Πληροφορική στην Ιστορική Έρευνα

Κ. Κερμανίδου

Πληροφορική στην Ιστορική Έρευνα

- **Αναγνώριση ιστορικού συγγραφέα**
  - **Ο προσδιορισμός του συγγραφέα ενός ιστορικού έργου, του οποίου η πατρότητα δεν είναι γνωστή**
- Επισημείωση ιστορικού κειμένου
  - Ο εμπλουτισμός του ιστορικού κειμένου με μεταδεδομένα (περαιτέρω πληροφορία όπως συγγραφέας, ιστορική πηγή προέλευσης κειμένου, διασύνδεση με άλλες ιστορικές πηγές κλπ)
- Αυτόματη κατηγοριοποίηση ιστορικών κειμένων
  - Ανά ιστορικό γεγονός, ανά ιστορική περίοδο, τοπολογική κατηγοριοποίηση κλπ
- Αυτόματη εξόρυξη πληροφορίας από ιστορικά κείμενα
  - Αναγνώριση συγκεκριμένων στοιχείων ιστορικών γεγονότων (χρόνος, τόπος, συμμετέχοντες, είδος γεγονότος κλπ)
- Κατανόηση ιστορικών κειμένων
  - Αυτόματη εξαγωγή συμπεράσματος, αυτόματη αναγνώριση αιτίων, αυτόματη αναπαράσταση νοήματος

Τι είναι η Αναγνώριση Συγγραφέα (Author Identification / Authorship Attribution)

- Η επιστημονική περιοχή που ασχολείται με τον προσδιορισμό του συγγραφέα ενός κειμένου, όταν δεν είναι ξεκάθαρο ποιός το έχει γράψει.
- Είναι χρήσιμη όταν δύο ή περισσότεροι δημιουργοί ισχυρίζονται ότι έχουν γράψει ένα κείμενο ή όταν κανείς δεν ισχυρίζεται την συγγραφή κάποιου έργου
  - είτε γιατί δεν μπορεί
  - είτε γιατί δεν θέλει

Γλώσσα και Αναγνώριση Συγγραφέα

- Η σύνταξη της γλώσσας για κάποιον ειδήμονα είναι μια διαδικασία υποσυνείδητη και 'αυτόματη'
- Η ιδιότητα αυτή της σύνταξης την καθιστά εγγενή διαδικασία, και άρα ενδεικτικό παράγοντα της ιδιαιτερότητας/ατομικότητας του καθενός μας

Η ταυτότητα του δημιουργού

- Η συγγραφή ενός κειμένου περιλαμβάνει πληροφορία από όλα τα επίπεδα γλωσσολογικής γνώσης: μορφολογία, σύνταξη, σημασιολογία και πραγματολογία.
- Κάθε ένα από αυτά τα επίπεδα κυβερνάται από τους δικούς του κανόνες. Οι κανόνες αυτοί δίνουν σε έναν συγγραφέα επιλογές.
- Το κείμενο σαν τελικό προϊόν είναι το αποτέλεσμα των επιλογών αυτών του συγγραφέα. Έτσι, κάθε κείμενο φέρει την «υπογραφή» ή την «ταυτότητα» του δημιουργού του.

Υποθέσεις

- Σε κάθε κείμενο αντιστοιχεί ένας και μόνο συγγραφέας
- Υπάρχει μια ομάδα υποψήφιων συγγραφέων για ένα κείμενο
- Ο συγγραφέας είναι συνεπής στις γλωσσολογικές επιλογές του
- Οι επιλογές αυτές είναι παρούσες και μπορούν να ανιχνευθούν σε όλα τα έργα του συγγραφέα.

Αναγνώριση Συγγραφέα: Ιστορικά Στοιχεία (1/3)

- Ο Augustus de Morgan (1851) πρώτος πρότεινε την έννοια της «στυλομετρίας» (stylometry)
- Στυλομετρία: μοντελοποίηση του ύφους ενός συγγραφέα/κειμένου με βάση τα γλωσσολογικά χαρακτηριστικά του
- Ο Mendenhall (1887) πρώτος πρότεινε την κατανομή του μήκους των λέξεων και άλλα στατιστικά χαρακτηριστικά για την Αναγνώριση Συγγραφέα
- Οι Yule (1938) και Morton (1965) πρώτοι πρότειναν την κατανομή του μήκους των προτάσεων.

Αναγνώριση Συγγραφέα: Ιστορικά Στοιχεία (2/3) - Στυλομετρία

- "People's unconscious use of everyday words comes out with a certain stamp",

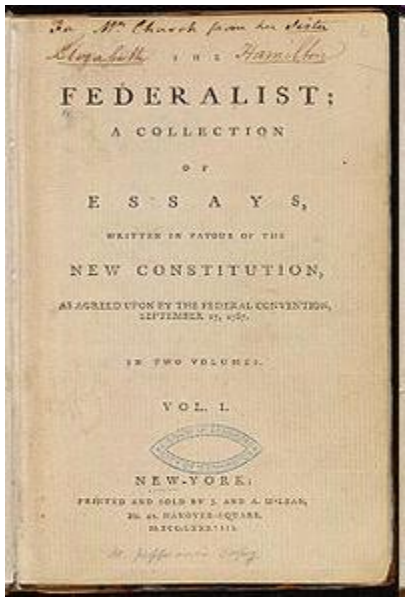
  David Holmes - stylometrist στο College of New Jersey

- "Rare words are noticeable words, which someone else might pick up or echo unconsciously. It's much harder for someone to imitate my frequency pattern of 'but' and 'in'.",

  John Burrows - emeritus καθηγητής Αγγλικών στο University of Newcastle στην Αυστραλία

Αναγνώριση Συγγραφέα: Ιστορικά Στοιχεία (3/3)

- Το πρόβλημα της συγγραφής των Federalist papers (Mosteller & Wallace, 1968)
- Federalist papers: 146 δοκίμια που γράφτηκαν το 1787-1788 για την υποστήριξη του καινούριου συντάγματος των ΗΠΑ από τους
  - A. Hamilton,
  - J. Lay,
  - J. Madison
- Τα 12 από αυτά είναι διφορούμενα: δεν ξέρουμε αν έχουν γραφεί από τον Hamilton ή από τον Madison



Που μπορεί να χρησιμοποιηθεί η Αναγνώριση Συγγραφέα;

- Στον προσδιορισμό του συγγραφέα ενός έργου
- Στον εντοπισμό αντιγραφής έργων – όπου προσδιορίζεται η εγκυρότητα του ισχυρισμού συγγραφής ενός έργου από κάποιον δημιουργό
- Σε εγκληματολογικές έρευνες: αναγνώριση συγγραφέων τρομοκρατικών μηνυμάτων κλπ
- Στην αναγνώριση των συγγραφέων σε μηνύματα email ή άλλο ηλεκτρονικό κείμενο, ή στον προσδιορισμό της προέλευσης μιας πληροφορίας.

