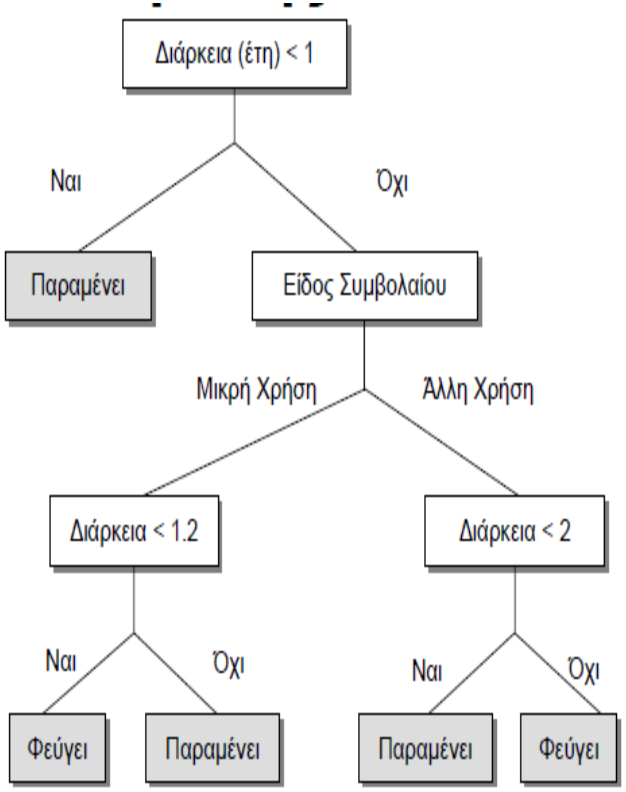
Επεξεργασία Φυσικής Γλώσσας & Μηχανική Μάθηση

Δέντρα απόφασης στην ΕΦΓ

Κάτια Κερμανίδου
kerman@ionio.gr

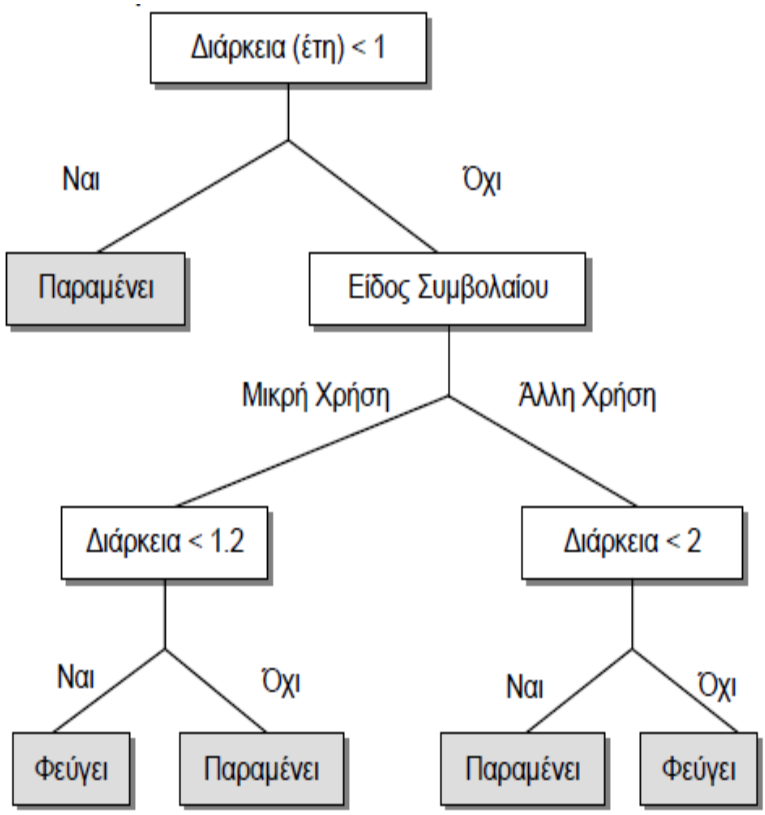
Δέντρα απόφασης

- Το μοντέλο που επάγεται είναι μια αφαιρετική δενδροειδής δομή που αναπαριστά τα δεδομένα.
- Έστω τα δεδομένα εταιρείας κινητής τηλεφωνίας που περιγράφουν περιπτώσεις συνδρομητών που παρέμειναν ή έφυγαν μετά τη λήξη του συμβολαίου τους, με βάση τη διάρκεια και το είδος αυτού.
- Μια αναπαράσταση σε δένδρο θα μπορούσε να έχει την μορφή του σχήματος δεξιά.
  - Κάθε κόμβος ορίζει μια συνθήκη ελέγχου της τιμής κάποιου χαρακτηριστικού των περιπτώσεων
  - Κάθε κλαδί που φεύγει από ένα κόμβο αντιστοιχεί σε μια διαφορετική διακριτή τιμή του χαρακτηριστικού που σχετίζεται με τον κόμβο.
  - Στα κλαδιά φύλλα έχουμε το τι συνέβη.



