Επεξεργασία Φυσικής Γλώσσας & Μηχανική Μάθηση

Εισαγωγή

Κάτια Κερμανίδου
kerman@ionio.gr

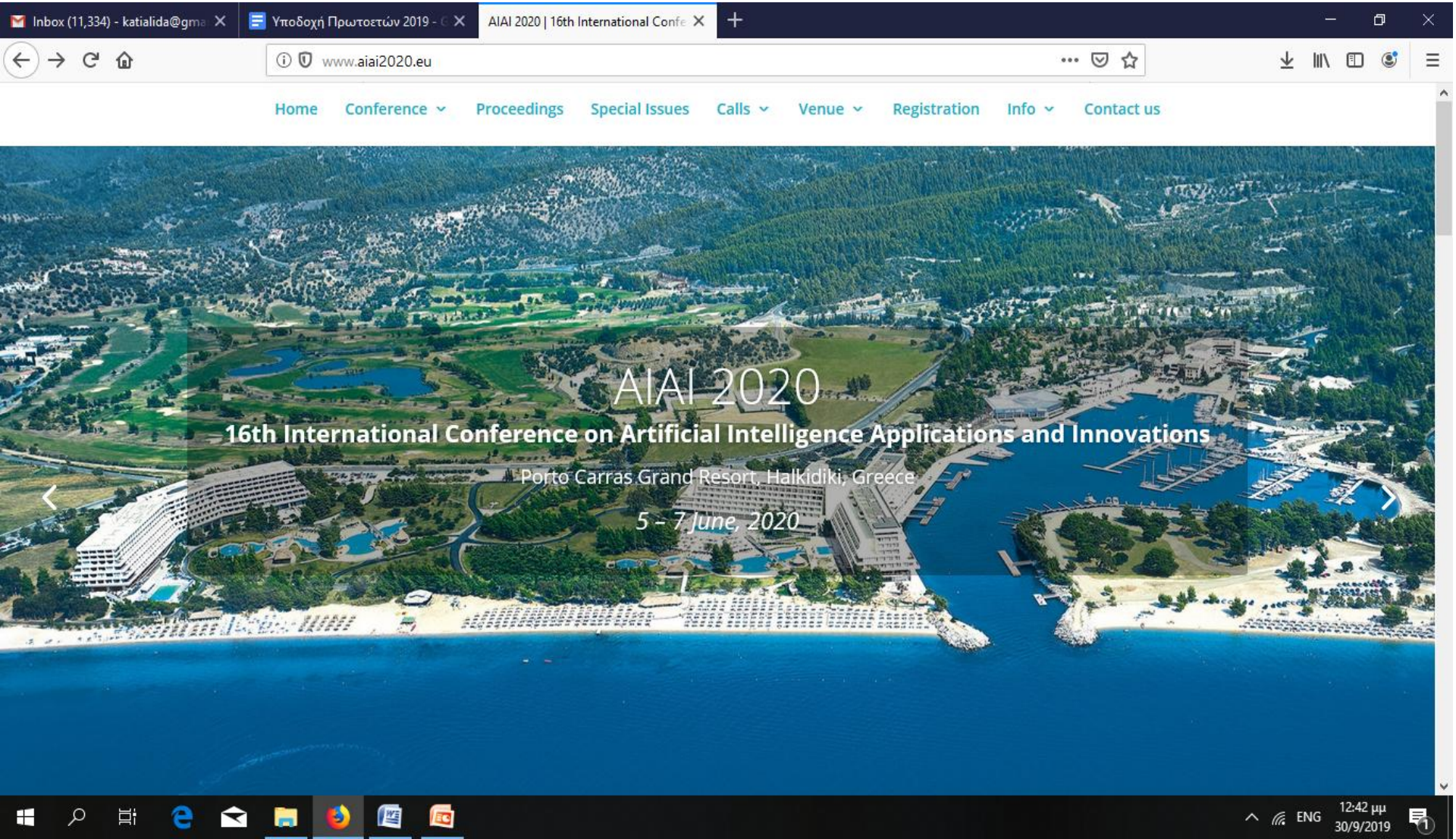
Χρονοδιάγραμμα μαθήματος

Α/Α	Εβδομάδα	Περιεχόμενο μαθήματος	Διδάσκων
1	23/10/19	Εισαγωγή	Κ. Κερμανίδου
2	30/10/19	Παρουσιάσεις φοιτητών – Βήμα Ι Πιθανοτικά Μοντέλα στην ΕΦΓ	Κ. Κερμανίδου
3	06/11/19	Reading Week	-
4	13/11/19	Παρουσιάσεις φοιτητών – Βήμα ΙΙ Δέντρα απόφασης στην ΕΦΓ	Κ. Κερμανίδου
5	20/11/19	Μάθηση Βασισμένη στην Μνήμη στην ΕΦΓ Αναπαραστάσεις Λέξεων	Κ. Κερμανίδου
6	27/11/19	Μηχανική Μετάφραση	Κ. Κερμανίδου
7	04/12/19	RapidMiner και Εξόρυξη Γλωσσικής Πληροφορίας	Μ. Μαραγκουδάκης
8	11/12/19	Reading Week	-
9	18/12/19	Βαθιά Μάθηση στην ΕΦΓ	Κ. Κερμανίδου
10	08/01/20	Παρουσιάσεις φοιτητών – Βήμα ΙΙΙ Εργαλεία Python για Βαθιά Μάθηση στην ΕΦΓ Ι	Κ. Κερμανίδου Δ. Μουρατίδου
11	15/01/20	Εργαλεία Python για Βαθιά Μάθηση στην ΕΦΓ ΙΙ	Δ. Μουρατίδου

