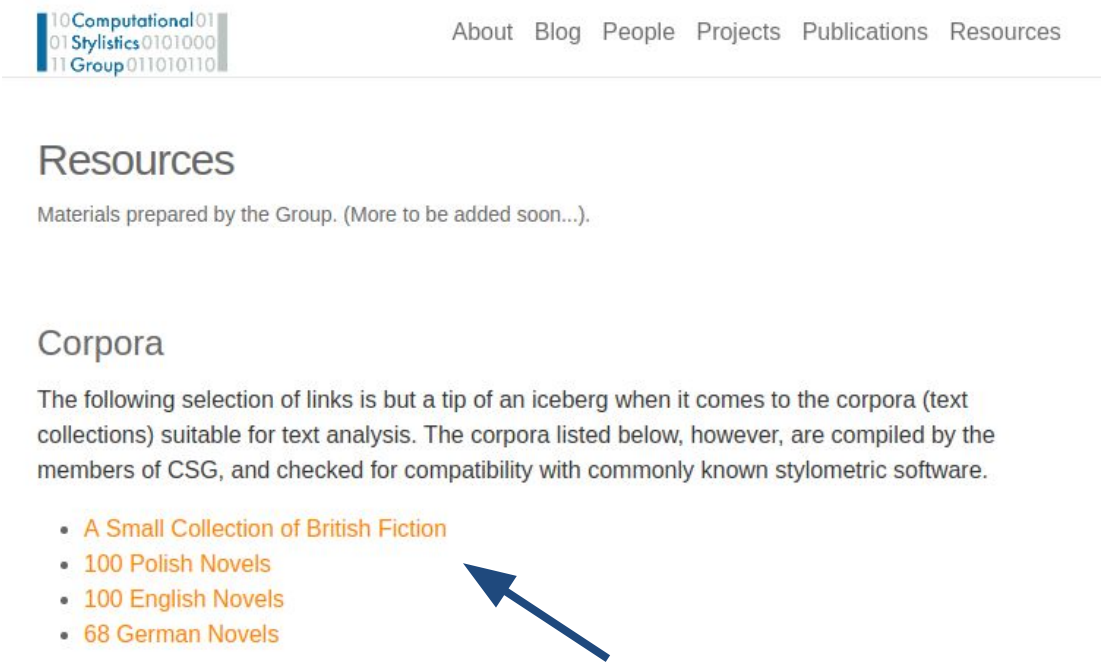
Analysis Workflow

1. Data preparation
    a. Naming Convention
    b. Directory
2. Language and Input Formatting
3. Text Characteristics (Features)
4. Summary Type (Statistics)
5. Output

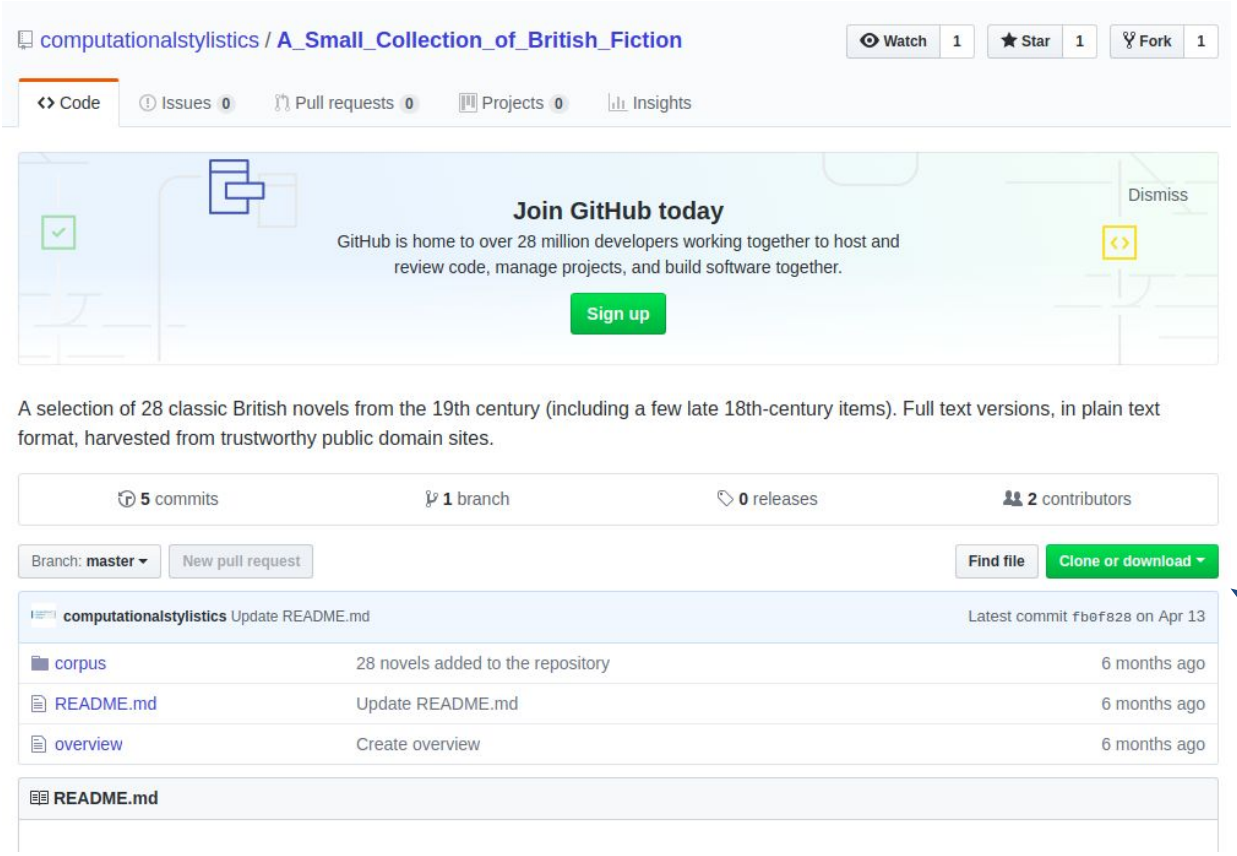
COMPUTATIONAL STYLISTICS GROUP

https://computationalstylistics.github.io/

Data preparation



Data preparation



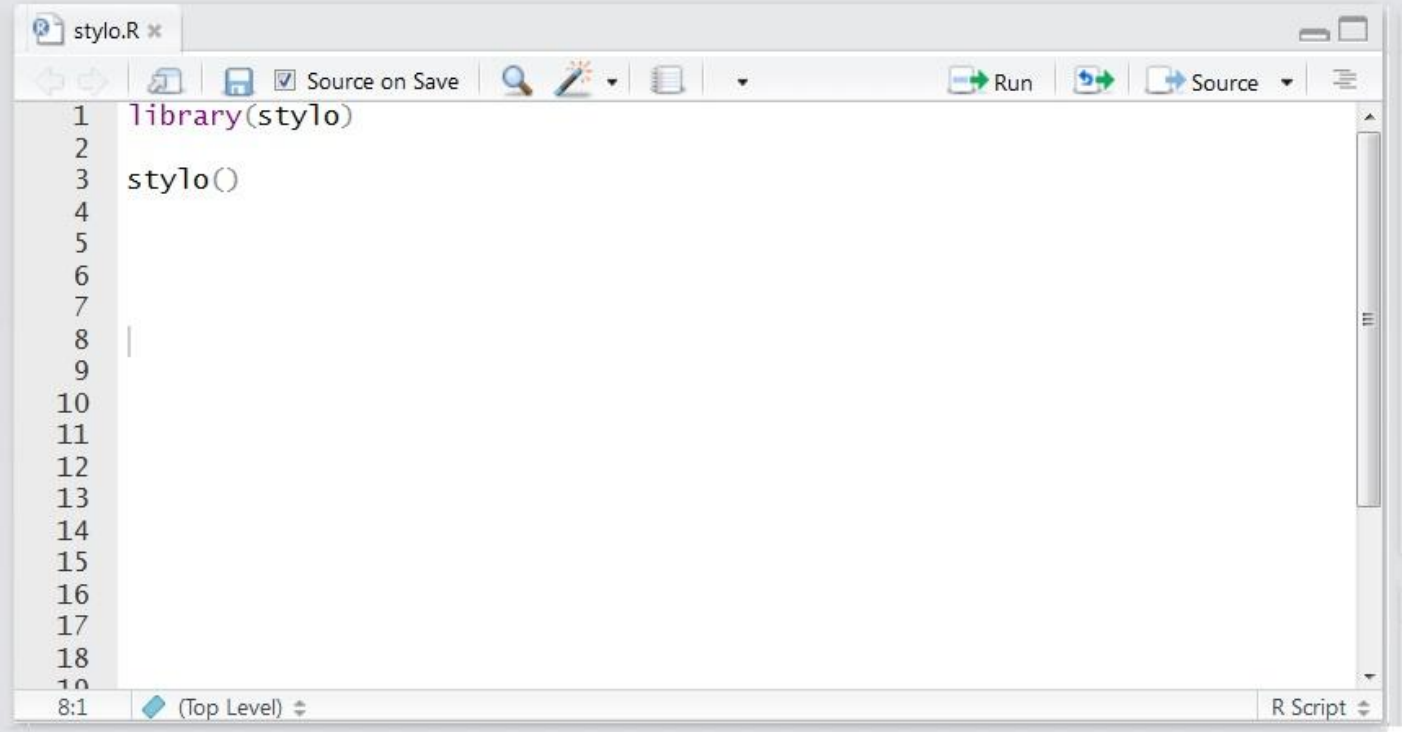
Data preparation

- https://github.com/computationalstylistics/100_english_novels
- https://github.com/computationalstylistics/A_Small_Collection_of_British_Fiction

Exercise - project preparation

- Create new project for the Exercises
- Name the project Stylometry_Excercise
- Create script with named exercise.R

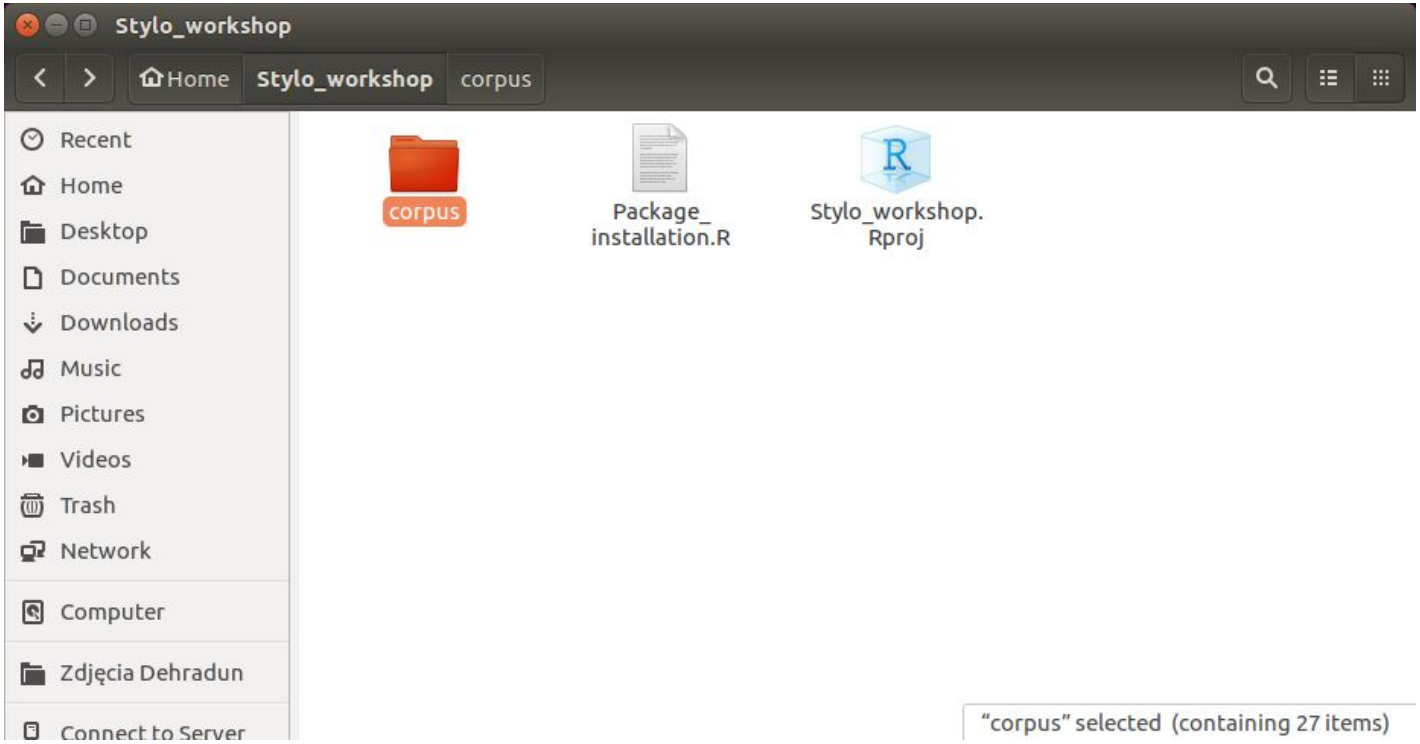
STYLO


stylo.R
Source on Save
Run
Source
library(stylo)
stylo()
8:1
(Top Level)
R Script

What is Stylo?

Stylo: Comprehensive collection of functions written in the programming language R, for performing variety of experiments in computational stylistics.

Data preparation for Stylo()

