«Τεχνογλωσσία VIII»
Εξαγωγή πληροφοριών από κείμενα

Σεμινάριο 1: Γενική Επισκόπηση

Ευάγγελος Καρκαλέτσης, Γεώργιος Πετάσης

Εργαστήριο Τεχνολογίας Γνώσεων & Λογισμικού,

Ινστιτούτο Πληροφορικής & Τηλεπικοινωνιών, Ε.Κ.Ε.Φ.Ε. "Δημόκριτος"

Τηλ.: 210-6503197, Fax: 210-6532175, {vangelis, petasis}@iit.demokritos.gr

Ακαδημαϊκό Έτος: 2013 – 2014

ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

Γενική Επισκόπηση και ιστορική αναδρομή

Τι είναι η γλωσσική τεχνολογία;

- Η ανάπτυξη υπολογιστικών μοντέλων επεξεργασίας πληροφορίας εκφρασμένης σε φυσική γλώσσα
- Γλώσσα: μέσο καταγραφής και ανταλλαγής πληροφορίας
- Φυσική γλώσσα: μέσο για την επικοινωνία μεταξύ ανθρώπων
  - Τεχνητή γλώσσα: μέσο για την επικοινωνία ανθρώπου - μηχανής

Γλωσσική τεχνολογία

- Αυτόματη **ανάλυση** («κατανόηση»;) και **παραγωγή** γραπτών ή προφορικών εκφράσεων φυσικής γλώσσας
  - Αυτόματη διόρθωση κειμένων, μηχανική μετάφραση, εξαγωγή πληροφορίας, αυτόματη παραγωγή περιλήψεων, συστήματα ερωταποκρίσεων, διαλογικά συστήματα, αυτόματη παραγωγή κειμένων, κ.α.
- Διάφορα μέσα περιέχουν φυσική γλώσσα
  - Γραπτός λόγος (κείμενα), προφορικός λόγος (ομιλία), εικόνα εγγράφου, κλπ.

Γιατί είναι σημαντική; (1)

- Πολύ μεγάλο μέρος της καταγεγραμμένης ανθρώπινης γνώσης είναι εκφρασμένο σε φυσική γλώσσα
  - Γνώση οργανισμών: νόμοι, κανονισμοί, πατέντες, αναφορές, πρακτικά, αλληλογραφία, εγχειρίδια, οδηγίες, κλπ.
  - Πληροφορία από/για χρήστες: ιστόχωροι οργανισμών, περιγραφές προϊόντων, ηλεκτρονική αλληλογραφία, ιστολόγια, επικοινωνία μέσω κοινωνικών δικτύων, φόρα συζητήσεων, κλπ.

Γιατί είναι σημαντική; (2)

- Η ραγδαία ανάπτυξη του παγκόσμιου ιστού έχει καταστήσει μεγάλους όγκους πληροφορίας άμεσα προσβάσιμους
  - Οδηγώντας στην υπερ-πληροφόρηση
- Η γλωσσική τεχνολογία έχει ήδη συμβάλει στην ανακάλυψη νέων τρόπων για την καλύτερη συμβίωσή μας με την τεχνολογία
  - Συστήματα που: αναγνωρίζουν ομιλία και γραφή, κατανοούν κείμενα αρκετά καλά ώστε να μπορούν να επιλέγουν πληροφορίες, μεταφράζουν από μια γλώσσα σε άλλες, συνθέτουν ομιλία και κείμενα, κλπ.

Όμως, δεν είναι εύκολη…

- Η φυσική γλώσσα είναι περίπλοκη…
  - Πολλαπλοί τρόποι έκφρασης της ίδιας πληροφορίας, ασάφεια, ελλιπής πληροφορία, διαφορετικό νόημα ανάλογα το περιβάλλον, δημιουργία νέων εκφράσεων, κλπ.
- επειδή απευθύνεται σε ανθρώπους
  - Οι οποίοι χρησιμοποιούν γνώση του κόσμου και εμπειρία για την κατανόηση της φυσικής γλώσσας
- Η μηχανή δυσκολεύεται σημαντικά
  - Περιορισμένη γνώση του κόσμου: εστίαση σε θεματικές περιοχές, χρήση οντολογιών

