Ανάκτηση Πληροφοριών και Εξόρυξη Δεδομένων 2016-2017

2η Εργασία, 04/04/2017
(Ομαδική εργασία των δύο ατόμων)

Ταξινόμηση και Συσταδοποίηση Κειμένων

Να επιλέξτε απο την διεύθυνση που σας δίνεται παρακάτω, ένα σετ δεδομένων για ανάλυση:

- https://github.com/niderhoff/nlp-datasets

Την Πέμπτη 06/04/2017 στις ώρες του εργαστηρίου να ενημερώσετε και να συζήτησετε με τον κ. Συμεωνίδη για το σετ των δεδομένων που επιλέξατε.

Με τη βοήθεια οποιασδήποτε γλώσσας (πχ. Python, R) να εφαρμόσετε στα δεδομένα τα εξής:

- Ανάλυση και περιγραφή του dataset (στατιστική ανάλυση).

- Προεπεξεργασία των δεδομένων, εξαγωγή και επιλογή χαρακτηριστικών (feature selection)
  - Πιο συγκεκριμένα, να προεπεξεργαστείτε τα δεδομένα και να γίνει η επιλογή των χαρακτηριστικών για την ταξινόμηση και την συσταδοποίηση.

- Αλγόριθμους Ταξινόμησης (classification)
  - Πιο συγκεκριμένα, να χρησιμοποιήσετε δύο αλγορίθμους ταξινόμησης (π.χ. SVM και Naive Bayes) με τις προκαθορισμένες ρυθμίσεις (ή εάν χρησιμοποιήσειτε κάποιες ρυθμίσεις να αναφέρετε ποιες ήταν στο παραδοτέο σας).
  - Να γίνει εξ'αρχής στο σύνολο δεδομένων διαχωρισμός σε δύο, το 80% των οποίων θα χρησιμοποιηθούν για την εκπαιδεύση των μοντέλων (train set) και το 20% που θα κρατηθεί ως σύνολο δεδομένων επικύρωσης (test set).
  - Να εξαχθούν οι ανάλογες μετρικές (πχ. accuracy, precision, recall, F1 score).

- Αλγόριθμους Συσταδοποίησης (clustering)
    - Πιο συγκεκριμένα, να εφαρμόσετε δύο αλγόριθμους συσταδοποίησης (π.χ. EM και SimpleKMeans) με τις προκαθορισμένες ρυθμίσεις στα δεδομένα που σας δίνονται.
    - Με βάση τις συστάδες που θα προκύψουν για το κάθε αλγόριθμο ξεχωριστά να προσπαθήσετε να βγάλετε κάποια χρήσιμα συμπεράσματα που μπορεί να αφορούν συστάδες (πχ. μέτρηση σύμπτωσης των clusters με τις κλάσεις) ή την συλλογή (πχ τι θεματολογία φαίνεται να προκύπτει για την συλλογή μέσω της συσταδοποίησης).

Παράδοση και Εξέταση:

Η παράδοση και εξέταση θα γίνει ανά ομάδα. Το παραδοτέο θα είναι ένα zip αρχείο που θα περιλαμβάνει μια αναφορά (σε μορφή PDF) με τις ρυθμίσεις, τα συμπεράσματα και τα αποτελέσματα σας, και επιπλέον τα αρχεία της επεξεργασίας, της ταξινόμησης και της συσταδοποίησης. Θα πρέπει να το ανεβάσετε στο eclass μέχρι τις 23:59 της Τετάρτης 24/05/2017. Η εξέταση θα πραγματοποιηθεί την Πέμπτη 25/05/2017 και ώρες 16:00-18:00 σύμφωνα με το πρόγραμμα που θα ανακοινωθεί στο eclass.

Βοηθητικοί σύνδεσμοι:
Παράδειγμα χρήσης Python:

- https://www.datacamp.com/community/tutorials/machine-learning-python#gs.Zd4EwL8
- http://www.kdnuggets.com/2015/11/seven-steps-machine-learning-python.html

Παράδειγμα χρήσης R:

- https://www.datacamp.com/community/tutorials/machine-learning-in-r#gs.Mu8Fq4Y
- http://machinelearningmastery.com/machine-learning-in-r-step-by-step/

Σχετική ύλη από το βιβλίο "Introduction to Information Retrieval":

- Ch.13 (από την αρχή μέχρι και όλο το 13.1, και τα 13.5-6)
- Ch.14 (εκτός από τα 14.3.1 και 14.6)
- Ch.16 (εκτός από 16.4.1 και 16.5)

Ο διδάσκων
Αυγερινός Αραμπατζής