Αναπαράσταση με Κανόνες Classification Rules

- Εναλλακτικά, το δένδρο μπορεί να αναπαρασταθεί και ως σύνολο κανόνων if-then, που ονομάζονται **κανόνες ταξινόμησης (classification rules).**
- π.χ. για τα δεδομένα της εταιρίας κινητής τηλεφωνίας προκύπτουν 5 κανόνες:
  - 1) **if Διάρκεια<1 then Παραμένει**
  - 2) **if Διάρκεια>1 and Είδος_Συμβολαίου=Μικρή Χρήση and Διάρκεια<1.2 then Φεύγει**
  - 3) **if Διάρκεια>1 and Είδος_Συμβολαίου=Μικρή Χρήση and Διάρκεια>1.2 then Παραμένει**
  - 4) **if Διάρκεια>1 and Είδος_Συμβολαίου=Άλλη_Χρήση and Διάρκεια<2 then Παραμένει**
  - 5) **if Διάρκεια>1 and Είδος_Συμβολαίου=Άλλη Χρήση and Διάρκεια>2 then Φεύγει**



Ο αλγόριθμος ID3

- Είναι ο πιο γνωστός αλγόριθμος μάθησης δένδρων ταξινόμησης.
- Είναι αναδρομικός
  - Βρες το χαρακτηριστικό οι τιμές του οποίου διαχωρίζουν βέλτιστα τα παραδείγματα εκπαίδευσης σε σχέση με την τιμή τους στην κλάση ταξινόμησης και τοποθέτησέ το στη ρίζα του δέντρου
  - Κάνε τον διαχωρισμό των παραδειγμάτων εκπαίδευσης βάσει των τιμών του χαρα/κου
  - Για κάθε τιμή του χαρα/κού δημιούργησε έναν καινούριο κόμβο στο δέντρο. Για κάθε κόμβο επανέλαβε την διαδικασία. Επέλεξε δηλ. το χαρακτηριστικό εκείνο που διαχωρίζει τα παραδείγματα του συγκεκριμένου κόμβου βέλτιστα.
  - Η διαδικασία τερματίζει όταν
    - Όλα τα παραδείγματα που ανήκουν σε αυτόν έχουν ίδια τιμή στην κλάση ταξινόμησης
    - Έχω ξεμείνει από χαρακτηριστικά προς έλεγχο
- Η επιλογή του πιο κατάλληλου χαρα/κου σε κάθε κόμβο πραγματοποιείται με κάποιο στατιστικό μέτρο

Καταλληλότητα χαρακτηριστικού

- Η απάντηση σε μια ερώτηση που έχει *n* δυνατές απαντήσεις προσφέρει πληροφορία ίση με

$$I(P(v_1),\ldots,P(v_n)) = \sum_{i=1}^{n} -P(v_i)\log_2 P(v_i)$$

- Επιλέγω εκείνο το χαρ/κο *A* που μου δίνει το μεγαλύτερο Κέρδος Πληροφορίας (Information Gain):

$$Κέρδος(A) = I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - Υπόλοιπο(A)$$

$$Υπόλοιπο(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I\left(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}\right)$$

όπου v ο αριθμός των τιμών του χαρακτηριστικού
p_i ο αριθμός θετικών παραδειγμάτων που έχουν τιμή i στο χαρακτηριστικό *A*
n_i ο αριθμός αρνητικών παραδειγμάτων που έχουν τιμή i στο χαρακτηριστικό *A*

p: το σύνολο των θετικών παραδειγμάτων στα δεδομένα εκπαίδευσης
n: το σύνολο των αρνητικών παραδειγμάτων στα δεδομένα εκπαίδευσης