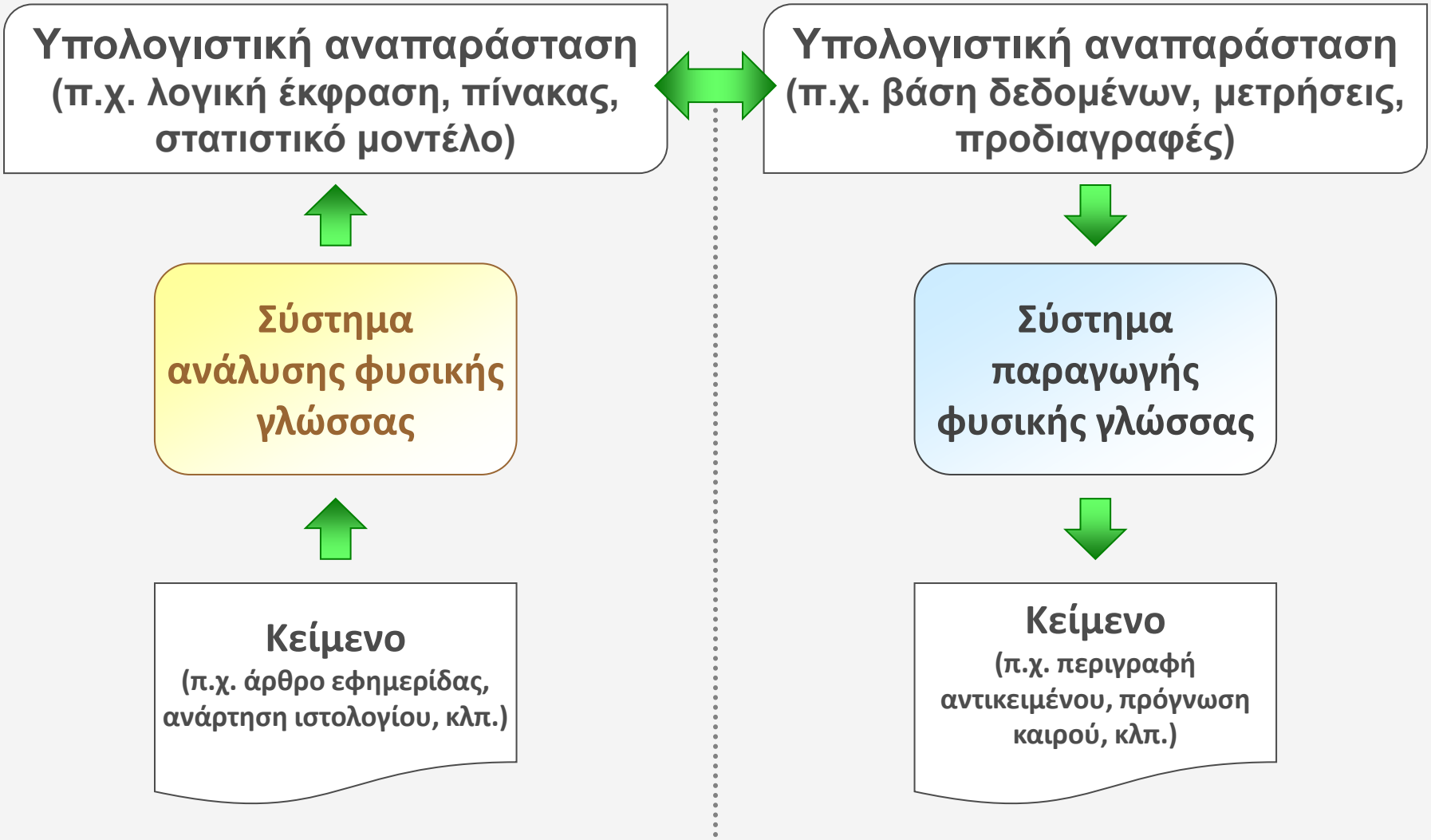
Ασάφεια

- Φωνολογική: «λύπη, λίπη, λείπει», «στον ώμο, στο νόμο»
- Μορφολογική: «η/την μητέρα»
- Συντακτική: «*Κάνε το δικό σου*.»
- Σημασιολογική: «τόνος», «σκοπός», «ρόκα»
- Πραγματολογική: «Ξέρεις τι ώρα είναι;»

