Π.Μ.Σ. Πληροφοριακά Συστήματα και Υπηρεσίες
Ειδίκευση: Προηγμένα Πληροφοριακά Συστήματα
Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Πειραιώς
Εργασία Μαθήματος: Αποθήκες Δεδομένων και Επιχειρηματική Ευφυία

Διερευνητική Αναλυτική με Δεδομένα από το Twitter

Αναπληρωτής Καθηγητής Χρήστος Δουλκερίδης

Στόχος της εργασίας είναι η εφαρμογή μεθόδων διερευνητικής αναλυτικής δεδομένων, όπως αυτές που διδαχθήκατε στο μάθημα, έτσι ώστε να παρουσιάσετε ενδιαφέροντα αποτελέσματα από ένα δοθέν σύνολο δεδομένων και έτσι να μπορεί να κατανοήσει κανείς καλύτερα το σύνολο δεδομένων.

Πιο συγκεκριμένα, η εργασία αφορά στην ανάλυση ενός μεγάλου συνόλου δεδομένων από το Twitter το οποίο σας δίνεται διαθέσιμο στον ΛΕΥΚΙΠΠΟ (περιοχή: *Έγγραφα*). Το σύνολο δεδομένων αφορά tweets που έχουν συλλεχθεί στην περιοχή της Ολλανδίας, επομένως το κείμενο μπορεί να είναι στα ολλανδικά. Σε πρώτη φάση, δίνεται ένα μικρό σύνολο 100 tweets για να δοκιμάσετε τον κώδικά σας και σε δεύτερη φάση θα γίνει διαθέσιμο ένα μεγαλύτερο σύνολο δεδομένων (με τον ίδιο ακριβώς μορφότυπο).

1 Ανάλυση Δεδομένων

Καλείστε να εφαρμόσετε τεχνικές ανάλυσης δεδομένων ώστε να εξάγετε χρήσιμη πληροφορία και συμπεράσματα. Ενδεικτικές τεχνικές και εργαλεία που μπορείτε να χρησιμοποιήσετε αποτελούν: στατιστική ανάλυση, οπτικοποιήσεις, ανάλυση χρονοσειράς, γεωγραφική ανάλυση, καθώς και μέθοδοι μάθησης όπως συσταδοποίηση ή ανίχνευση ανωμαλιών.

Πιο συγκεκριμένα, ορισμένες *ενδεικτικές* ιδέες για ανάλυση δεδομένων ακολουθούν:

- εντοπισμός ασυνήθιστης δραστηριότητας βάσει πλήθους tweets και χρονικής περιόδου
- ανακάλυψη ανθρώπινης δραστηριότητας βάσει κειμένου ή hashtags και χρονικής περιόδου
- ανάλυση θέσης των tweets βάσει γεωγραφικών συντεταγμένων και χρονικής περιόδου
- ομαδοποίηση των tweets σε ομάδες βάσει διαφόρων κριτηρίων
- πρόβλεψη πλήθους tweets ανά χρονική περίοδο (π.χ. ανά ώρα)
- ανάλυση κειμένου για τον εντοπισμό των κύριων θεμάτων συζήτησης (και με χρήση μετάφρασης στα αγγλικά, αν χρειαστεί)
- ανάλυση συναισθήματος βάσει του κειμένου
- ...

Τα αποτελέσματα της ανάλυσης δεδομένων θα αποτυπωθούν σε τεχνική αναφορά (pdf) όπου **απαραιτήτως θα εξηγούνται με σαφήνεια τα βήματα που ακολουθήθηκαν**. Είναι απαραίτητο και αποτελεί κομμάτι της αξιολόγησης **η χρήση κατάλληλων οπτικοποιήσεων με μορφή διαγραμμάτων**. Επιπλέον, αναμένεται **να διατυπωθούν συμπεράσματα που εξήχθησαν** από το δοθέν σύνολο δεδομένων.

2 Διαδικαστικά Θέματα – Αξιολόγηση

Θα χρησιμοποιηθεί το πρότυπο της ACM με όριο τις 8 σελίδες (δίστηλο): `https://www.acm.org/publications/proceedings-template`.

- Η εργασία θα εκπονηθεί σε ομάδες μέχρι 2 ατόμων
- Η παράδοση των εργασιών (αρχείο pdf και κώδικας) θα γίνει μέσω του ΛΕΥΚΙΠΠΟΥ
- Ημερομηνία παράδοσης βάσει προγράμματος εξεταστικής Ιουνίου
- Η εργασία προσμετράται στον τελικό βαθμό του μαθήματος με συντελεστή 50%