ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ
ΕΡΓΑΣΤΗΡΙΟ 6

ΡΗΧΗ ΣΥΝΤΑΚΤΙΚΗ ΑΝΑΛΥΣΗ (SHALLOW PARSING)

- Ανοίξτε το αρχείο με όνομα "corpus for chunking.csv" με το UltraEdit.
- Το αρχείο αυτό είναι ένα κομμάτι από ένα αγγλικό σώμα κειμένων, το οποίο είναι χειρωνακτικά επισημειωμένο με πληροφορία για το εάν μια λέξη ανήκει σε μια φράση, και σε τι είδους φράση ανήκει (ονοματική, ρηματική, προθετική, επιθετική, επιρρηματική κλπ).
- Συγκεκριμένα, τα χαρακτηριστικά σε κάθε παράδειγμα είναι τα εξής:

Λέξη-2	**ΜΤΛ-2**	**Λέξη-1**	**ΜΤΛ-1**	**Λέξη εστίασης**	**ΜΤΛ**	**Λέξη+1**	**ΜΤΛ+1**	**Λέξη+2**	**ΜΤΛ+2**	**Επισημείωση φράσης**
Marshall	NNP	Hahn	NNP	Jr	NNP	has	VBZ	made	VBN	I-NP

Επομένως για κάθε λέξη της οποίας ψάχνω την επισημείωση φράσης (chunk tag) (λέξη εστίασης), λαμβάνω υπόψη τις δύο λέξεις που προηγούνται με τα μέρη του λόγου τους (ΜΤΛ), τις δύο που έπονται με τα ΜΤΛ τους και την ίδια την λέξη εεστίασης με το ΜΤΛ της. Αυτό, το [-2, +2], ονομάζεται **παράθυρο συμφραζομένων** (context window).

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

Chunk	**Εξήγηση**
I-NP	στο εσωτερικό (inside) μιας ονοματικής φράσης
I-VP	στο εσωτερικό (inside) μιας ρηματικής φράσης
I-PP	στο εσωτερικό (inside) μιας προθετικής φράσης
I-ADJP	στο εσωτερικό (inside) μιας επιθετικής φράσης
I-ADVP	στο εσωτερικό (inside) μιας επιρρηματικής φράσης
I-SBAR	στο εσωτερικό (inside) μιας δευτερεύουσας πρότασης
B-NP	αρχή (beginning) μιας ονοματικής φράσης
B-VP	αρχή (beginning) μιας ρηματικής φράσης
B-PP	αρχή (beginning) μιας προθετικής φράσης
B-ADJP	αρχή (beginning) μιας επιθετικής φράσης
B-ADVP	αρχή (beginning) μιας επιρρηματικής φράσης
B-SBAR	αρχή (beginning) μιας δευτερεύουσας πρότασης
B-PRT	αρχή (beginning) μορίου (η πρόθεση σε ένα phrasal verb)
O	έξω (outside) από φράση

- 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;

- Τρέξτε τον αλγόριθμο δέντρων αποφάσεων (C4.5). Επιλέξτε στον κατάλογο trees τον αλγόριθμο J48. Συγκρίνετε τα αποτελέσματα με αυτά του IB1. Τι παρατηρείτε;

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

- Αλλάξτε το παράθυρο συμφραζομένων σε [-2, +1] 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;

- Χρησιμοποιείστε σαν χαρακτηριστικά μόνο τα ΜΤΛ, και όχι τις λέξεις στο παράθυρο συμφραζομένων και τρέξτε πάλι τα πειράματα. Τι παρατηρείτε;

- Ανοίξτε το αρχείο corpus for chunking with chunktags for stacking. csv. Το αρχείο αυτό περιέχει και πληροφορία επισημείωσης φράσης για τις λέξεις του παραθύρου συμφραζομένων της λέξης εστίασης. Τρέξτε πειράματα με τον IB1 και τον C4.5 με αυτό το αρχείο δεδομένων. Τι παρατηρείτε σε σχέση με τα προηγούμενα αποτελέσματα; Γιατί;