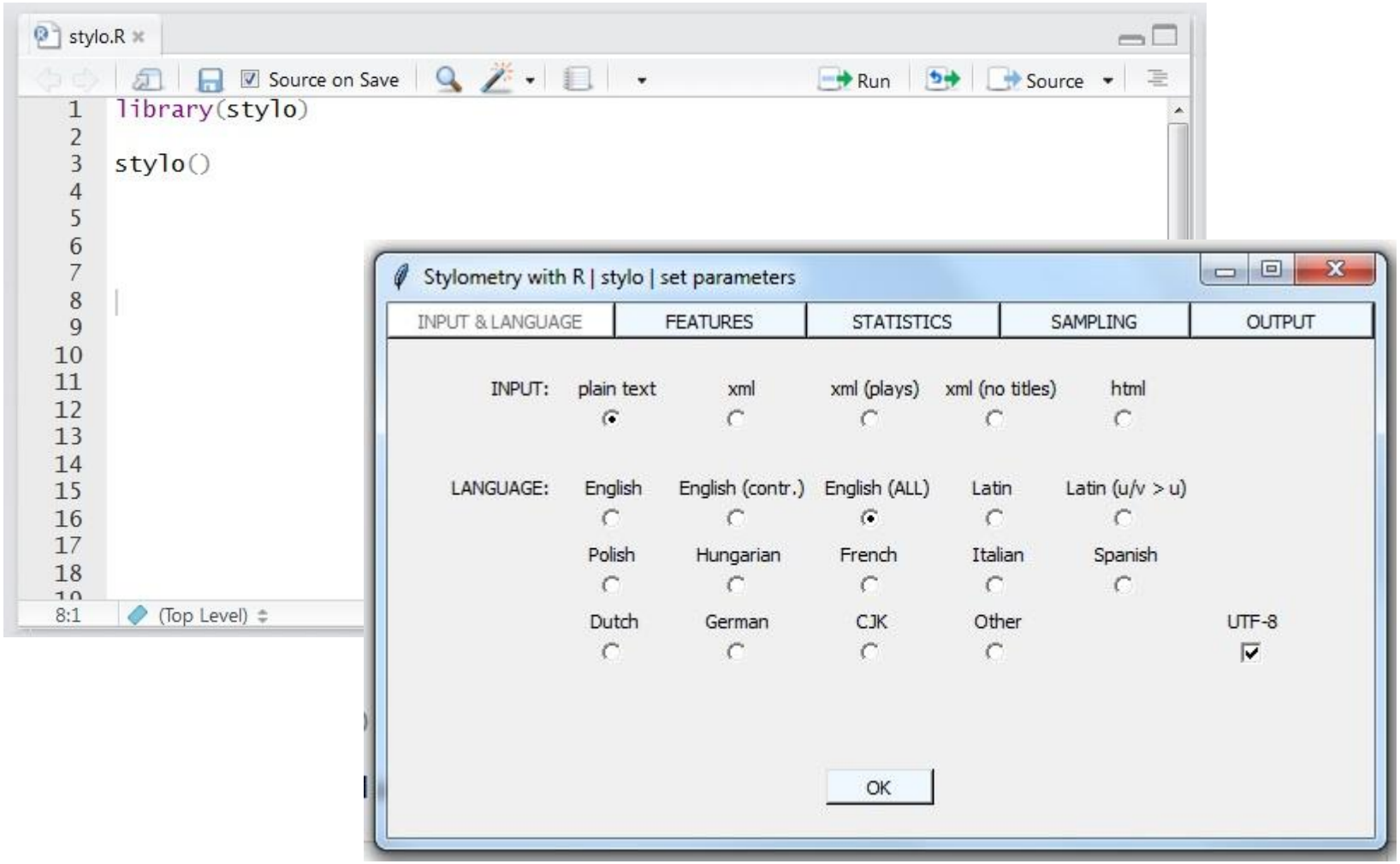


Corpus preparation - labels

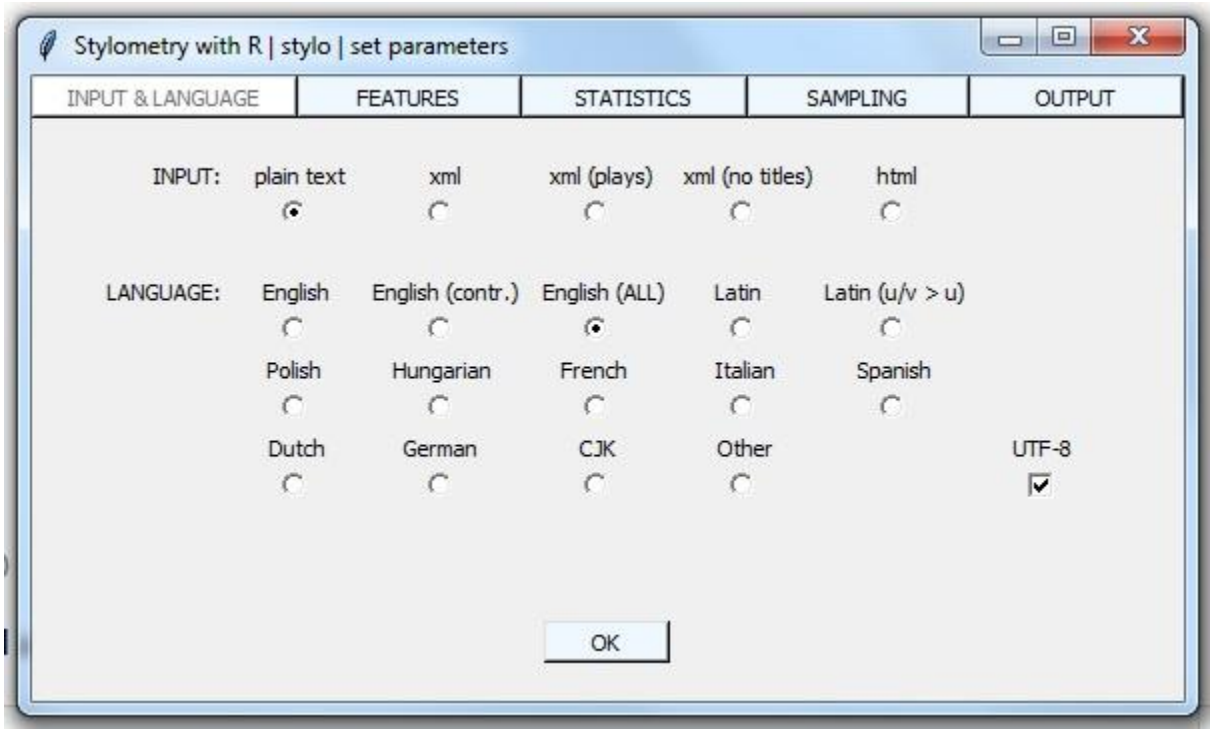
Colors on graphs are assigned according to filenames: the sequence of letters before “ - ” (underscore) is assumed to be the label of the author (genre, etc.).

- ABronte_Agnes.txt
- Austen_Emma.txt
- Austen_Pride.txt
- Conrad_Lord.txt
- Dickens_Pickwick.txt
- ...

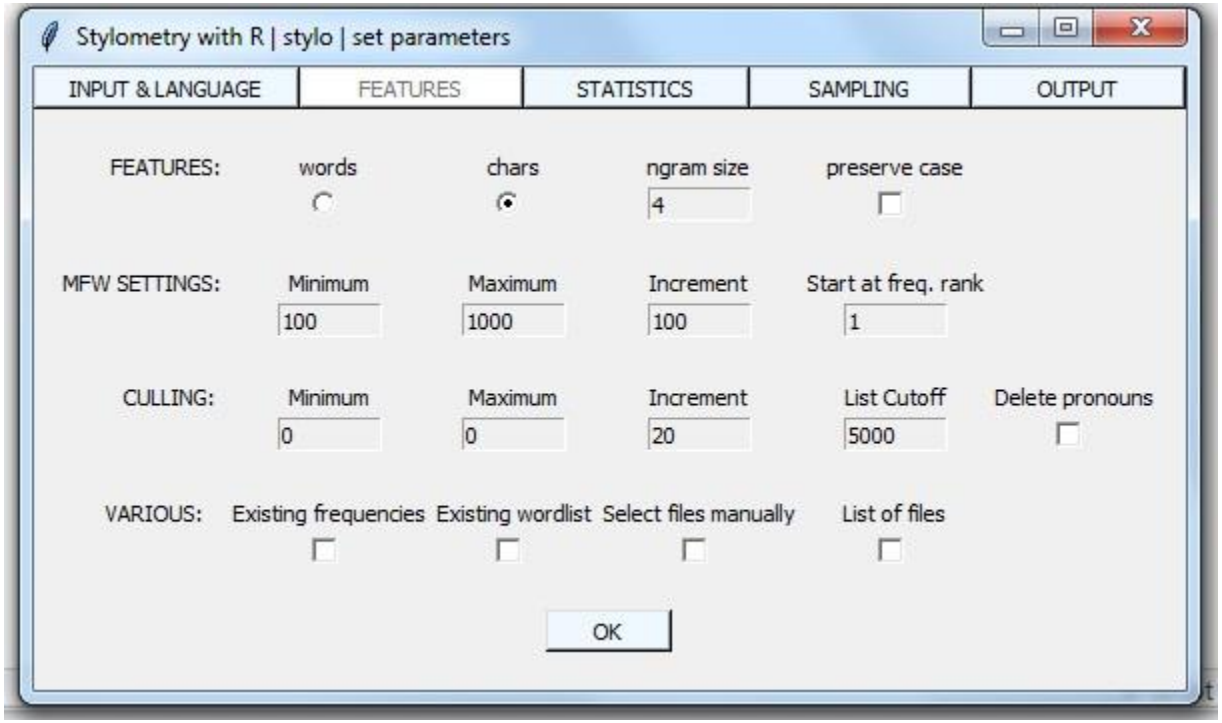
LIVE DEMO


stylo.R
Source on Save
Run
Source
1 library(stylo)
3 stylo()
8:1
(Top Level)
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

INPUT & LANGUAGE



FEATURES


Stylometry with R | stylo | set parameters
INPUT & LANGUAGE
FEATURES
STATISTICS
SAMPLING
OUTPUT
FEATURES:
words
chars
ngram size
4
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
OK

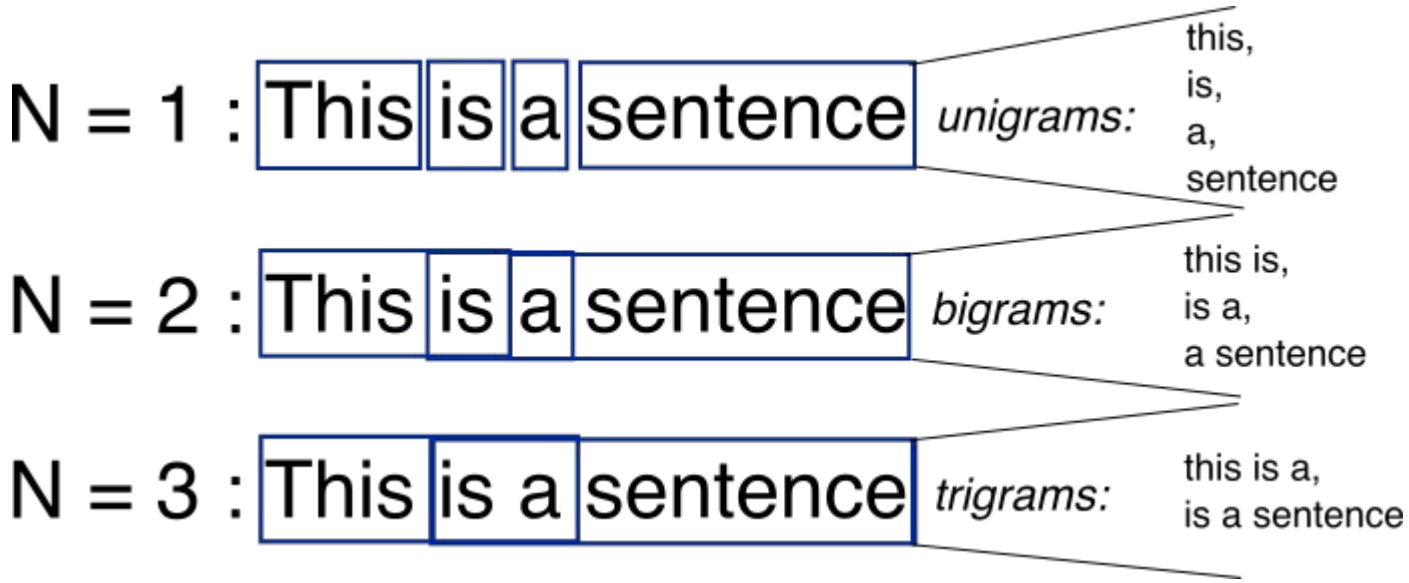
FEATURES

- 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.

FEATURES

	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

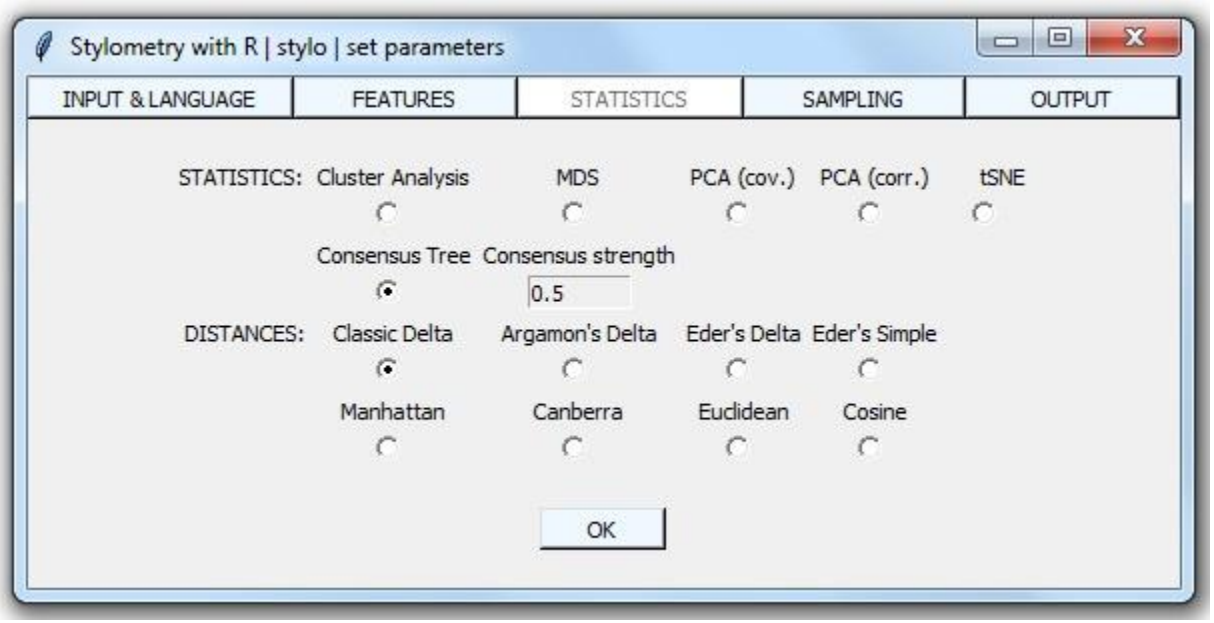
Culling

- 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.

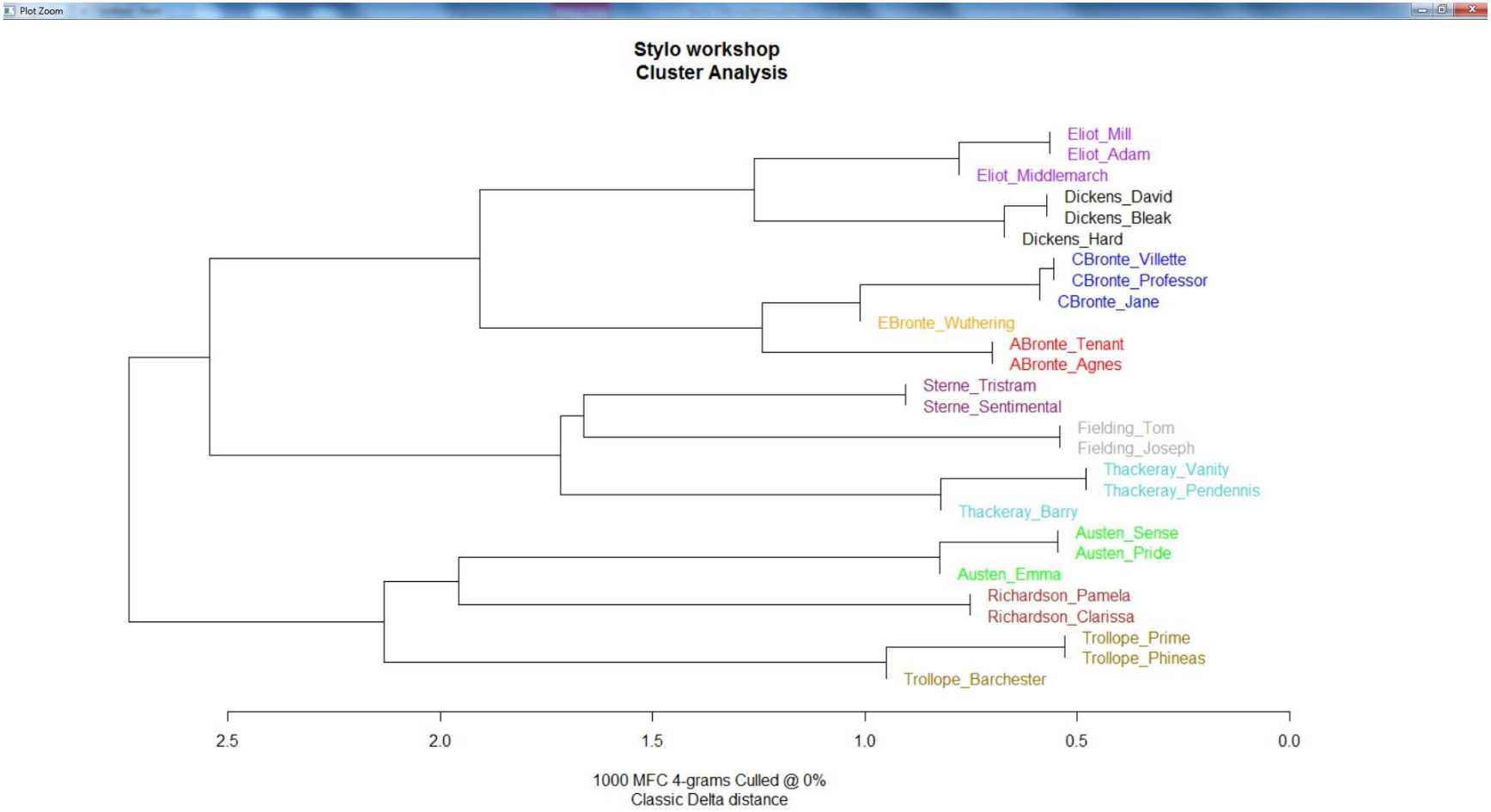
Culling

	Word Frequency								
Book Number	The	Big-Data	Analytics	Tree	newbie	book	for	Girl	honest
1	120	0	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

STATISTICS


Stylometry with R | stylo | set parameters
INPUT & LANGUAGE
FEATURES
STATISTICS
SAMPLING
OUTPUT
STATISTICS:
Cluster Analysis
MDS
PCA (cov.)
PCA (corr.)
tSNE
Consensus Tree
Consensus strength
0.5
DISTANCES:
Classic Delta
Argamon's Delta
Eder's Delta
Eder's Simple
Manhattan
Canberra
Euclidean
Cosine
OK