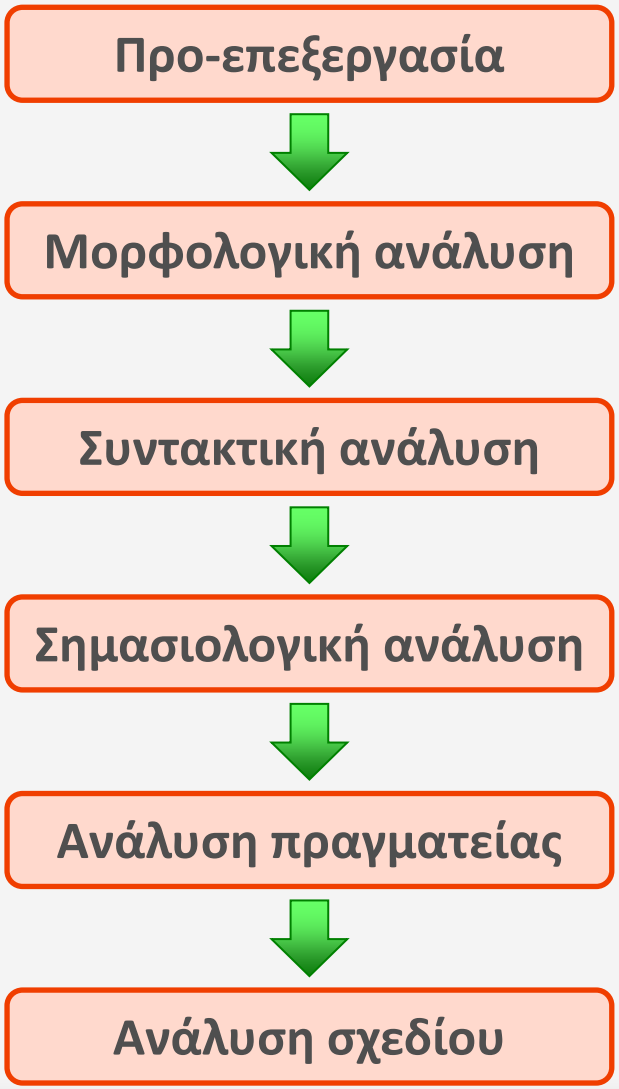
Πολλοί συναφείς όροι

- **Επεξεργασία φυσικής γλώσσας** (natural language processing)
  - Κυρίως γραπτή γλώσσα, κύριος στόχος η δημιουργία υπολογιστικών συστημάτων, υποτομέας της ΤΝ
- **Υπολογιστική γλωσσολογία** (computational linguistics)
  - Κυρίως γραπτή γλώσσα, κύριος στόχος η δημιουργία υπολογιστικών μοντέλων γλωσσολογικών θεωριών
    - Θεωρητική προσέγγιση, πλέον συνώνυμο της ΕΦΓ
- **Γλωσσική τεχνολογία** (human language technology)
  - Λιγότερο καθιερωμένος όρος, συνήθως περιλαμβάνει και τεχνολογίες αναγνώρισης και σύνθεσης φωνής, έμφαση στη δημιουργία χρήσιμων υπολογιστικών συστημάτων

Ανάλυση και παραγωγή



Επίπεδα ανάλυσης



Αναγνώριση λέξεων, προτάσεων, …

Πληροφορίες για τις λέξεις, όπως θέμα, κατάληξη, πρόσωπο, αριθμό, γένος, …

Συντακτική δομή περιόδων, ρόλοι των λέξεων, κλπ.

Αναγνώριση του νοήματος των προτάσεων

Αναφορικές εκφράσεις, σχέσεις μεταξύ προτάσεων

Σκοποί του χρήστη, σχέδια δράσεως, …

Σύντομη ανασκόπηση (1)

- Η γλωσσική τεχνολογία είναι τόσο παλιά, όσο και οι Η/Υ
  - Η έρευνα ξεκίνησε την δεκαετία του 1950
- Μεγάλη επιρροή άσκησε η έρευνα του Noam Chomsky
  - Διατύπωσε θεωρίες σχετικά με την εκμάθηση της γλώσσας από τα παιδιά
  - Όρισε μια ιεραρχία γλωσσών, αποτελούμενη από 4 κατηγορίες τυπικών γραμματικών (formal languages)
    - Κανονικές, ανεξάρτητες από τα συμφραζόμενα, εξαρτημένες από τα συμφραζόμενα, απεριόριστες

Σύντομη ανασκόπηση (2)

- 1950 – 1965: Πρώτα βήματα, έμφαση στην μηχανική μετάφραση
- 1965 – 1980: Έμφαση στην σημασιολογία
- 1980 – 1990: Έμφαση στην σύνταξη, την (στατιστική) μηχανική μετάφραση, την αναγνώριση ομιλίας
- 1990 – σήμερα: Έμφαση στις εργασίες χαμηλού επιπέδου, την συνεργασία ΕΦΓ και μηχανικής μάθησης, την εξαγωγή πληροφορίας, την αξιολόγηση

