

# Υφομετρία στην R

## Το πακέτο *stylo*

Γιώργος Μικρός

ΕΚΠΑ ~ University of Massachusetts, Boston

# Εγκατάσταση του `stylo`

- Τρέχουμε την R
- Πληκτρολογούμε `install.packages("stylo")`
- Διαλέγουμε τον διακομιστή της R (R server)
- Πατάμε `OK`

# Μερικές βασικές συναρτήσεις της R

- Ενεργοποίηση του πακέτου (package): `library(stylo)`
- Ορισμός του καταλόγου εργασίας (working directory):  
`setwd("path/to/my/stuff")`
- Για να εντοπίσετε τον ενεργοποιημένο κατάλογο εργασίας: `getwd()`
- Για να δείτε τα υπάρχοντα αρχεία στον κατάλογο εργασίας:  
`list.files()`
- Για να πάρετε βοήθεια: `help(function)`, π.χ. `help(stylo)`
- Για να κλείσετε την R: `q()`

# Βασικές συναρτήσεις: `stylo()`

- Υπολογίζει αποστάσεις (διαφορές) μεταξύ κειμένων αντιπροσωπευόμενες ως σειρές (rows) των συχνοτήτων των πιο συχνών λέξεων.
- Εν συνεχεία κάνει γραφήματα αυτών των αποστάσεων:
  - Γραφήματα Ανάλυσης Συστάδων (Cluster Analysis plots) και ειδικότερα τα δενδρογράμματα (dendrograms).
  - Γραφήματα Πολυδιάστατης Απεικόνισης (Multidimensional Scaling plots) και ειδικότερα γραφήματα σκεδασμού (scatterplots).
  - Γραφήματα Ανάλυσης Πρωτευουσών Συνιστωσών (Principal Components Analysis)
  - Γραφήματα Αναδειγματοληπτικών Δένδρων Συναίνεσης (Bootstrap Consensus Trees)
  - Γραφήματα Αναδειγματοληπτικών Δικτύων Συναίνεσης (Bootstrap Consensus Networks)
- Τα γραφήματα μπορούν να απεικονιστούν στην οθόνη και να αποθηκευτούν σε μορφή αρχείου εικόνας (π.χ. PNG).

# Βασικές συναρτήσεις: `stylo.network()`

- Είναι μια τροποποιημένη έκδοση της συνάρτησης `stylo()`.
- Παράγει τα Αναδειγματοληπτικά Δίκτυα Συναίνεσης (Bootstrap Consensus Networks).
- Δημιουργεί αλληλεπιδραστικές οπτικοποιήσεις σε ένα web browser. Για να λειτουργήσει πρέπει να εγκαταστήσετε ένα επιπλέον πακέτο της R πρώτα που ονομάζεται `networkD3`. Πληκτρολογήστε:  
`install.packages("networkD3")`

# Βασικές συναρτήσεις: `classify()`

- Εκπαιδεύει ένα μοντέλο για μια προκαθορισμένη ομάδα κειμένων χαρακτηρισμένη ως προς κάποιο χαρακτηριστικό τους, π.χ. τον συγγραφέα.
- Εν συνεχεία υπολογίζει αποστάσεις (διαφορές) μεταξύ των κειμένων, αντιπροσωπευόμενες ως σειρές (rows) των συχνοτήτων των πιο συχνών λέξεων.
- Στο τέλος συγκρίνει τα εκπαιδευμένα μοντέλα με τα κείμενα προς έλεγχο χρησιμοποιώντας:
  - Τον ταξινομητή Delta
  - Τον ταξινομητή k-NN
  - Τις Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines – SVM)
  - Τον ταξινομητή Naïve Bayes
  - Τον ταξινομητή Nearest Shrunked Centroids που υποστηρίζει δεδομένα υψηλής διαστασιμότητας (high-dimensional datasets).
- Η συνάρτηση παράγει μία αναφορά με την απόδοση του ταξινομητή.

