ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ
ΕΡΓΑΣΤΗΡΙΟ 5

ΣΤΟΧΑΣΤΙΚΗ ΕΠΙΣΗΜΕΙΩΣΗ ΜΕΡΩΝ ΤΟΥ ΛΟΓΟΥ (POS TAGGING)

- Ανοίξτε το αρχείο με όνομα "corpus for pos tagging-words only.csv" με το UltraEdit.
- Το αρχείο αυτό είναι ένα κομμάτι από ένα αγγλικό σώμα κειμένων, το οποίο είναι χειρωνακτικά επισημειωμένο με πληροφορία μέρους του λόγου (ΜΤΛ).
- Συγκεκριμένα, τα χαρακτηριστικά σε κάθε παράδειγμα είναι τα εξής:

Λέξη-3	Λέξη-2	Λέξη-1	Λέξη εστίασης	Λέξη+1	Λέξη+2	Λέξη+3	ΜΤΛ
--	--	--	The	cafeteria	remains	closed	DT

Επομένως για κάθε λέξη της οποίας ψάχνω το ΜΤΛ (λέξη εστίασης) λαμβάνω υπόψη τις τρεις λέξεις που προηγούνται και τις τρεις που έπονται. Αυτό, το [-3, +3], ονομάζεται **παράθυρο συμφραζομένων** (context window).

- Οι επισημειώσεις (tags) των ΜΤΛ επεξηγούνται στον παρακάτω πίνακα:

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NP	Proper noun, singular
15.	NPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PP	Personal pronoun
19.	PP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present

```
33.     WDT     Wh-determiner
34.     WP      Wh-pronoun
35.     WP$     Possessive wh-pronoun
36.     WRB     Wh-adverb
```

- Ανοίξτε το Weka Experimenter. Στο Open File ανοίξτε το παραπάνω αρχείο.
- Βρείτε πόσες διαφορετικές τιμές παίρνει κάθε χαρακτηριστικό.
- Διαλέξτε τον αλγόριθμο IB1 για ταξινόμηση.
- Τρέξτε τον αλγόριθμο. Τι αποτελέσματα βγάζετε;
- Ποιά κλάση ταξινόμησης εμφανίζει τα καλύτερα αποτελέσματα; Γιατί, κατά την γνώμη σας;

- Ποιά κλάση ταξινόμησης εμφανίζει τα χειρότερα αποτελέσματα; Γιατί, κατά την γνώμη σας;

- Αλλάξτε το παράθυρο συμφραζομένων σε [-2, +2] και επαναλάβετε τα παραπάνω βήματα. Τι παρατηρείτε σε σχέση με το μεγαλύτερο παράθυρο;

- Αλλάξτε το παράθυρο συμφραζομένων σε [-1, +1]. Τι παρατηρείτε σε σχέση με τα προηγούμενα παράθυρα;

- Ανοίξτε το αρχείο με όνομα “corpus for pos tagging-prev tags and words.csv” με το UltraEdit.
- Σε αυτό το σετ δεδομένων, αντί για τις λέξεις που προηγούνται της λέξης εστίασης, χρησιμοποιούνται οι επισημειώσεις ΜΤΛ τους.

ΜΤΛ-3	**ΜΤΛ-2**	**ΜΤΛ-1**	**Λέξη εστίασης**	**Λέξη+1**	**Λέξη+2**	**Λέξη+3**	**ΜΤΛ**
--	--	DT	cafeteria	remains	closed	PERIOD	NN

- Τρέξτε τον αλγόριθμο IB1 στο καινούριο σώμα δεδομένων.
- Πώς είναι τα αποτελέσματα σε σχέση με το προηγούμενο σώμα δεδομένων; Γιατί, κατά την γνώμη σας;

- Τρέξτε και εδώ τον αλγόριθμο με διαφορετικά παράθυρα συμφραζομένων. Τι παρατηρείτε;

- Τρέξτε τον αλγόριθμο C4.5 (δέντρα αποφάσεων) με τα δύο σετ δεδομένων. Τι παρατηρείτε σε σχέση με τον IB1;

- Τι συμπεράσματα βγάζετε από το δέντρο αποφάσεων που προκύπτει; Ποιό είναι το πιο σημαντικό χαρακτηριστικό για την ταξινόμηση; Ποιό το λιγότερο σημαντικό;