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 συνάρτηση 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.

Βασικές συναρτήσεις: `rolling.classify()`

- Είναι σχεδιασμένη για να εντοπίζει ίχνη διαφορετικών συγγραφέων σε ένα κείμενο το οποίο έχει παραχθεί συνεργατικά
- Η βασική μεθοδολογία είναι αυτή του σειριακά μετακινούμενου «παράθυρου».
- Ένας αλγόριθμος (π.χ. SVM) εκπαιδεύεται στο ύφος κάποιων συγγραφέων και μετά κάνει προβλέψεις στο αμφισβητούμενο κείμενο σε διαδοχικά κομμάτια του.
- Παράγει ένα γράφημα όπου αποτυπώνεται σε κάθε σημείο του κειμένου ποιος είναι ο πιο πιθανός συγγραφέας και πόσο «ισχυρή» υφομετρικά είναι η παρουσία του.

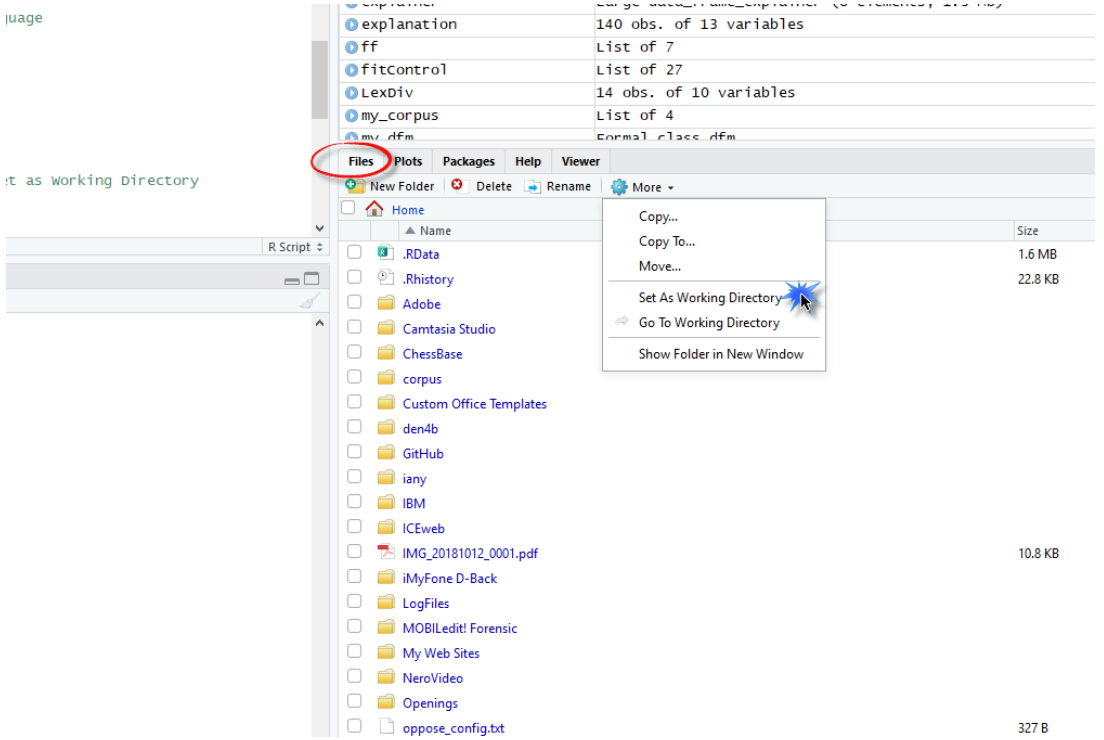
Προετοιμάζοντας το corpus

- 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()`

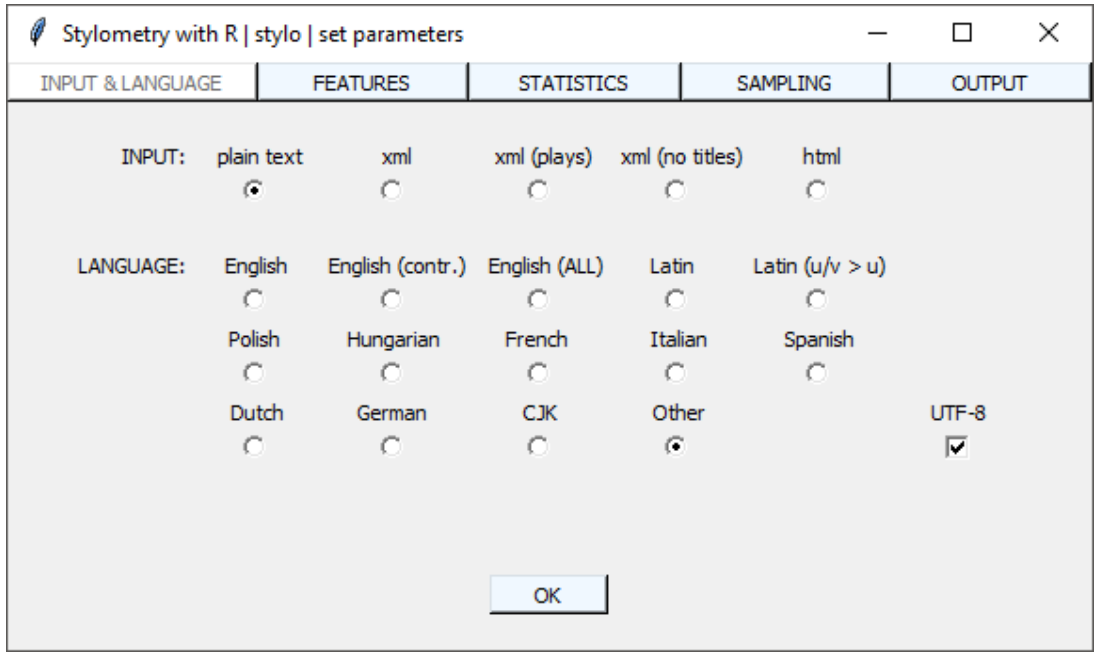
Ορισμός του ενεργού φακέλου στο R studio