Εφαρμογή σε text classification

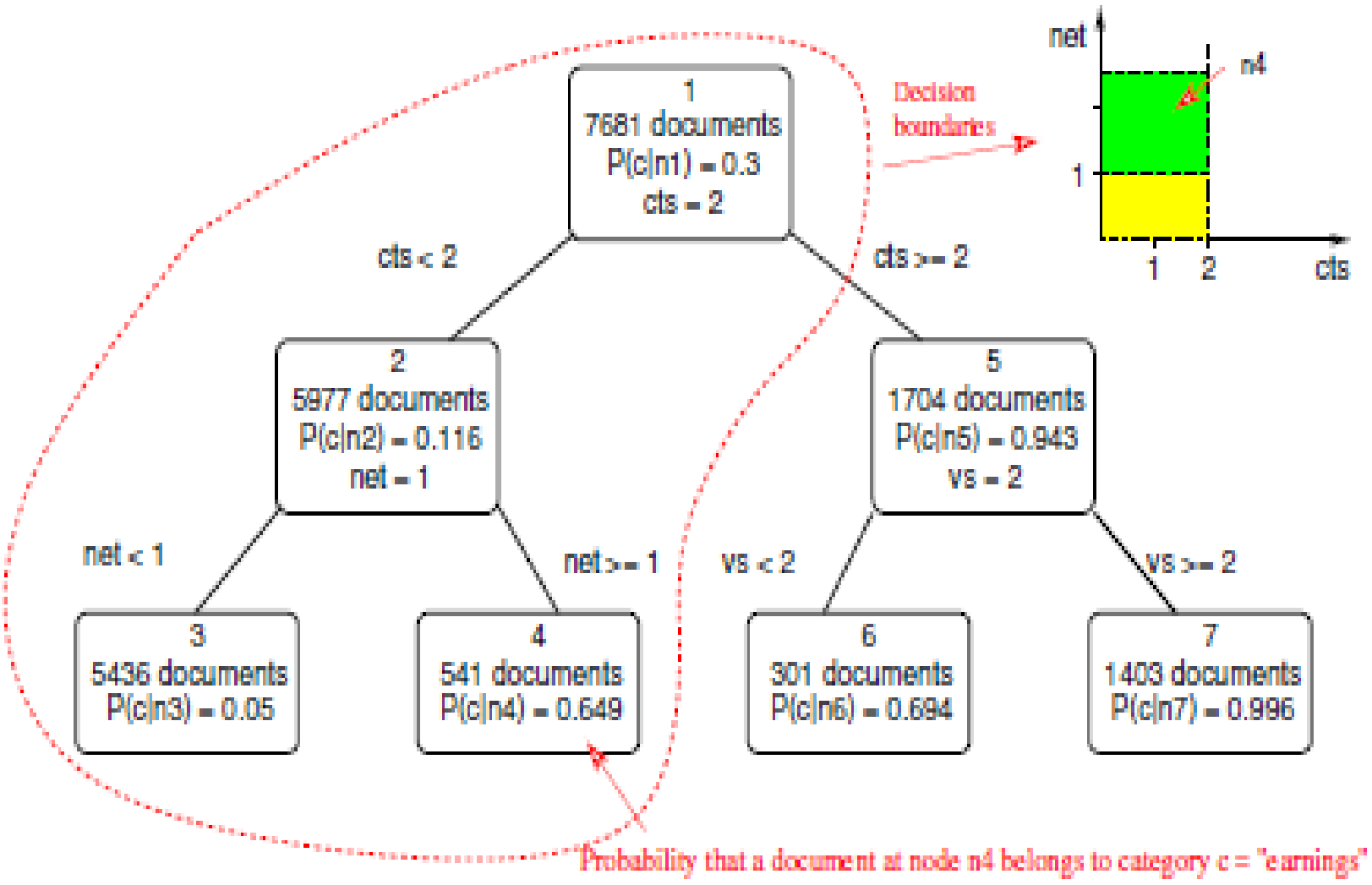
Task: classify REUTERS texts as belonging to category "earnings" (or not).

```
[...]<title> Cobanco Inc year net</title>
<body>Shr 34 cts vs 1.19 dlrs
Net 807,000 vx 2,858,000
Assets 510.2 mln vs 479 mln
Deposits 472 mln vs 440 mln
Loans 299.2 mln vs 327 mln
Note: 4th qtr not available. Year includes 1985
extraordinary gain from tax carry forward of 132,000 dlrs,
or five cts per shr</body>...
```

$\mathcal{T}$ =<vs, mln, cts, ; , &, 000, loss, ', ", 3, profit, dlrs, 1, pct, is, s, that, net, it, at>