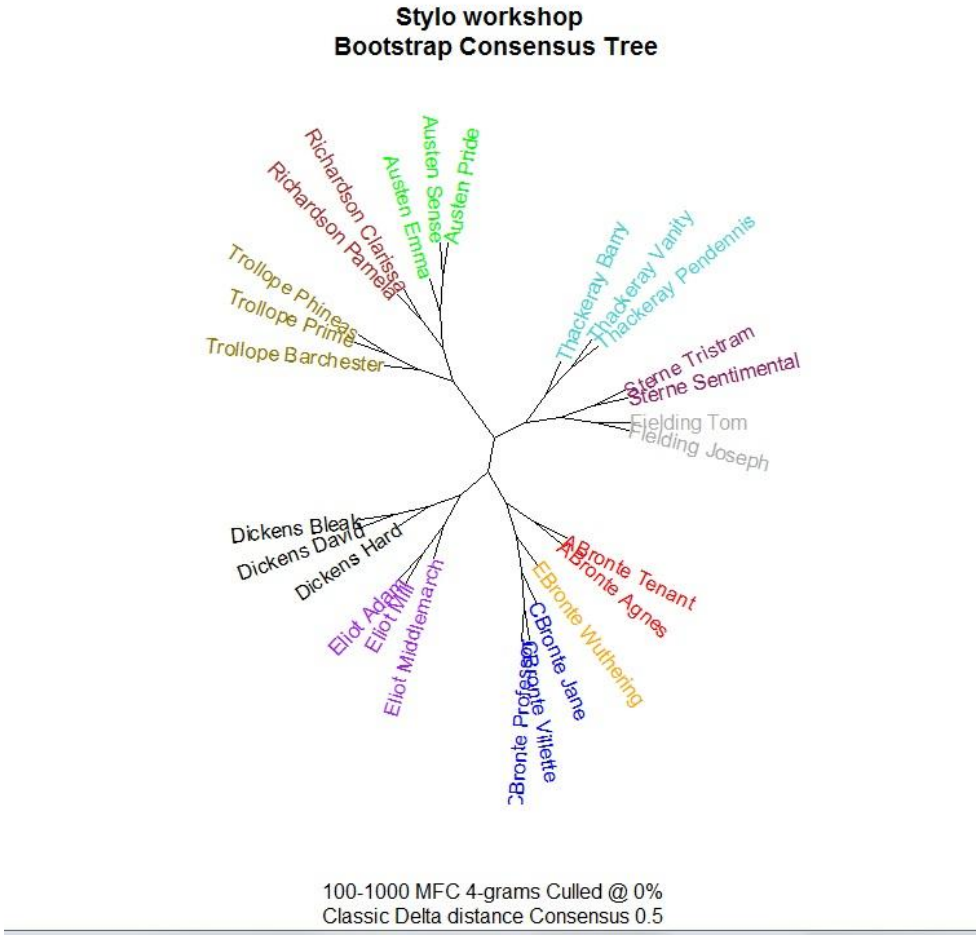
CLUSTER ANALYSIS



Cluster Analysis

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

Consensus tree

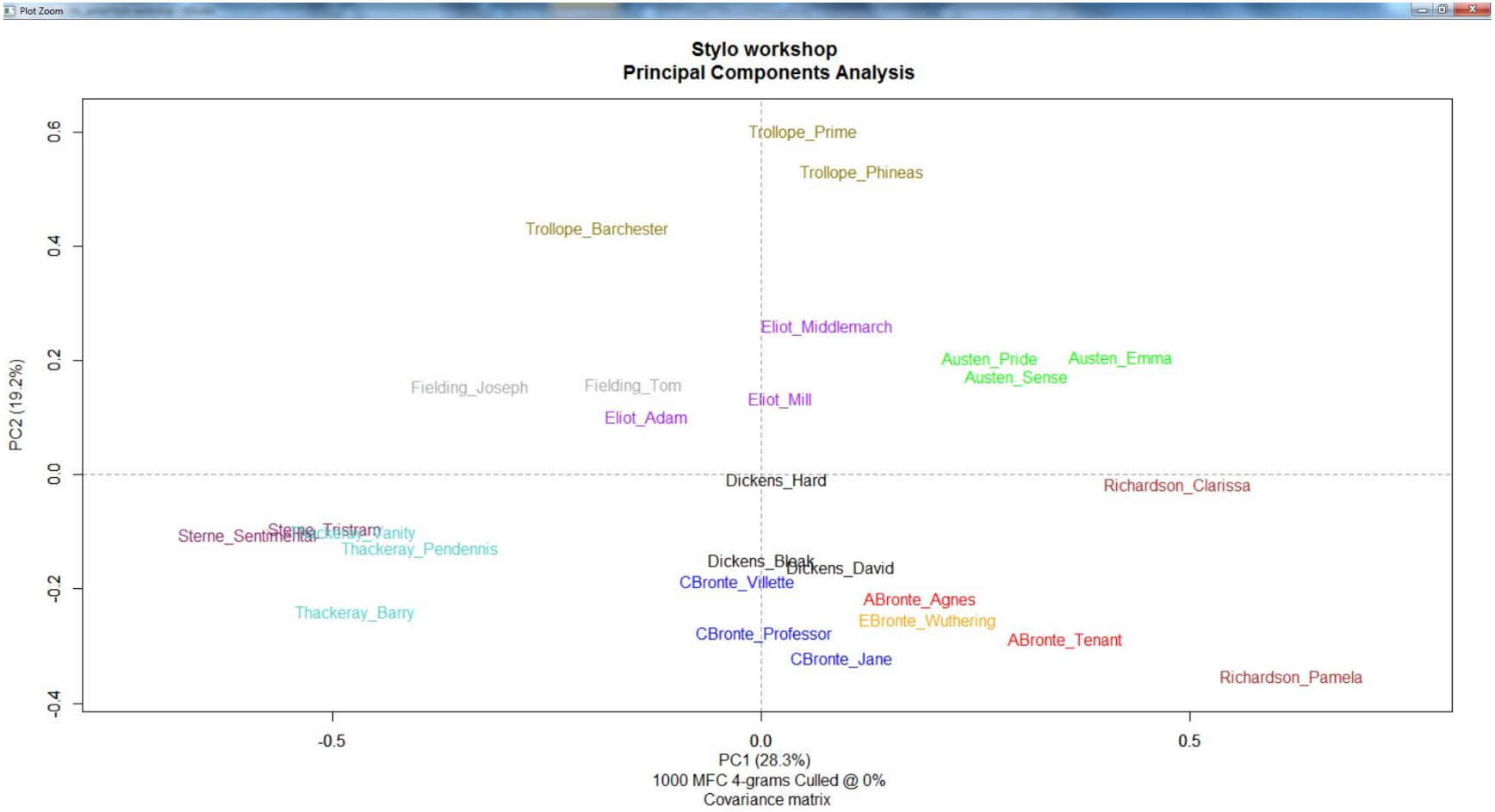


Consensus Tree

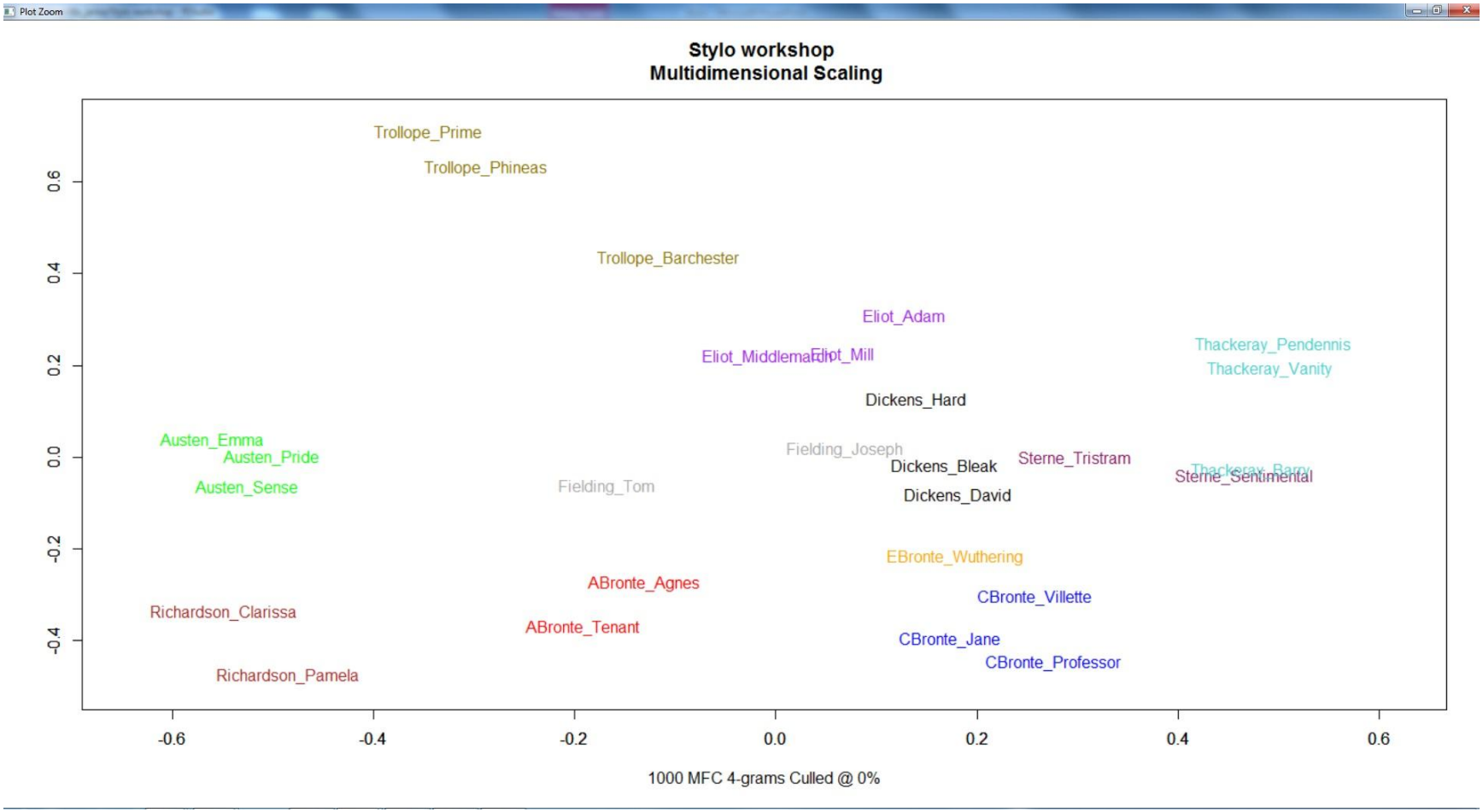
Uses many trees to discover unchanged patterns for different parameters

It is more robust but harder to interpret.

PCA (principal component analysis)



MDS (multidimensional scaling)

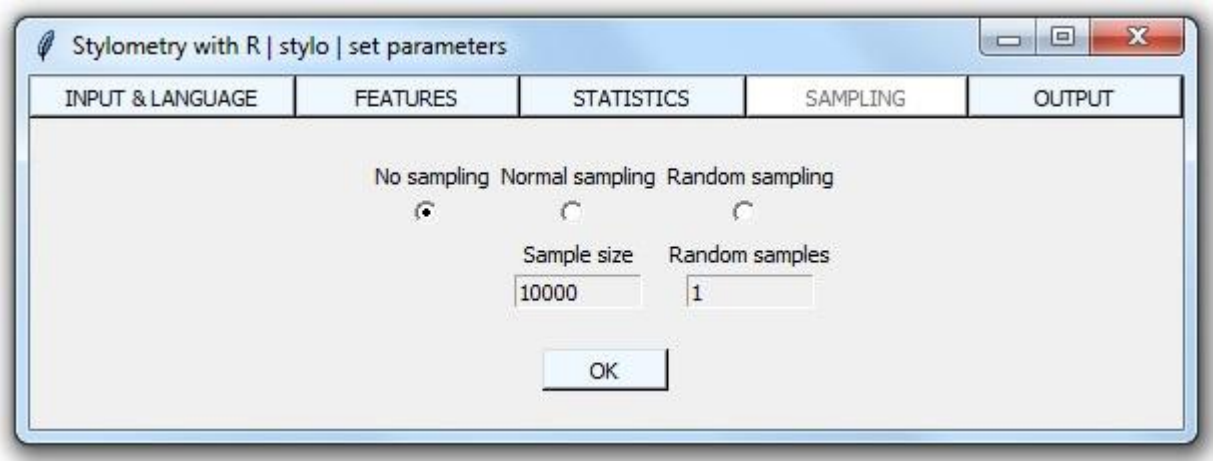


Dimension reduction

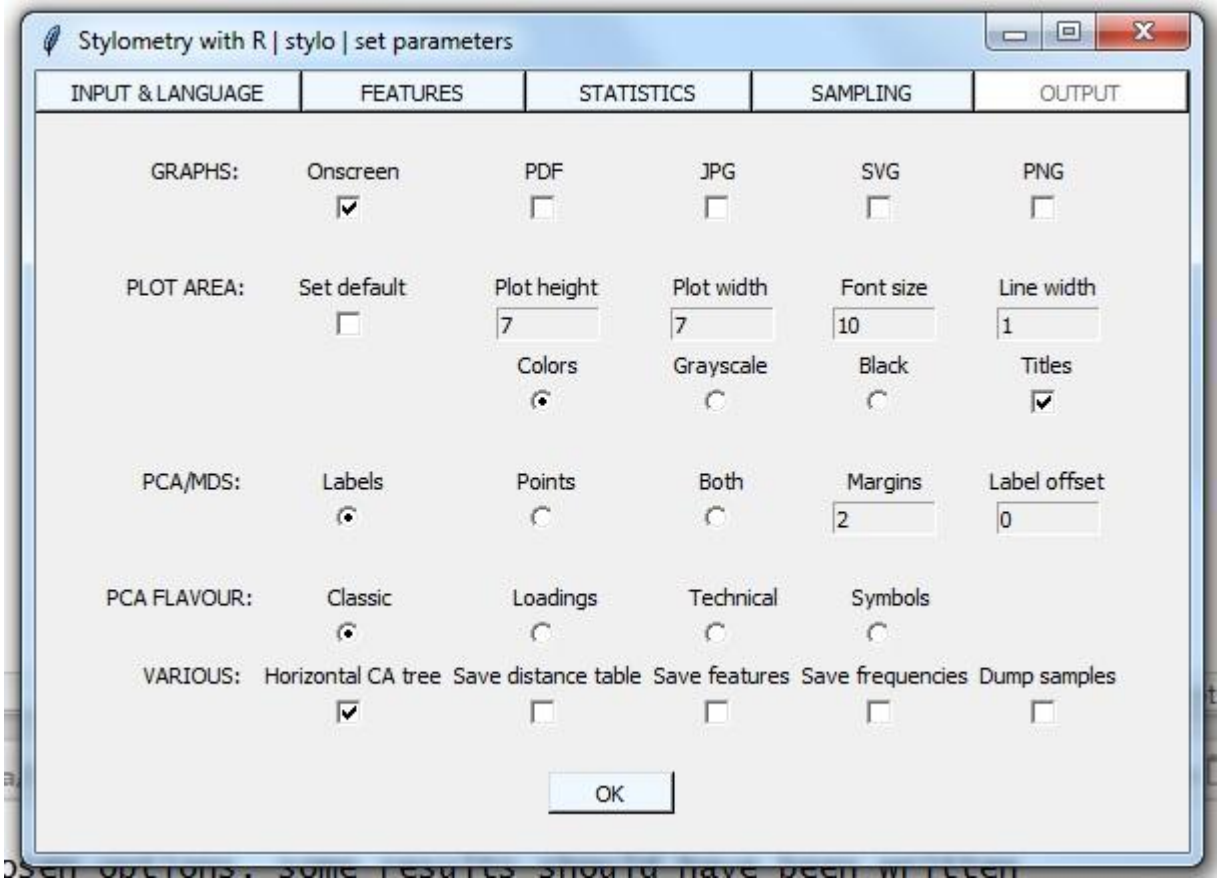
Methods to reduce high dimensions into easy to interpret 2D or 3D space.

It is easy to interpret and visualization gives good intuitions on the similarity between texts.