Παραδείγματα εφαρμογών (1)

- Υποστήριξη συγγραφής
  - Ορθογραφική/συντακτική διόρθωση, συστήματα υπαγόρευσης
- Μετάφραση κειμένων
- Κατηγοριοποίηση και φιλτράρισμα κειμένων
  - Μηχανές αναζήτησης
- Εξαγωγή πληροφορίας – παραγωγή περίληψης
- Διεπαφές σε φυσική γλώσσα
  - Σε βάσεις δεδομένων/μηχανές αναζήτησης, διαλογικά συστήματα, αυτοματοποιημένες υπηρεσίες μέσω τηλεφώνου

Παραδείγματα εφαρμογών (2)

- Δρομολόγηση αλληλογραφίας
- Αυτόματος υποτιτλισμός
- Αντιστοίχηση βιογραφικών με αγγελίες ευρέσεως εργασίας
- Εξαγωγή στοιχείων για τον συγγραφέα από τον τρόπο γραφής (stylometry)
  - Εντοπισμός αντιγράφων, αυθεντικότητα κειμένων, κλπ.
- Εντοπισμός συναισθήματος

Κοινές εργασίες

- Αναγνώριση ομιλίας, OCR
- Καθάρισμα κειμένων (π.χ. από HTML)
- Αναγνώριση λέξεων/προτάσεων
- Αναγνώριση μερών του λόγου
- Ρηχή συντακτική ανάλυση
- Αναγνώριση ονομάτων οντοτήτων
- Εξαγωγή συσχετίσεων
- Αναγνώριση συναισθήματος/πολικότητας
- Αποσαφήνιση έννοιας λέξεων

Πλατφόρμες

- Υποδομές ενθυλάκωσης εργαλείων ΕΦΓ
- GATE - http://gate.ac.uk/
  - Η πρώτη δημοφιλής πλατφόρμα – Java
- Ellogon - http://www.ellogon.org/
  - Η πρώτη UNICODE πλατφόρμα – C/C++/Tcl/…
- NLTK - http://nltk.org/
  - Η «εκπαιδευτική» πλατφόρμα – Python
- Apache UIMA - http://uima.apache.org/
  - Η «ανερχόμενη» πλατφόρμα – Java/C++

Ανασκόπηση

- Γλωσσική τεχνολογία: γλωσσικές δυνατότητες που ενσωματώνονται σε συστήματα πληροφορικής και επικοινωνιακής τεχνολογίας
  - Αφορά την ανάλυση και παραγωγή φυσικής γλώσσας
- Είναι σημαντική και ταυτόχρονα δύσκολη
- Έξι επίπεδα ανάλυσης
- Σύντομη ανασκόπηση
- Ενδεικτικές εφαρμογές/κοινές εργασίες
- Πλατφόρμες ΕΦΓ

ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Οι διαφάνειες αυτής της ενότητας βασίζονται
στα κεφάλαια 1, 2 και 3 του βιβλίου:

«*Η τεχνολογία της πληροφορίας στην επεξεργασία φυσικής γλώσσας*»,
Κ. Φράγγος και Αν. Κουτσούκος, εκδόσεις ΜΥΡΜΙΔΟΝΕΣ, 2010.

Μοντέλα και αλγόριθμοι (1)

- Τα ποικίλα είδη γνώσης της ΕΦΓ μπορούν να αναπαρασταθούν από ένα μικρό σύνολο τυπικών μεθόδων (formal methods) ή θεωριών
  - Προέρχονται από τον χώρο της επιστήμης υπολογιστών, των μαθηματικών και της γλωσσολογίας

Μοντέλα και αλγόριθμοι (2)

- Σημαντικές μέθοδοι:
  - Μηχανές καταστάσεων (state machines)
  - Συστήματα τυπικών κανόνων (formal rule systems)
  - Λογική
  - Θεωρία πιθανοτήτων
  - Μηχανική μάθηση (machine learning)

Μηχανές καταστάσεων

- Τυπικά μοντέλα που αποτελούνται:
  - Καταστάσεις
  - Μεταβάσεις μεταξύ καταστάσεων
  - Αναπαράσταση εισόδου
- Διάφοροι τύποι:
  - (Μη) ντετερμινιστικά αυτόματα πεπερασμένων καταστάσεων
  - Finite state transducers
  - Αυτόματα με βάρη
  - Αυτόματα με πιθανότητες (Markov models)