$\vec{d_j}$ =< 5, 5, 3, 3, 3, 4, 0, 0, 0, 4, 0, 3, 2, 0, 0, 0, 0, 3, 2, 0>

Το δέντρο


1
7681 documents
P(c|n1) = 0.3
cts = 2
Decision boundaries
net
n4
1
1
2
cts
cts < 2
cts >= 2
2
5977 documents
P(c|n2) = 0.116
net = 1
5
1704 documents
P(c|n5) = 0.943
vs = 2
net < 1
net >= 1
vs < 2
vs >= 2
3
5436 documents
P(c|n3) = 0.05
4
541 documents
P(c|n4) = 0.649
6
301 documents
P(c|n6) = 0.694
7
1403 documents
P(c|n7) = 0.996
Probability that a document at node n4 belongs to category c = "earnings"

Υπολογίζοντας τις πιθανότητες στους κόμβους

- One can assign probabilities to a leaf node (i.e. the probability that a new document d belonging to that node should be filled under category c) as follows (using add-one smoothing):

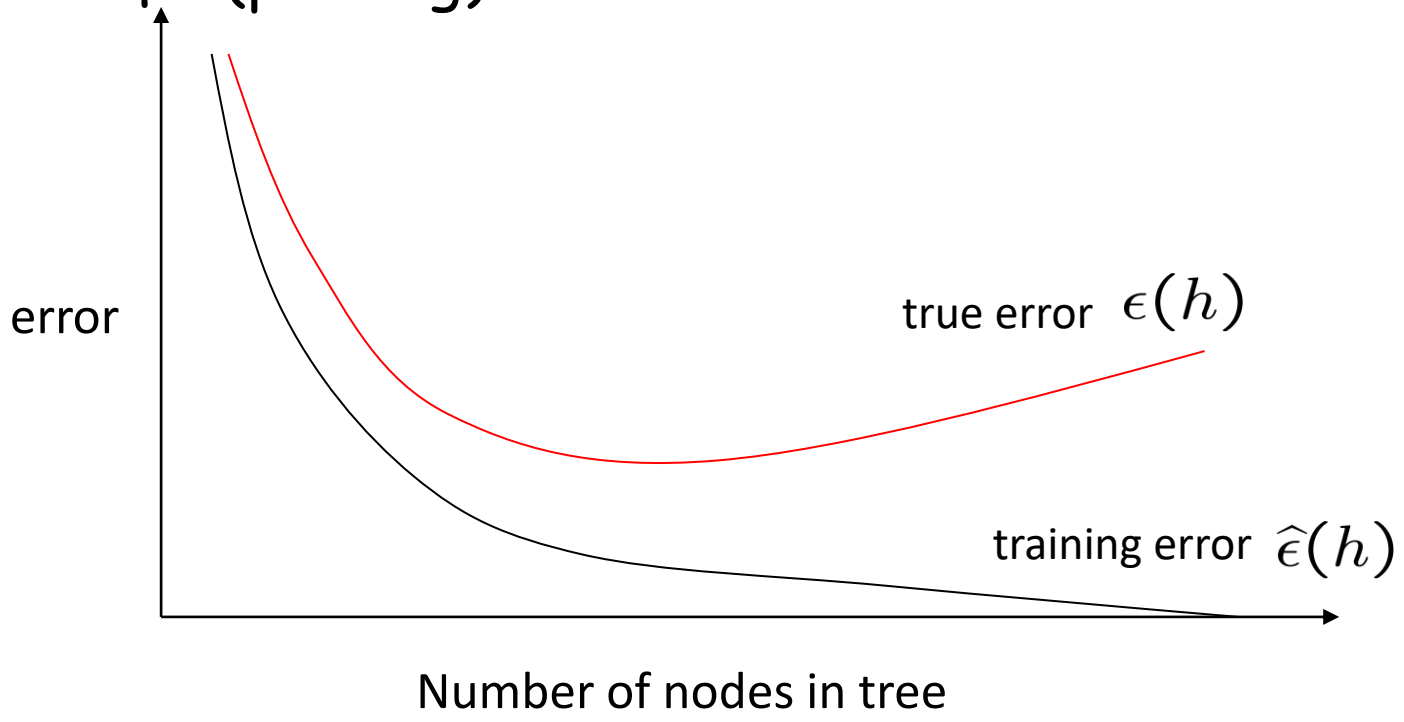
$$P(c|d_n) = \frac{|D_{cn}| + 1}{|D_{cn}| + |D_{\bar{c}n}| + 1 + 1} \tag{5}$$

where

- $P(c|d_n)$ is the probability that a document d_n which ended up in node n belongs to category c,
- $|D_{cn}|$ ($|D_{\bar{c}n}|$) number of (training) documents in node n which have been assigned category c ($\bar{c}$)