SAMPLING


Stylometry with R | stylo | set parameters
INPUT & LANGUAGE
FEATURES
STATISTICS
SAMPLING
OUTPUT
No sampling
Normal sampling
Random sampling
Sample size
Random samples
10000
1
OK

OUTPUT


Stylometry with R | stylo | set parameters
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

Graphs

- Onscreen: the graph on the screen
- PDF: a PDF file with your graph
- JPG: graph in JPEG format
- SVG: a SVG vector file
- PNG: graph in PNG format

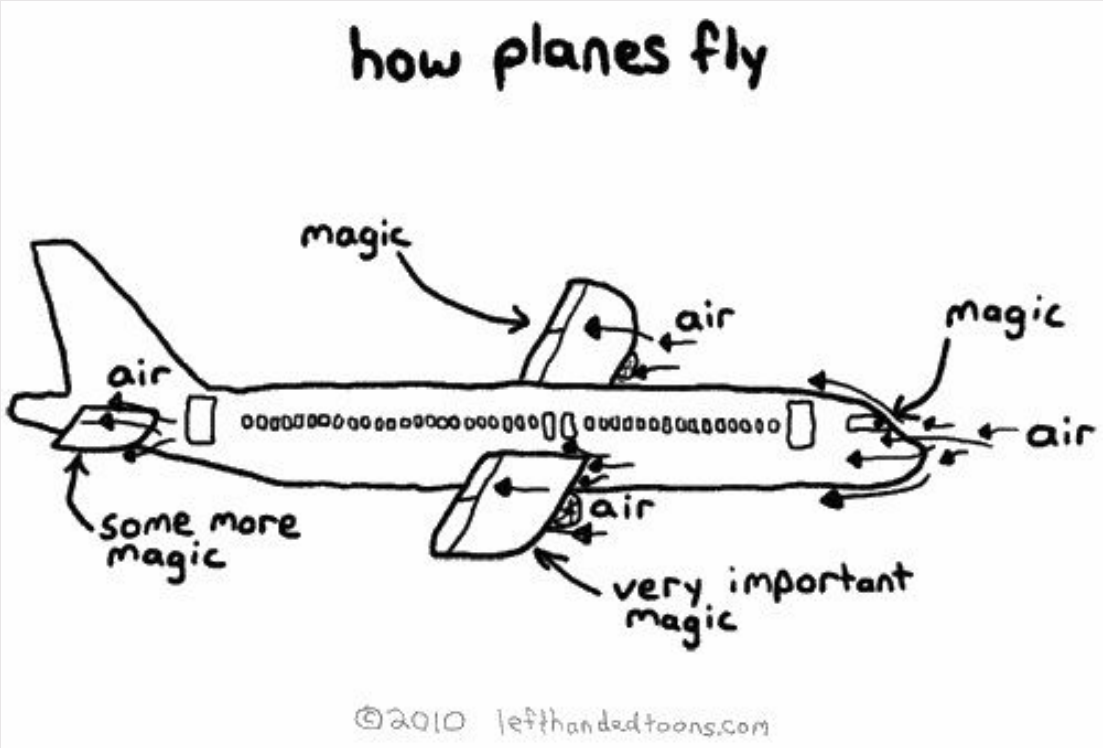
Lets get it to work

```
> library(stylo)
> stylo()
using current directory...
Performing no sampling (using entire text as sample)
loading ABronte_Agnes.txt       ...
loading ABronte_Tenant.txt      ...
loading Austen_Emma.txt ...
loading Austen_Pride.txt        ...
loading Austen_Sense.txt        ...
loading CBronte_Jane.txt        ...
loading CBronte_Professor.txt   ...
loading CBronte_Villette.txt    ...
loading Dickens_Bleak.txt       ...
loading Dickens_David.txt       ...
loading Dickens_Hard.txt        ...
loading EBronte_Wuthering.txt   ...
loading Eliot_Adam.txt  ...
loading Eliot_Middlemarch.txt   ...
loading Eliot_Mill.txt  ...
loading Fielding_Joseph.txt     ...
loading Fielding_Tom.txt        ...
loading Richardson_Clarissa.txt ...
```

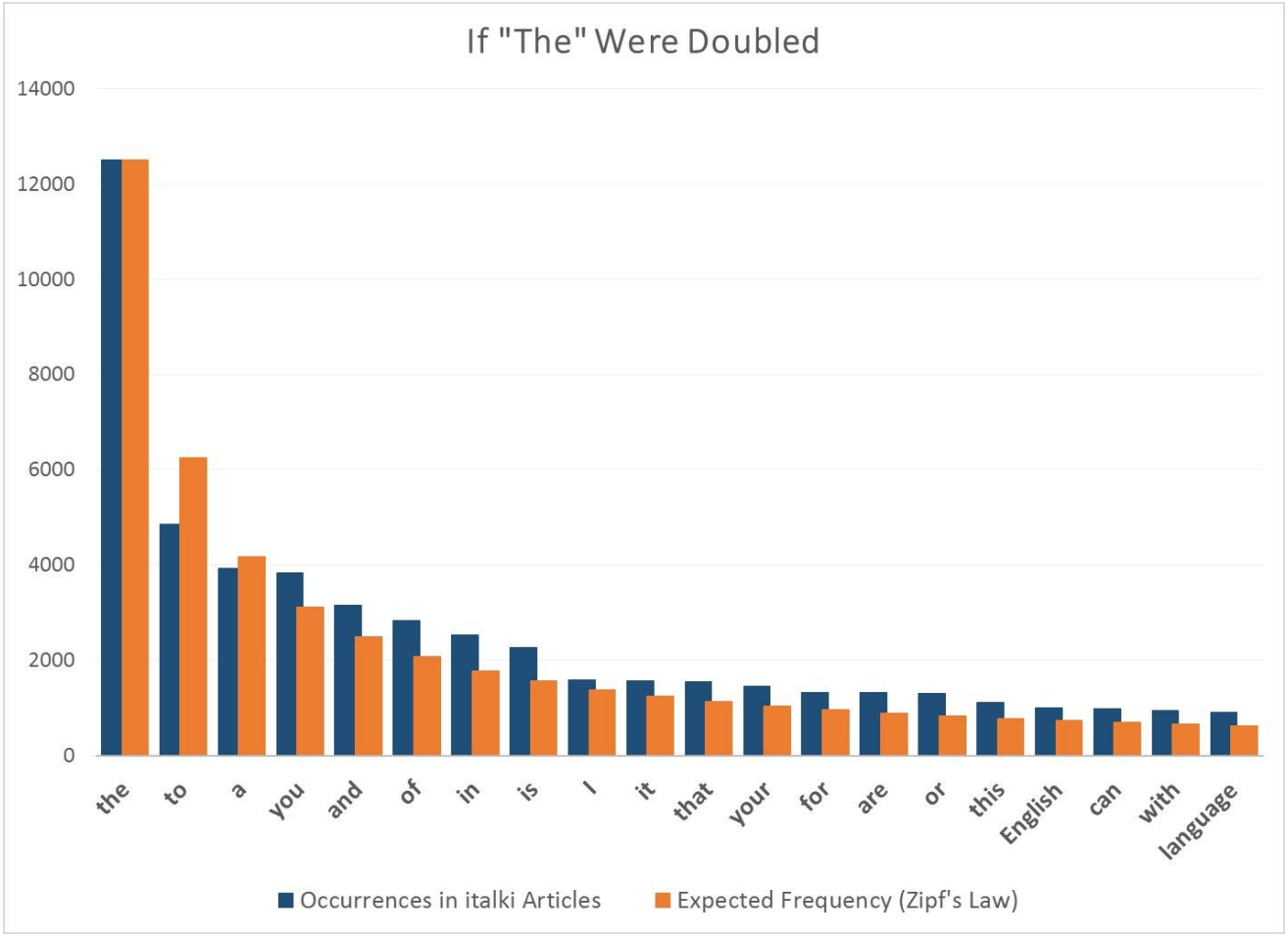
Stylo - Exercise

- Use 100 English Novels for stylo() analysis
- Use Eder’s and Manhattan Distance
- Cluster analysis, PCA, Consensus treen
- Use different values for culling and MFW

Intuitions how it is done



Zpif Law I


If "The" Were Doubled
14000
12000
10000
8000
6000
4000
2000
0
the
to
a
you
and
of
in
is
I
it
that
your
for
are
or
this
English
can
with
language
Occurrences in italki Articles
Expected Frequency (Zipf's Law)

Zpif Law II

1.	i	15.	po	3600.	przystani
2.	się	16.	za	3601.	racji
3.	w	17.	tak	3602.	twardo
4.	nie	18.	już	...	...
5.	na	19.	jej	80725.	dzieciobójstwo
6.	z	20.	od	80726.	dzieciskami
7.	do	...	...	80727.	dziecisków
8.	to	3593.	ludzki	80728.	dzieciński
9.	a	3594.	mgnienie	80729.	dziedzicami
10.	że	3595.	pacierz	80730.	dziedzicowe
11.	ale	3596.	pióra	80731.	dziedzicowym
12.	jak	3597.	podała	80732.	dziedzictwem
13.	o	3598.	postawie	80733.	dziedziczko
14.	co	3599.	pragnęła	80734.	dziedzicznej

Zpif Law III

The more frequent word the more meanings it has

Term document frequency - novel into numbers

	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

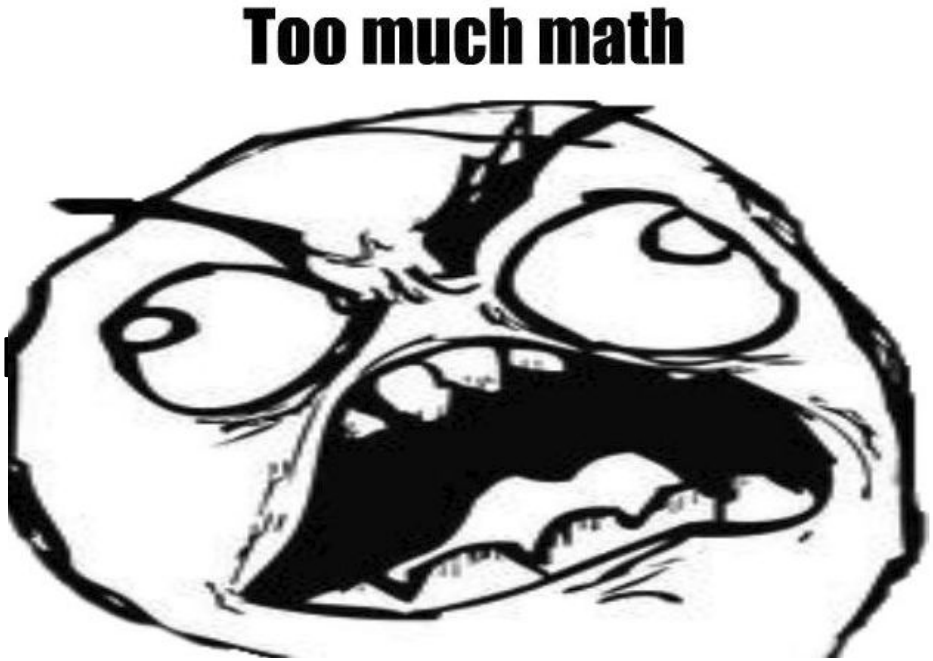
Z Score

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

$$Z = \frac{X - \bar{X}}{S}$$

Z Score WHY SO MUCH MATH?

- X - frequency of term
- mean(X) -
- S - standar



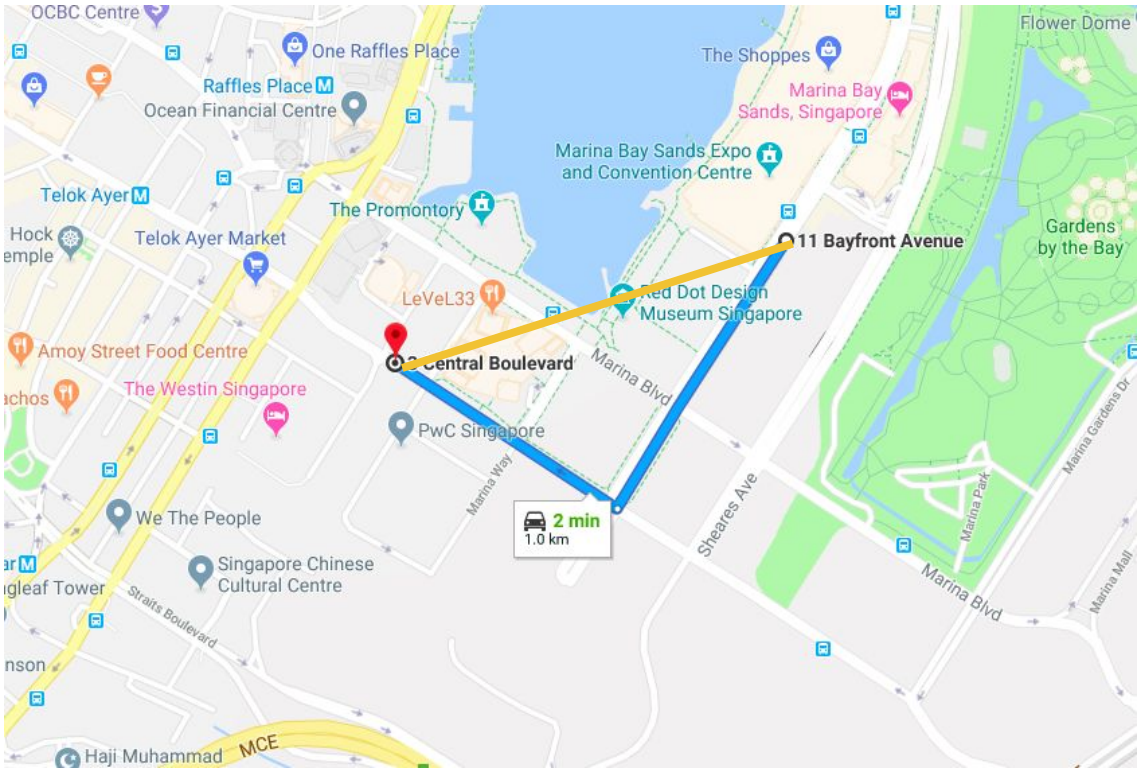
$$Z = \frac{}{S}$$

Term document frequency - novel into numbers

	"EBronte_Wuthering"	"Eliot_Mill_rep"
"the"	3.9190127861008	4.19596902207898
"and"	4.0682268396635	2.97705517340901
"to"	2.98771127938188	2.92943383520131
"of"	1.90548061503632	2.43013131944778
"i"	3.07775424273868	1.32810621001491
"a"	1.98866316213736	2.31709076915677
"in"	1.25631372683538	1.49405935831449
"that"	1.03249264649133	1.41805762662947
"he"	1.67479911843651	1.10106306219635
"it"	1.12425071391207	1.15493770744144
"you"	1.48956787953109	1.16792534513445
"was"	0.966461140029671	1.46471691760065
"her"	1.33006320158476	1.38390494973303
"his"	1.22115409352462	0.948097551589783
"as"	0.802668701923489	0.968781567174948
"my"	0.948452547358311	0.372312280532974
"for"	0.722058810918352	0.856222040502189
"with"	0.691186937767449	1.0063014093992
"not"	0.803526253955459	0.671027947472221
"be"	0.618295015050038	0.683053537928712
"had"	0.590853350027013	1.0014911732166
"she"	1.10023925701692	0.946654480735004
"have"	0.542830436236719	0.538265428832556
"me"	0.91243536201559	0.417047477031122
"but"	0.589138245963074	0.771080860070229
"is"	0.560839028908079	0.427148973014575
"at"	0.672320793064119	0.682091490692193
"him"	0.794093181603794	0.513733224301313
"so"	0.307861179477065	0.347780076001732
"this"	0.253835401462984	0.381932752898167

Distances

- Time
- Manhattan
- Euclidean



Distance

- dictionary (the, dog, cat, to be, sun)
- text 1 (12, 4, 4, 5, 1)
- text 2 (21, 1, 7, 5, 2)
- text 3 (17, 3, 6, 3, 2)

d(text1, text2)=|12-21| + |4-1| + |4-7| + |5-5| + |1-2|

= 9 + 3 + 3 + 0 + 1 =

= 16

Distance - Exercise

- dictionary (the, dog, cat, to be, sun)
- text 1 (12, 4, 4, 5, 1)
- text 2 (21, 1, 7, 5, 2)
- text 3 (17, 3, 6, 3, 2)

Exercise:

Calculate d(text2, text3).