Διαδικαστικά

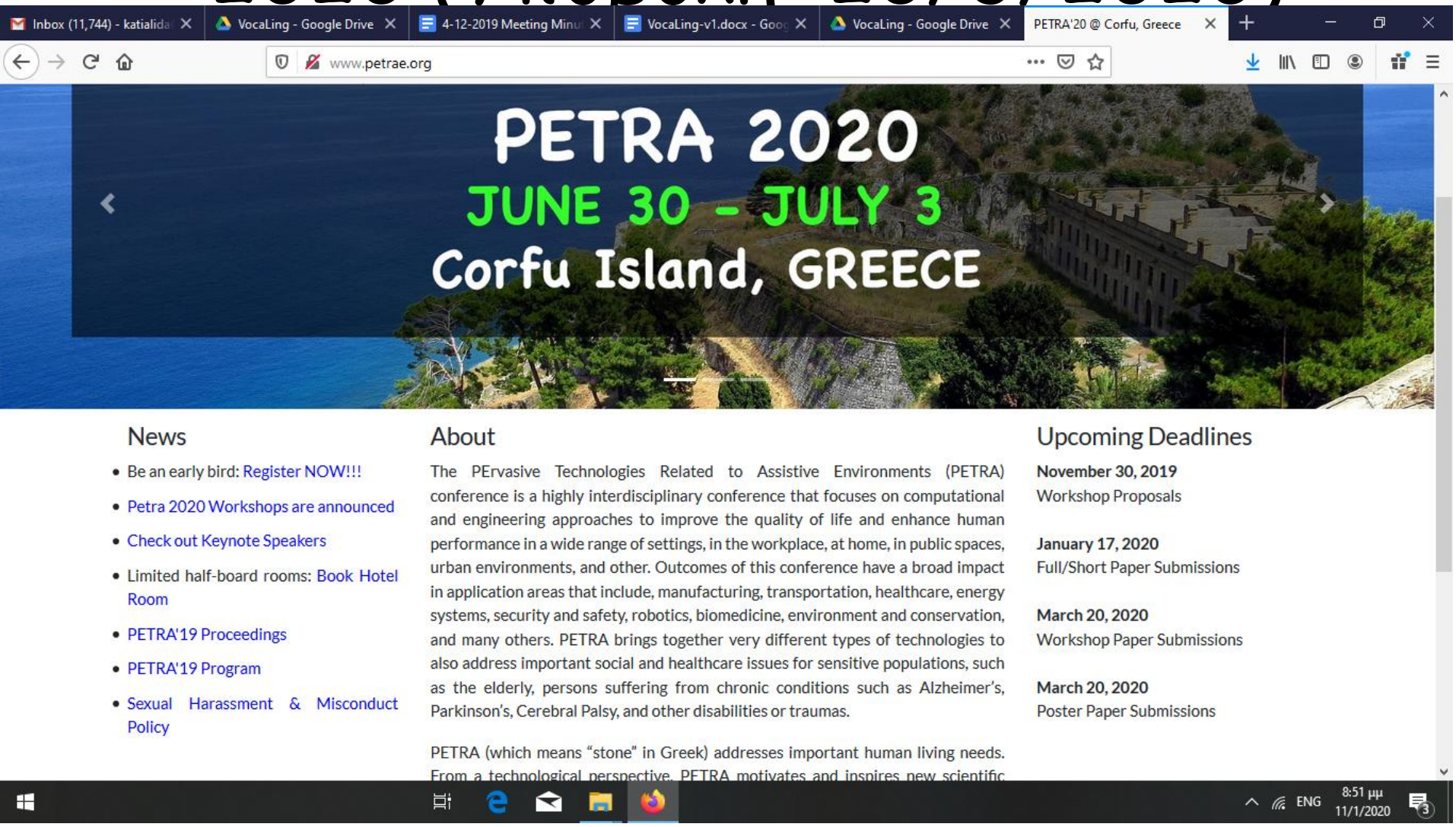
- Ομαδική εργασία
  - Ομάδες 3-4 ατόμων
  - Ο βαθμός της εργασίας πιάνει το 60% του τελικού βαθμού
  - Η καλύτερη εργασία πάει για δημοσίευση σε διεθνές συνέδριο (ακολουθούν προτάσεις για συνέδρια)
- Γραπτές εξετάσεις
  - Ο βαθμός του γραπτού πιάνει το 40% του τελικού βαθμού.

1η Πρόταση για δημοσίευση: AIAI 2020 (Υποβολή: 29/2/2020)

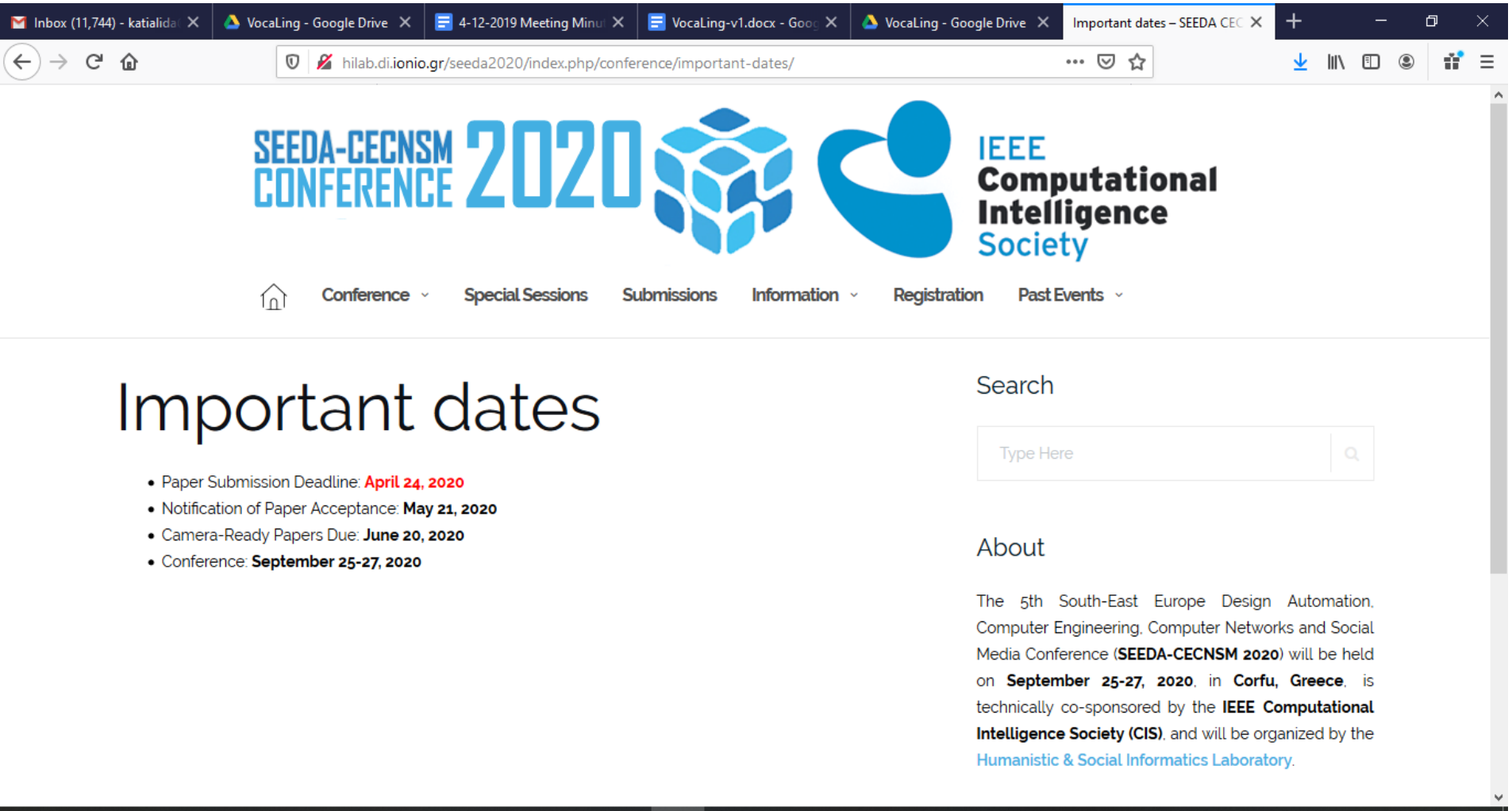


2η Πρόταση για δημοσίευση: SMAP 2020 (Υποβολή: 31/3/2020)

3η Πρόταση για δημοσίευση: PETRA 2020 (Υποβολή: 20/3/2020)



4η Πρόταση για δημοσίευση: SEEDA 2020 (Υποβολή: 24/4/2020)



Υλικό (καλύψτε τυχόν κενά!)