Το πρόβλημα της υπερπροσαρμογής (Overfitting)

- Το επαγόμενο δένδρο έχει την τάση να αποδίδει καλύτερα στην ταξινόμηση των παραδειγμάτων εκπαίδευσης, παρά των παραδειγμάτων ελέγχου.
- Χαμηλή ικανότητα γενίκευσης
- Κλάδεμα (pruning)



Μετα-κλάδεμα (Post-pruning)
Reduced Error Pruning

[1] Χωρίζω τα δεδομένα εκπαίδευσης (S_{full}) σε δυο τμήματα:
Ένα μικρότερο σετ εκπαίδευσης S
Ένα σετ επικύρωσης (validation set) V

[2] Κατασκευάζω ένα δέντρο απόφασης T, χρησιμοποιώντας το S.

[3] Κλαδεύω χρησιμοποιώντας το V

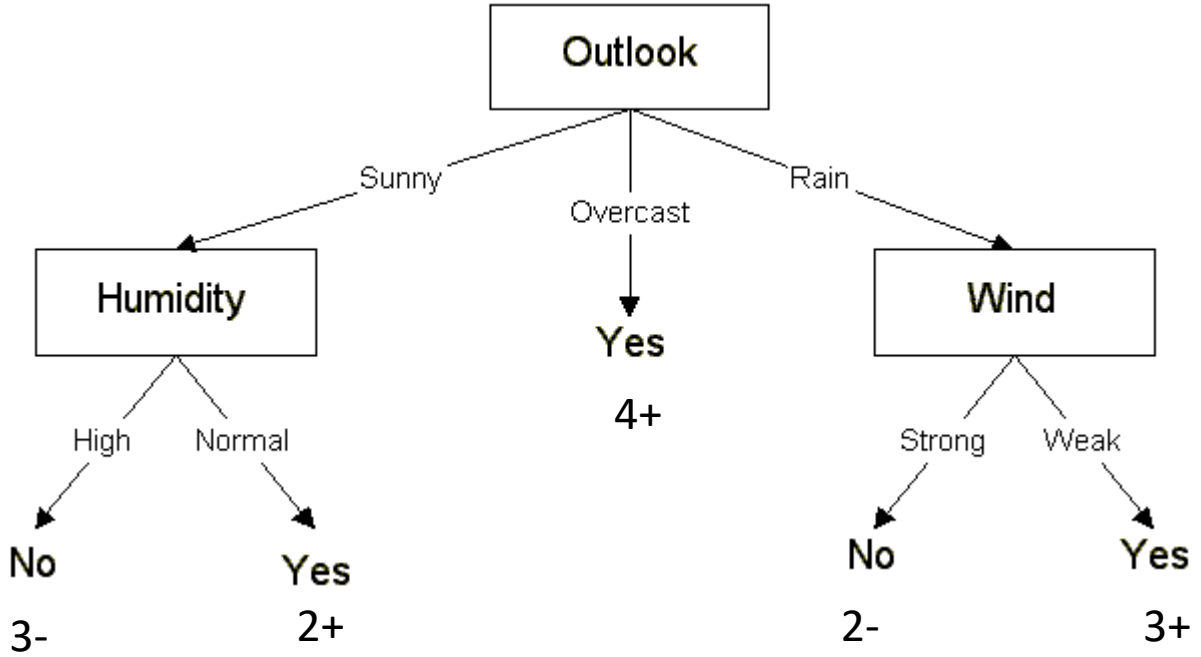
επανάληψη

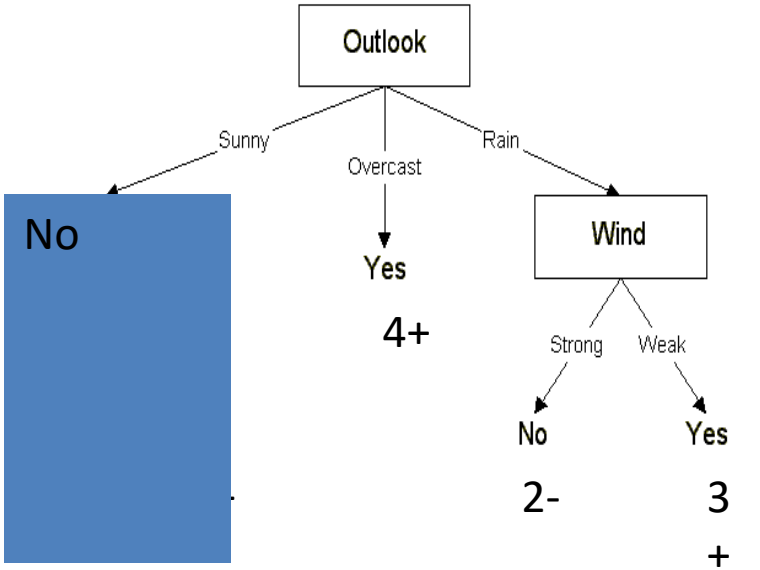
Για κάθε κόμβο u στο T μετράω στο υπο-δέντρο που τον έχει ρίζα πόσα παραδείγματα στα φύλλα του ταξινομούνται σε κάθε τιμή της κλάσης.