Distances

- Non-standard words weights adjusted
- Non-standard words weights adjusted
- More importance to MFV
- Modification of Manhattan distance
- All words equal weights

- Classic Delta:

$$\delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{f_i(A) - f_i(B)}{\sigma_i}\right|$$

- Argamon's Delta:

$$\delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\sqrt{f_i(A)^2 - f_i(B)^2}}{\sigma_i}\right|$$

- Eder's Delta:

$$\delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left(\left|\frac{f_i(A) - f_i(B)}{\sigma_i}\right| \times \frac{n - n_i + 1}{n}\right)$$

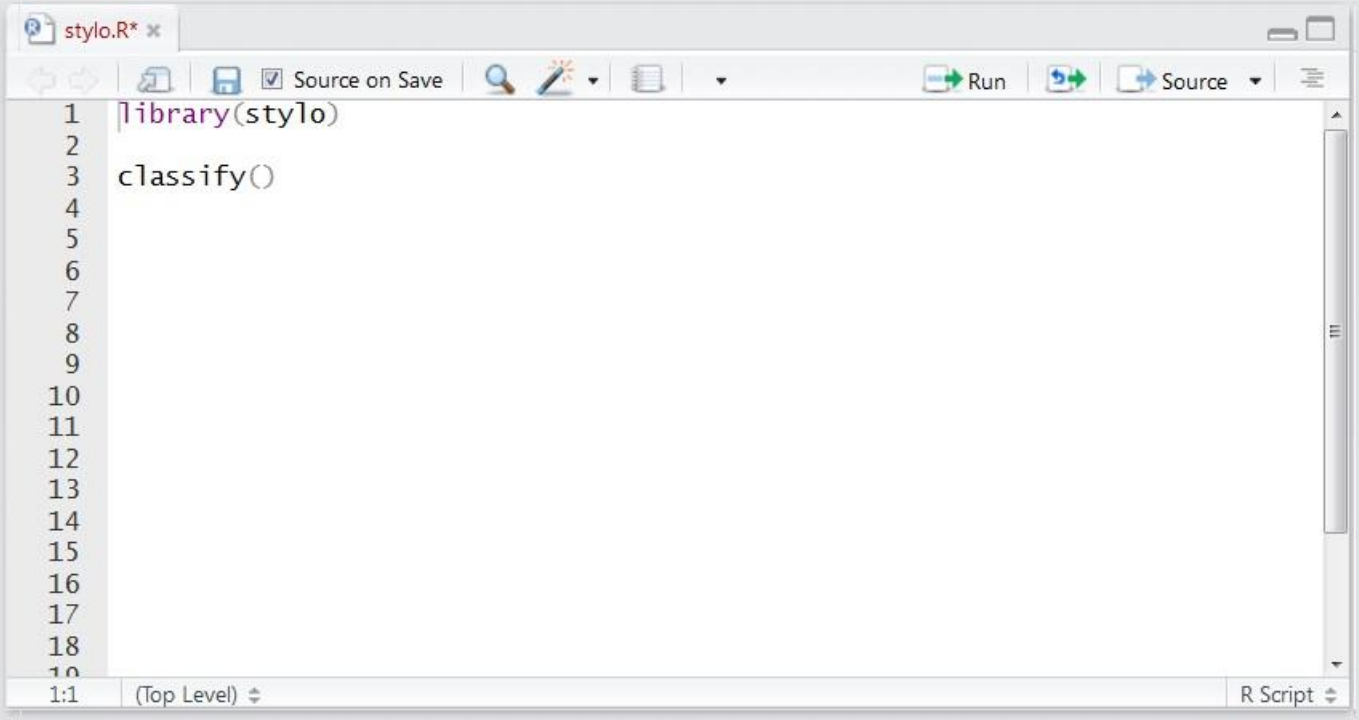
- Eder's Simple:

$$\delta_{(AB)} = \sum_{i=1}^{n}\left|\sqrt{f_i(A)} - \sqrt{f_i(B)}\right|$$

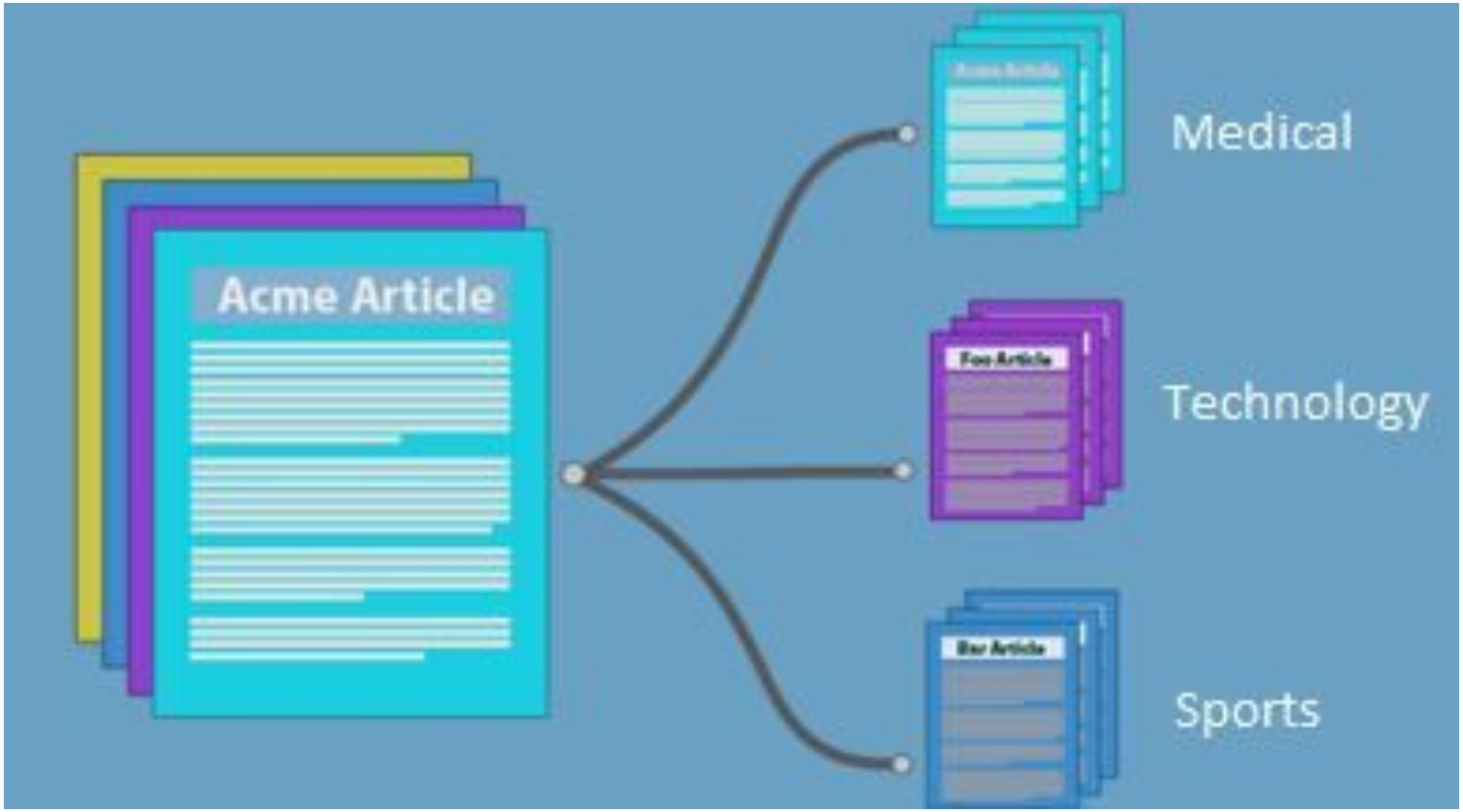
- Manhattan:

$$\delta_{(AB)} = \sum_{i=1}^{n}|f_i(A) - f_i(B)|$$

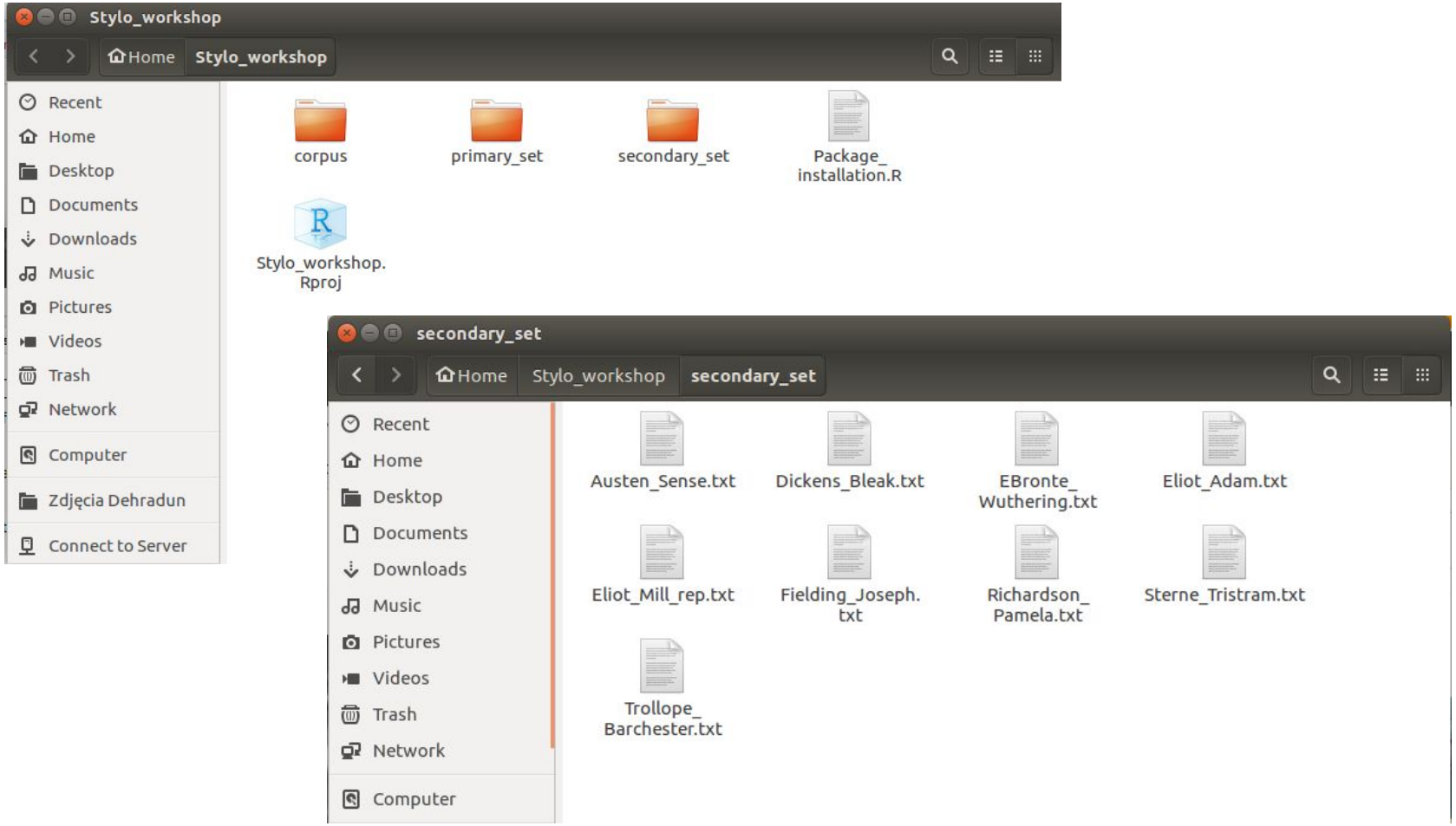
CLASSIFY


stylo.R*
Source on Save
Run
Source
1 library(stylo)
3 classify()
1:1
(Top Level)
R Script

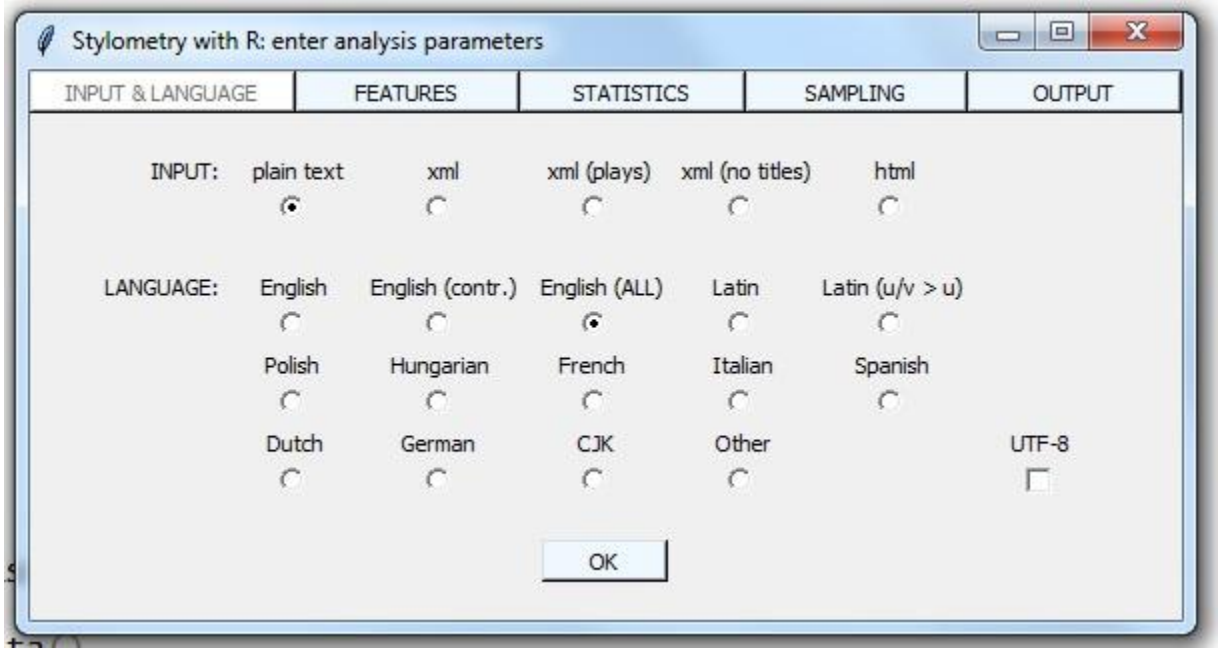
What is classification?



