Εξαμηνιαία εργασία Θεωρίας Πληροφορίας
Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων –
Πανεπιστήμιο Αιγαίου

Κωνσταντίνος Μαλιάτσος – Επίκουρος Καθηγητής

Οδηγίες:

Μπορείτε να χρησιμοποιήσετε οποιαδήποτε γλώσσα προγραμματισμού προτιμάτε. Θα πρέπει επίσης να υποβάλετε μαζί με την εργασία πολύ σύντομες οδηγίες σχετικά με τον τρόπο εκτέλεσης/compiling του κώδικα. Αν αποφασίσετε να χρησιμοποιήσετε ένα εξειδικευμένο πακέτο-βιβλιοθήκη, πρέπει να δώσετε λεπτομέρειες σχετικά με τον τρόπο σύνδεσής/χρήσης του από το project σας.

Χρησιμοποιήστε εκτενώς σχόλια στον κώδικα σας. Η αποτελεσματικότητα του κώδικα δεν βαθμολογείται. Ωστόσο, η παρουσίασή του ναι. Αν ο κώδικας είναι ευκολοδιάβαστος, τότε ο βαθμός θα είναι υψηλότερος.

Παραδίδεται:

- Ο κώδικας μέσω του eclass.
- Μια παράγραφος ανά ερώτημα με εξήγηση/περιγραφή του τι κάνατε και γιατί.

Όταν χρησιμοποιείτε εξωτερικές πηγές, χρησιμοποιήστε αναφορές!

Εργασία:

Κατεβάστε το αρχεία από τον παρακάτω σύνδεσμο:

https://eclass.icsd.aegean.gr/modules/document/file.php/ICSD549/%CE%95%CF%81%CE%B3%CE%B1%CF%83%CE%AF%CE%B1%202023/files.zip

Ο φάκελος εκτός από την εκφώνηση περιέχει τρία αρχεία. Αρχικά εργάζεστε με τα hamlet.txt και matrix.txt. Τα αρχεία περιέχουν κείμενα από χαρακτήρες ASCII (χαρακτήρες του λατινικού αλφαβήτου, αριθμητικοί, το κενό, σημεία στίξης), αλλά και ειδικών χαρακτήρων (tab, end-of-line, end-of-file κλπ.).

Για να επιλύσετε αυτό το θέμα, θα χρειαστεί να γράψετε ένα πρόγραμμα.

Συμβουλή: Για να μπορέσετε να επικυρώσετε τη λειτουργικότητα του κώδικα σας, καλό θα ήταν να δημιουργήσετε ένα «μικρότερο» παράδειγμα. Π.χ. ένα κείμενο 3 χαρακτήρων όπου μπορείτε να ελέγξετε την ορθότητα του αποτελέσματος. Με αυτόν τον τρόπο μπορείτε να βρείτε πιθανά σφάλματα.

Ερώτημα 0: (Προπαρασκευαστικές εργασίες)

- Ανοίχτε και «διαβάστε» το αρχείο hamlet.txt και matrix.txt μέσω του προγράμματος σας.
- Αφαιρέστε όλους τους χαρακτήρες πλην των λατινικών χαρακτήρων, των (μονών-απλών) κενών, και των «.» και «,». Όποιον χαρακτήρα αφαιρείται από τους υπόλοιπους, τον αντικαταστείτε με το κενό.
- Μετατρέψτε όλα τα κεφαλαία γράμματα των κειμένων σε πεζά.

- Εκτιμήστε το μέγεθος του συνόλου χαρακτήρων. Ποιος ο πληθυσμός χαρακτήρων?
- Διαβάστε το αρχείο και υπολογίστε τον αριθμό των συμβόλων που περιέχονται στο αρχείο.
- Ιδανικά χρησιμοποιήστε το σύνολο των χαρακτήρων. Αν έχετε περιορισμούς μνήμης, περιορίστε τον αριθμό των εγγραφών που επεξεργάζεστε.