Η Αναγνώριση Συγγραφέα

- Είναι έντονα διεπιστημονική περιοχή. Συνδυάζει
  - Γλωσσολογία
  - Λογοτεχνία
  - Στατιστική
  - Τεχνητή Νοημοσύνη

Γλωσσολογικά Χαρακτηριστικά και Αναγνώριση Συγγραφέα (1/6)

- Απλές μετρήσεις
  - Μέσο μήκος λέξεων (σε αριθμό χαρακτήρων)
  - Μέσο μήκος προτάσεων (σε αριθμό λέξεων)
- Λεξιλογικά χαρακτηριστικά
  - Πλούτος λεξιλογίου (Vocabulary richness)
    - το ποσοστό των διαφορετικών λέξεων σε ένα κείμενο προς τον συνολικό αριθμό των λέξεων του κειμένου
  - Άπαξ λεγόμενα (Hapax Legomena)
    - το ποσοστό των λέξεων που εμφανίζονται μια μόνο φορά μέσα σε ένα κείμενο προς τον συνολικό αριθμό των λέξεων του κειμένου
  - Χρήση λειτουργικών λέξεων (function words)
    - Λειτουργικές λέξεις είναι οι λέξεις σε ένα κείμενο που δεν έχουν σημασιολογικό περιεχόμενο (π.χ. άρθρα, προθέσεις, αντωνυμίες κλπ)
    - Υπολογισμός της συχνότητας εμφάνισης λειτουργικών λέξεων μέσα στο κείμενο

Γλωσσολογικά Χαρακτηριστικά και Αναγνώριση Συγγραφέα (2/6)

- Συχνότητες χαρακτήρων
  - Συχνότητα αλφαριθμητικών χαρακτήρων
    - Υπολογισμός της συχνότητας των χαρακτήρων γραμμάτων και αριθμών μέσα στο κείμενο
  - Συχνότητα χαρακτήρων
    - Υπολογισμός της συχνότητας όλων των χαρακτήρων μέσα στο κείμενο
  - Συχνότητα διγράμμων χαρακτήρων (bigrams)
    - Υπολογισμός της συχνότητας ζευγών χαρακτήρων μέσα στο κείμενο
  - Συχνότητα τριγράμμων χαρακτήρων (trigrams)
    - Υπολογισμός της συχνότητας ακολουθιών τριών χαρακτήρων μέσα στο κείμενο
  - Συχνότητα τετραγράμμων χαρακτήρων (tetragrams)
    - Υπολογισμός της συχνότητας ακολουθιών τεσσάρων χαρακτήρων μέσα στο κείμενο

Γλωσσολογικά Χαρακτηριστικά και Αναγνώριση Συγγραφέα (3/6)

- Μορφολογικά χαρακτηριστικά
  - Μέσος αριθμός συλλαβών ανά λέξη στο κείμενο
  - Ορθογραφικά Λάθη
    - κατηγοριοποίηση των ορθογραφικών λαθών σε κατηγορίες:
      - λάθη με διπλά σύμφωνα
      - λάθη με διπλά σύμφωνα στις καταλήξεις
      - λάθη σε φωνήεντα πριν από ένρινα σύμφωνα
  - Γραμματικά λάθη
    - ελλιπής πρόταση
    - ασυμφωνία υποκειμένου-ρήματος
    - παράλειψη ρήματος
    - λανθασμένος χρόνος ρήματος
  - Χρήση μερών του λόγου (part-of-speech)
    - Υπολογισμός της συχνότητας εμφάνισης των διαφόρων μερών του λόγου μέσα στο κείμενο
  - Χρήση παθητικής φωνής
    - Υπολογισμός της συχνότητας χρήσης της παθητικής φωνής στα ρήματα του κειμένου

Γλωσσολογικά Χαρακτηριστικά και Αναγνώριση Συγγραφέα (4/6)

- Συντακτικές Δομές
  - Ανάλυση των δομών των φράσεων σε μια πρόταση, π.χ.
    - ανάλυση της δομής των ρηματικών φράσεων
    - ανάλυση της δομής των προθετικών φράσεων
  - Υπολογισμός των συχνοτήτων των διαφορετικών δομών
- Χρήση σημείων στίξης
  - Κατηγοριοποίηση των σημείων στίξεων ως προς την συντακτική τους λειτουργία
    - τελεία που υποδηλώνει τέλος πρότασης
    - κόμμα που χωρίζει κύρια από δευτερεύουσα πρόταση
    - κόμμα σε παρατακτική σύνδεση
  - Για κάθε κατηγορία υπολογίζεται η συχνότητά της

Γλωσσολογικά Χαρακτηριστικά και Αναγνώριση Συγγραφέα (5/6)

- Υφολογικά χαρακτηριστικά
  - Χρήση αρχαϊσμών
    - Υπολογισμός της συχνότητας χρήσης αρχαϊσμών μέσα στο κείμενο
  - Χρήση μεταφορών
    - Υπολογισμός της συχνότητας χρήσης μεταφορών μέσα στο κείμενο
  - Χρήση σλόγκαν
    - Υπολογισμός της συχνότητας χρήσης σλόγκαν μέσα στο κείμενο

Γλωσσολογικά Χαρακτηριστικά και Αναγνώριση Συγγραφέα (6/6)

- Σημασιολογικά χαρακτηριστικά
  - Αναγνώριση λέξεων ή φράσεων που εμπεριέχουν συγκεκριμένη σημασιολογική πληροφορία, όπως είναι
    - η εισαγωγή επεξηγηματικής πρότασης (π.χ. «δηλαδή», «με άλλα λόγια»)
    - η περαιτέρω διευκρίνιση (π.χ. «ειδικότερα»)
    - η αιτιότητα (π.χ. «επειδή»)
    - η επαύξηση (π.χ. «επιπλέον», «περαιτέρω»)
    - η αντίθεση (π.χ. «αλλά», «αντίθετα»)
    - η χρονική συσχέτιση (π.χ. «μετά», «πριν»)
  - Χρήση σημασιολογικού θησαυρού (WordNet) για τον υπολογισμό της αμφισημίας, της πολυσημίας – κατά πόσο ο συγγραφέας χρησιμοποιεί αμφίσημες λέξεις κλπ

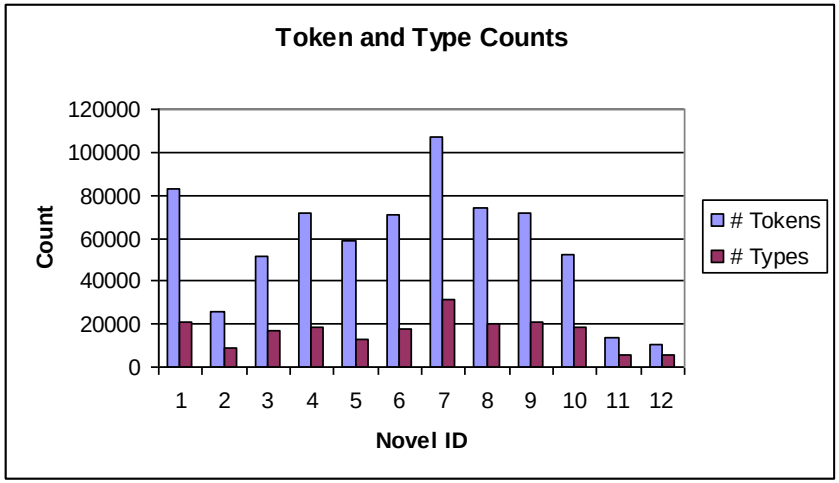
Χαρακτηριστικά Εξαρτώμενα από την Εφαρμογή