# Βασικές συναρτήσεις: `oppose()`

- Είναι σχεδιασμένη για να συγκρίνει δύο κείμενα ή δύο ομάδες κειμένων.
- Κόβει τα κείμενα σε ισομεγέθη δείγματα.
- Βρίσκει τις πιο χαρακτηριστικές λέξεις των δύο κειμένων ή των δύο ομάδων κειμένων.
- Παράγει ένα διάγραμμα χρήσης των λέξεων αυτών στα δύο κείμενα ή στις δύο ομάδες κειμένων.

# Προετοιμάζοντας το corpus

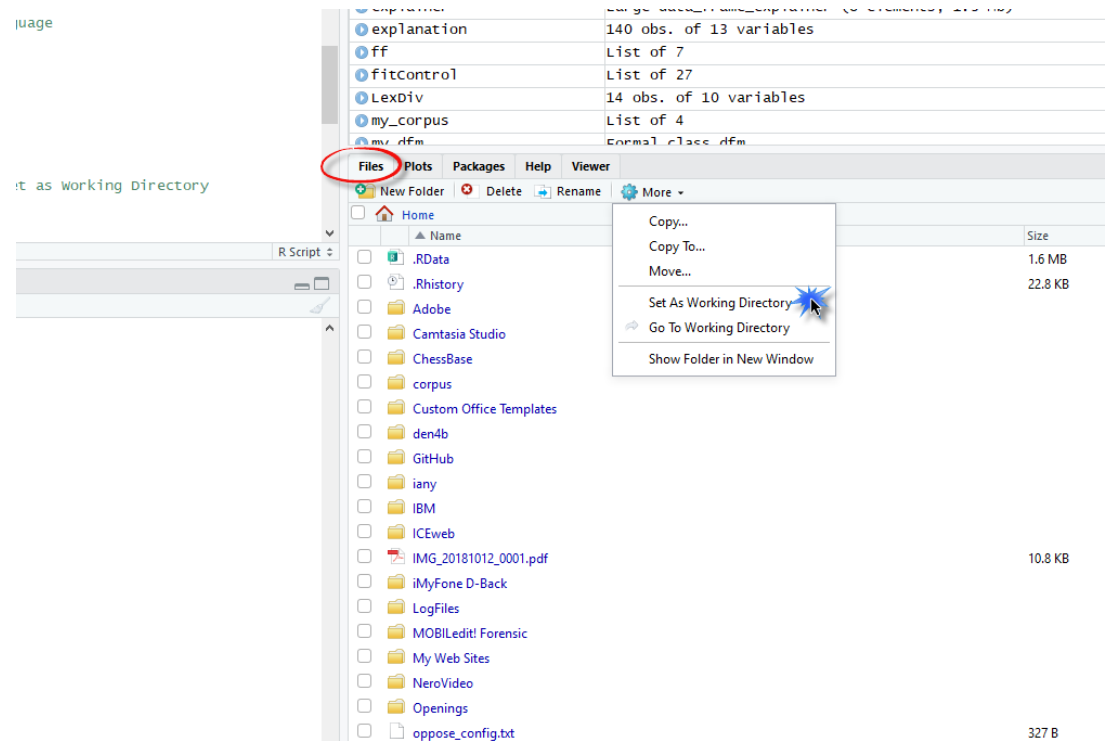
- Πριν ξεκινήσετε την R, ...
- Στον κατάλογο (folder) που θα δουλέψετε, δημιουργήστε έναν υποκατάλογο (subfolder) που θα το ονομάσετε `corpus`.
- Βάλτε τα κείμενά σας (σε μορφή απλού κειμένου txt) εκεί, π.χ. :
  - `Roidis_Diigimata.txt`
  - `Vikelas_Diigimata .txt`
  - κ.λ.π.
- Τα αρχεία σας θα πρέπει να είναι κωδικοποιημένα σε UTF-8.



# Εκτέλεση του `stylo()`

## Ορισμός του ενεργού φακέλου στο R studio

1. Ενεργοποιήστε το πακέτο
  - `library(stylo)`
2. Πλοηγηθείτε στον κατάλογο σας:
  - geeks:  
`setwd("the/path/to/my/favourite/folder")`
  - Rstudio: Βρείτε τον κατάλογο σας στο **Files** και μετά ακολουθείστε το **More > Set as Working Directory**
3. Πληκτρολογήστε `stylo()` και μετά ENTER



## Επιλογές στο `style()`

- INPUT: Δηλώνετε το format των κειμένων που θα αναλύσετε
- LANGUAGE: Για τα ελληνικά επιλέγετε Other και UTF-8 (θυμηθείτε ότι τα κείμενα σας πρέπει να είναι σε UTF-8).
- **ΜΗΝ** πατήσετε το OK ακόμα!