- Το μάθημα προϋποθέτει
  - βασική εξοικείωση με έννοιες, και τεχνικές Επεξεργασίας Φυσικής Γλώσσας
    - https://e-class.ionio.gr/courses/DCS153/
    - https://www.cs.vassar.edu/~cs366/docs/Manning_Schuetze_StatisticalNLP.pdf
    - https://mitpress.mit.edu/books/introduction-natural-language-processing
  - βασική εξοικείωση με έννοιες και αλγορίθμους Μηχανικής Μάθησης
    - http://ciml.info/
    - http://profsite.um.ac.ir/~monsefi/machine-learning/pdf/Machine-Learning-Tom-Mitchell.pdf
    - ftp://ftp.ingv.it/pub/manuela.sbarra/Data%20Mining%20Practical%20Machine%20Learning%20Tools%20and%20Techniques%20-%20WEKA.pdf
    - http://alex.smola.org/drafts/thebook.pdf?utm_source=twitterfeed&utm_medium=twitter
    - https://www.amazon.com/Introduction-Deep-Learning-MIT-Press/dp/0262039516

Περισσότερο υλικό

- https://www.researchgate.net/publication/228686410_Machine_learning_for_natural_language_processing/link/0912f5126527f24567000000/download
- http://www.ai.mit.edu/courses/6.891-nlp/
- https://www.scss.tcd.ie/kevin.koidl/cs4062/
- http://u.cs.biu.ac.il/~89-680/

Μηχανική Μάθηση (Machine Learning)

- **Η δημιουργία μοντέλων ή προτύπων από ένα σύνολο δεδομένων, από ένα υπολογιστικό σύστημα, ονομάζεται μηχανική μάθηση.**
- Simon (1983)
  - "η μάθηση σηματοδοτεί προσαρμοστικές αλλαγές σε ένα σύστημα με την έννοια ότι αυτές του επιτρέπουν να κάνει την ίδια εργασία, ή εργασίες της ίδιας κατηγορίας, πιο αποδοτικά και αποτελεσματικά την επόμενη φορά".
- Carbonell (1987)
  - "... η μελέτη υπολογιστικών μεθόδων για την απόκτηση νέας γνώσης, νέων δεξιοτήτων και νέων τρόπων οργάνωσης της υπάρχουσας γνώσης".
- Mitchell (1997)
  - "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".
    - Task T: playing chess
    - Performance measure P: percent of games won against opponents
    - Training Experience E: playing practice games against itself
- Witten & Frank (2000),
  - Κάτι μαθαίνει όταν αλλάζει τη συμπεριφορά του κατά τέτοιο τρόπο ώστε να αποδίδει καλύτερα στο μέλλον"

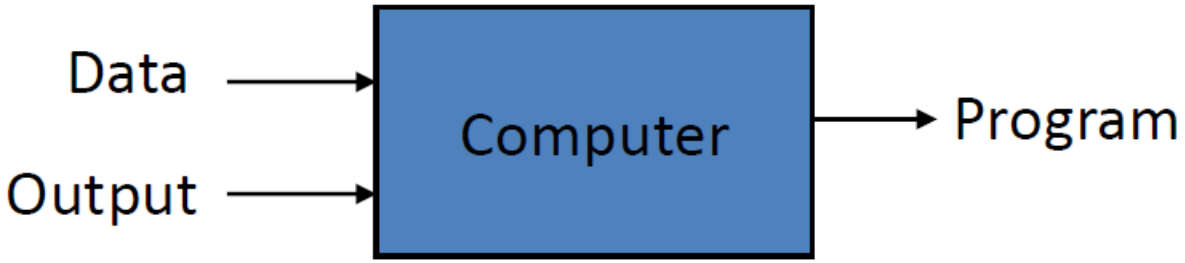
Γιατί Μηχανική Μάθηση;

- Αλγόριθμοι για την εξαγωγή δομικών περιγραφών (προτύπων) από δεδομένα
- Οι περιγραφές αυτές
  - Μπορούν να χρησιμοποιηθούν για την πρόβλεψη (prediction) ενός αποτελέσματος σε καινούρια (άγνωστα) δεδομένα
  - μπορούν να χρησιμοποιηθούν για την ερμηνεία του τρόπου με τον οποίο επιλέγεται η πρόβλεψη
- Αντί να απαιτείται η συγγραφή κώδικα για να αναγνωριστούν αυτά τα πρότυπα χειρωνακτικά, οι αλγόριθμοι μάθησης τρέχουν πάνω στα δεδομένα και εξάγουν τα πρότυπα αυτά αυτόματα.

Παραδοσιακός Προγραμματισμός - Μηχανική Μάθηση

Traditional Programming

Machine Learning



Η ΜΜ σήμερα

- Μεγάλη εξέλιξη σε αλγορίθμους
- Διάθεση μεγάλου όγκου δεδομένων
- Μεγάλη διαθέσιμη υπολογιστική ισχύς
- Πληθώρα εφαρμογών

Είδη Μηχανικής Μάθησης

- Επιβλεπόμενη Μάθηση
  - Επαγωγή (induction)
- Μη επιβλεπόμενη Μάθηση

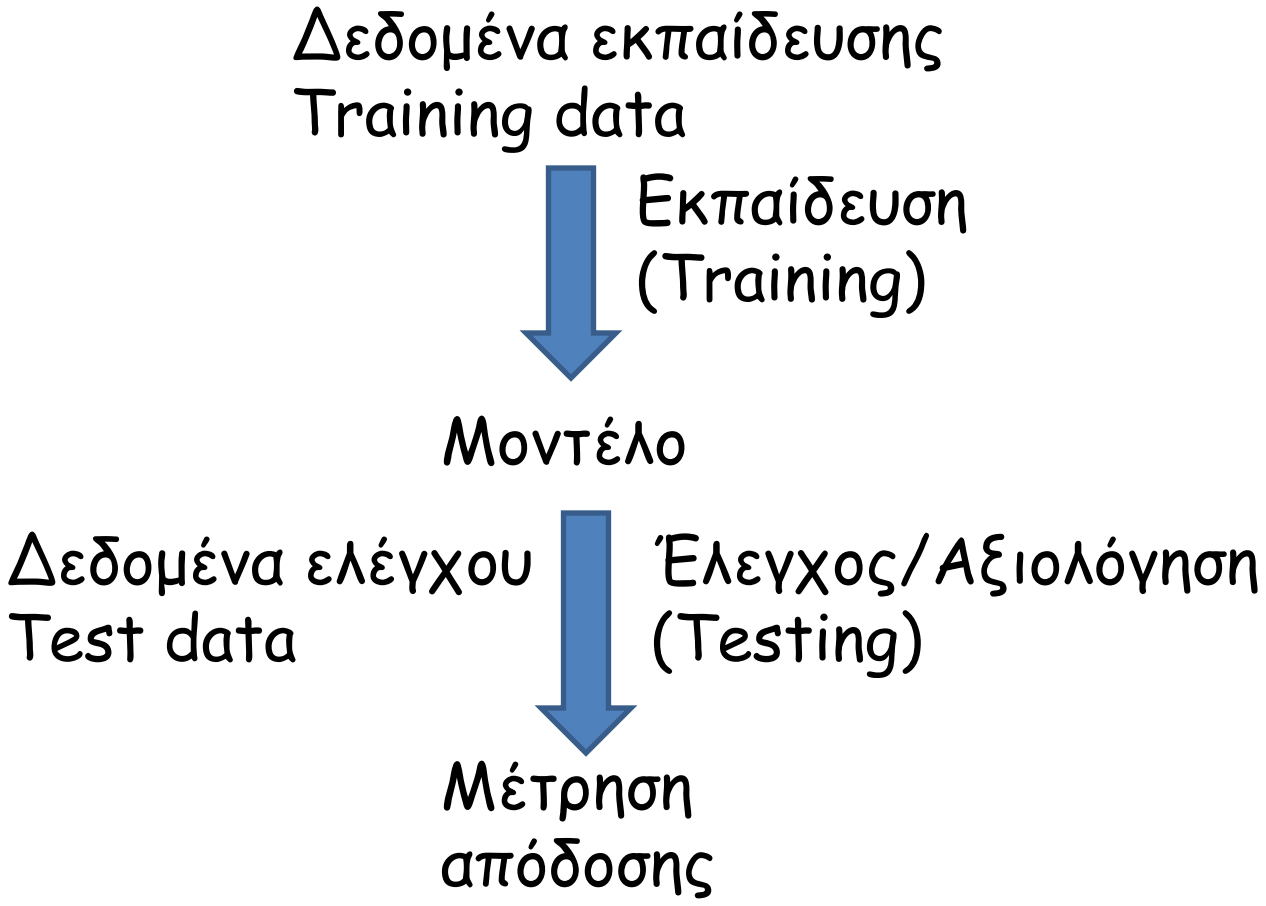
Επιβλεπόμενη Μάθηση: Επαγωγή (Induction)