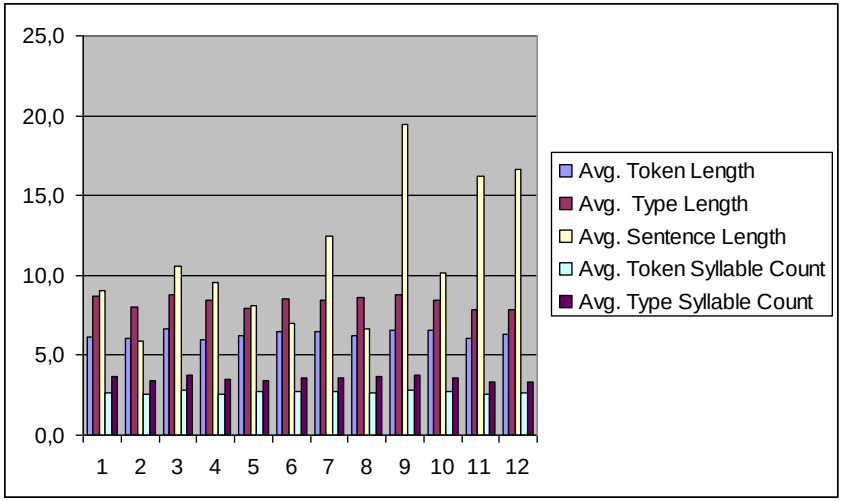
- Σε εξειδικευμένες εφαρμογές παίζουν ρόλο και πιο εξειδικευμένα χαρακτηριστικά
- Στην αναγνώριση του συγγραφέα ενός ηλεκτρονικού μηνύματος
  - η χρήση χαιρετισμών
  - η χρήση υπογραφών
  - το μήκος των παραγράφων
  - η χρήση στοίχισης
- Στην αναγνώριση του συγγραφέα ενός κειμένου σε html μορφή
  - η κατανομή των html tags
  - συχνότητα χρήσης χρωμάτων και μεγέθους χαρακτήρων
- Σε συγκεκριμένες γλώσσες
  - μέτρηση συχνότητας καταλήξεων των ρημάτων στην δημοτική και στην καθαρεύουσα

Παράδειγμα

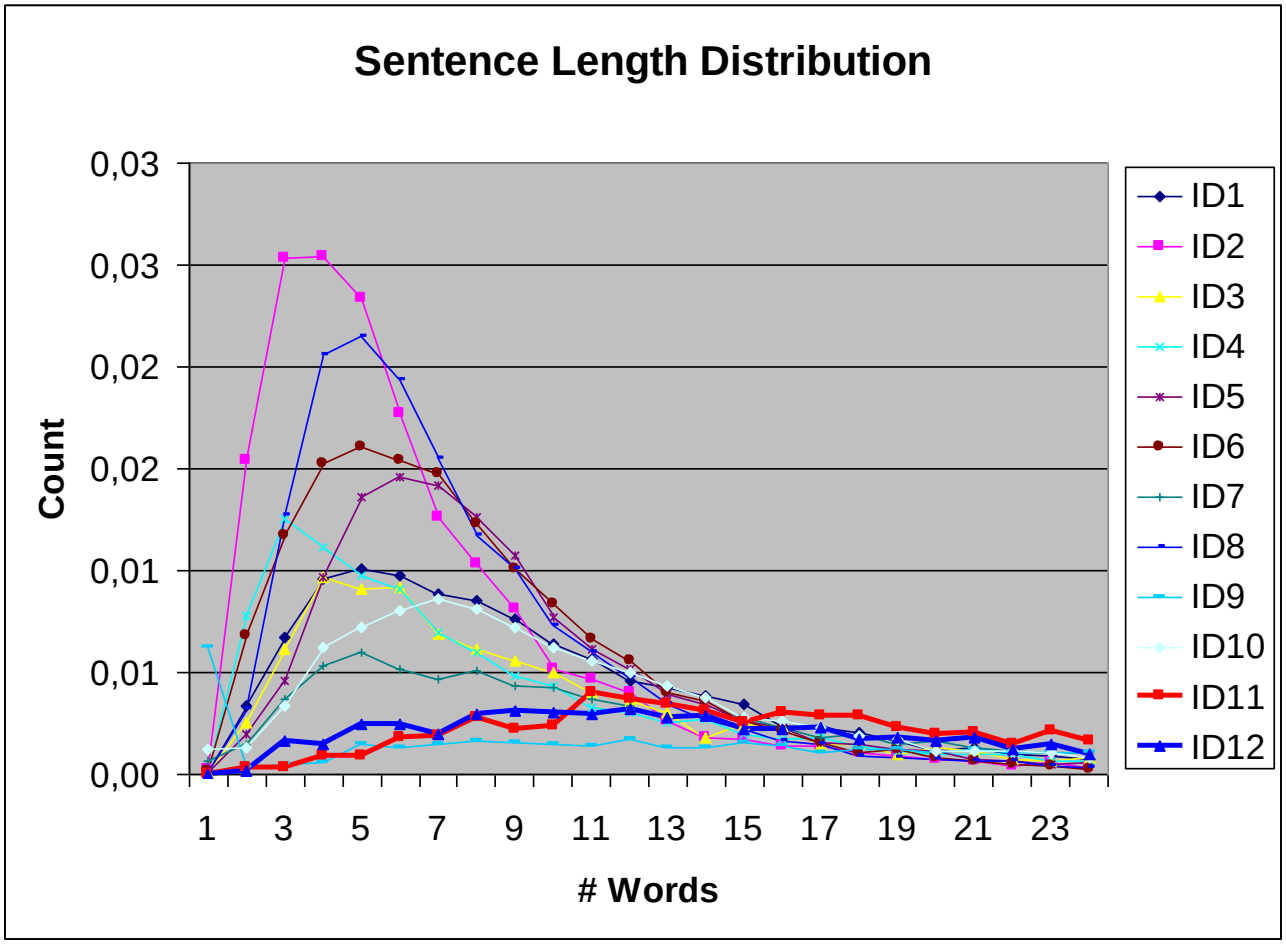
ID	Novel	Author	Year	# Tokens	# Types	Avg. Token Lengt	Avg. Type Length	Avg. Sentence Length	Avg. Token Syllable Count	Avg. Type Syllable Count
1	Yarın Yarın	Pınar Kür	1976	83325	21121	6,1790	8,7150	9,0400	2,6460	3,6920
2	Hakkari'de Bir Mevsim	Ferid Edgü	1977	25636	8791	6,0390	8,0350	5,9150	2,5500	3,3820
3	Ölüm İlişkileri	Selim İleri	1979	51550	17273	6,6600	8,8220	10,6050	2,8100	3,7240
4	Sessiz Ev	Orhan Pamuk	1983	71386	18140	6,0050	8,4180	9,5780	2,5440	3,5360
5	Sevgili Arsız Ölüm	Latife Tekin	1983	58627	12944	6,1980	7,9100	8,0880	2,6920	3,3860
6	Issızlığın Ortasında	Mehmet Eroğlu	1984	70478	17831	6,4810	8,5410	6,9730	2,7720	3,6070
7	Haco Hanım Vay!..	Atilla İlhan	1984	106742	31329	6,4690	8,4380	12,4160	2,7390	3,5450
8	Devrimciler	Kaan Arslanoğlu	1988	73885	19945	6,2340	8,5920	6,6610	2,6660	3,6280
9	Kılıç Yarası Gibi	Ahmet Altan	1998	71643	20867	6,5540	8,7900	19,4520	2,7970	3,7270
10	Boğazkesen: Fatih'in Romanı	Nedim Gürsel	1995	52021	18137	6,5690	8,4630	10,1760	2,7720	3,5610
11	Gazete Yazıları*	Cem Behar	2002	13797	5251	6,0990	7,8770	16,1940	2,5760	3,2980
12	**Gizli Kalmış Bir İstanbul Masalı**	**Nurten Ay?**	**1991**	**10559**	**5539**	**6,2850**	**7,8340**	**16,6020**	**2,6730**	**3,3050**

Παράδειγμα (συν)