Δηλωτικά μοντέλα

- Συστήματα τυπικών κανόνων:
  - Κανονικές γραμματικές
  - Γραμματικές ανεξάρτητες από συμφραζόμενα
  - Γραμματικές με χαρακτηριστικά (feature augmented grammars)
  - Πιθανοτικές παραλλαγές

  - Συνήθως χρησιμοποιούνται στον χειρισμό γνώσης:
    - Φωνολογίας
    - Μορφολογίας
    - Σύνταξης

Λογική

- Επίσης δημοφιλές μοντέλο, κυρίως για την σημασιολογική/πραγματολογική ανάλυση, καθώς και την επεξεργασία λόγου
  - First order logic
  - Predicate calculus
  - Επαγωγή/απαγωγή
- Το κυρίαρχο μοντέλο για την αξιοποίηση οντολογιών

Θεωρία πιθανοτήτων

- Το κυρίαρχο μοντέλο αναπαράστασης γλωσσολογικής γνώσης
- Όλα τα προηγούμενα μοντέλα μπορούν να εμπλουτιστούν με πιθανότητες
- Μπορεί να λύσει πολλά είδη προβλημάτων ασάφειας
  - Σχεδόν κάθε πρόβλημα ΕΦΓ μπορεί να δοθεί σαν: «δεδομένων Ν επιλογών για μια ασαφή είσοδο, επέλεξε την πιο πιθανή»
- Εκμάθηση πιθανοτικών μοντέλων από σώματα κειμένων (μηχανική μάθηση)

Στατιστική ανάλυση

- Στατιστική συμπερασματολογία
  - Κλάδος της στατιστικής
  - Ασχολείται με μεθόδους μεταφοράς πληροφοριών από δείγμα στον γενικό πληθυσμό
- Περιλαμβάνει:
  - Εκτιμητική: εκτίμηση παραμέτρων πληθυσμού με βάση αντίστοιχες παραμέτρους του δείγματος
  - Έλεγχο υποθέσεων: επιβεβαίωση/απόρριψη ισχυρισμών για τις τιμές παραμέτρων του πληθυσμού
  - Διατύπωση στατιστικών μοντέλων εκτίμησης τιμής/διαστήματος εμπιστοσύνης εξαρτημένων μεταβλητών, με βάση τιμές ανεξάρτητων μεταβλητών

Στατιστικός έλεγχος

- Η διαδικασία της γενίκευσης από ένα δείγμα στον πληθυσμό δεν είναι συχνά δίχως σφάλματα
  - Σφάλμα τύπου I (α): η πιθανότητα απόρριψης μιας υπόθεσης Η, ενώ είναι ορθή
    - Ονομάζεται και στάθμη σημαντικότητας του ελέγχου
  - Σφάλμα τύπου II (β): η πιθανότητα αποδοχής μιας υπόθεσης Η, ενώ είναι λανθασμένη
    - Η πιθανότητα $\gamma = 1 - \beta$ ονομάζεται ισχύς ελέγχου
- Υπάρχουν διάφορες τεχνικές ελέγχου
    - Απαρίθμηση μερικών στις σελ. 31-32

Μέτρα αποτίμησης

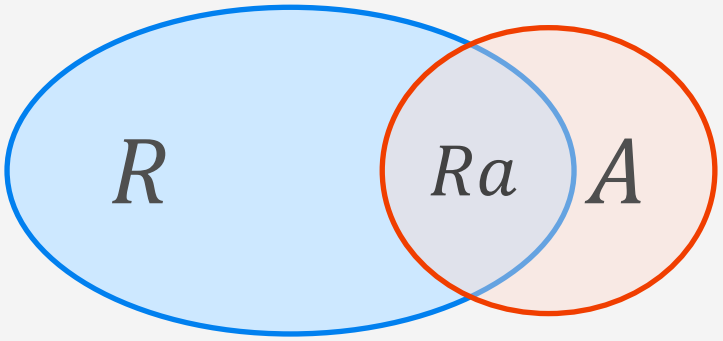
- Η αποτίμηση/εκτίμηση της αποδοτικότητας συστημάτων ΕΦΓ είναι σημαντική

- Κυρίαρχα μέτρα αποτίμησης:
  - Ακρίβεια (precision)
    - Πόσες από τις απαντήσεις που έδωσε ένα σύστημα είναι σωστές
  - Ανάκληση (recall)
    - Πόσα ερωτήματα προς το σύστημα απαντήθηκαν σωστά
  - F-measure: συνδυασμός ακρίβειας και ανάκλησης

Παράδειγμα (1)