- Έννοια (concept) είναι αυτό που καλείται ο αλγόριθμος μάθησης να μάθει
- Επαγωγή είναι η διαδικασία δημιουργίας ενός γενικευμένου μοντέλου περιγραφής ή ορισμού μιας έννοιας από ένα σύνολο ειδικών παραδειγμάτων της έννοιας
- Προτάθηκε από τον Αριστοτέλη σαν την αντίστροφη διαδικασία από τον συλλογισμό
- Το παράδειγμα του Αρειανού
- Πιο μαθηματικοποιημένα
  - Ένα **παράδειγμα-example (στιγμιότυπο-instance ή παρατήρηση-observation ή δεδομένο-data)** είναι μια δυάδα $(x,f(x))$, όπου f είναι μια συνάρτηση, το x είναι η είσοδος, και το $f(x)$ η έξοδος της συνάρτησης όταν εφαρμόζεται στο x
  - Επαγωγή: Δοθέντος ενός συνόλου παραδειγμάτων μιας συνάρτησης f, βρες μια συνάρτηση h που προσεγγίζει την f
  - Η συνάρτηση h ονομάζεται **υπόθεση και αποτελεί το γενικευμένο μοντέλο που περιγράφει τα παραδείγματα.**

Επιβλεπόμενη Μάθηση: Έννοιες & Είδη

- Το σύστημα μάθησης καλείται να μάθει επαγωγικά την συνάρτηση στόχο (target function) που περιγράφει τα δεδομένα
- Η τιμή της συνάρτησης-στόχου ονομάζεται και εξαρτημένη μεταβλητή ενώ οι υπόλοιπες μεταβλητές που αναπαριστούν το παράδειγμα ονομάζονται ανεξάρτητες μεταβλητές.
- Κατά την εκπαίδευση η τιμή της συνάρτησης στόχου των παραδειγμάτων εκπαίδευσης είναι γνωστή, και καθοδηγεί τη διαδικασία μάθησης
- Κατά τον έλεγχο, η απόδοση μετράται σε καινούρια άγνωστα παραδείγματα (ελέγχου), για τα οποία ο αλγόριθμος δεν γνωρίζει την τιμή της συνάρτησης στόχου
- Δύο μορφές επιβλεπόμενης μάθησης
  - Ταξινόμηση (ή Κατηγοριοποίηση - Classification)
    - Η συνάρτηση-στόχος παίρνει διακριτές τιμές
  - Παλινδρόμηση (ή Παρεμβολή - Regression)
    - Η συνάρτηση-στόχος παίρνει αριθμητικές τιμές

Επιβλεπόμενη Μάθηση: Φάσεις



Επιβλεπόμενη Μάθηση: Εκπαίδευση, Επικύρωση και Αξιολόγηση

- Ένας αλγόριθμος ταξινόμησης (classifier) επάγει ένα πρώτο μοντέλο από τα δεδομένα εκπαίδευσης
- Οι παράμετροι του αλγορίθμου συντονίζονται με επαναληπτικά πειράματα πάνω σε ένα σώμα επικύρωσης (validation set)
- Η απόδοση του αλγορίθμου αξιολογείται με εφαρμογή του πάνω στο σετ αξιολόγησης (test set)
- **Τα σετ εκπαίδευσης και αξιολόγησης πρέπει να είναι τελείως διαφορετικά**

Επιβλεπόμενη Μάθηση: Μέτρα Αξιολόγησης

- Πίνακας σύγχυσης (confusion matrix)

		Predicted class	
		Yes	No
Actual class	Yes	True positive	False negative
	No	False positive	True negative

- True Positive (TP): Ένα παράδειγμα ελέγχου ταξινομείται σωστά ως θετικό
- True Negative (TN): Ένα παράδειγμα ελέγχου ταξινομείται σωστά ως αρνητικό
- False Positive (FP): Ένα παράδειγμα ελέγχου ταξινομείται λάθος ως θετικό
- False Negative (FN): Ένα παράδειγμα ελέγχου ταξινομείται λάθος ως αρνητικό

Επιβλεπόμενη Μάθηση: Μέτρα αξιολόγησης

- Ορθότητα (Accuracy)
  - Accuracy = (TP+TN) / (TP+TN+FP+FN)
- Ακρίβεια (Precision)
  - Precision = TP / (TP+FP)
- Ανάκληση (Recall)
  - Recall = TP / (TP+FN)
- Μέτρο f (f-measure)
  - $F=(1+\beta^2) \times (Precision \times Recall) / (\beta^2 \times Precision + Recall)$
  - Το β είναι θετικός πραγματικός αριθμός
  - Στην ανάκληση αποδίδεται β φορές περισσότερη σημασία από ότι στην ακρίβεια)
  - Αν β=1, τότε $F= 2 \times Precision \times Recall / (Precision + Recall)$

Επιβλεπόμενη Μάθηση: Αξιολόγηση σε περισσότερες από δυο τιμές εξόδου

System	A	B	C	D	
	32	1	1	6	A
	3	25	10	2	B
	0	5	30	5	C
	4	10	6	20	D
					Expert

System	A	B	C	D	
	32	1	1	6	A
	3	25	10	2	B
	0	5	30	5	C
	4	10	6	20	D
					Expert

E.g: for class A → TP FP FN TN

Επιβλεπόμενη Μάθηση: Αξιολόγηση με μέσες τιμές

- Micro-averaging

$$\pi^{\mu} = \frac{\sum_{i=1}^{|\mathcal{C}|} TP_i}{\sum_{i=1}^{|\mathcal{C}|} (TP_i + FP_i)}$$

$$\rho^{\mu} = \frac{\sum_{i=1}^{|\mathcal{C}|} TP_i}{\sum_{i=1}^{|\mathcal{C}|} (TP_i + FN_i)}$$