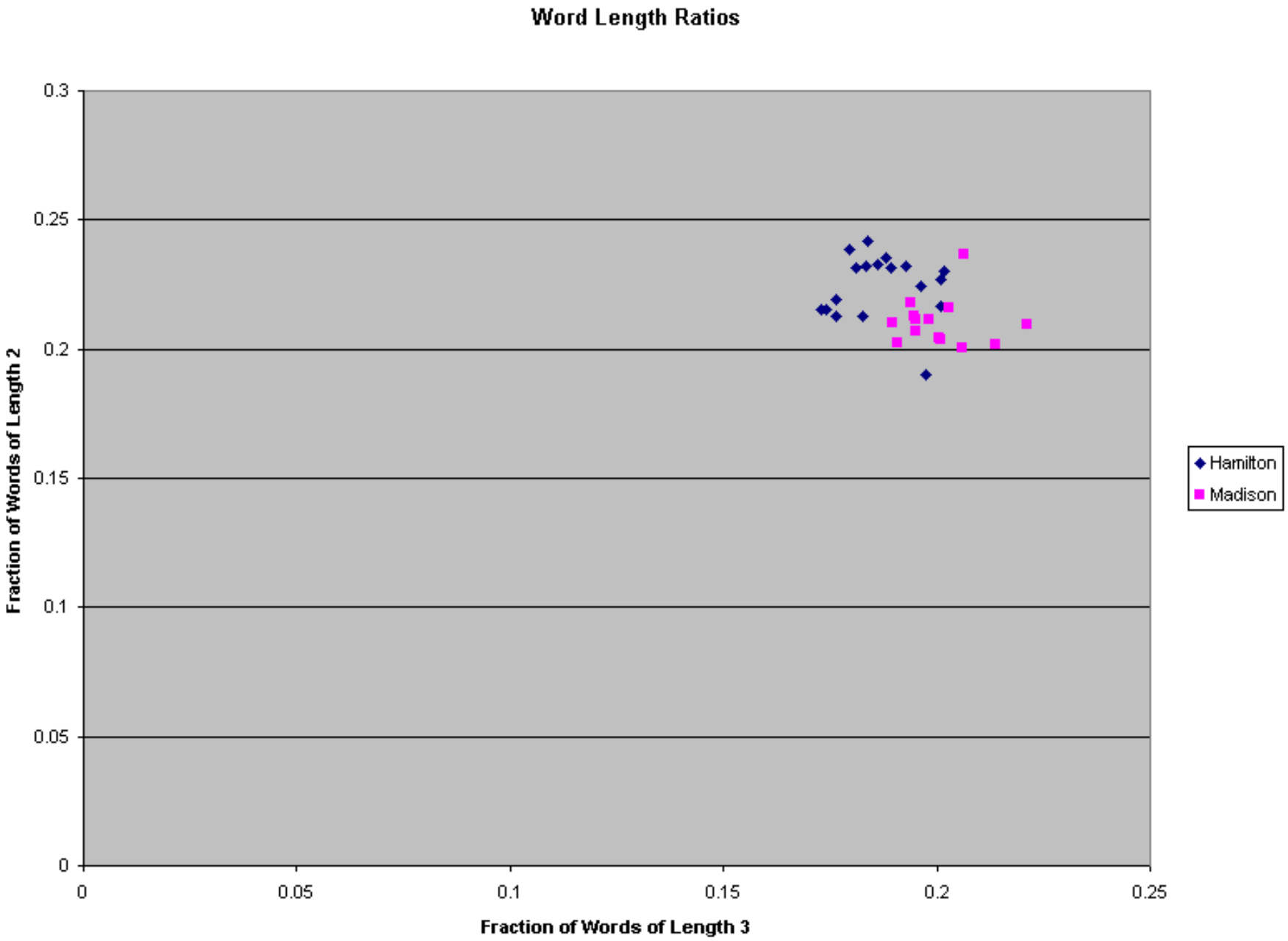




Παράδειγμα (συν)



Federalist papers



Διάκριση μεταξύ των έργων του Shakespeare και του Marlowe

- (Merriam & Matthews, 1994)
- T10: συχνότητα εμφάνισης των *but, by, for, no, not, so, that, the, to, with*
- Χρησιμοποιούν τα εξής χαρακτηριστικά (*x* είναι οποιαδήποτε λέξη):
  - 'no'/T10
  - (of *x* and)/ 'of'
  - 'so'/T10
  - (the *x* and)/ 'the'
  - 'with'/T10

Shakepeare vs Marlowe (Tearle et al. 2008)

(1) average sentence length (words per sentence),
(2) possessive apostrophes per sentence,
(3) possessive apostrophes per sentence, averaged by sentence length,
(4) quotation marks per sentence,
(5) quotation marks per sentence, averaged by sentence length,
(6) dashes per sentence,
(7) dashes per sentence, averaged by sentence length,
(8) semicolons per sentence,
(9) semicolons per sentence, averaged by sentence length,
(10) commas per sentence,
(11) commas per sentence, averaged by sentence length,
(12) average word length,
(13) *no*/T10,
(14) *so*/T10,
(15) *with*/T10,
(16) *of X and/of*,
(17) *the X and/the* (where *the* is either *the* or *th'*),
(18) *no*/(*no* + *not*),
(19) *upon*/(*on* + *upon*),
(20) type-token ratio (number of different words/number of words),
(21) Yule's Characteristic $K = 100D(1 - 1/N)$, where D is Simpson's Index,
(22) entropy: $\sum_i -p_i \log(p_i)$ where $p_i =$ (number of occurrences of word i)/(total number of words),
(23) word-frequency distribution characteristic A,
(24) word-frequency distribution characteristic X,
(25) hapax legomena,
(26) hapax dislegomena,
(27) mixture measure (weighted average of number of times words on either side of a word from V_1 are used),
(28) average percentage of sentence position of T10 words,
(29) average spacing of the letter *e*,
(30) average spacing of the letter *m*,
(31) average spacing of the letter *o*, and
(32) average spacing of the letter *t*.

spacing of a letter: the average number of (nonpunctuation) characters between consecutive occurrences of the letter

Περισσότερες πληροφορίες: Holmes, «Authorship Attribution», 1994.

Shakepeare vs Marlowe (Tearle et al. 2008)