Stylometry with R | stylo | set parameters

INPUT & LANGUAGE	FEATURES	STATISTICS	SAMPLING	OUTPUT
<p>INPUT: plain text <input checked="" type="radio"/> xml <input type="radio"/> xml (plays) <input type="radio"/> xml (no titles) <input type="radio"/> html <input type="radio"/></p> <p>LANGUAGE: English <input checked="" type="radio"/> English (contr.) <input type="radio"/> English (ALL) <input type="radio"/> Latin <input type="radio"/> Latin (u/v &gt; u) <input type="radio"/></p> <p>Polish <input type="radio"/> Hungarian <input type="radio"/> French <input type="radio"/> Italian <input type="radio"/> Spanish <input type="radio"/></p> <p>Dutch <input type="radio"/> German <input type="radio"/> CJK <input type="radio"/> Other <input checked="" type="radio"/> UTF-8 <input checked="" type="checkbox"/></p>				

OK

# Επιλογές στο `stylo()`

- FEATURES: τα γλωσσικά χαρακτηριστικά που θα μετρηθούν (χαρακτήρες ή λέξεις).
  - ngram size: 1 για μονά χαρακτηριστικά, 2 για δι-γράμματα κ.λ.π.
- MFW SETTINGS: Ο αριθμός των πιο συχνών λέξεων (ή άλλων χαρακτηριστικών) που θα χρησιμοποιηθούν στην ανάλυση
  - Στις περισσότερες περιπτώσεις `Minimum = Maximum`

Stylometry with R | stylo | set parameters

INPUT & LANGUAGE	FEATURES	STATISTICS	SAMPLING	OUTPUT
<b>FEATURES:</b>				
	words <input checked="" type="radio"/>	chars <input type="radio"/>	ngram size <input type="text" value="1"/>	preserve case <input type="checkbox"/>
<b>MFW SETTINGS:</b>				
	Minimum <input type="text" value="1000"/>	Maximum <input type="text" value="1000"/>	Increment <input type="text" value="100"/>	Start at freq. rank <input type="text" value="1"/>
<b>CULLING:</b>				
	Minimum <input type="text" value="0"/>	Maximum <input type="text" value="0"/>	Increment <input type="text" value="20"/>	List Cutoff <input type="text" value="5000"/>
				Delete pronouns <input type="checkbox"/>
<b>VARIOUS:</b>				
	Existing frequencies <input type="checkbox"/>	Existing wordlist <input type="checkbox"/>	Select files manually <input type="checkbox"/>	List of files <input type="text"/>
<input type="button" value="OK"/>				

# Επιλογές στο `stylo()`

- CULLING: προαιρετικά, για να φιλτράρει κάποιες λέξεις που δεν θέλουμε να αναλύσουμε.
  - Παραδείγματα:
    - 0 – όλες οι λέξεις θα χρησιμοποιηθούν
    - 20 – μία λέξη για να διατηρηθεί στη λίστα με τα χαρακτηριστικά που θα χρησιμοποιηθούν στην ανάλυση θα πρέπει να εμφανίζεται το λιγότερο στο 20% των κειμένων του corpus.
    - 100 - ένα ακραίο φίλτρο. Όλες οι λέξεις που δεν εμφανίζονται σε όλα τα κείμενα απομακρύνονται.
- DELETE PRONOUNS: προαιρετικά απομακρύνει τις προσωπικές αντωνυμίες. Η λίστα με τις προσωπικές αντωνυμίες επιλέγεται με βάση την επιλεγμένη γλώσσα.