- Macro-averaging

$$\pi^{M} = \frac{\sum_{i=1}^{|\mathcal{C}|} \pi_i}{|\mathcal{C}|} \qquad \rho^{M} = \frac{\sum_{i=1}^{|\mathcal{C}|} \rho_i}{|\mathcal{C}|}$$

- όπου τα π_i και ρ_i είναι τα σκορ ανα τιμή εξόδου

Επιβλεπόμενη Μάθηση: Αξιολόγηση με μέσες τιμές

Category	Sport		Politics		World	
Judgment: Expert and System	E	S	E	S	E	S
"Brazil beat Venezuela"	T	F	F	F	F	T
"US defeated Afghanistan"	F	T	T	T	T	F
"Elections in Wicklow"	F	F	T	T	F	F
"Elections in Peru"	F	F	F	T	T	T
Precision (local):	$\pi = 0$		$\pi = 0.67$		$\pi = 0.5$	
Recall (local):	$\rho = 0$		$\rho = 1$		$\rho = 0.5$	
$\pi^{\mu} =$	$\frac{0+2+1}{0+2+1+1+1+1} = .5$					
$\pi^{M} =$	$\frac{0+.67+.5}{3} = .39$					

Μέτρα απεικόνισης – ROC curve

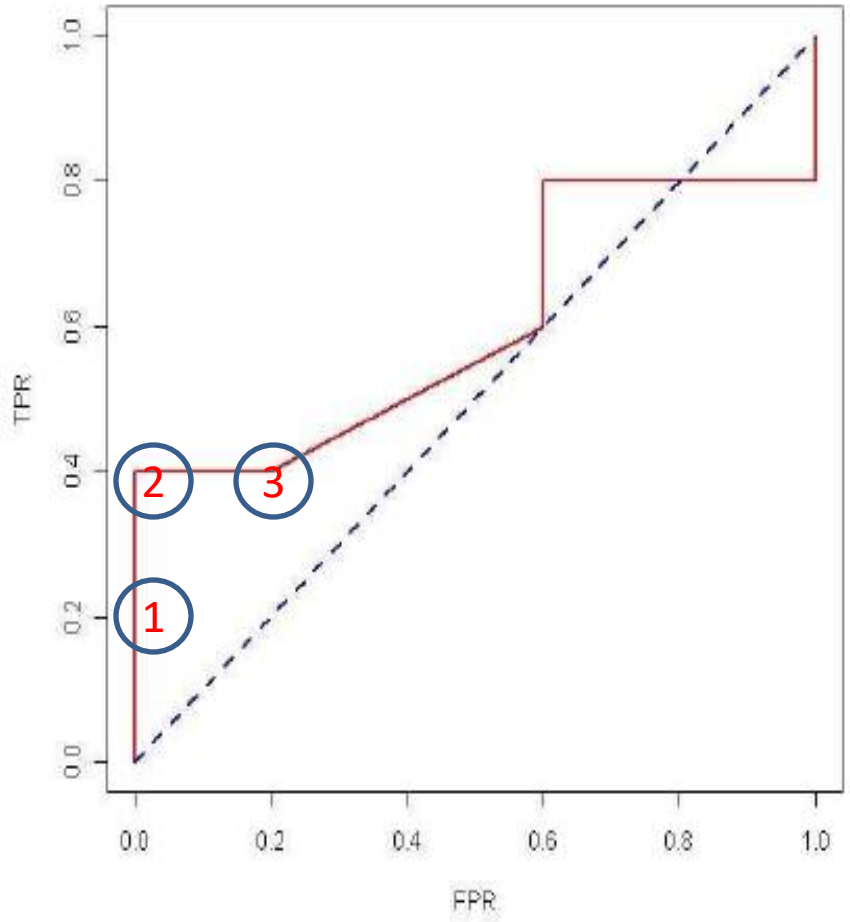
- Εφαρμόζω έναν αλγόριθμο ταξινόμησης που να υπολογίζει τη εκ των υστέρων πιθανότητα (posterior probability) κάθε παραδείγματος ελέγχου να είναι θετικό (P(+|παράδειγμα)), όπως ο Naïve Bayes. Ταξινομώ αυτές τις πιθανότητες από την μεγαλύτερη στην μικρότερη.

Instance	P(+|A)	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

- Θέτω κατώφλι σε κάθε μοναδική τιμή του P(+|παράδειγμα).
- Υπολογίζω TP, FP, TN, FN σε κάθε κατώφλι.
- Αξονας x: FP rate (FPR =FP/(FP+TN))
- Αξονας y: TP rate (TPR =TP/(TP+FN))

ROC curve – Παράδειγμα

Instance	P(+)	True class	FPR	TPR
1	0.95	+	0	1/5
2	0.93	+	0	2/5
3	0.87	-	1/5	2/5
4	0.85	-		
5	0.85	-		
6	0.85	+	3/5	3/5
7	0.76	-	4/5	3/5
8	0.53	+	4/5	4/5
9	0.43	-	1	4/5
10	0.25	+	1	1



Σημείο 1: TP=1, TN=5, FP=0, FN=4
FPR=FP/(FP+TN)=0/(0+5)
TPR=TP/(TP+FN)=1/(1+4)
Σημείο 2: TP=2, TN=5, FP=0, FN=3
FPR=FP/(FP+TN)=0/(0+5)
TPR=TP/(TP+FN)=2/(2+3)
Σημείο 3: TP=2, TN=4, FP=1, FN=3
FPR=FP/(FP+TN)=1/(1+4)
TPR=TP/(TP+FN)=2/(2+3)

Γιατί Επεξεργασία Φυσικής Γλώσσας;

- Η ψηφιακή πληροφορία κυρίως εμφανίζεται σε μορφή κειμένου
  - Διαδίκτυο: Wikipedia, κοινωνικά δίκτυα
  - Νέα, έρευνα, νομικά κείμενα, πατέντες
  - Αναφορές επιχειρήσεων
  - Δημόσια έγγραφα
  - …
- Ο υπολογιστής δεν μπορεί να αντιμετωπίσει το αδόμητο κείμενο
  - Πρέπει να μετατραπεί το κείμενο σε ονοματικά/αριθμητικά χαρακτηριστικά

Επιβλεπόμενη Μάθηση: Σχεδίαση συστήματος

- Εμπειρία (Δεδομένα εκπαίδευσης)
  - Πώς θα έχει το σύστημα πρόσβαση σε παραδείγματα εκπαίδευσης;
  - Σώματα κειμένων
- Συνάρτηση-στόχος
  - Ποια είναι η έννοια που θέλω το σύστημα να μάθει;
  - Στην ΕΦΓ η συν/ση στόχος καθορίζεται από την επισημείωση των παραδειγμάτων εκπαίδευσης
- Αναπαράσταση
  - Πώς θα αναπαρασταθούν οι έννοιες προς μάθηση;
  - Πρέπει να είναι ενιαία η αναπαράσταση
  - Πχ δυαδικά διανύσματα; Μοντέλο bag-of-words;
- Επαγωγή
  - Ποιος αλγόριθμος θα χρησιμοποιηθεί για την μάθηση των επιθυμητών εννοιών;
  - Παραδοχή: Μια προσέγγιση της συνάρτησης στόχου που αποδίδει καλά πάνω σε ένα επαρκώς μεγάλο σετ παραδειγμάτων θα αποδίδει καλά και σε καινούρια παραδείγματα