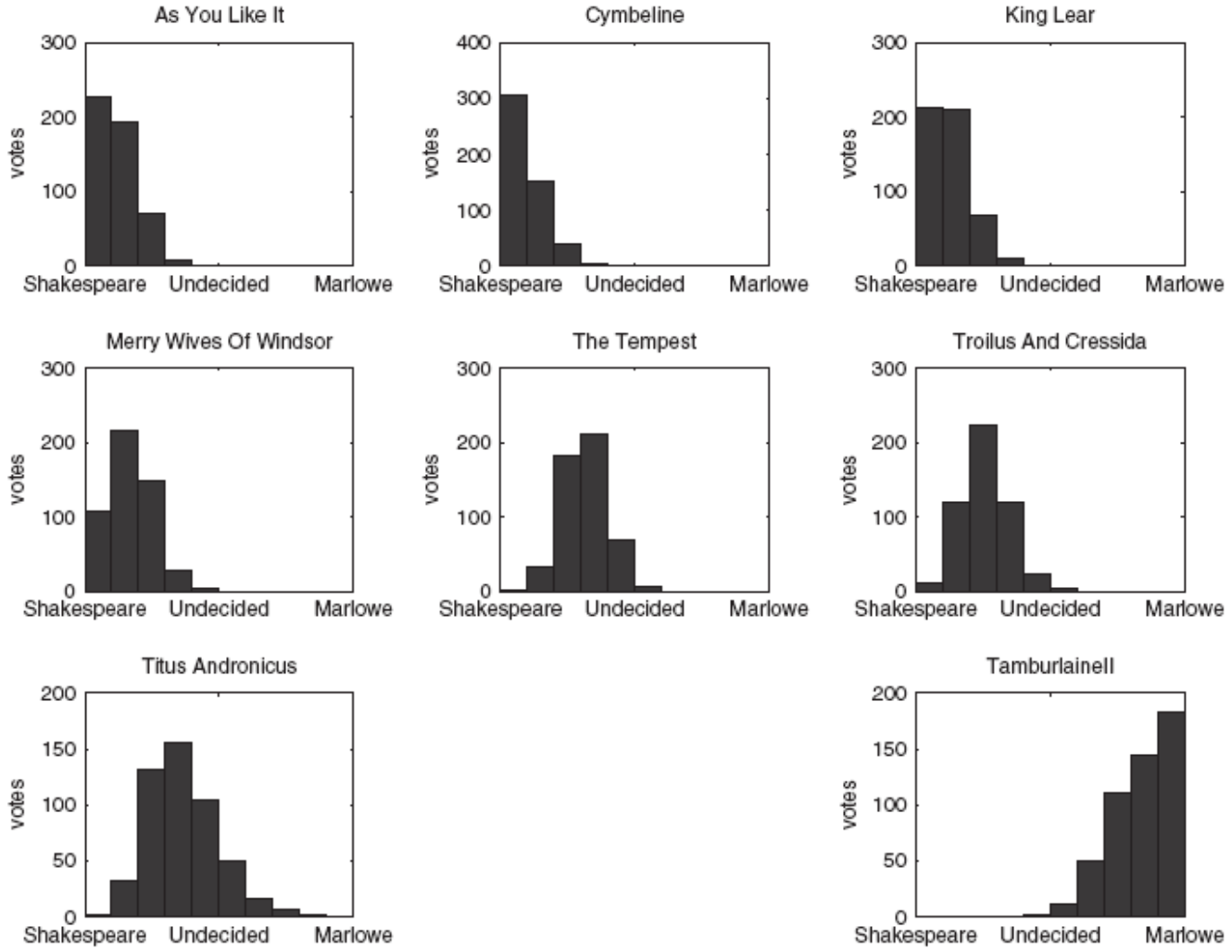


Fig. 5 Histogram of the predictions for 'questionable' works from a committee of 500 5-input networks, with the input metrics fixed

Είναι τα άρθρα Quintus Curtius Snodgrass του M. Twain;

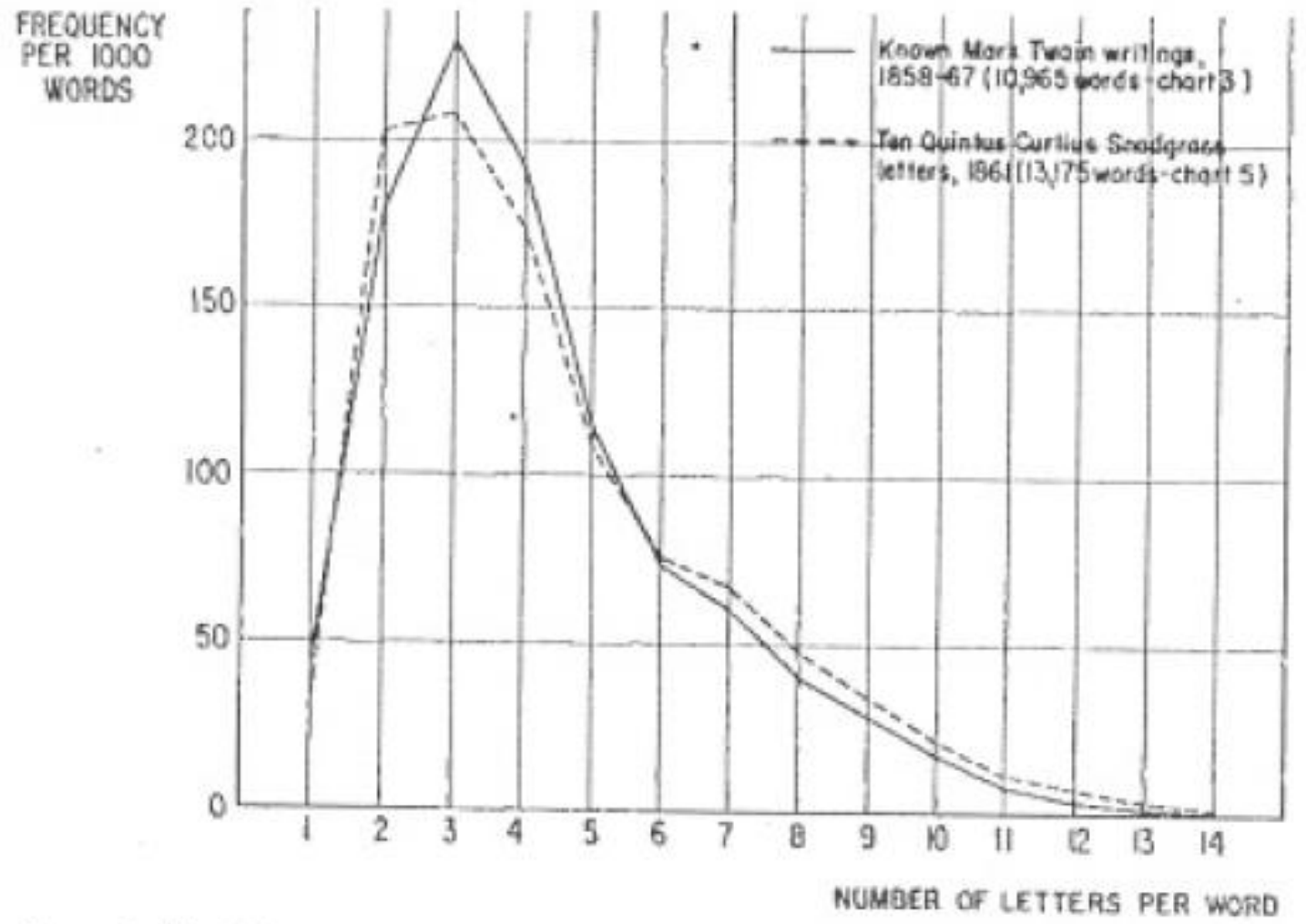


CHART 6. Word frequencies for known Mark Twain Writingsand Quintus Curtius Snodgrass Letters.

Histograms of word length in Mark Twain and Quintus Curtius Snodgrass

Αλγόριθμοι για την Αναγνώριση Συγγραφέα

- Στατιστικές τεχνικές
  - Συνεμφάνιση
  - Στατιστική Σημαντικότητα
- Τεχνικές Μηχανικής Μάθησης
  - **k-πλησιέστεροι γείτονες**
  - Naïve Bayes
  - Νευρωνικά Δίκτυα

Μηχανική Μάθηση (Machine Learning)

- Ένα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από την εμπειρία E σε σχέση με μια κατηγορία εργασιών T και μια μετρική απόδοσης P, αν η απόδοση του σε εργασίες της T, όπως μετριούνται από την P, βελτιώνονται με την εμπειρία E". Tom Mitchell (1997)

  - Task T: playing chess
  - Performance measure P: percent of games won against opponents
  - Training Experience E: playing practice games against itself

Μηχανική Μάθηση (Machine Learning)