- Ανάκτηση πληροφορίας
  - Υποθέτουμε ένα σύνολο εγγράφων
  - Υποθέτουμε ένα σύνολο ερωτημάτων
    - Κάθε ερώτημα πρέπει να απαντηθεί με ένα σύνολο εγγράφων, που ικανοποιούν το ερώτημα
  - Έστω ερώτημα q, και R το σύνολο των σχετικών εγγράφων
  - Έστω ένα ελεγχόμενο σύστημα ΕΦΓ, επεξεργάζεται το ερώτημα q, και επιστρέφει το σύνολο εγγράφων A

Παράδειγμα (2)



Έστω:

$|R|$: ο αριθμός των εγγράφων στο σύνολο $|R|$

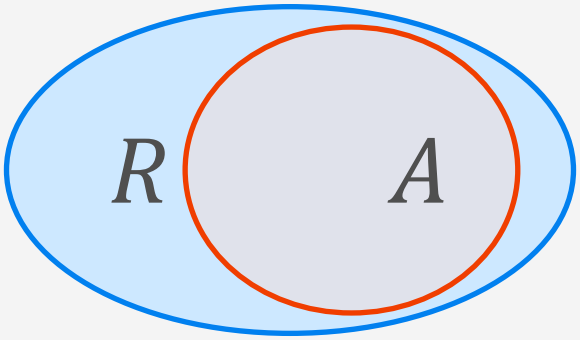
$|A|$: ο αριθμός των εγγράφων στο σύνολο $|A|$

$|Ra|$: ο αριθμός των εγγράφων στην τομή $R \cap A$

$$\text{Ακρίβεια} = \frac{|Ra|}{|A|} \quad \text{Ανάκληση} = \frac{|Ra|}{|R|}$$

$$F_1 = 2 \cdot \frac{\alpha\kappa\rho\iota\beta\varepsilon\iota\alpha \cdot \alpha\nu\alpha\kappa\lambda\eta\sigma\eta}{\alpha\kappa\rho\iota\beta\varepsilon\iota\alpha + \alpha\nu\alpha\kappa\lambda\eta\sigma\eta}$$

Παράδειγμα (3)



Μέση ακρίβεια, ανάκληση

- Είδαμε την περίπτωση ενός ερωτήματος
- Αν έχουμε ένα σύνολο ερωτημάτων N, υπολογίζουμε τον μέσο όρο:

$$\bar{P} = \sum_{i=1}^{N} \frac{P_i}{N}$$

$$\bar{R} = \sum_{i=1}^{N} \frac{R_i}{N}$$

Αξιολόγηση ταξινομητών

- Ταξινομητής: κατηγοριοποίηση σε n κατηγορίες

- Συνολικά για n κατηγορίες:
  - Macro averaging (ίσο βάρος σε όλες τις κατηγορίες):

$$Macro\ P = \sum_{i=1}^{N} \frac{P_i}{N}, Macro\ R = \sum_{i=1}^{N} \frac{R_i}{N}$$

  - Micro averaging (σημαντικότερες οι πολυπληθέστερες):

$$Micro\ P = \frac{\sum_{i=1}^{n} |Ra|_i}{\sum_{i=1}^{n} |A|_i}, Micro\ R = \frac{\sum_{i=1}^{n} |Ra|_i}{\sum_{i=1}^{n} |R|_i}$$

ΣΥΝΟΛΑ ΧΑΡΑΚΤΗΡΩΝ ΚΑΙ UNICODE

Τι είναι αυτά, και γιατί με αφορούν;

Σύνολα χαρακτήρων; (1)

- Τι είναι αυτά, και γιατί με αφορούν;
  - Σχετίζονται με την επεξεργασία φυσικής γλώσσας;
- Έχετε λάβει ποτέ e-mail με το εύγλωττο θέμα «??? ???? ?????????»;
- Χρειάστηκε να γράψετε ποτέ κώδικα που θα διαχειρίζεται e-mails στα Ιαπωνικά;
- Έχετε αναρωτηθεί τι κάνει αυτή η «μυστηριώδης» ετικέτα «Content-Type» στην HTML;

Σύνολα χαρακτήρων; (2)

- Δεν βλέπετε κανένα λάθος στο ακόλουθο:

  Απλό κείμενο ≡ ASCII ≡ Χαρακτήρες 8 δυφίων;

Ιστορική αναδρομή (1)