Μοντέλο Bag of words

- Ένα κείμενο είναι μια σακούλα από λέξεις
- Δεν ενδιαφέρει η διάταξη των λέξεων
- Συνήθως δεν ενδιαφέρει η συχνότητα εμφάνισης των λέξεων
- Δεν υπάρχει καμία μορφοσυντακτική πληροφορία σχετικά με την δομή των προτάσεων του κειμένου
- Δεν υπάρχει καμία σημασιολογική πληροφορία σχετικά με τις λέξεις και τις προτάσεις του κειμένου
- Αν τα διανύσματα αναπαράστασης δυο κειμένων είναι κοντά, τότε τα κείμενα είναι όμοια. Όσο μεγαλύτερη απόσταση έχουν, τόσο πιο ανόμοια είναι.
- Οπότε το «I love the film» έχει ίδια απόσταση με το «I hate the film» και με το «I like the film»

Επιβλεπόμενη Μάθηση και ΕΦΓ: Υπερπροσαρμογή

- Υπερπροσαρμογή (Overfitting)
  - Ανικανότητα του ταξινομητή να γενικεύσει σε καινούρια δεδομένα
  - Ο ταξινομητής αποδίδει καλά στα δεδομένα εκπαίδευσης, αλλά όχι καλά σε άγνωστα παραδείγματα
  - Απαιτείται μεγάλη ποσότητα δεδομένων εκπαίδευσης για να αντιμετωπιστεί
  - ΕΦΓ: σε πολλές εφαρμογές υπάρχει ανεπάρκεια επισημειωμένων δεδομένων
    - K-fold cross validation
    - όσο πιο λεπτομερής η αναπαράσταση των δεδομένων, τόσο περισσότερα δεδομένα χρειάζονται για αποδοτική μάθηση της συνάρτησης στόχου
    - Όσο πιο λεπτομερής η αναπαράσταση, τόσο μεγαλύτερος ο κίνδυνος υπερπροσαρμογής
    - Όσο λιγότερο λεπτομερής η αναπαράσταση, τόσο μεγαλύτερος ο κίνδυνος να χαθούν κριτήρια που είναι σημαντικά για την ταξινόμηση

Το πρόβλημα της διαστατικότητας

- Ο χώρος του μικρόκοσμου που θέλω να κάνω εξόρυξη είναι τόσων διαστάσεων όσο είναι και το πλήθος των ανεξάρτητων μεταβλητών μου
- Όσο αυξάνουν οι διαστάσεις, τα δεδομένα γίνονται ολοένα και πιο αραιά στον πολυδιάστατο χώρο
- Η απόσταση και η πυκνότητα μεταξύ των σημείων που είναι κρίσιμες για την ομαδοποίηση ή την ανίχνευση εξαιρέσεων χάνουν τη σημασία τους

Μείωση της διαστατικότητας

- Μείωση χρόνου και χώρου για τους αλγόριθμους εξόρυξης δεδομένων
- Επιτρέπει την οπτικοποίηση των δεδομένων
- Βοηθάει στην αφαίρεση άσχετων χαρακτηριστικών ή στη μείωση του θορύβου
- Εμπειρικός κανόνας: η υπερπροσαρμογή αποφεύγεται εάν ο αριθμός των παραδειγμάτων εκπαίδευσης είναι 50-100 φορές μεγαλύτερος του αριθμού των χαρακτηριστικών

- Τεχνικές
  - Principle Component Analysis (PCA)
  - Singular Value Decomposition (SVD)
  - Άλλες

Text Mining

- Μη δομημένα δεδομένα (unstructured data) είναι αυτά που δεν ακολουθούν κάποιο προκαθορισμένο μοντέλο
  - Βάση δεδομένων – δομημένα δεδομένα
  - Κείμενο – μη δομημένα δεδομένα
- Εξόρυξη δεδομένων (Data Mining)
  - Η εξαγωγή πληροφορίας από δομημένα δεδομένα
  - Επισκέψεις χρηστών σε ιστοσελίδα
  - Δείκτες χρηματιστηριακών τιμών
- Εξόρυξη κειμενικών δεδομένων (Text Mining)
  - Η εξαγωγή πληροφορίας από κειμενικά δεδομένα
  - Συνήθως με στόχο την εξαγωγή πληροφορίας υψηλού επιπέδου (θέμα, συναίσθημα, προφίλ συγγραφέα κλπ)

ΕΦΓ: Βασικές έννοιες

- Type
  - Κάθε διαφορετική «λέξη»
  - Στην ΜΜ συχνά χρησιμοποιούνται σαν χαρακτηριστικά μάθησης
- Token
  - Κάθε «λέξη»
  - Λέξεις, σύμβολα, ακρωνύμια, συντομογραφίες, σημεία στίξης
- Στην παρακάτω πρόταση πόσα types και πόσα tokens υπάρχουν;
  - "One could, for instance, use a set to store all types in a document, and a list to store all tokens."

Το κείμενο ως διάνυσμα χαρακτηριστικών-τιμών (feature value vectors)

- Επιλέγεται ένα σετ από types του κειμένου σαν σετ χαρακτηριστικών
  - Τα χαρακτηριστικά μπορεί να είναι
    - καταλήξεις
    - ν-γραμμα (βάσει γραμματικών ή βάσει συχνότητας εμφάνισης)
    - φράσεις ('White House', 'Barack Obama')
    - μεταδεδομένα των types
    - αποστάσεις μεταξύ types
    - συντακτικές συσχετίσεις μεταξύ types
    - σημασιολογικές συσχετίσεις μεταξύ types
- Κάθε κείμενο αναπαρίσταται σαν διάνυσμα με τιμές (βάρη) σε αυτά τα χαρακτηριστικά
- Τα βάρη καθορίζουν πόσο το συγκεκριμένο χαρακτηριστικό συνεισφέρει στην σημασιολογική αναπαράσταση του κειμένου
  - δυαδικά (παρουσία/απουσία του type στο κείμενο)
  - συχνότητα εμφάνισης
  - tfidf

Το κείμενο ως διάνυσμα χαρακτηριστικών-τιμών: Παράδειγμα σε Relation Extraction

- Στόχος η ανεύρεση σχέσεων ανάμεσα σε οντότητες σε μια πρόταση
- Οντότητες
  - Κύρια ονόματα
  - Ονοματικές φράσεις
- Παράδειγμα
  - Κάθε παράδειγμα είναι ένα ζευγάρι οντοτήτων
- Έξοδος
  - Είδος σχέσης ανάμεσα στις οντότητες
- Αρχιτεκτονική ταξινόμησης
  - Δυαδική έξοδος για κάθε είδος σχέσης
    - Το είδος που θα αποδοθεί με μεγαλύτερο σκορ στο παράδειγμα κερδίζει
  - Ονοματική έξοδος με διαφορετική τιμή για κάθε είδος σχέσης