- Το σύστημα μάθησης καλείται να μάθει επαγωγικά την **έννοια μάθησης** (το πότε ένα έργο είναι του συγγραφέα Α και πότε του Β)
- Η έννοια μάθησης λέγεται και συνάρτηση στόχος (target function)
- Επαγωγή είναι η διαδικασία δημιουργίας ενός γενικευμένου μοντέλου περιγραφής ή ορισμού μιας έννοιας από ένα σύνολο ειδικών παραδειγμάτων της έννοιας
- Η τιμή της συνάρτησης-στόχου ονομάζεται και εξαρτημένη μεταβλητή ενώ οι υπόλοιπες μεταβλητές που αναπαριστούν το παράδειγμα ονομάζονται ανεξάρτητες μεταβλητές.
- Ένα πείραμα μάθησης διακρίνεται σε δυο φάσεις:
- **Εκπαίδευση:** Κατά την εκπαίδευση, με επαγωγή, χρησιμοποιούνται τα παραδείγματα εκπαίδευσης για να εξαχθεί το γενικό μοντέλο (ο κανόνας) που διέπει την έννοια μάθησης.
- Η τιμή της συνάρτησης στόχου των παραδειγμάτων εκπαίδευσης είναι γνωστή, και καθοδηγεί τη διαδικασία μάθησης
- **Έλεγχος:** Κατά τον έλεγχο, αξιολογείται το μοντέλο που προέκυψε. Η απόδοσή του μετράται σε καινούρια παραδείγματα (ονομάζονται παραδείγματα ελέγχου), τα οποία ο αλγόριθμος τα βλέπει πρώτη φορά, και για τα οποία ο αλγόριθμος δεν γνωρίζει την τιμή της συνάρτησης στόχου, και καλείται να την προβλέψει.

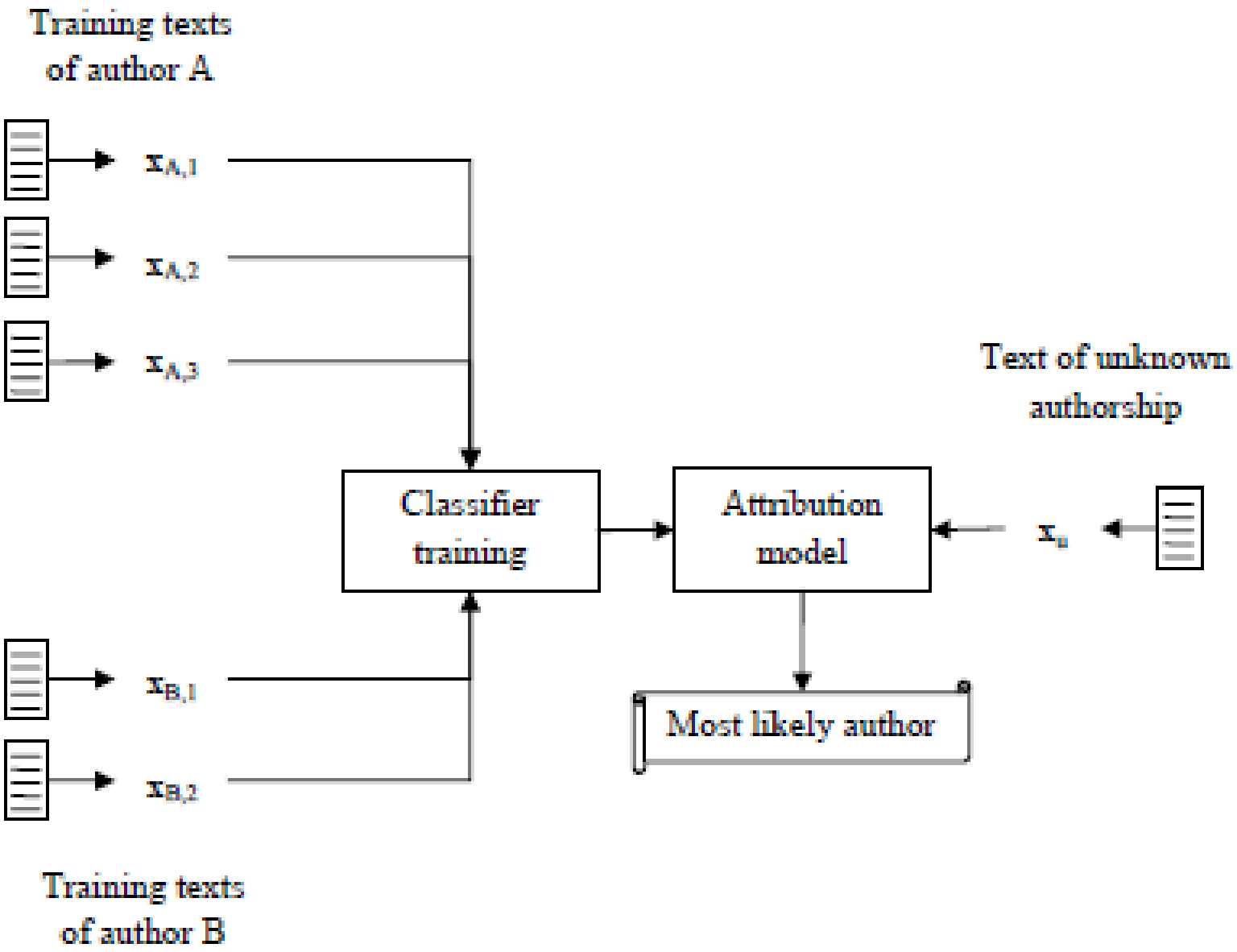


FIG 2. Typical architecture of instance-based approaches.

Κατανομή των συγγραφέων στα κείμενα εκπαίδευσης και ελέγχου

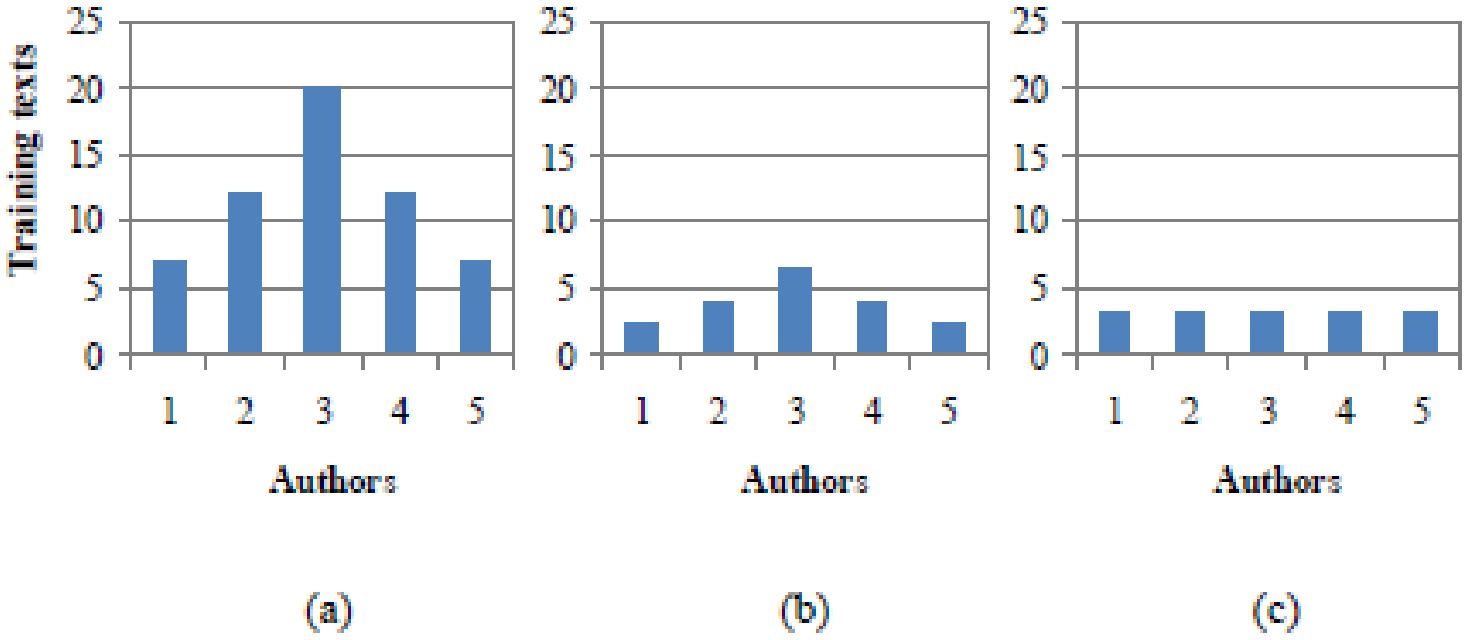


FIG. 3. Different distributions of training and test texts over 5 candidate authors: (a) an imbalanced distribution of training texts, (b) an imbalanced distribution of test texts imitating the distribution of training texts, (c) a balanced distribution of test texts.