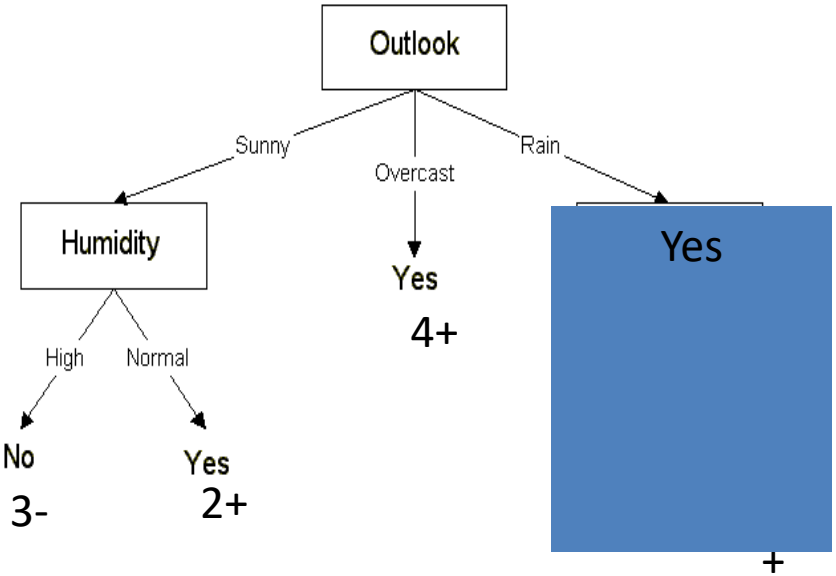
Υπολογίζω την κλάση που έχει πλειοψηφική τιμή (c) στο υπο-δέντρο.

Αφαιρώ το υπο-δέντρο, και στη θέση του τοποθετώ ένα φύλλο με τιμή c στην κλάση ταξινόμησης.

Αν το μικρότερο δέντρο T’ μειώνει το σφάλμα στο V, σε σχέση με το T, τότε T = T’


Outlook
Sunny
Overcast
Rain
Humidity
Yes
4+
Wind
High
Normal
Strong
Weak
No
Yes
No
Yes
3-
2+
2-
3+


Outlook
Sunny
Overcast
Rain
No
Yes
4+
Wind
Strong
Weak
No
Yes
2-
3
+


Outlook
Sunny
Overcast
Rain
Humidity
Yes
4+
Yes
High
Normal
No
Yes
3-
2+
+

ΔΑ στην Αποσαφήνιση έννοιας λέξεων
Word Sense Disambiguation

```
RIV y be? Then he ran down along the bank, toward a narrow, muddy path.
FIN  four bundles of small notes the bank cashier got it into his head
RIV ross the bridge and on the other bank you only hear the stream, the
RIV beneath the house, where a steep bank of earth is compacted between
FIN op but is really the branch of a bank. As I set foot inside, despite
FIN raffic police also belong to the bank. More foolhardy than entering
FIN require a number.  If you open a bank account, the teller identifies
RIV  circular movement, skirting the bank of the River Jordan, then turn
```