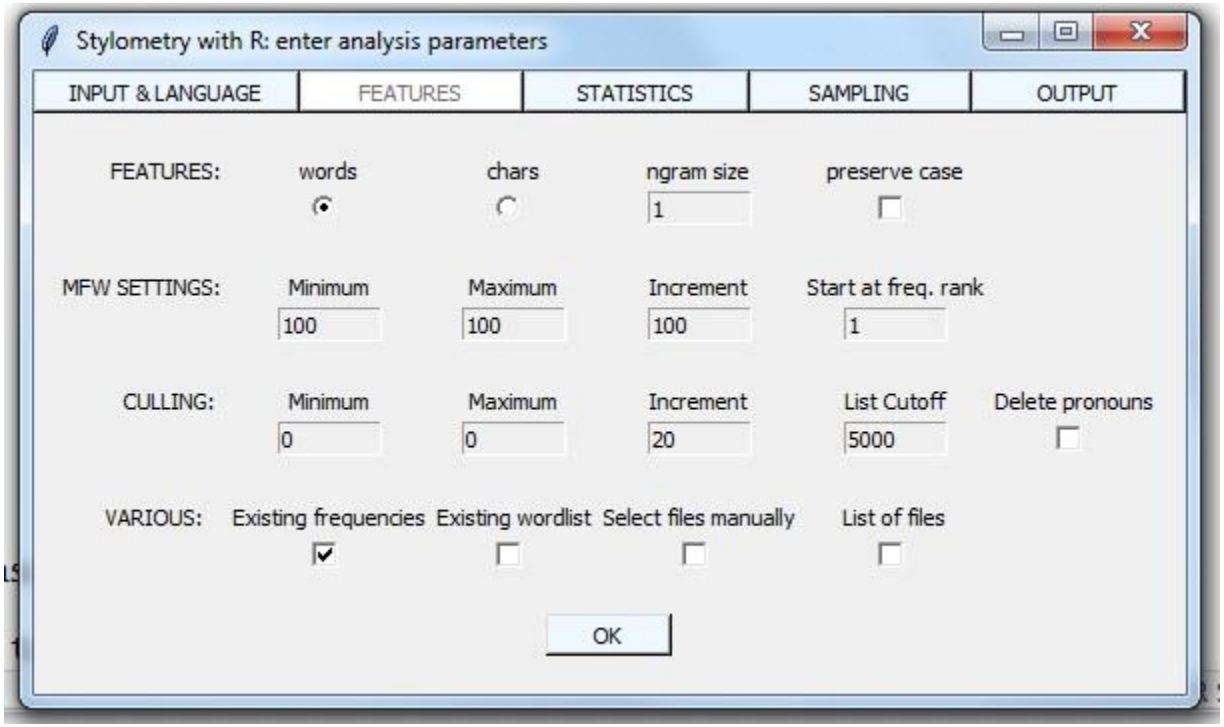
CORPUS PREPARATION FOR CLASSIFY



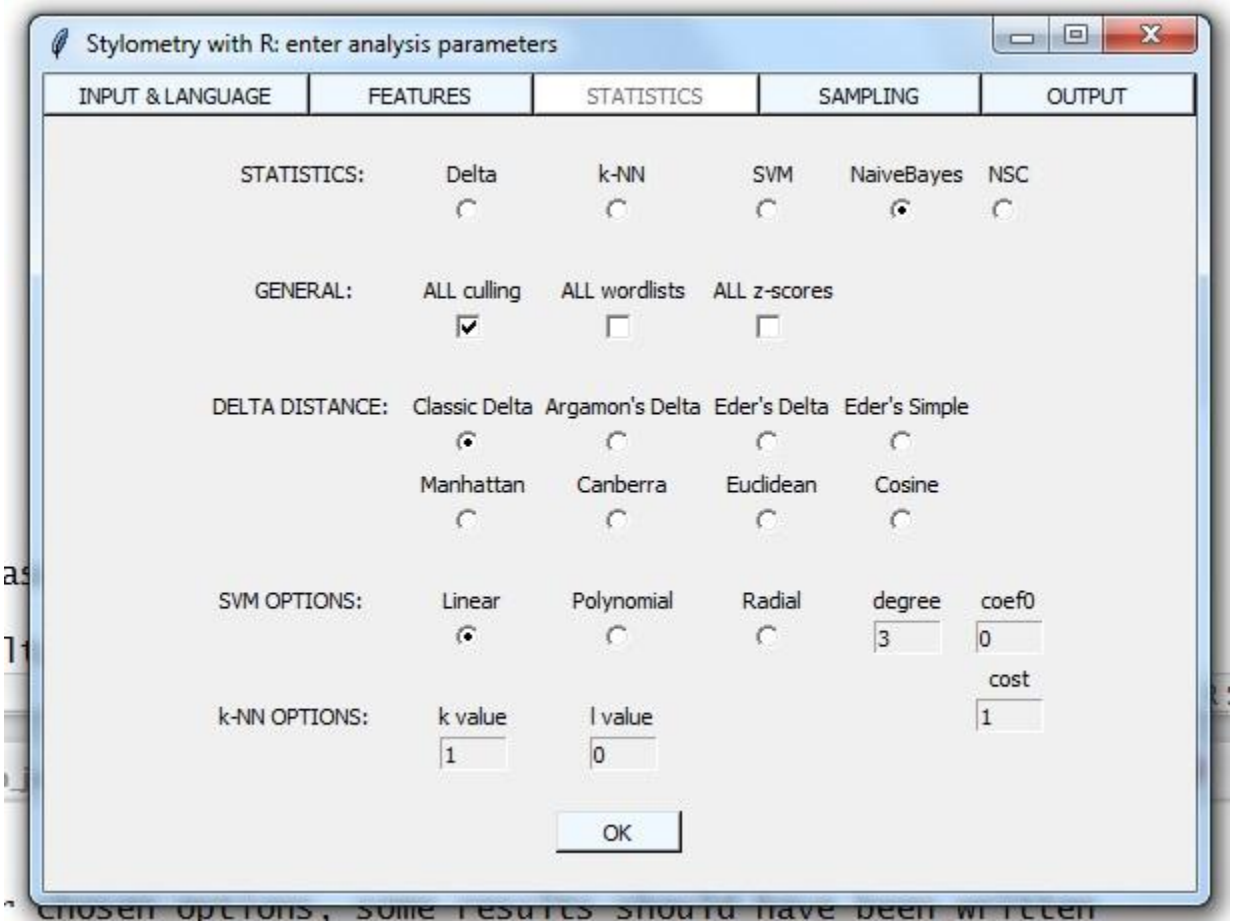
INPUT&LAGUAGE



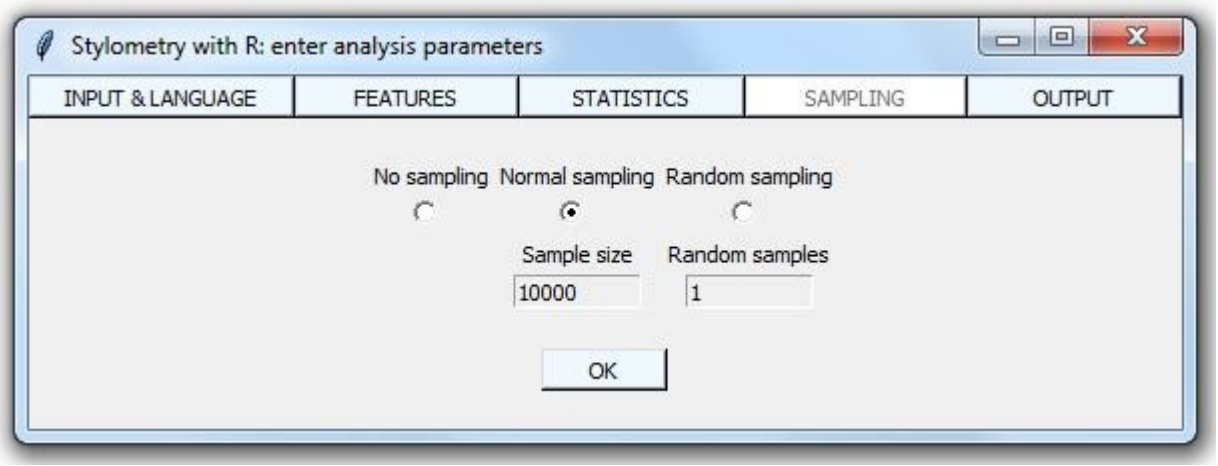
FEATURES


Stylometry with R: enter analysis parameters
INPUT & LANGUAGE
FEATURES
STATISTICS
SAMPLING
OUTPUT
FEATURES:
words
chars
ngram size
preserve case
1
MFW SETTINGS:
Minimum
Maximum
Increment
Start at freq. rank
100
100
100
1
CULLING:
Minimum
Maximum
Increment
List Cutoff
Delete pronouns
0
0
20
5000
VARIOUS:
Existing frequencies
Existing wordlist
Select files manually
List of files
OK

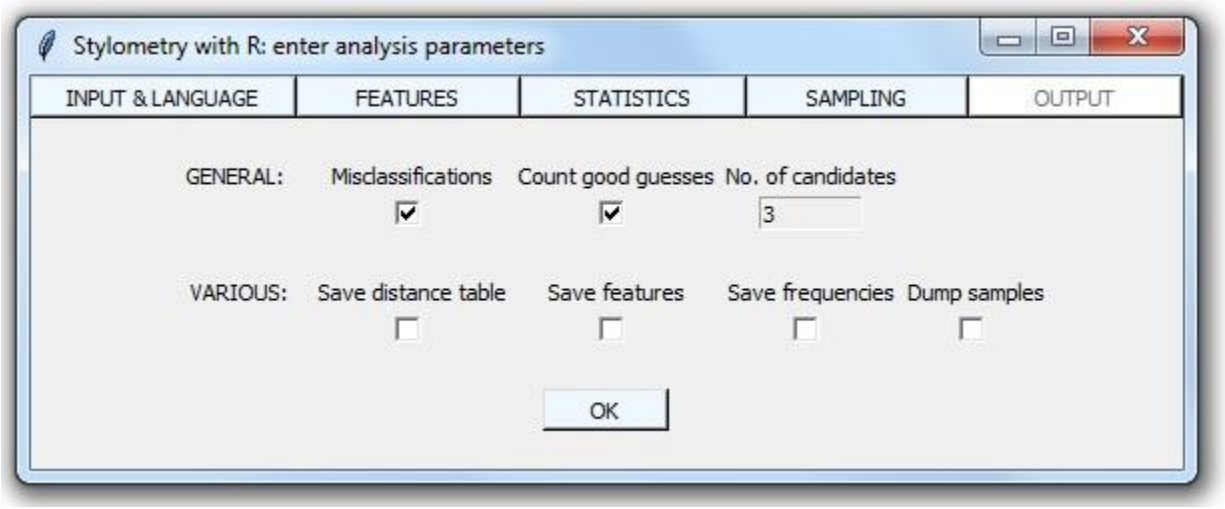
STATISTICS


Stylometry with R: enter analysis parameters
INPUT & LANGUAGE
FEATURES
STATISTICS
SAMPLING
OUTPUT
STATISTICS:
Delta
k-NN
SVM
NaiveBayes
NSC
GENERAL:
ALL culling
ALL wordlists
ALL z-scores
DELTA DISTANCE:
Classic Delta
Argamon's Delta
Eder's Delta
Eder's Simple
Manhattan
Canberra
Euclidean
Cosine
SVM OPTIONS:
Linear
Polynomial
Radial
degree
coef0
3
0
cost
1
k-NN OPTIONS:
k value
l value
1
0
OK

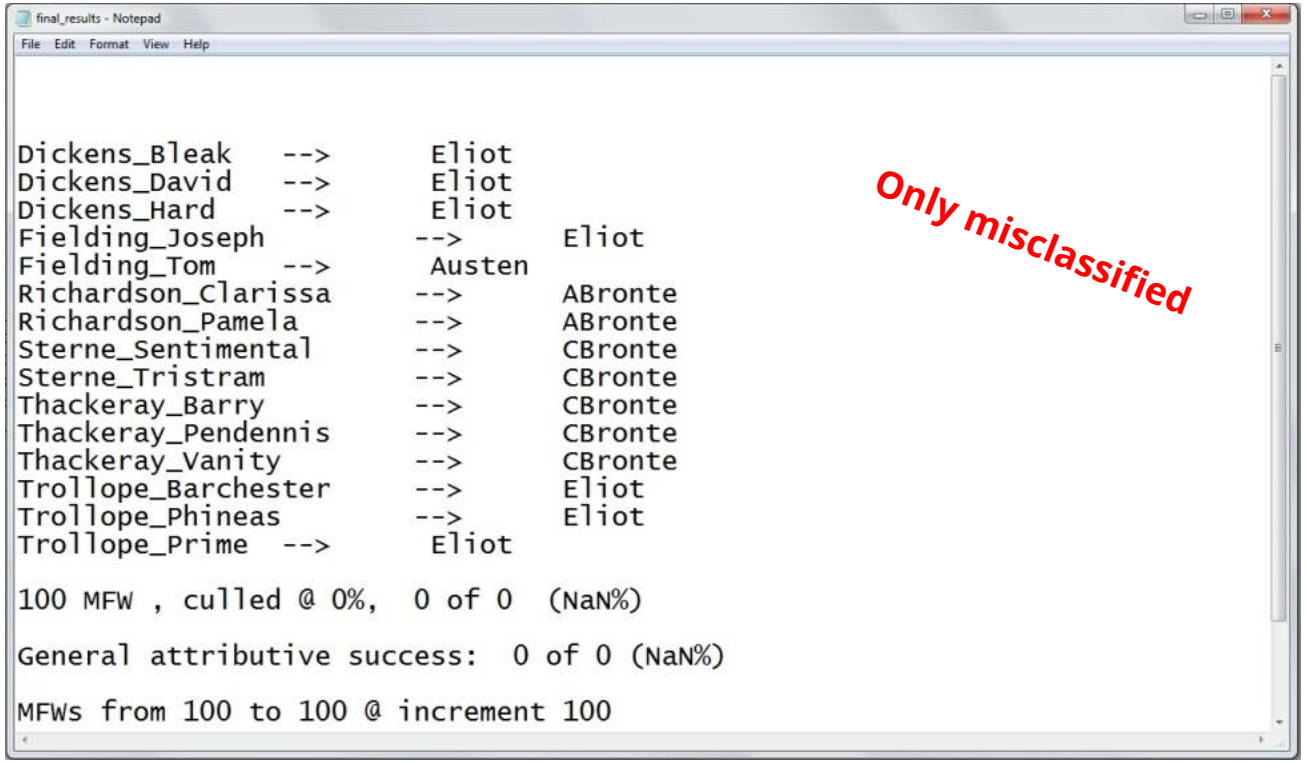
SAMPLING


Stylometry with R: enter analysis parameters
INPUT & LANGUAGE
FEATURES
STATISTICS
SAMPLING
OUTPUT
No sampling
Normal sampling
Random sampling
Sample size
Random samples
10000
1
OK

OUTPUT


Stylometry with R: enter analysis parameters
INPUT & LANGUAGE
FEATURES
STATISTICS
SAMPLING
OUTPUT
GENERAL:
Misclassifications
Count good guesses
No. of candidates
3
VARIOUS:
Save distance table
Save features
Save frequencies
Dump samples
OK

RESULTS FOR CLASSIFY



WAY FORWARD

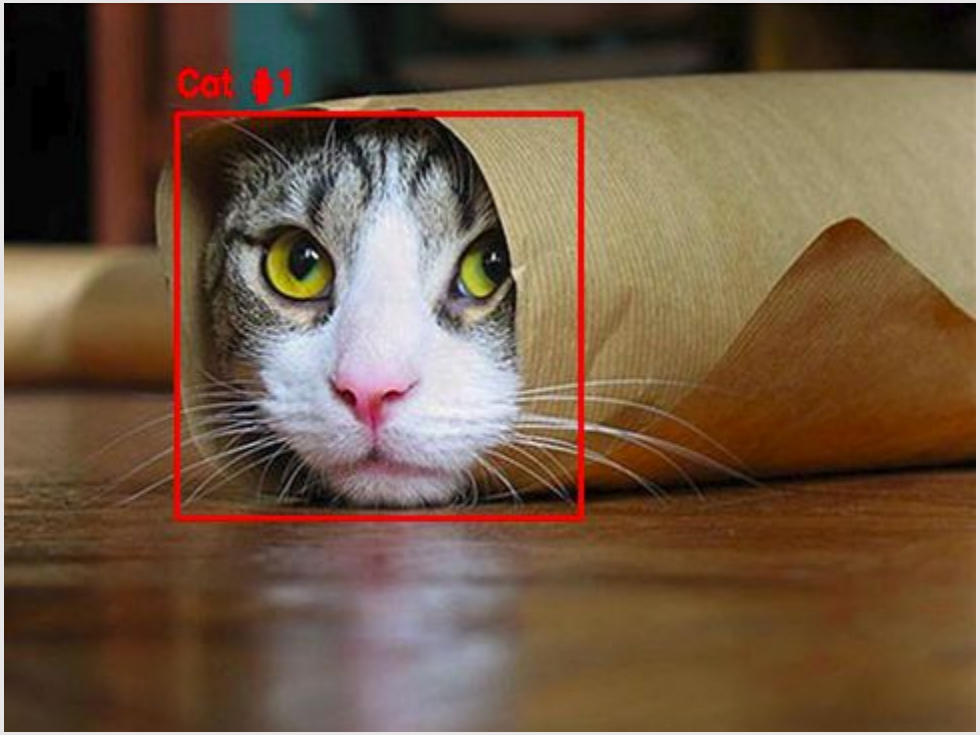
Stylo_workflow.R | Wordcloud.R | Creating_graph.R

Source on Save | Run | Source

```r
library(stylo)
# Load corpus
raw.corpus <- load.corpus(files = 'all', corpus.dir = 'corpus', encoding = 'UTF-8')
# Summary of corpus files
summary(raw.corpus)
# Tokenization
tokenized.corpus <- txt.to.words.ext(raw.corpus, language = 'English.all', preserve.case = FALSE)
# Austin's Pride and Prejudice famous sentence
tokenized.corpus$Austen_Pride.txt[8:30]

# English pronouns
stylo.pronouns(language = 'English')

# Deleting unuseful words
corpus.no.pronouns <- delete.stop.words(tokenized.corpus, stop.words = stylo.pronouns(language = 'English'))

# Counting frequent features
frequent.features <- make.frequency.list(corpus.no.pronouns, head = 3000)
freqs <- make.table.of.frequencies(corpus.no.pronouns, features = frequent.features, relative = FALSE)
freqs <- make.table.of.frequencies(corpus.no.pronouns, features = frequent.features)

# Culling (words in at least 80% of samples)
culled.freqs <- perform.culling(freqs, culling.level = 80)

# Simple cluster analysis with stylo
stylo(frequencies = culled.freqs, gui = FALSE)

# More options in stylo
stylo(frequencies = culled.freqs, gui = FALSE, analysis.type = 'PCR', custom.graph.title = "PCA analysis for novels")
```