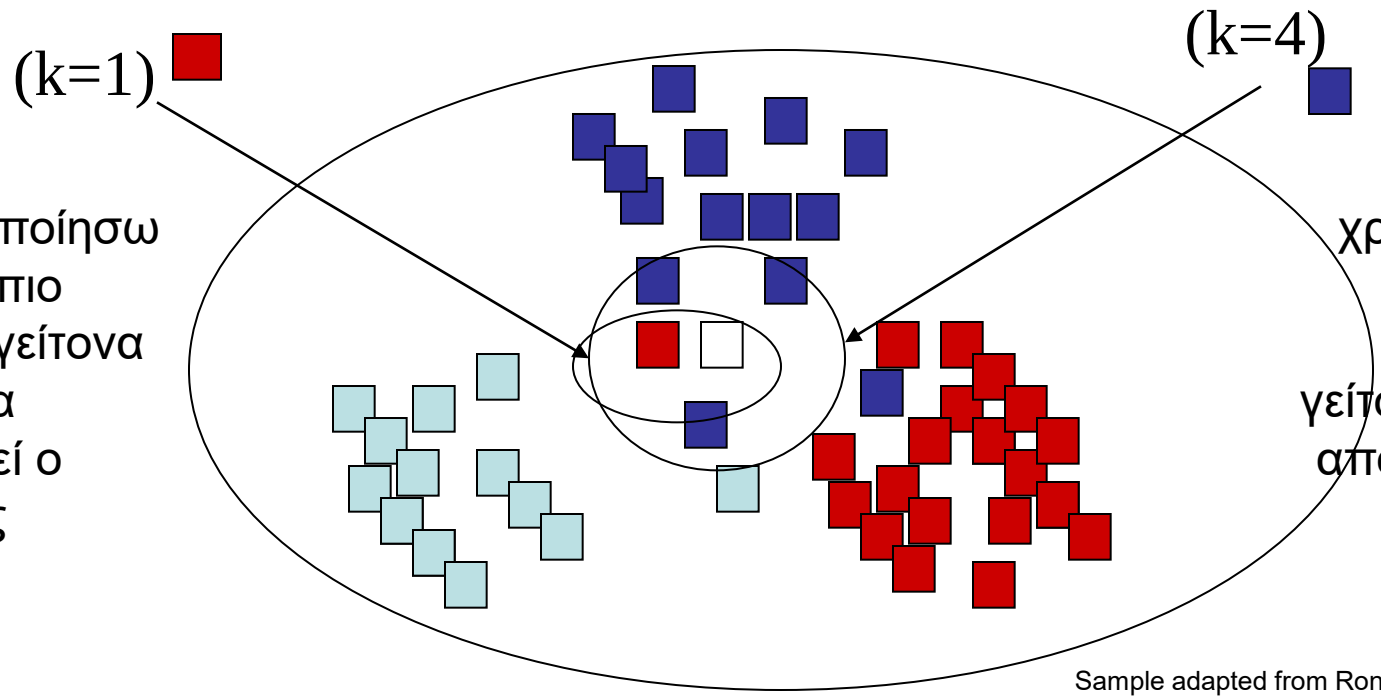
Ο αλγόριθμος των πλησιέστερων γειτόνων

- Ονομάζεται και
  - Memory-based Learning
  - Case-based Learning
  - Lazy Learning
- Τα παραδείγματα εκπαίδευσης διατηρούνται αυτούσια...
  - ...σε αντίθεση με άλλες μεθόδους μηχανικής μάθησης οι οποίες κωδικοποιούν τα παραδείγματα εκπαίδευσης σε μια συμπαγή περιγραφή.
- Όταν ένα τέτοιο σύστημα κληθεί να αποφασίσει για την συγγραφέα ενός νέου έργου, εξετάζει εκείνη τη στιγμή τη σχέση του έργου με τα ήδη αποθηκευμένα παραδείγματα εκπαίδευσης.
- Κάνει την παραδοχή ότι τα διάφορα παραδείγματα μπορεί να αναπαρασταθούν ως σημεία σε κάποιον n-διάστατο Ευκλείδειο χώρο Rn όπου n ο αριθμός των χαρακτηριστικών (ανεξάρτητων μεταβλητών).
- Κάθε νέο παράδειγματοποθετείται στο χώρο αυτό ως νέο σημείο και η τιμή του η προσδιορίζεται με βάση το χαρακτηρισμό των k γειτονικών σημείων.

Αλγόριθμος πλησιέστερων γειτόνων

Τα μπλε, κόκκινα, και σιέλ παραδείγματα είναι εκπαίδευσης, ξέρω τον συγγραφέα τους.

- Έστω ότι θέλω να προβλέψω αν ο συγγραφέας του άσπρου άγνωστου (ελέγχου) παραδείγματος είναι ο μπλε, ο κόκκινος ή ο σιέλ.
- Υπολογίζω την απόσταση του άγνωστου εγγράφου από όλα τα γνωστά έγγραφα (τα παραδείγματα εκπαίδευσης) και επιλέγω τα k παραδείγματα εκπαίδευσης που είναι πιο κοντά στο άσπρο (k κοντινότεροι γείτονες).
- Στον συγγραφέα που ανήκει η πλειοψηφία των k κοντινότερων γειτόνων, σε αυτόν τον συγγραφέα θα αποδοθεί και το άσπρο παραδειγμα.



Αν χρησιμοποίησω τον ένα πιο κοντινό γείτονα (k=1), θα αποδοθεί ο κόκκινος

Αν χρησιμοποίησω τους 4 πιο κοντινούς γείτονες (k=4), θα αποδοθεί ο μπλε (πλειοψηφία)

Sample adapted from Rong Jin's slides

Ο αλγόριθμος των Πλησιέστερων Γειτόνων: Απόσταση

- Για τον υπολογισμό της απόστασης μπορώ να χρησιμοποιήσω οποιοδήποτε μέτρο απόστασης ανάμεσα σε δυο σημεία στον χώρο n διαστάσεων (n: αριθμός ανεξάρτητων χαρακτηριστικών)
- Συνήθως χρησιμοποιείται η Ευκλείδεια απόσταση (Euclidean distance)
- Αν έχω δυο παραδείγματα, το (1) και το (2), και *n* χαρακτηριστικά, η ΕΑ μεταξύ τους είναι
  - $EA(1,2) = ((a_1^{(1)} - a_1^{(2)})^2 + (a_2^{(1)} - a_2^{(2)})^2 + \ldots + (a_n^{(1)} - a_n^{(2)})^2)^{1/2}$
  - $a_k^{(1)}$: η τιμή του χαρα/κού *k* στο παράδειγμα (1)
  - $a_k^{(2)}$: η τιμή του χαρα/κού *k* στο παράδειγμα (2)
- Άλλη συνάρτηση απόστασης: Manhattan ή City-block distance
  - $M(1,2) = |a_1^{(1)} - a_1^{(2)}| + |a_2^{(1)} - a_2^{(2)}| + \ldots + |a_n^{(1)} - a_n^{(2)}|$