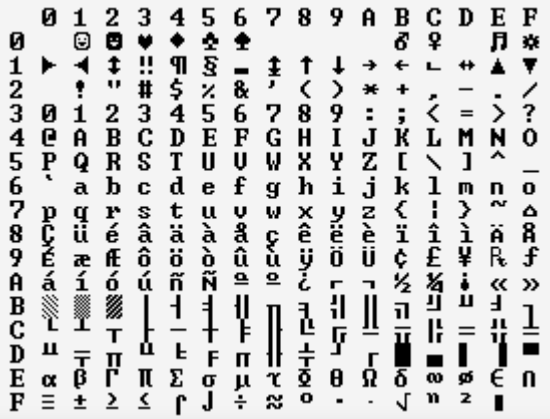
- Το 1963 δημοσιεύθηκε το πρότυπο ASCII (American Standard Code for Information Interchange)
  - Αφορούσε του αγγλικούς χαρακτήρες
  - Κωδικοποιούσε χαρακτήρες με έναν αριθμό από το 32 έως το 127
    - Π.χ. το κενό είναι το 32, το “Α” το 65, κλπ.
  - Χρησιμοποιούσε 7 δυφία (bits)
  - Οι κωδικοί κάτω από το 32 αναφέρονται σαν «μη εκτυπώσιμοι», και αποτελούν χαρακτήρες ελέγχου
    - Ο χαρακτήρας 7 παράγει ένα «μπιπ»

ASCII

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2	SPC	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

Ιστορική αναδρομή (2)

- Επειδή όμως το byte χωρά 8 δυφία, αρκετοί σκέφτηκαν:
  - «Μμμ, μπορούμε να χρησιμοποιήσουμε τους κωδικούς 128-255 για άλλους σκοπούς…»
    - Το πρόβλημα ήταν ότι πολλοί είχαν την ίδια ιδέα, την ίδια στιγμή, για διαφορετικούς σκοπούς
  - IBM-PC: “OEM character set” ή ASCII-DOS
    - Μια **κωδικοσελίδα** που πρόσθεσε μερικούς τονισμένους χαρακτήρες, και χαρακτήρες σχεδίασης



Ιστορική αναδρομή (3)

- Η εξάπλωση των υπολογιστών ανά την υφήλιο, έφερε πληθώρα κωδικοσελίδων…
- … Τις οποίες ανέλαβε να οργανώσει ο οργανισμός ANSI
  - Οι κωδικοί < 128 είναι ίδιοι (ASCII)
  - Οι κωδικοί > 127 διαφέρουν, ανάλογα την κωδικοσελίδα (code page)
  - Η κωδικοσελίδα 737 περιέχει τα ελληνικά σε περιβάλλον DOS

Ιστορική αναδρομή (4)

- Τα ασιατικά αλφάβητα ωστόσο, έχουν χιλιάδες «χαρακτήρες», οι οποίοι δεν χωρούν σε 8 δυφία
- Η λύση δόθηκε μέσω του DBCS (double byte character set)
  - Ένας χαρακτήρας μπορεί να ήταν 1 ή δύο bytes
    - Ήταν εύκολο να προσπεράσεις μια συμβολοσειρά από την αρχή προς το τέλος, αλλά όχι το αντίστροφο
    - Χρήση συναρτήσεων όπως AnsiNext() & AnsiPrev() για μετακινήσεις μέσα σε συμβολοσειρές

Ιστορική αναδρομή (5)

- Ωστόσο, για πολύ καιρό «προσποιούμασταν» ότι ένα byte ήταν ένας χαρακτήρας, και τα πάντα δούλευαν όσο:
  - Ήμασταν στο ίδιο λειτουργικό σύστημα
  - Τα πάντα ήταν σε μια γλώσσα
- Μέχρι που ήρθε ο παγκόσμιος ιστός…
  - Όπου τα πάντα μπερδεύτηκαν

Ευτυχώς όμως, είχε εφευρεθεί το **UNICODE**!

UNICODE (1)

- Το UNICODE προσπαθεί να δημιουργήσει **ένα** σύνολο χαρακτήρων για **όλες τις γλώσσες** του κόσμου
  - Μύθος: οι χαρακτήρες του UNICODE έχουν μήκος 16 δυφία, οπότε περιγράφονται μόνο 65536 χαρακτήρες
    - Αυτό δεν ισχύει
  - Κάθε γράμμα αντιστοιχίζεται με ένα code point
    - “A” → U+0391, ΰ → U+03B0 (charmap.exe)
    - Hello → U+0048 U+0065 U+006C U+006C U+006F

UNICODE (2)

- Μια συμβολοσειρά είναι ένα σύνολο από code points
- Αναπαράσταση στην μνήμη/δίσκο;
  - Κωδικοποιήσεις (encodings)
    - Hello → U+0048 U+0065 U+006C U+006C U+006F
    - Big endian: 00 48 00 65 00 6C 00 6C 00 6F (BOM: FE FF)
    - Little-endian: 48 00 65 00 6C 00 6C 00 6F 00 (BOM: FF FE)
      - BOM: Byte Order Marker
    - Encoding: UCS-2 (2 bytes) or UTF-16 (16 δυφία)
      - Υπάρχει και το UCS-4, ή UTF-32!