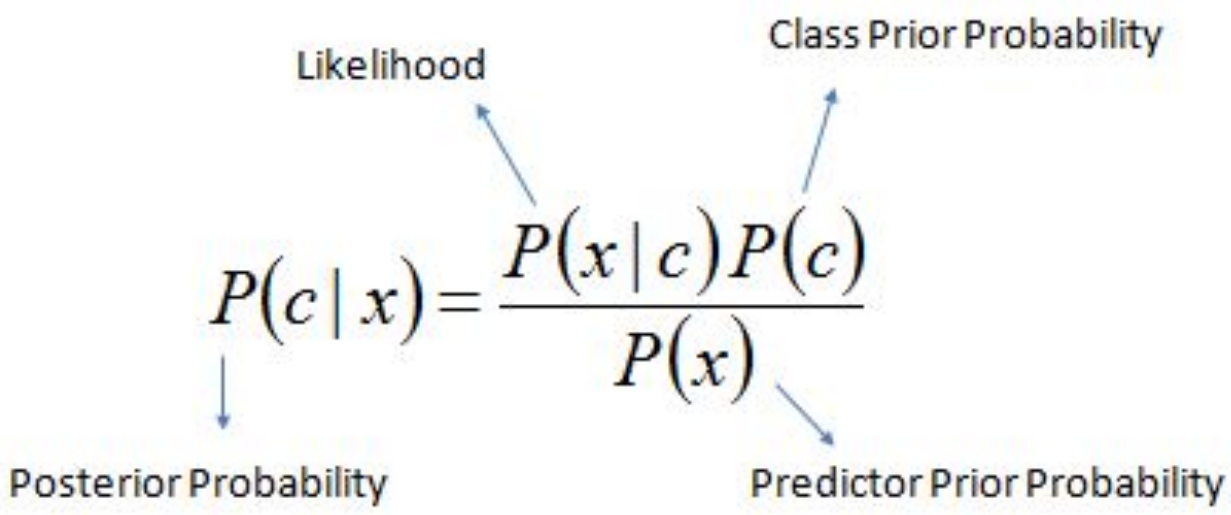
1:1 (Top Level) R Script

More intuitions on classification



Naive Bayes

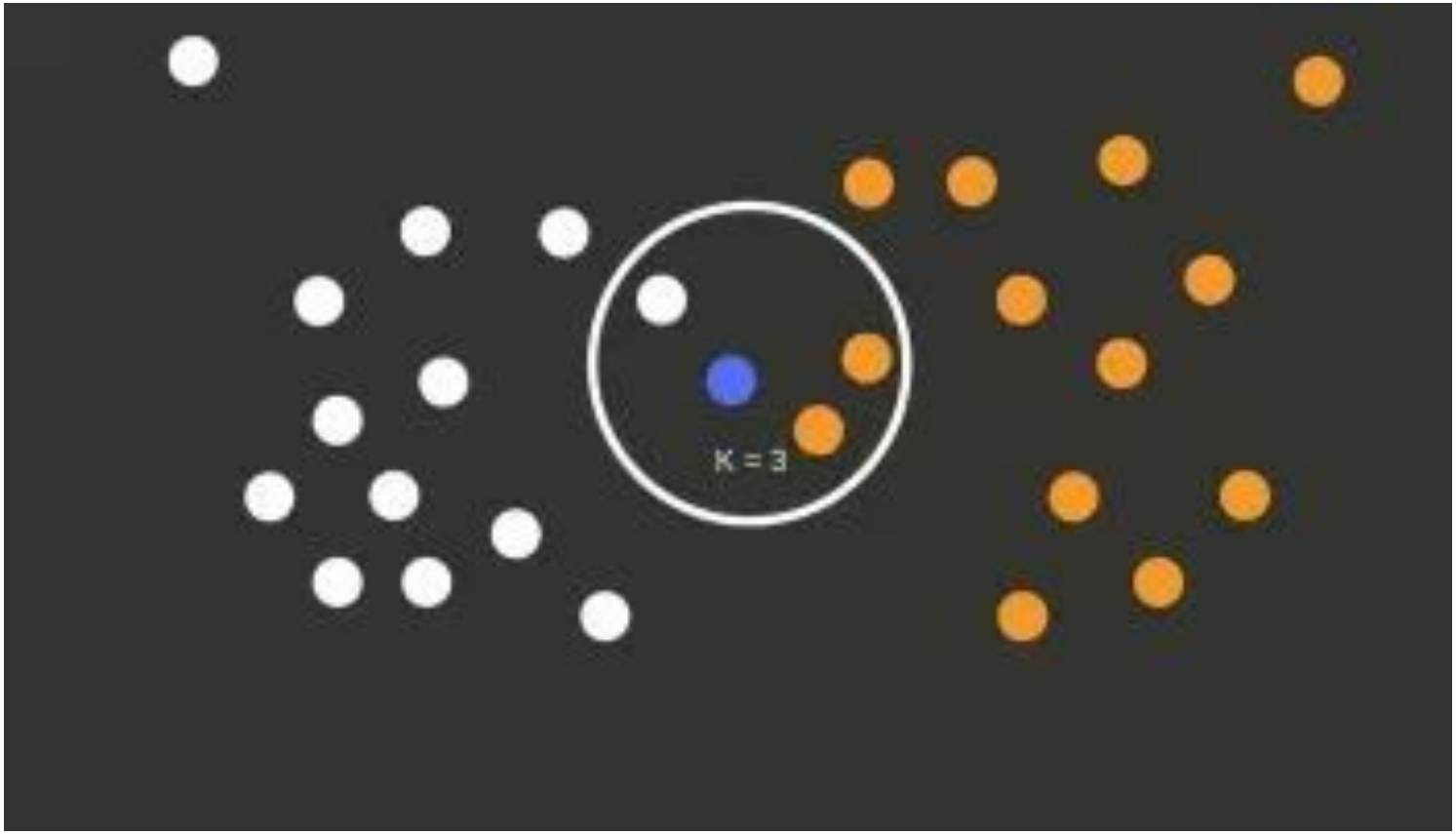
What is the probability that this is ABronte given there are such words used?



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

K-nn

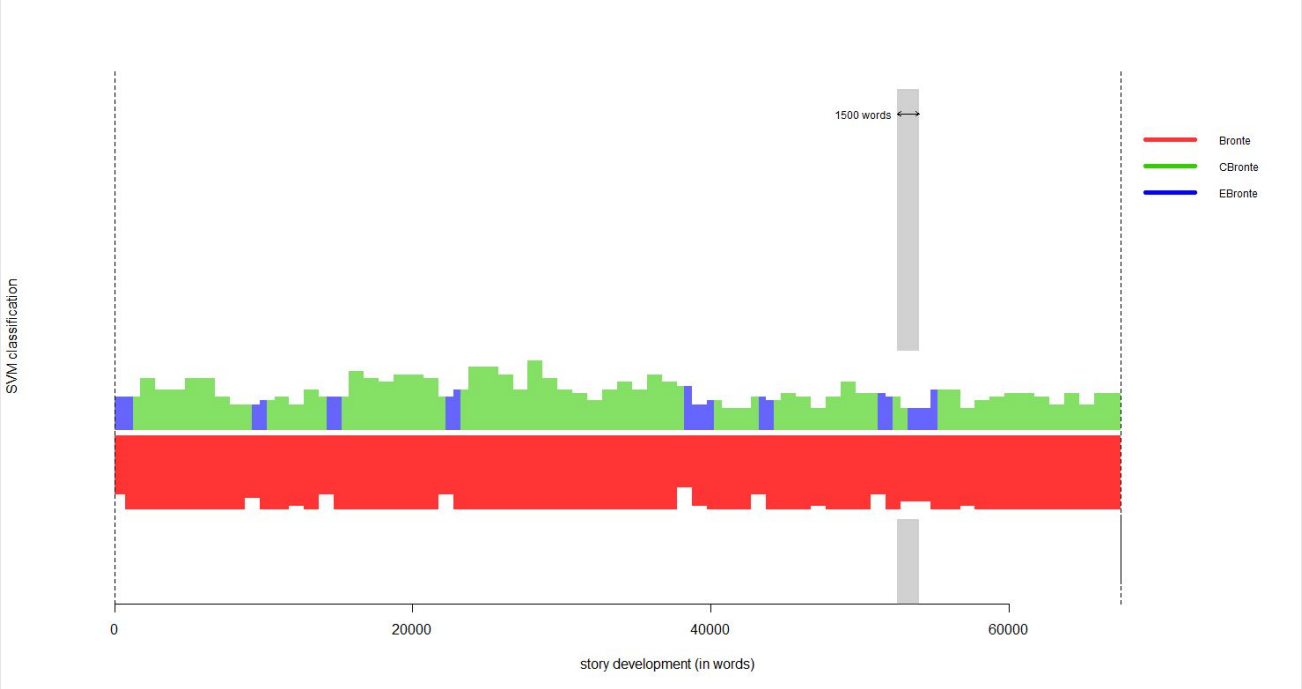
What are the closest observations to ABronte?



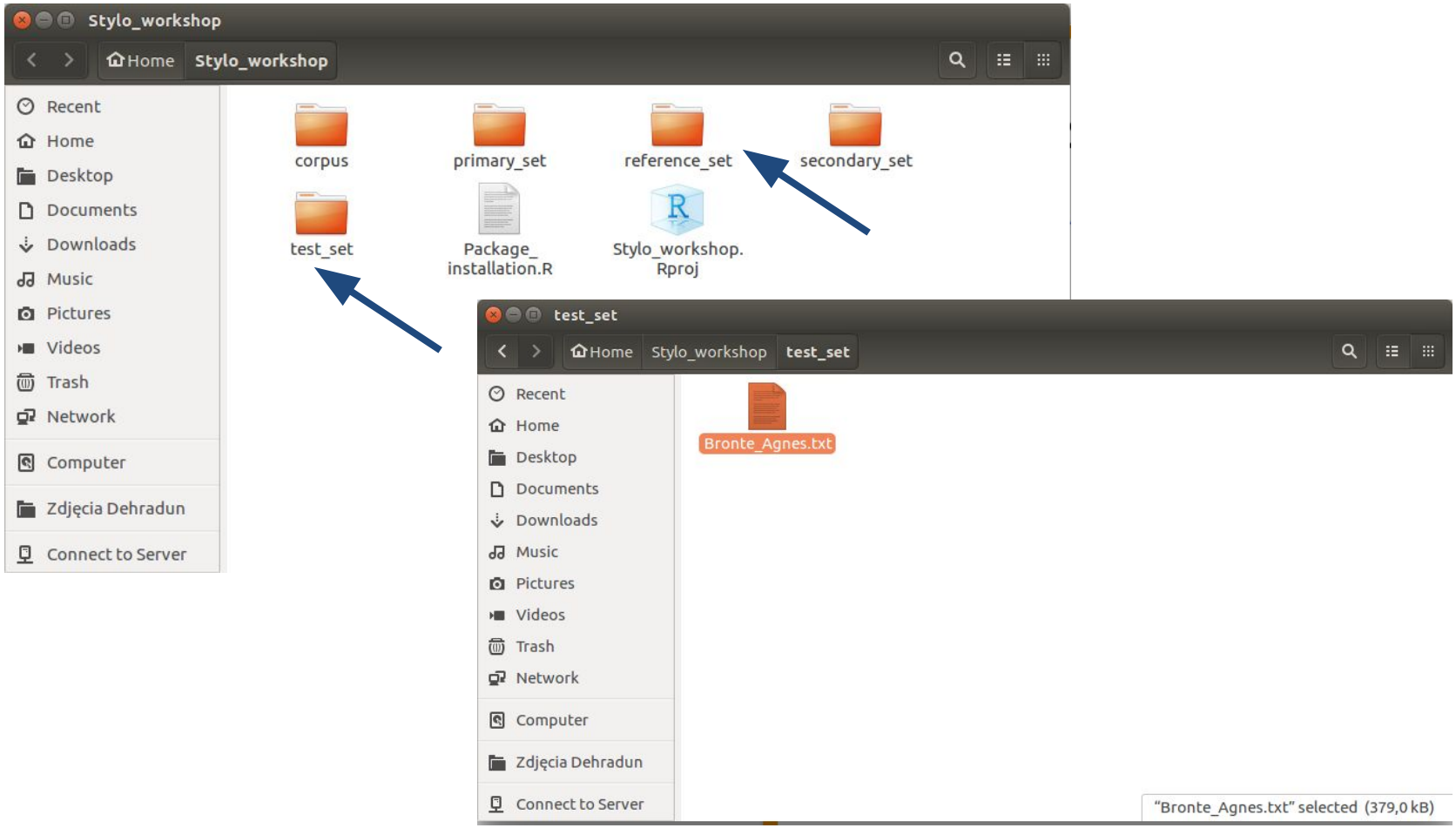
Classify - Exercise

- Remember about correct data set up
- Use 100 English Novels for classify() analysis
- KNN (for different k values), Naive Bayes,
- Use different values for culling and MFW

ROLLING CLASSIFY



Corpus prep for Rolling.classify



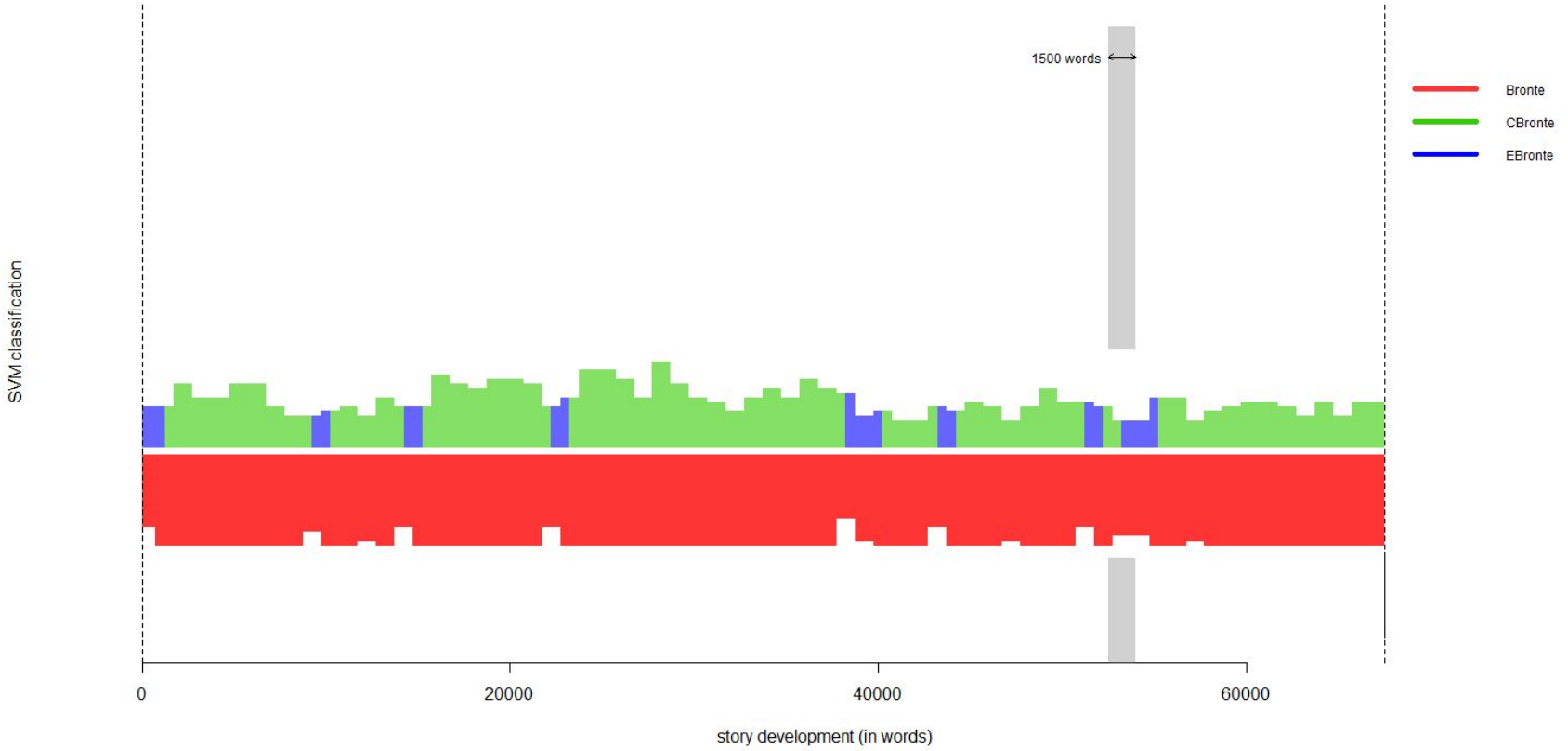
How to use it

```
rolling.classify(slice.size = 1500, mfw = 100,
slice.overlap = 500, plot.legend = T,
classification.method = "svm")
```

Rolling.classify

```
> rolling.classify(slice.size = 500, mfw = 100, slice.overlap = 100,plot.legend = T)
using current directory...
The subcorpora will be loaded from text files...
loading Bronte_Wuthering.txt    ...
slicing input text into tokens...
turning words into features, e.g. char n-grams (if applicable)...
loading Bronte_Wuthering.txt    ...
slicing input text into tokens...
Bronte_Wuthering
        - text length (in words): 116612
        - nr. of samples: 291
        - nr. of words dropped at the end of the text: 212
turning words into features, e.g. char n-grams (if applicable)...
loading Bronte_Agnes.txt        ...
loading Bronte_Jane.txt ...
loading Bronte_Professor.txt    ...
loading Bronte_Tenant.txt       ...
loading Bronte_Villette.txt     ...
loading CBronte_Jane.txt        ...
loading CBronte_Professor.txt   ...
loading CBronte_Villette.txt    ...
loading EBronte_Wuthering.txt   ...
slicing input text into tokens...
turning words into features, e.g. char n-grams (if applicable)...
```