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



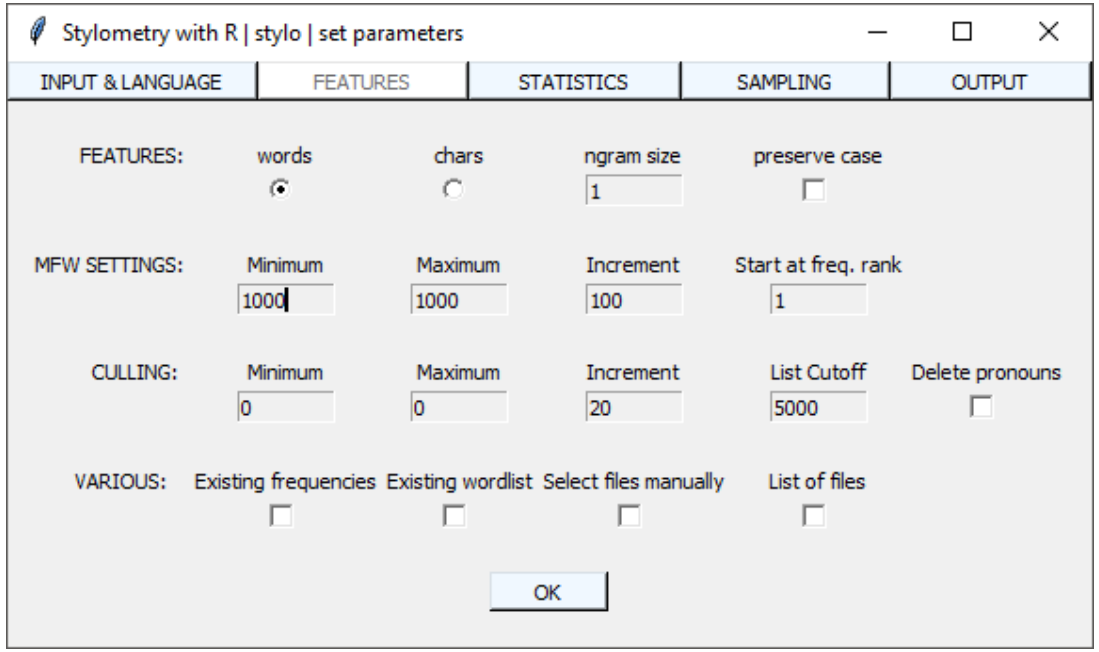
Επιλογές στο `stylo()`

- 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!



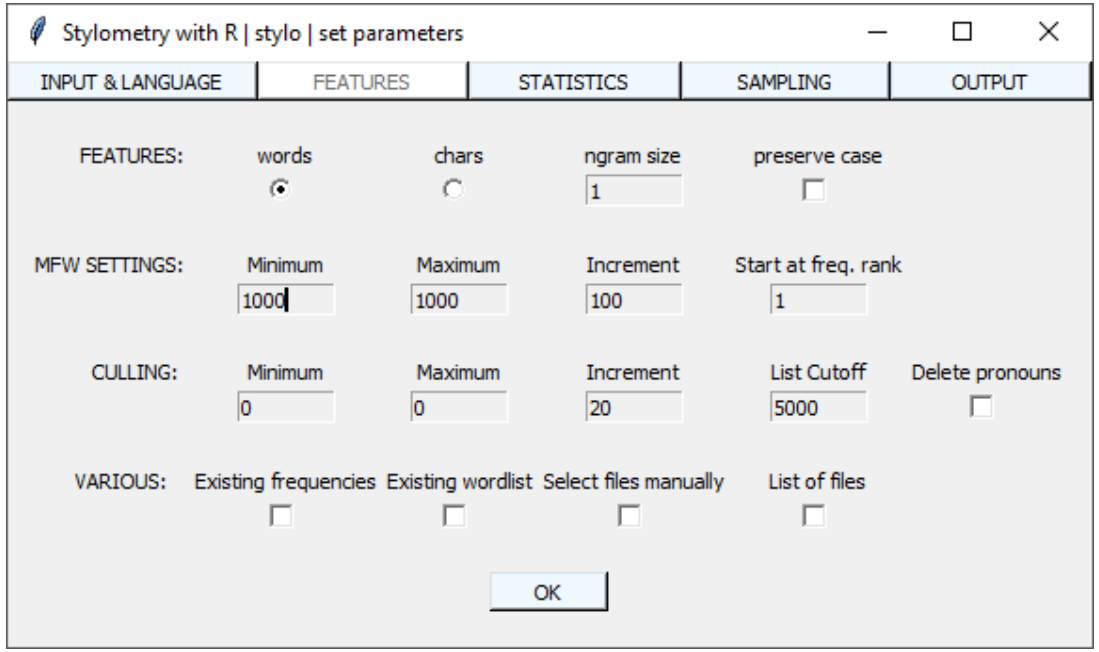
Επιλογές στο `stylo()`

- FEATURES: τα γλωσσικά χαρακτηριστικά που θα μετρηθούν (χαρακτήρες ή λέξεις).
  - ngram size: 1 για μονά χαρακτηρισικά, 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`