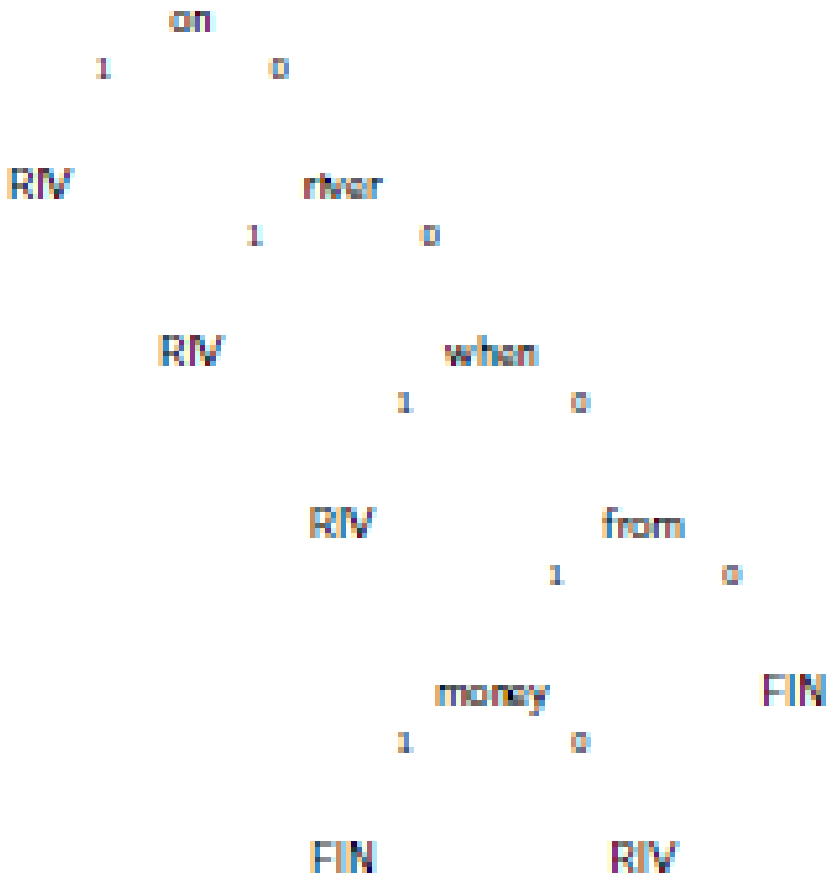
- Στόχος ενός συστήματος WSD είναι να μάθει να διακρίνει την έννοια της λέξης 'bank' σαν τράπεζα ή σαν όχθη ποταμού

ΔΑ στην Αποσαφήνιση έννοιας λέξεων Word Sense Disambiguation

- Είναι εργασία επιβλεπόμενης μάθησης όπου
  - Οι έξοδοι είναι οι έννοιες των λέξεων
  - Τα χαρακτηριστικά εισόδου είναι το συμφραστικό περιβάλλον της λέξης (λέξεις που προηγούνται και έπονται της αμφίσημης λέξης)
- E.g.: For $T = \{along, cashier, stream, muddy, \ldots\}$, we could have:
  $d_1 = (along = 1, cashier = 0, stream = 0, muddy = 1, \ldots)$ and
- $f(d_1) = \texttt{RIV}$

Το δέντρο

Using the algorithm above (slide 9) we get this decision tree



Trained on small training set and $\mathcal{T} = \{$small, money, on, to, river, from, in, his, accounts, when, by, other, estuary, some, with$\}$

Random Forests

L. Breiman. Random forests. Machine Learning, 45(1):5-32, 2001.

- Επιλέγω τον αριθμό K των δέντρων που θέλω να σχηματίσω
- Επιλέγω τον αριθμό m των χαρακτηριστικών που θα χρησιμοποιηθούν για τον σχηματισμό των δέντρων
- Εκπαίδευση
  - Για κάθε δέντρο
    - κάνω δειγματοληψία με επανατοποθέτηση N παραδειγμάτων εκπαίδευσης (όπου N ο συνολικός αριθμός παραδειγμάτων εκπαίδευσης)
    - Σε κάθε κόμβο χρησιμοποιώ m τυχαία επιλεγμένα χαρακτηριστικά για να πραγματοποιήσω την βέλτιστη διακλάδωση
    - Δεν κλαδεύω το δέντρο που προκύπτει
- Αξιολόγηση
  - Τρέχω κάθε ένα από τα K δέντρα στο παράδειγμα αξιολόγησης
  - Εφαρμόζω πλειοψηφική ψήφο

An Evaluation of Authorship Attribution Using Random Forests, Mahmoud Khonji ; Youssef Iraqi ; Andrew Jones, 2015

Dataset	Authors num.	Docs num.	Words avg.	Chars avg.
Problem A (learn)	3	6	3266.00	22887.17
Problem A (test)	3	3	3972.17	21799.00
Problem C (learn)	8	16	6013.81	33790.69
Problem C (test)	8	8	4057.63	22766.50
Problem I (learn)	14	28	51649.35	489148.32
Problem I (test)	14	14	86052.21	500788.93