Συμβουλή: Εργαστείτε με ASCII. Όλες οι γλώσσες προγραμματισμού, έχουν μετατροπή χαρακτήρων σε ASCII. Τα κεφαλαία λατινικά έχουν δεκαδικό κωδικό από 65 έως 90. Τα πεζά λατινικά γράμματα έχουν από 97 έως 122. Κατά συνέπεια, π.χ. η μετατροπή κεφαλαίου σε πεζό γίνεται με πρόσθεση 22 (A+22=65+22=97=a)!. Στο ASCII αλφάβητο θα βρείτε και τα ειδικά σύμβολα για κάθε γλώσσα.

Σημείωση: Υπάρχει ο χαρακτήρας EOF στο τέλος του κάθε αρχείου που σηματοδοτεί το τέλος του εγγράφου. Αποκλείστε το από την ανάλυση σας καθώς είναι με 100% πιθανότητα το τελευταίο σύμβολο.

Ερώτημα 1:

- Καταμετρήστε πόσες φορές εμφανίζεται κάθε χαρακτήρας στο αρχείο hamlet.txt.
- Εκτιμήστε την κατανομή πιθανότητας $p_{hamlet}(x_n)$ θεωρώντας τους χαρακτήρες ως τυχαίες μεταβλητές ανεξάρτητες μεταξύ τους – βασιζόμενοι στο αρχείο hamlet.txt.
- Υπολογίστε την εντροπία της πηγής $H_{hamlet}(x_n)$ που προκύπτει από το αρχείο hamlet.txt
- Καταμετρήστε πόσες φορές εμφανίζεται κάθε χαρακτήρας στο αρχείο matrix.txt.
- Εκτιμήστε την κατανομή πιθανότητας $p_{matrix}(x_n)$ θεωρώντας τους χαρακτήρες ως τυχαίες μεταβλητές ανεξάρτητες μεταξύ τους – βασιζόμενοι στο αρχείο hamlet.txt.
- Υπολογίστε την εντροπία της πηγής $H_{matrix}(x_n)$

Ερώτημα 2:

- Για το αρχείο hamlet.txt:
    - Υπολογίστε την κατανομή σε ζεύγη γειτονικών χαρακτήρων $p(x_n, x_{n+1})$ θεωρώντας την ακόλουθη διαδικασία: Καταμετρήστε τα ζεύγη σειριακά αλλά προσοχή! Η λέξη George δεν περιέχει μόνο τα ζεύγη Ge-or-ge, αλλά Ge-eo-or-rg-ge!
    - Επίσης το κενό και τα δυο εναπομείναντα σημεία στίξης συμμετέχει στις καταμετρήσεις, σαν να είναι γράμμα.
    - Καταμετρήστε τον αριθμό των ζευγών που προέκυψαν.
    - Χρησιμοποιήστε το για να καταμετρήστε το ποσοστό εμφάνισης του κάθε ζεύγους. Συνεχίστε επιλέγοντας ένα άλλο ζεύγος κ.ο.κ.
    - Υπολογίστε την από κοινού εντροπία $H(x_n, x_{n+1})$
    - Είναι μικρότερη του $2H(x_n)$; Αν ναι γιατί;
    - Υπολογίστε την υπό συνθήκη εντροπία $H(x_n|x_{n+1})$.
- Επαναλάβετε για το αρχείο matrix.txt

Ερώτημα 3:

- Φτιάξτε έναν σταθερού μήκους κώδικα (ανά χαρακτήρα στο κείμενο) και για τα δυο αρχεία (λογικά δεν θα έχει κάποια διαφορά και μπορείτε να χρησιμοποιήσετε τον ίδιο κώδικα και για τα δυο αρχεία).
- Φτιάξτε έναν σταθερού μήκους κώδικα (ανά ζεύγος χαρακτήρων στο κείμενο).