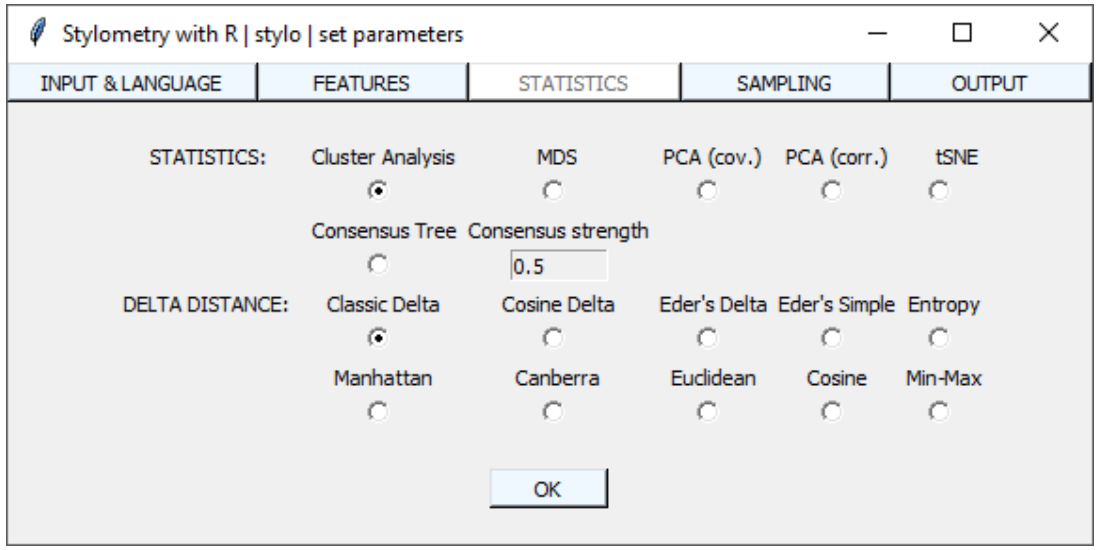
Επιλογές στο `stylo()`

- 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.



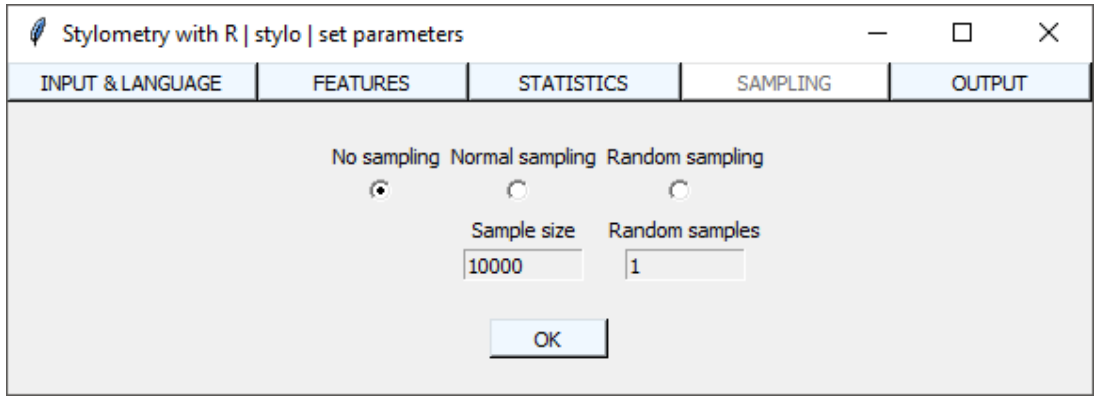
Επιλογές στο `stylo()`

- 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.



