Ποσοτική Ανάλυση Γλωσσικών Δεδομένων

2η εργασία (Inferential Statistics)

1) Αξιοποιείστε τα αποτελέσματα της 1ης εργασίας (δηλαδή τα αποτελέσματα του QUITA) για να εκτιμήσετε ένα διάσημα εμπιστοσύνης για τη μέση τιμή των ακόλουθων δεικτών: TTR, Entropy, και Average Token Length.

2α) Επιβεβαιώστε τα αποτελέσματα του ερωτήματος (1) για εφαρμόζοντας τα κατάλληλα tests.

2β) Παρουσιάστε 2 ενδεικτικά διαγράμματα που αναπαριστούν γραφικά ένα αποτέλεσμα απόρριψης και ένα αποτέλεσμα μη απόρριψης της NULL Hypothesis.

3) Συνεργαστείτε με μια άλλη ομάδα που έχει επιλέξει διαφορετικό κείμενο κατά την 1η εργασία έτσι ώστε να σας είναι διαθέσιμα τα αποτελέσματά της από το QUITA. Ελέγξτε αν οι μέσες τιμές των δεικτών TTR, Entropy, και Average Token Length διαφέρουν (στατιστικά) σημαντικά. Θεωρείστε ότι τα δείγματα είναι ανεξάρτητα. (Ως δείγμα 1 θεωρείστε τα δικά σας αποτελέσματα και ως δείγμα 2 τα αποτελέσματα της άλλης ομάδας).

Οδηγίες

- Τα τμήματα κώδικα θα πρέπει να υλοποιηθούν σε python.

- Είναι απαραίτητος ο σχολιασμός των επιλογών σας και των αποτελεσμάτων (π.χ. Ποια κατανομή επιλέχθηκε και γιατί; Ποιο test εφαρμόστηκε και γιατί;)

- Η εργασία θα πρέπει να παραδοθεί σε αρχείο που θα επιτρέπει την εκτέλεση του κώδικα που αναπτύξατε και να περιέχει τα σχόλια σας (π.χ. ipnyb).

- Η εργασία θα πρέπει να περιέχει τα αρχεία με τις τιμές των δεικτών και της 1ης εργασίας σας και της 1ης εργασίας της άλλης ομάδας.

- Η εργασία μπορεί να γίνει ατομικά ή σε ομάδες δύο μελών.