Stylometry with R | stylo | set parameters

INPUT & LANGUAGE	FEATURES	STATISTICS	SAMPLING	OUTPUT
<b>FEATURES:</b>				
	words <input checked="" type="radio"/>	chars <input type="radio"/>	ngram size <input type="text" value="1"/>	preserve case <input type="checkbox"/>
<b>MFW SETTINGS:</b>				
	Minimum <input type="text" value="1000"/>	Maximum <input type="text" value="1000"/>	Increment <input type="text" value="100"/>	Start at freq. rank <input type="text" value="1"/>
<b>CULLING:</b>				
	Minimum <input type="text" value="0"/>	Maximum <input type="text" value="0"/>	Increment <input type="text" value="20"/>	List Cutoff <input type="text" value="5000"/> Delete pronouns <input type="checkbox"/>
<b>VARIOUS:</b>				
	Existing frequencies <input type="checkbox"/>	Existing wordlist <input type="checkbox"/>	Select files manually <input type="checkbox"/>	List of files <input type="text"/>
<input type="button" value="OK"/>				

# Επιλογές στο `stylo()`

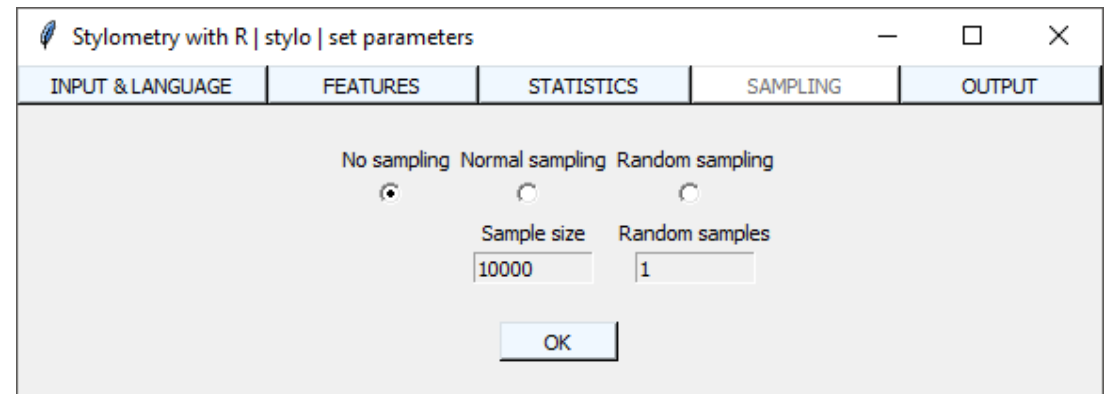
- STATISTICS: Ανάλυση Συστάδων (Cluster Analysis), Ανάλυση Πολυδιάστατης Κλιμάκωσης (MDS) κ.λ.π.
- DISTANCES: Επιλογή για το πώς θα μετρηθούν οι αποστάσεις μεταξύ των κειμένων
  - Classic Delta: Ίσως η καλύτερη επιλογή για να ξεκινήσετε μια ανάλυση
  - Cosine Delta: Μια ακόμα καλύτερη επιλογή.
  - Eder's Delta: Μια καλή επιλογή για γλώσσες με πλούσιο κλιτικό σύστημα.

Stylometry with R | stylo | set parameters

INPUT & LANGUAGE	FEATURES	STATISTICS	SAMPLING	OUTPUT
<p>STATISTICS: Cluster Analysis <input checked="" type="radio"/> MDS <input type="radio"/> PCA (cov.) <input type="radio"/> PCA (corr.) <input type="radio"/> tSNE <input type="radio"/></p> <p>Consensus Tree <input type="radio"/> Consensus strength <input type="text" value="0.5"/></p> <p>DELTA DISTANCE: Classic Delta <input checked="" type="radio"/> Cosine Delta <input type="radio"/> Eder's Delta <input type="radio"/> Eder's Simple <input type="radio"/> Entropy <input type="radio"/></p> <p>Manhattan <input type="radio"/> Canberra <input type="radio"/> Euclidean <input type="radio"/> Cosine <input type="radio"/> Min-Max <input type="radio"/></p> <p>OK</p>				

# Επιλογές στο `stylo()`

- SAMPLING: επιλογές για να κόψετε τα κείμενα σε μικρότερα δείγματα
  - No sampling: τα κείμενα θα αναλυθούν ολόκληρα.
  - Normal sampling: τα κείμενα θα χωριστούν σε ισομεγέθη τμήματα.
  - Random sampling: θα συλλεχθούν με τυχαίο τρόπο  $N$  λέξεις από κάθε κείμενο.
  - Random samples: Η τυχαία επιλογή λέξεων θα επαναληφθεί  $n$  φορές.