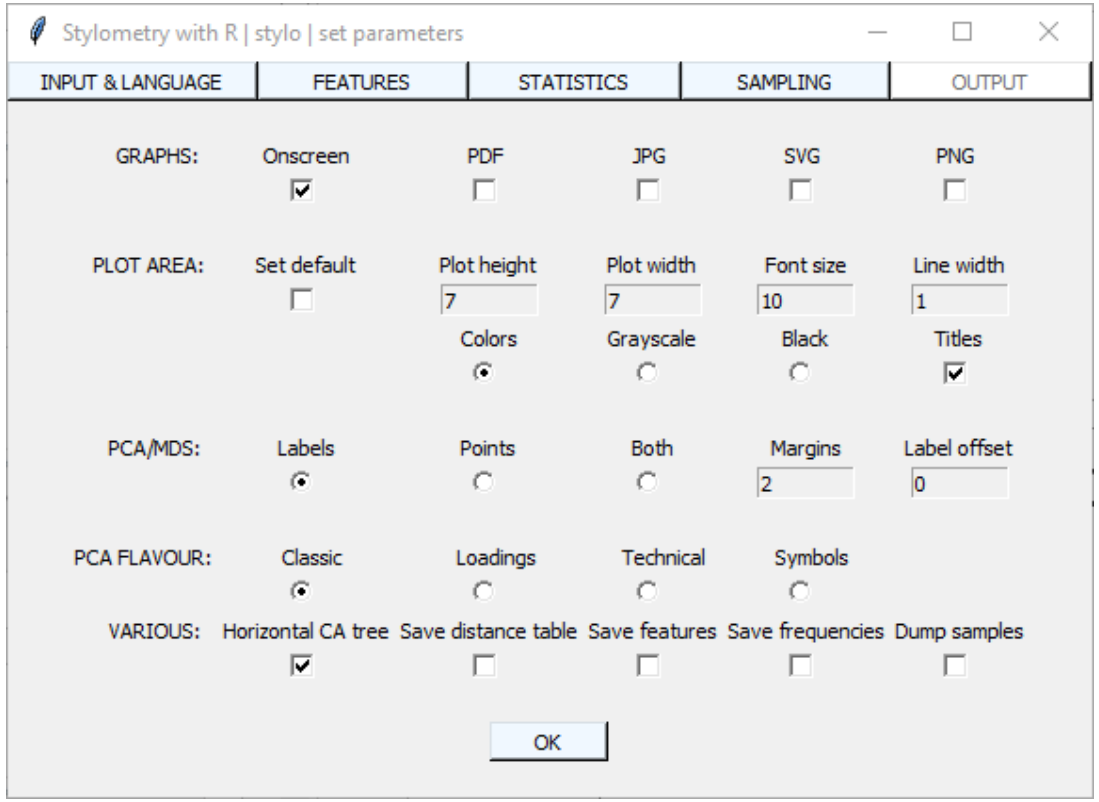
Επιλογές στο `stylo()`

- 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.



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.