UNICODE (3)

- Μια θαυμάσια ιδέα: UTF-8
  - Μια ακόμα κωδικοποίηση, που χρησιμοποίει bytes 8 δυφίων
  - Κάθε χαρακτήρας από το 0-127, κωδικοποιείται με 1 byte
  - Χαρακτήρες > 127, κωδικοποιούνται σε 2, 3, …, 6 bytes
  - Οι αγγλικοί χαρακτήρες αναπαριστώνται με τον ίδιο τρόπο όπως στο ASCII
    - Και φυσικά καταλαμβάνουν τον ίδιο χώρο στην μνήμη

UNICODE (4)

- Και φυσικά είναι δυνατή η μετατροπή σε εκατοντάδες κωδικοσελίδες:
  - ISO-8859-1 (Latin-1), ISO-8859-15, Windows-1252 (Αγγλικά)
  - ISO-8859-7, Windows-1253 (Ελληνικά)

Τι κρατάμε από όλα αυτά;

Ότι δεν έχει **καμιά σημασία**
να έχουμε μια συμβολοσειρά,
αν δεν ξέρουμε την **κωδικοποίησή** της!

C++

- Για να χρησιμοποιήσουμε UCS-2:

  char → wchar_t

  str*() → wcs*()

  strlen() → wcslen()

- Literal strings:

  wchar_t *str = L”Hello”;

C++11

- char: size enough for UTF-8
- wchar_t : undefined size, semantics
- Adds support for 2 more encodings:
  - char16_t, char32_t
    - u8"This is a Unicode Character: \u2018."
    - u"This is a bigger Unicode Character: \u2018."
    - U"This is a Unicode Character: \U00002018."

Java

- Στην Java, τα πάντα αναπαριστώνται σε UCS-2
  - Ο τύπος char είναι 16 δυφίων
- Αυτό σημαίνει ότι χωράνε μόνο οι πρώτοι 65,536 χαρακτήρες του UNICODE
  - Οι υπόλοιποι ονομάστηκαν «συμπληρωματικοί» (supplementary characters)
    - Αναπαριστώνται σαν δυάδες από char

```
String newString(int codePoint) {
   return new String(Character.toChars(codePoint));
}
```

Περισσότερα εδώ: http://docs.oracle.com/javase/tutorial/i18n/text/usage.html

Python (1)

- Η python υποστηρίζει επίσης UNICODE
  - Αν και δεν είναι ο εγγενής της τύπος

```
>>> import sys
>>> import codecs
>>> sys.stdin.encoding
cp1253
>>> sys.stdout.encoding
cp1253
>>> sys.stdout = codecs.getwriter("cp1253" )(sys.stdout)
```

Python (2)

```
>>> a = unicode("απλό τεστ στα ελληνικά", "cp1253" )
>>> a
u"\u03b1\u03c0\u03bb\u03cc \u03c4\u03b5\u03c3\u03c4 \u03c3\u03c4\u03b1
\u03b5\u03bb\u03bb\u03b7\u03bd\u03b9\u03ba\u03ac"
>>> print a
απλό τεστ στα ελληνικά
>>> import nltk
>>> b = nltk.word_tokenize(a)
>>> print b
[u"\u03b1\u03c0\u03bb\u03cc", u"\u03c4\u03b5\u03c3\u03c4", u"\u03c3\u03c4\u03b1",
"\u03b5\u03bb\u03bb\u03b7\u03bd\u03b9\u03ba\u03ac"]
>>> for item in b:
   print item
απλό
τεστ
στα
ελληνικά
```

Tcl/Tk

- Η Tcl χρησιμοποιεί εγγενή αναπαράσταση σε UTF-8:

```
set fd [open file.txt]
fconfigure $fd –encoding utf-8
puts $fd “Καλημέρα Κόσμε!”
close $fd
```

```
> string is upper Λ
> 1
> string tolower ΚΑΛΗΜΕΡΑ
> καλημερα
```