Παραδείγματα σχέσεων

Relations in ACE dataset

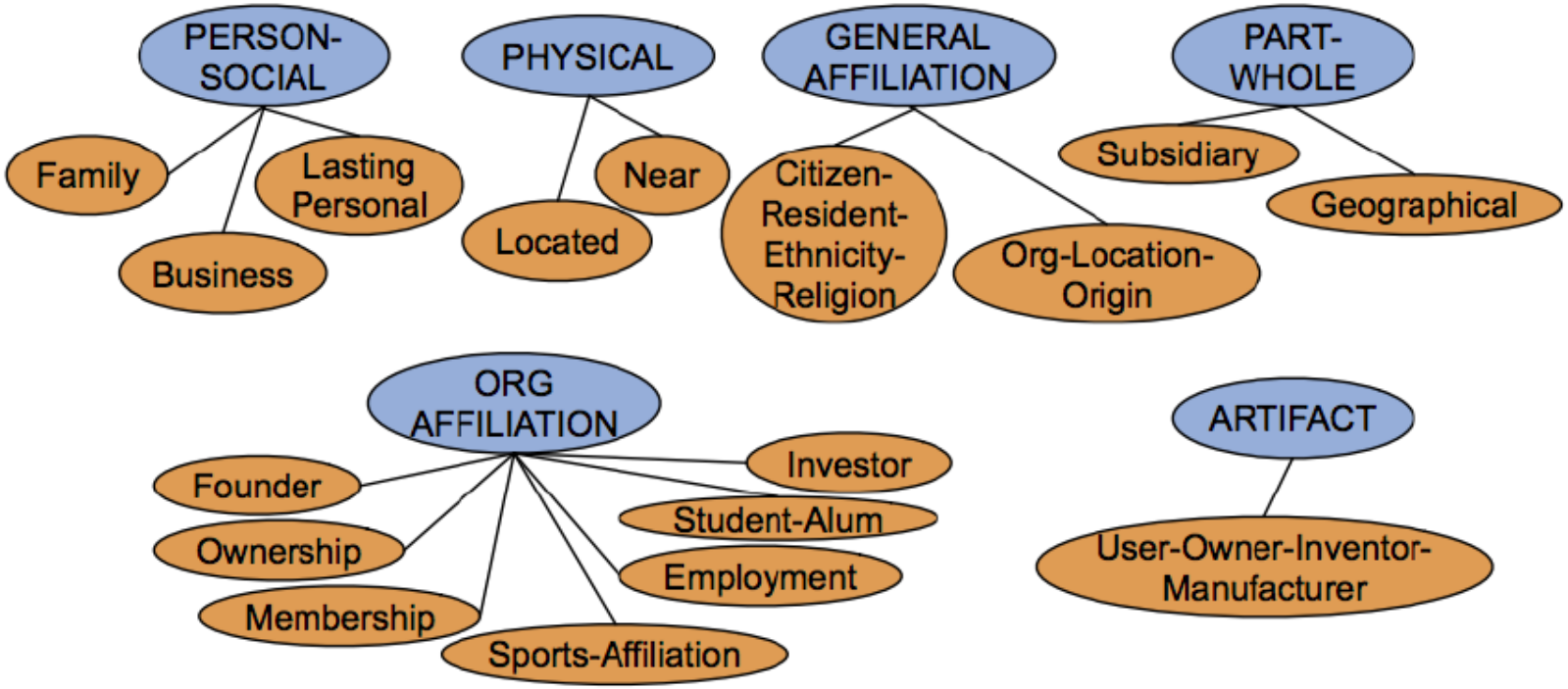
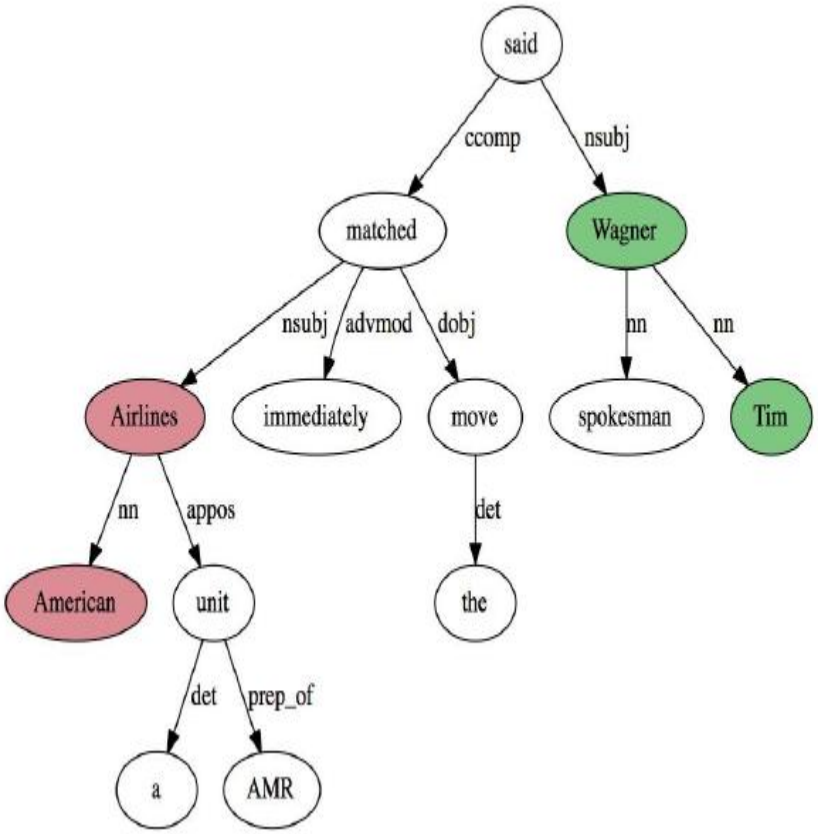


Figure 21.8: https://web.stanford.edu/~jurafsky/slp3/21.pdf

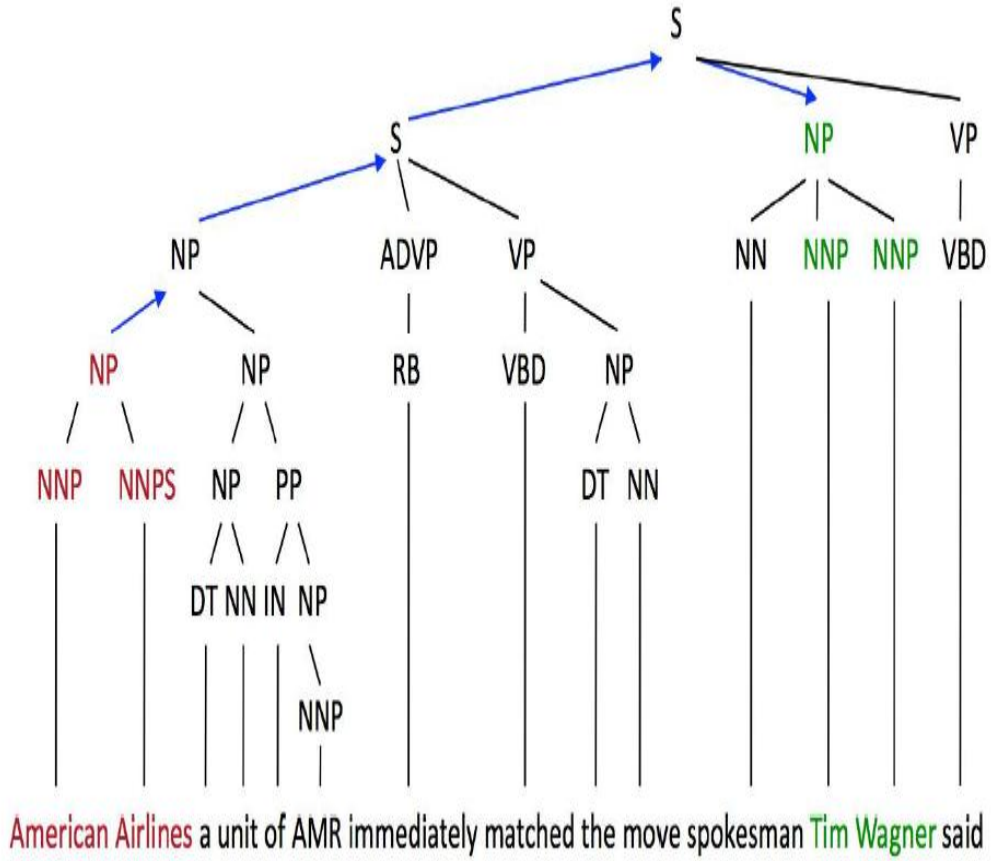
Το κείμενο ως διάνυσμα χαρακτηριστικών-τιμών: Παράδειγμα σε Relation Extraction (Exploring Various Knowledge in Relation Extraction, Zhou et al., ACL 2005)