Results of rolling classify



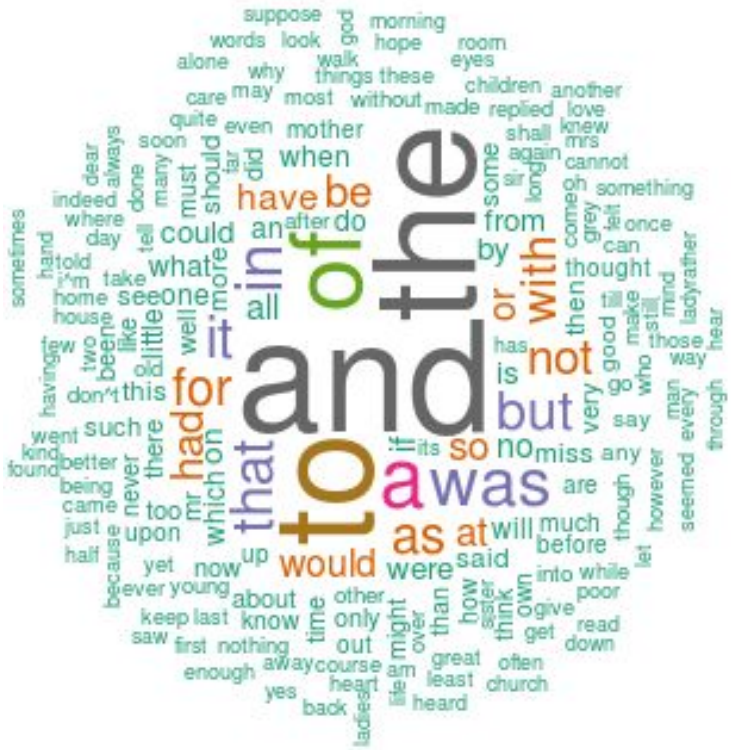
Rolling.classify - Exercise

- Remember about correct data set up
- Use 100 English Novels for rolling.classify()
- Choose 2 authors for reference set and only one novel
- Change values slice.size, mfw, and slice.overlap

Way Forward

- Workflow using code
- Word clouds
- Graph data visualizations
- …

Visualizations



Download R files

https://github.com/jandziak/Stylo_workshop

Workflow using code

- # Load corpus
- # Tokenization
- # Deleting useless words
- # Creating features
- # Culling
- # Simple cluster analysis with stylo
- # More options in stylo

Load corpus

```
# Load corpus
raw.corpus <- load.corpus(files = 'all', corpus.dir
= 'corpus', encoding = 'UTF-8')

# Summary of corpus files
summary(raw.corpus)
```

Tokenization

```
# Tokenization
tokenized.corpus <-
txt.to.words.ext(raw.corpus, language =
'English.all', preserve.case = FALSE)

# Austin's Pride and Prejudice famous sentence
tokenized.corpus$Austen_Pride.txt[8:30]
```

Deleting useless words

```
# English pronouns
stylo.pronouns(language = 'English')

# Deleting useless words
corpus.no.pronouns <-
delete.stop.words(tokenized.corpus,
stop.words = stylo.pronouns(language =
'English'))
```

Creating features

```
# Counting frequent features
frequent.features <-
make.frequency.list(corpus.no.pronouns, head
= 3000)
freqs <-
make.table.of.frequencies(corpus.no.pronouns,
features = frequent.features, relative = FALSE)
freqs <-
make.table.of.frequencies(corpus.no.pronouns,
features = frequent.features)
```

Culling

```
# Culling (words in at least 80% of samples)
culled.freqs <- perform.culling(freqs,
culling.level = 80)
```

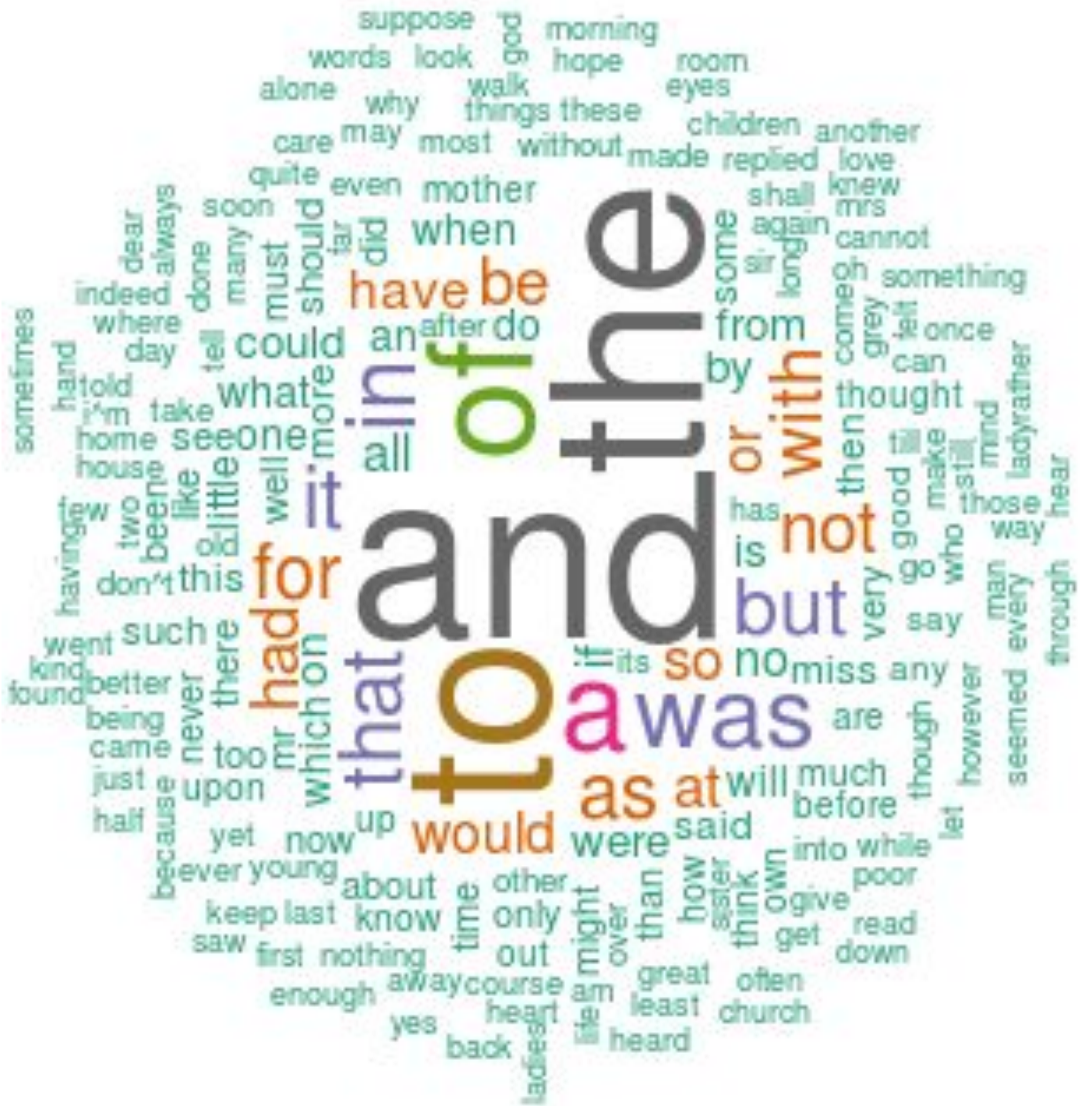
Simple cluster analysis with stylo

```
# Simple cluster analysis with stylo
stylo(frequencies = culled.freqs, gui = FALSE)
```

More options in stylo

```
# More options in stylo
stylo(frequencies = culled.freqs, gui = FALSE,
analysis.type = 'PCR', custom.graph.title = "PCA
analysis for novels")
```

Word clouds


suppose
god
morning
words
look
hope
room
alone
walk
eyes
why
things
these
children
another
care
may
most
without
made
replied
love
quite
even
mother
shall
knew
mrs
soon
should
far
did
when
again
cannot
dear
always
many
must
the
some
sir
long
indeed
done
have
be
oh
something
where
after
do
from
come
grey
left
once
day
could
an
can
sometimes
hand
told
tell
what
in
of
by
with
thought
i'm
take
more
all
then
till
make
still
mind
ladyrather
home
see
one
or
house
little
well
it
has
not
good
those
hear
few
two
been
like
old
way
having
don't
this
for
and
is
go
who
man
every
through
very
but
say
went
such
there
had
on
that
if
its
so
no
miss
any
kind
found
better
being
never
which
mr
to
a
was
are
however
seemed
came
too
though
just
upon
as
at
will
much
before
half
because
up
would
let
yet
now
said
into
while
ever
young
were
about
other
how
own
poor
keep
last
know
time
only
than
sister
think
give
read
saw
first
nothing
out
might
over
get
down
away
course
great
often
enough
am
least
church
yes
heart
life
heard
back
ladies

Workflow using code

- # Prepare environment
- # Plot Word cloud

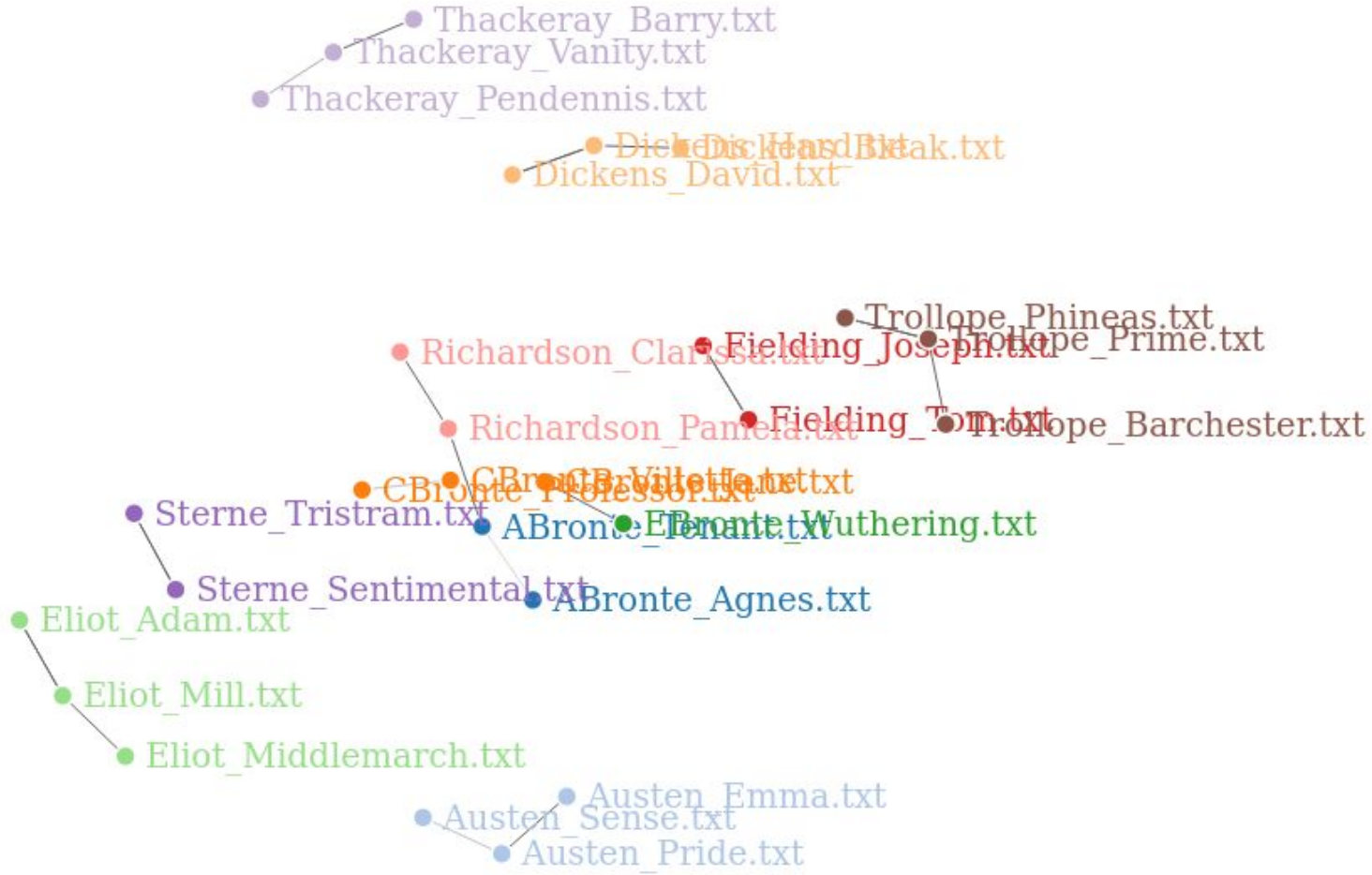
Prepare Environment

```
# Install missing packages
install.packages("wordcloud") # word-cloud generator
install.packages("RColorBrewer") # color palettes
library("wordcloud")
library("RColorBrewer")
```

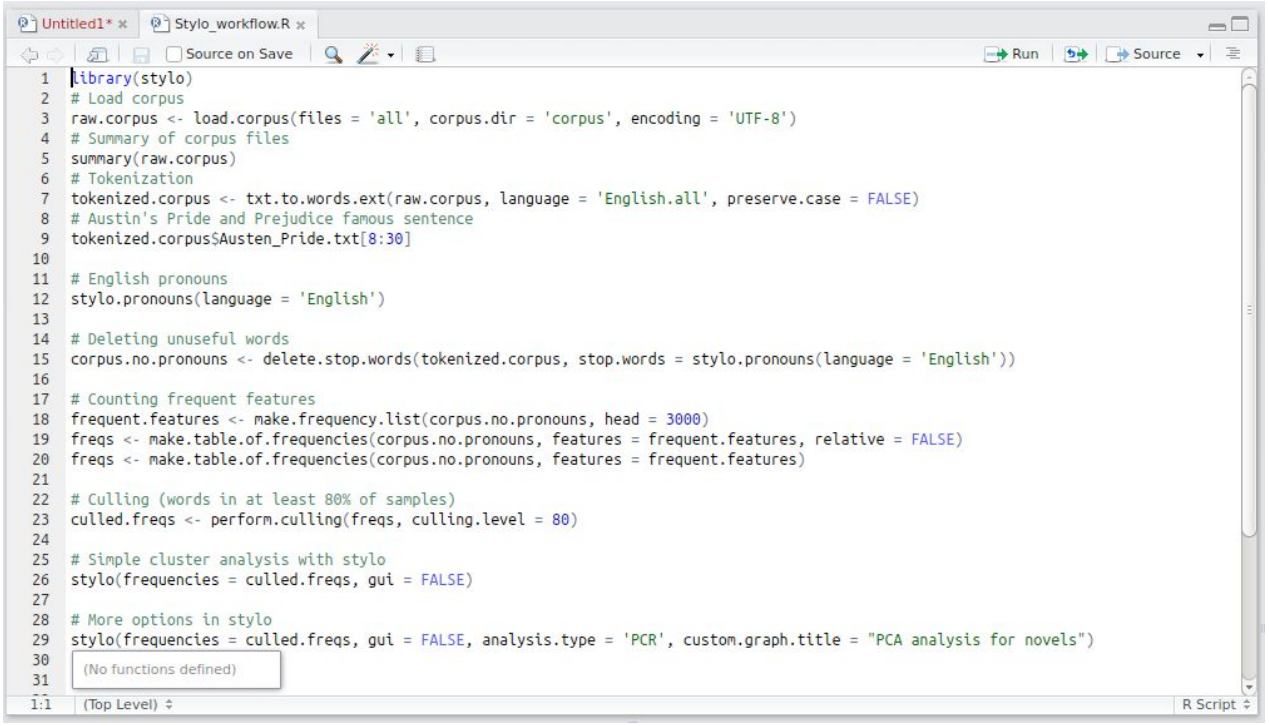
Plot Word cloud

```
# Ensure reproducibility
set.seed(1234)

wordcloud(words = names(df), freq =
as.numeric(df[1,1:2543]), min.freq = 0.01,
        max.words=200, random.order=FALSE,
rot.per=0.35,
        colors=brewer.pal(8, "Dark2"))
```

Graph data visualizations



Just go to the code



Way forward - Exercise

- Apply stylo code analysis on new data
- Use 100 English Novels
- Produce new word cloud using different novel
- Create graph for novels