Παράδειγμα με Ευκλείδεια

- Έστω πως έχω τέσσερα παραδείγματα εκπαίδευσης, δυο έργα του συγγραφέα Α και δυο του συγγραφέα Β. Οι ιστορικοί ερευνητές χρησιμοποιούν για την διάκριση ανάμεσα στους δυο συγγραφείς δυο ανεξάρτητες μεταβλητές (χαρακτηριστικά): το μέσο μήκος πρότασης και το μέσο μήκος λέξης

	Μέσο μήκος πρότασης	Μέσο μήκος λέξης	Συγγραφέας
1	9.2	5.3	A
2	10.6	6.1	A
3	10	6.9	B
4	7.4	6	B

- Έστω πως έρχεται ένα έργο με ΜΜΠ=8.8 και ΜΜΛ=6.2.
- $EA(1,5)=((9.2-8.8)^2+(5.3-6.2)^2)^{1/2}=0.985$
- $EA(2,5)=((10.6-8.8)^2+(6.1-6.2)^2)^{1/2}=1.80$
- $EA(3,5)=((10-8.8)^2+(6.9-6.2)^2)^{1/2}=1.39$
- $EA(4,5)=((7.4-8.8)^2+(6-6.2)^2)^{1/2}=1.41$
- Σε ποιόν συγγραφέα θα αποδώσει το άγνωστο έργο ο 1-ΝΝ. Σε ποιόν ο 3-ΝΝ;

Παράδειγμα με Manhattan

- Έστω πως έχω τέσσερα παραδείγματα εκπαίδευσης, δυο έργα του συγγραφέα Α και δυο του συγγραφέα Β. Οι ιστορικοί ερευνητές χρησιμοποιούν για την διάκριση ανάμεσα στους δυο συγγραφείς δυο ανεξάρτητες μεταβλητές (χαρακτηριστικά): το μέσο μήκος πρότασης και το μέσο μήκος λέξης

	Μέσο μήκος πρότασης	Μέσο μήκος λέξης	Συγγραφέας
1	9.2	5.3	A
2	10.6	6.1	A
3	10	6.9	B
4	7.4	6	B

- Έστω πως έρχεται ένα έργο με ΜΜΠ=8.8 και ΜΜΛ=6.2.
- M(1,5)=|9.2-8.8|+|5.3-6.2|=1.3
- M(2,5)=|10.6-8.8|+|6.1-6.2|=1.9
- M(3,5)=|10-8.8|+|6.9-6.2|=1.9
- M(4,5)=|7.4-8.8|+|6-6.2|=1.6
- Σε ποιόν συγγραφέα θα αποδώσει το άγνωστο έργο ο 1-ΝΝ. Σε ποιόν ο 3-ΝΝ;

Αξιολόγηση Αναγνώρισης Συγγραφέα

- Πολλές παράμετροι πρέπει να ληφθούν υπόψη κατά την αξιολόγηση ενός συστήματος αναγνώρισης συγγραφέα
  - το μέγεθος του σώματος κειμένων εκπαίδευσης (μήκος και πλήθος κειμένων)
  - το μέγεθος του σώματος κειμένων ελέγχου
  - ο αριθμός των υποψήφιων συγγραφέων
  - η κατανομή των κειμένων εκπαίδευσης ως προς τους συγγραφείς (ισορροπημένο ή μη)
  - η δυνατότητα να αντιμετωπίζει περισσότερες φυσικές γλώσσες

Αξιολόγηση – Πίνακας Σύγχυσης

Η απόδοση ενός συστήματος αναγνώρισης συγγραφέα μπορεί να μετρηθεί μόνο αν ο ερευνητής ξέρει τον πραγματικό συγγραφέα των άγνωστων έργων (παραδειγμάτων ελέγχου) – ή, αλλιώς, την **αλήθεια.**

Έστω πως χρησιμοποιώ για αξιολόγηση 10 παραδείγματα ελέγχου. Ο παρακάτω πίνακας λέγεται **πίνακας σύγχυσης** (confusion matrix)

		Πραγματικός Συγγραφέας	
		A	**B**
Προβλεπόμενος Συγγραφέας	**A**	4	2
	B	1	3

4: τα παραδείγματα που στην πραγματικότητα είναι του Α, και ο k-NN σωστά τα απέδωσε στον Α

3: τα παραδείγματα που στην πραγματικότητα είναι του Β, και ο k-NN σωστά τα απέδωσε στον Β

2: τα παραδείγματα που στην πραγματικότητα είναι του Β, και ο k-NN λανθασμένα τα απέδωσε στον Α

1: τα παραδείγματα που στην πραγματικότητα είναι του Α, και ο k-NN λανθασμένα τα απέδωσε στον Β

Μέτρα Αξιολόγησης

Accuracy (Ορθότητα) = το ποσοστό των σωστών ταξινομήσεων = (4+3)/(4+3+2+1) = 0.7

Η ορθότητα δεν είναι από μόνη της αντικειμενικό μέτρο αξιολόγησης.

Ακρίβεια (Precision) για τον A = από το σύνολο των έργων που αποδόθηκε από τον k-NN στον A, τι ποσοστό είναι πράγματι του A = 4/(4+2)
Ανάκληση (Recall) για τον A = από το σύνολο των έργων που είναι πράγματι του A, πόσα βρήκε ο k-NN σωστά ότι είναι του A = 4/(4+1)

Ακρίβεια (Precision) για τον B = από το σύνολο των έργων που αποδόθηκε από τον k-NN στον B, τι ποσοστό είναι πράγματι του B = 3/(3+1)
Ανάκληση (Recall) για τον B = από το σύνολο των έργων που είναι πράγματι του B, πόσα βρήκε ο k-NN σωστά ότι είναι του B = 3/(3+2)

Μέτρο f (f-measure) = 2*Precision*Recall / (Precision + Recall)

Βιβλιογραφία

- Thomas V.N Merriam and Robert A.J. Matthews. Neural computation in stylometry: An application to the works of Shakespeare and Marlowe. Literary and Linguistic Computing, 9(1), 1994.
- Frederick Mosteller and David L. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, Mass., 1964.
- Ephratt, Michal. Authorship attribution - the case of lexical innovations. http://www.cs.queensu.ca/achallc97/papers/p006.html
- Gerritsen, Corey M. Authorship Attribution Using Lexical Attraction. http://genesis.csail.mit.edu/papers/Gerritsen2003.pdf
- Holmes, David I. Stylometry: Its Origins, Development and Aspirations. http://www.cs.queensu.ca/achallc97/papers/s004.html
- Pfleeger, Charles P. and Shari Lawrence Pfleeger. Security in Computing. Pg 342.

Δικτυογραφία

- http://www.scss.tcd.ie/undergraduate/bacsll/bacsll_web/mccombe0102.pdf
- http://www.math.neu.edu/~Malioutov/marlowe7.pdf
- http://www.britannica.com/bps/additionalcontent/18/31637683/Other-Applications-of-Authorship-Attribution
- https://www.aaai.org/Papers/FLAIRS/2006/Flairs06-151.pdf
- http://www.icsd.aegean.gr/lecturers/Stamatatos/papers/survey.pdf