Ερώτημα 4:

- Για το αρχείο hamlet.txt:
    - Θεωρώντας ότι οι χαρακτήρες είναι ανεξάρτητοι και χρησιμοποιώντας την κατανομή του Ερωτήματος 1, υλοποιήστε την κωδικοποίηση Shannon-Fano, ή Shannon ή Huffman.
    - Εφαρμόστε τον κωδικοποιητή και προσδιορίστε το μέγεθος του συμπιεσμένου αρχείου.
    - Ποιο το μέσο μήκος συμβόλου του κώδικα?
- Επαναλάβετε για το αρχείο matrix.txt:

Ερώτημα 5:

- Για το αρχείο hamlet.txt:
    - Επαναλάβετε την υλοποίηση Shannon-Fano (ή Shannon ή Huffman) θεωρώντας τα ζεύγη χαρακτήρων ως σύμβολα του κώδικα.
    - Εφαρμόστε τον κωδικοποιητή και προσδιορίστε το μέγεθος του συμπιεσμένου αρχείου.
    - Ποιο το μέσο μήκος συμβόλου του κώδικα?
    - Ποιος κώδικας αποδίδει καλύτερα? (Ερώτημα 4 vs. 5)
- Επαναλάβετε για το αρχείο matrix.txt:

Ερώτημα 6:

- Εφαρμόστε τους κώδικες που φτιάξατε στο νέο αρχείο pulp_fiction.txt, αφού έχετε κάνει μετατροπή από κεφαλαία σε πεζά και αφαιρέσει όλους τους υπόλοιπους χαρακτήρες εκτός από τα πεζά λατινικά, τα κενά, τα κόμματα και τις τελείες.
- Ποιος κώδικας από όλους αποδίδει καλύτερα στο νέο αρχείο? Μπορείτε να δώσετε μια λογική εξήγηση για τα αποτελέσματα?
- Σύμφωνα με τη Θεωρία Πληροφορίας, και την μέχρι τώρα ανάλυση, η γραφή του Tarantino είναι πιο «συγγενής» με του Shakespeare ή των Wachowski's?
- Πόσο χειρότερα ή καλύτερα συμπεριφέρεται ο κώδικας στο νέο αρχείο σε σχέση με τα αρχεία αναφοράς και γιατί?

Ερώτημα 7:

- Θέλουμε να μεταδώσουμε το αρχείο pulp_fiction.txt πάνω από θορυβώδες κανάλι σηματοθορυβικού λόγου 7 dB.
- Θεωρώντας ότι για bit 0 αποστέλλεται η στάθμη σήματος -1 (π.χ. σε Volt) και για το bit 1 αποστέλλεται η στάθμη σήματος 1 (BPSK διαμόρφωση), φτιάξτε το σύστημα μετάδοσης και προσθέστε το θόρυβο σύμφωνα με την εκφώνηση της άσκησης.
- Εξομοιώστε τη διαδικασία λήψης με χρήση hard απόφασης (επιλογή του συμβόλου με την μικρότερη απόσταση). Υπάρχουν σφάλματα μετάδοσης? Αν ναι πόσα?

Ερώτημα 8:

- Εφαρμόστε κωδικοποίηση καναλιού Hamming με μήκος κωδικοποιημένης λέξης τα 15 bits.
- Εισάγετε θόρυβο και λήψη με hard απόφαση.
- Πραγματοποιήστε αποκωδικοποίηση Hamming ανά 15bits και μετρήστε τα σφάλματα.
- Ποιο το όφελος και ποιο το κόστος συγκριτικά με το ερώτημα 7.

Ερώτημα 9:

- Φτιάξτε ένα πρωτόκολλο, όπου σε περίπτωση λανθασμένης λήψης γίνεται επανάληψη της αποστολής του ίδιου Hamming block.
- Θεωρώντας κανάλι ταχύτητας 1 Msymbol/sec (για BPSK είναι ίδιο με Mbps) και εύρος ζώνη 1MHz, ποιος ο επιτεύξιμος ρυθμός μετάδοσης πληροφορίας και ποια η χωρητικότητα όπως υπολογίζεται θεωρητικά.