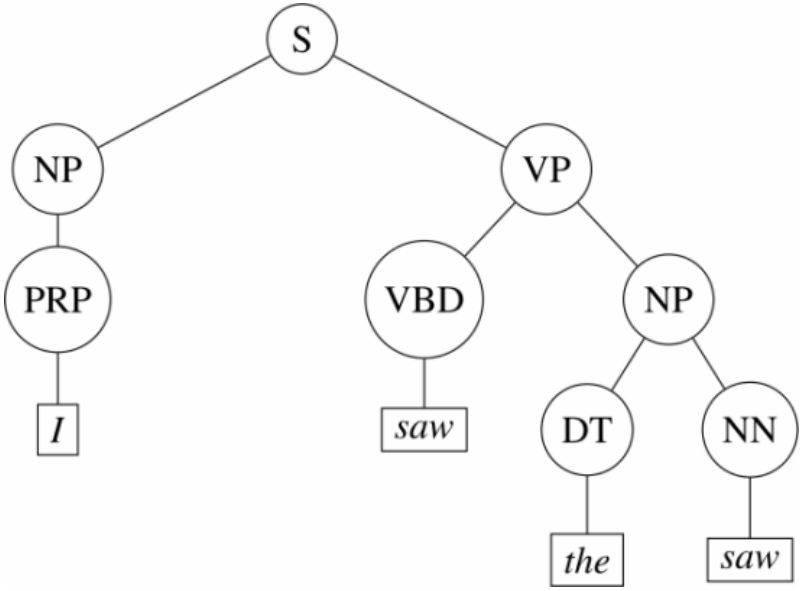
- Authorship Attribution: Η αναγνώριση συγγραφέα βάσει των γλωσσολογικών χαρακτηριστικών του κειμένου του
- Στυλομετρία: Κάθε συγγραφέας έχει εγγενή μοναδικό τρόπο να χρησιμοποιεί την γλώσσα

An Evaluation of Authorship Attribution Using Random Forests, Mahmoud Khonji ; Youssef Iraqi ; Andrew Jones, 2015

- Χαρακτηριστικά:
  - Η συχνότητα εμφάνισης για όλους τους 256 ASCII χαρ/ρες
  - Η συχνότητα εμφάνισης 361 λειτουργικών λέξεων
  - Η συχνότητα εμφάνισης των 2000 πιο συχνών n-γράμμων στο σετ εκπαίδευσης
  - Η συχνότητα εμφάνισης των 2000 πιο συχνών συντακτικών κανόνων που εμφανίζονται στο σετ εκπαίδευσης
  - Συχνότητα εμφάνισης μεγέθους λέξεων, για μεγέθη από 1 έως 50 χαρακτήρες
  - Συχνότητα εμφάνισης για 106 σχήματα λέξεων που εμφανίζονται στο σετ εκπαίδευσης
    - Πχ σχήμα λέξης: λέξη που ξεκινά με κεφαλαίο και ακολουθείται από 3 πεζά γράμματα

Συντακτικοί Κανόνες

- Οι κανόνες ελεύθερης σύνταξης που αναγνωρίζονται στο δίπλα συντακτικό δέντρο είναι
  - S -> NP VP
  - NP -> PRP
  - PRP -> I
  - VP -> VBD NP
  - VBD -> saw
  - NP -> DT NN
  - DT -> the
  - NN -> saw



An Evaluation of Authorship Attribution Using Random Forests, Mahmoud Khonji ; Youssef Iraqi ; Andrew Jones, 2015

- Χαρακτηριστικά (συνέχεια):
  - Ο συνολικός αριθμός χαρακτήρων του έργου
  - Ο συνολικός αριθμός λέξεων
  - Ο αριθμός των λέξεων που εμφανίζονται μια φορά (hapax legomena)
  - Το μέτρο K του Yule για τον πλούτο λεξιλογίου,

    $$K = C\left[-\frac{1}{N} + \sum_{m=1}^{m_{max}} V(m,N)\left(\frac{m}{N}\right)^2\right]$$

  - Όπου
    - N ο συνολικός αριθμός λέξεων στο κείμενο
    - V(m,N) ο αριθμός λέξεων που εμφανίζονται m φορές στο κείμενο
    - mmax η μέγιστη συχνότητα λέξης στο κείμενο
    - $C=10^4$

An Evaluation of Authorship Attribution Using Random Forests, Mahmoud Khonji ; Youssef Iraqi ; Andrew Jones, 2015

- Αποτελέσματα

Dataset	Accuracy
A	100%
C	87,5%
I	92,9%