The screenshot shows a window titled "Stylometry with R | stylo | set parameters" with a standard macOS-style title bar (minimize, maximize, close buttons). The window has five tabs: "INPUT & LANGUAGE", "FEATURES", "STATISTICS", "SAMPLING", and "OUTPUT". The "SAMPLING" tab is currently selected. Inside this tab, there are three radio buttons for sampling methods: "No sampling", "Normal sampling", and "Random sampling". The "No sampling" radio button is selected. Below these radio buttons, there are two input fields: "Sample size" with the value "10000" and "Random samples" with the value "1". At the bottom of the dialog is an "OK" button.

# Επιλογές στο `stylo()`

- OUTPUT: Οι περισσότερες επιλογές είναι προφανείς. Σιγουρευτείτε ότι το Onscreen είναι επιλεγμένο έτσι ώστε να δείτε τα αποτελέσματά σας στην οθόνη.
- PCA flavor: Επιλέξτε “loadings” για να εξετάσετε την διακριτική δύναμη συγκεκριμένων χαρακτηριστικών (αλλά πρώτα επιλέξτε PCA στην καρτέλα STATISTICS).
- Horizontal CA tree: Χρησιμοποιήστε αυτή την επιλογή για να τοποθετήσετε τα δενδρογράμματα οριζόντια.

Stylometry with R | stylo | set parameters

	INPUT & LANGUAGE	FEATURES	STATISTICS	SAMPLING	OUTPUT
GRAPHS:	Onscreen <input checked="" type="checkbox"/>	PDF <input type="checkbox"/>	JPG <input type="checkbox"/>	SVG <input type="checkbox"/>	PNG <input type="checkbox"/>
PLOT AREA:	Set default <input type="checkbox"/>	Plot height 7	Plot width 7	Font size 10	Line width 1
		Colors <input checked="" type="radio"/>	Grayscale <input type="radio"/>	Black <input type="radio"/>	Titles <input checked="" type="checkbox"/>
PCA/MDS:	Labels <input checked="" type="radio"/>	Points <input type="radio"/>	Both <input type="radio"/>	Margins 2	Label offset 0
PCA FLAVOUR:	Classic <input checked="" type="radio"/>	Loadings <input type="radio"/>	Technical <input type="radio"/>	Symbols <input type="radio"/>	
VARIOUS:	Horizontal CA tree <input checked="" type="checkbox"/>	Save distance table <input type="checkbox"/>	Save features <input type="checkbox"/>	Save frequencies <input type="checkbox"/>	Dump samples <input type="checkbox"/>

OK

# Αναδειγματοληπτικά Δίκτυα Συναίνεσης (Bootstrap Consensus Networks)

- Εκτελέστε την συνάρτηση `stylo.network()`
- Κάντε τις επιλογές όπως και στο `stylo()`
- Ένας web browser θα ξεκινήσει αυτόματα και θα εμφανιστεί το δίκτυο των κειμένων.



# Εκτέλεση του `oppose()`

- Πρέπει να δημιουργήσετε δύο νέους καταλόγους:
  - `primary_set`
  - `secondary_set`
  - `test_set` (προαιρετικά)
- Εκτέλεση της συνάρτησης. Για τα ελληνικά προσδιορίζουμε την κωδικοποίηση (UTF-8) και την γλώσσα (Other):
  - `library(stylo)`
  - `oppose(encoding = "UTF-8", corpus.lang = "Other")`
- Η συνάρτηση δημιουργεί:
  - `Words_preferred.txt` που είναι χαρακτηριστικές των κειμένων που βρίσκονται στο `primary_set`
  - `Words_avoided.txt` που είναι χαρακτηριστικές των κειμένων που βρίσκονται στο `secondary_set`
  - Γράφημα λεξικών συχνοτήτων