- Χαρακτηριστικά
  - Των δυο οντοτήτων
    - κατηγορίες
    - Κύριοι όροι
  - Των λέξεων που περιβάλλουν τις δυο οντότητες
    - Παράθυρο συμφραζομένων
  - Της συντακτικής δομής που κυβερνά τις δυο οντότητες
    - Συντακτική συσχέτιση
    - Μονοπάτι συντακτικού δέντρου ανάμεσα στις δυο οντότητες
    - Μήκος μονοπατιού δέντρου

Οι συντακτικές δομές



Δέντρο γραμματικής εξαρτήσεων
Dependency Grammar



Συντακτικό δέντρο

Το κείμενο ως διάνυσμα χαρακτηριστικών-τιμών: Παράδειγμα σε Relation Extraction (Zhou et al., 2005)

American Airlines, a unit of AMR, immediately matched the move, spokesman **Tim Wagner** said

Template	Features
Headwords of M1 and M2	Airlines, Wagner, Airlines-Wagner
Unigrams, Bigrams	American, Airlines, Tim, Wagner, American Airlines, Tim Wagner
Words in particular positions	M1-1: NONE, M1+1: a, M2-1: spokesman, M2+1: said
Unigrams between M1 and M2	a, unit, of, AMR, immediately, matched, the, move, spokesman
Named entity types	M1: ORG, M2: PER, ORG-PER
Entity Level (NAME, NOMINAL, PRONOUN)	M1: NAME, M2: NAME
Number of entities between arguments	1 (AMR)
Syntactic chunk sequence from M1 to M2	NP NP PP ADVP VP NP NP
Constituency paths between M1 and M2	NP_{UP} NP_{UP} S_{UP} VP_{DOWN} S_{DOWN} NP_{DOWN}
Dependency tree paths	Airlines <--(nsubj) matched -->(ccomp) said <—(nsubj) Wagner

Μείωση της διαστατικότητας

- Stemming
  - αποκοπή της κατάληξης με στόχο την ομαδοποίηση των types που μοιράζονται κοινή μορφολογική ρίζα
  - cluster, clustering, clustered,… -> cluster
- Lemmatization
  - Λαμβάνει υπόψη της την μορφολογική ανάλυση των λέξεων προς ομαδοποίηση ώστε να βρει κοινό λήμμα
- Διαγραφή λειτουργικών λέξεων (function words)
  - προθέσεις, άρθρα, σύνδεσμοι, …
- Διαγραφή λέξεων βάσει της συχνότητας εμφάνισης
- Επιλογή χαρακτηριστικών (feature selection)
- Latent Semantic Indexing
- Μπορεί να γίνει
  - για κάθε τιμή της εξόδου ξεχωριστά (local)
  - για όλες τις τιμές μαζί (global)

Latent Semantic Indexing

- Η τεχνική LSI χρησιμοποιεί Singular Value Decomposition για να διασπάσει τον αρχικό term-document matrix.
- Εύρωστη (δεν ενοχλείται από ανορθογραφίες κλπ)
- Υπολογιστικά ακριβή
- Καλή αν κάθε όρος συνεισφέρει από λίγο στην σωστή ταξινόμηση
- Όχι καλή αν ένας μεμονωμένος όρος είναι ιδιαίτερα σημαντικός για την ταξινόμηση

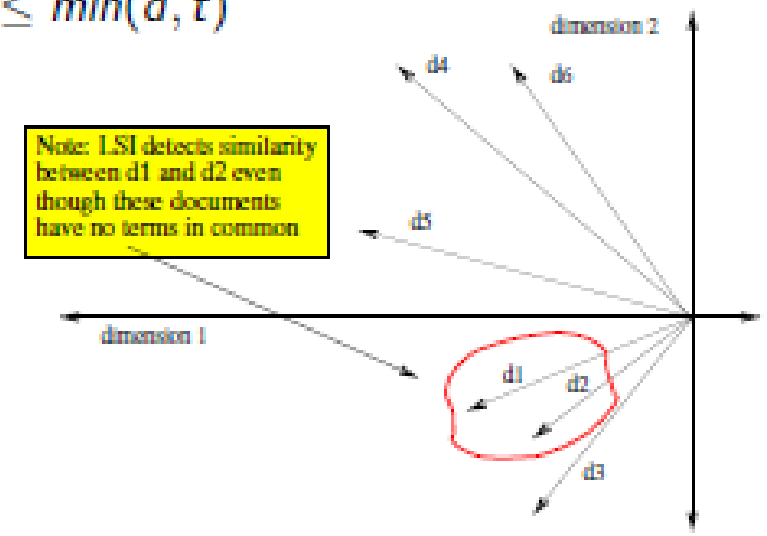
$A =$

	d_1	d_2	d_3	d_4	d_5	d_6
cosmonaut	1	0	1	0	0	0
astronaut	0	1	0	0	0	0
moon	1	1	0	0	0	0
car	1	0	0	1	1	0
truck	0	0	0	1	0	1

Table: Sample term-document incidence matrix

$$A_{t \times d} = T_{t \times n} S_{n \times n} (D_{d \times n})^T$$

where $n = \leq min(d, t)$



Για την επόμενη φορά

- Μελετάτε το άρθρο
  - «*A Survey on Natural Language Processing for Fake News Detection*» των Oshikawa, Qian και Wang
- Ετοιμάζετε σε ppt διαφάνειες για να παρουσιάσετε το περιεχόμενο του άρθρου
  - Το πρόβλημα της αναγνώρισης ψευδών ειδήσεων (1)
  - Τα datasets μιας ή λίγων προτάσεων (2.1.1)
  - Τα datasets αναρτήσεων σε κοινωνικά δίκτυα (2.1.2)
  - Τα datasets ολόκληρων άρθρων (2.2)
  - Τις μορφές εξόδου των εργασιών μάθησης (3.2)
  - Τις τεχνικές προεπεξεργασίας των κειμένων (4.1)
  - Τις τεχνικές μηχανικής μάθησης (4.4)
  - Τα αποτελέσματα (5)
  - Τα συμπεράσματα για την είσοδο (6.1)
  - Τα συμπεράσματα για τα μοντέλα (6.2)
  - Τις σχετικές προκλήσεις (7)