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

Επιλογές στο `oppose()`

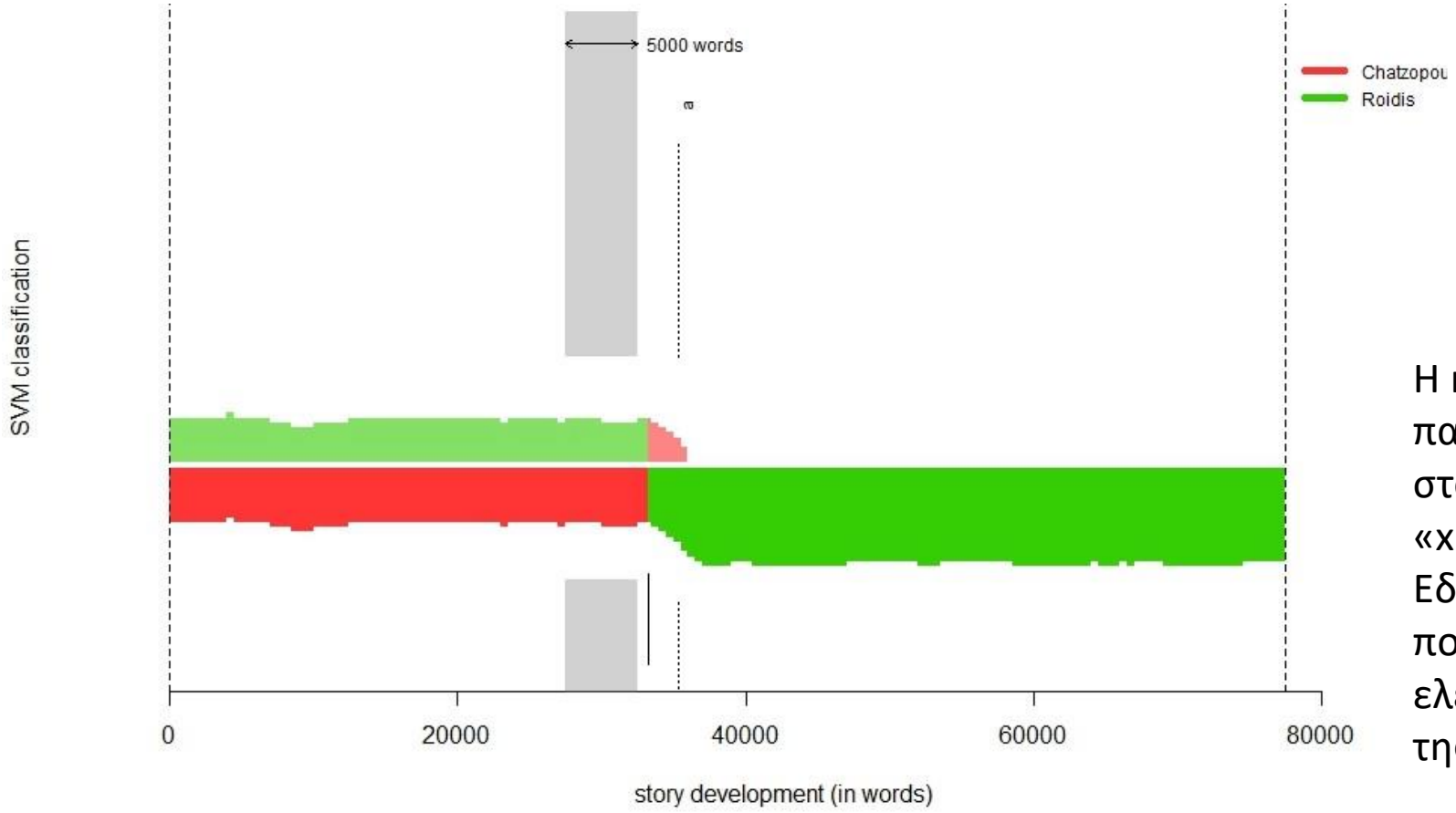
- Slice length: μέγεθος (σε λέξεις) των δειγμάτων (5000)
- Slice overlap: (0)
- Method: (Craig’s Zeta)
- Visualization: Τύπος του γραφήματος (Markers)

Εκτέλεση του `rolling.classify()`

```
rolling.classify(write.png.file = FALSE, classification.method =
"svm", mfw=1000, training.set.sampling = "normal.sampling", slice.size
= 5000, slice.overlap = 4500, encoding = "UTF-8", corpus.lang =
"Other")
```

- write.png.file: Επιλογή για να σώσουμε το γράφημα σε αρχείο εικόνας.
- classification.method: Η μέθοδος ταξινόμησης που θα επιλέξουμε. Διαθέσιμες επιλογές είναι η `svm`, `nsc`, `delta`, `knn`, `nb`.
- mfw= ο αριθμός των πιο συχνών γλωσσικών χαρακτηριστικών (default λέξεις).
- training.set.sampling: επιλογή για τον χωρισμό των κειμένων του training set σε μικρότερα διαδοχικά κομμάτια. Διαθέσιμες επιλογές `normal.sampling`, `none`, `random`.
- slice.size: μέγεθος (σε λέξεις) του κάθε τμήματος στο οποίο ο αλγόριθμος θα ελέγξει την πατρότητα.
- slice.overlap: μέγεθος (σε λέξεις) αλληλοεπικάλυψης του παραθύρου που θα μετακινηθεί στο κείμενο.

Το γράφημα του `rolling.classify()`



Η κάθετη γραμμή (δείκτης a) παράγεται αν βάλουμε μέσα στο κείμενο τη λέξη «xmilestone”.
Εδώ η λέξη μπήκε στο σημείο που άλλαξε ο συγγραφέας για ελέγξουμε οπτικά την ακρίβεια της διάκρισης του αλγόριθμου.