ΣΥΜΠΙΕΣΗ ΔΕΔΟΜΕΝΩΝ – ΕΡΓΑΣΙΑ 1

ΚΩΔΙΚΟΠΟΙΗΣΗ ΠΗΓΗΣ

Σε ένα διαγωνισμό συμπίεσης δεδομένων καλείστε να συμπιέσετε τις λατινικές μεταφράσεις κειμένων της αρχαίας ελληνικής γραμματείας. Το κείμενο που θα χρησιμοποιηθεί για την αποτίμηση των αποτελεσμάτων των προτεινόμενων αλγορίθμων είναι μία μετάφραση των Φυσικών του Αριστοτέλη την οποία μπορείτε να βρείτε στο αρχείο (http://classics.mit.edu/Aristotle/physics.mb.txt)

Στην προσέγγιση σας αγνοήστε τους αριθμούς, παρενθέσεις και σημεία στίξης, θεωρήστε ότι κεφαλαίοι και πεζοί χαρακτήρες αντιστοιχούν στο ίδιο σύμβολο αλλά μην αγνοήσετε τα κενά. Θεωρήστε επίσης ότι οι χαρακτήρες αλλαγών γραμμής αντιστοιχούν σε κενά.

ΕΡΩΤΗΜΑΤΑ

1. (10%)Χρησιμοποιήστε τις πιθανότητες εμφάνισης του Πίνακα 1 για να υπολογίσετε την εντροπία του λατινικού αλφάβητου

2. (15%)Να θεωρήσετε ένα συνεχώς αυξανόμενο πλήθος συμβόλων (τουλάχιστον 5) από το κείμενο ώστε να καλυφθεί όλο το κείμενο (π.χ. N=100,1000,10000,100000, μέγεθος κειμένου) και να υπολογίσετε από κάθε σύνολο συμβόλων την εντροπία του λατινικού αλφάβητου. Τι συμπεραίνετε σε σχέση με την εντροπία που υπολογίσατε στο προηγούμενο ερώτημα.

ΠΙΝΑΚΑΣ 1:ΣΥΧΝΟΤΗΤΕΣ ΛΑΤΙΝΙΚΟΥ ΑΛΦΑΒΗΤΟΥ

a	0,065	g	0,016	m	0,020	s	0,052	y	0,015
b	0,012	h	0,049	n	0,056	t	0,073	z	0,001
c	0,022	i	0,056	o	0,060	u	0,023	κενό	0,191
d	0,035	j	0,001	p	0,014	v	0,008		
e	0,104	k	0,005	q	0,001	w	0,017		
f	0,020	l	0,033	r	0,050	x	0,001		

3. (15%)Να υπολογίσετε για τους δυαδικούς κώδικες (Huffman και Αριθμητικό) τις πιθανότητες εκπομπής των συμβόλων 0/1. Ειδικά για τον κώδικα Huffman να υπολογιστεί η πιθανότητα εκπομπής των συμβόλων τόσο με χρήση των κωδικών λέξεων του δέντρου όσο και με βάση τις πιθανότητες από την κωδικοποιημένη ακολουθία της πηγής.
4. (20%)Να χρησιμοποιήσετε την τεχνική RLE για να κωδικοποιήσετε περαιτέρω το αποτέλεσμα που προκύπτει στο ερώτημα 3. Υπάρχει βελτίωση στην απόδοση και γιατί;
5. (40%)Θεωρήστε ότι πλέον του αλφαβήτου που περιέχεται στον Πίνακα 1 υπάρχουν και τα ακόλουθα σύμβολα που συνοψίζονται στον πίνακα 2 τα οποία εφόσον εμφανίζονται στο κείμενο κωδικοποιούνται κατά προτεραιότητα σε σχέση με τα γράμματα που τα αποτελούν.

a. (15%)Να υπολογίσετε τις πιθανότητες εμφάνισης των νέων συμβόλων (Πίνακας 1 και Πίνακας 2) με βάση τη στατιστική όλου του κειμένου σας.
b. (25%)Να χρησιμοποιήσετε το νέο σύνολο συμβόλων για την κωδικοποίηση του κειμένου με δυαδικό αριθμητικό κώδικα και συγκρίνετε τα αποτελέσματα της μεθόδου κωδικοποίησης σε σχέση με την επέκταση της πηγής με L=2. Χρησιμοποιήστε τον πλεονασμό ως μέτρο σύγκρισης μεταξύ των δύο επιλογών. Ποια μέθοδο θα επιλέγατε;

ΠΙΝΑΚΑΣ 2: ΣΥΜΠΛΗΡΩΜΑΤΙΚΑ ΣΥΜΒΟΛΑ ΛΑΤΙΝΙΚΟΥ ΑΛΦΑΒΗΤΟΥ

and	for	of	the	with	ch	gh	sh	th

ΣΗΜΕΙΩΣΕΙΣ - ΔΙΑΔΙΚΑΣΙΑ ΠΑΡΑΔΟΣΗΣ

Για κάθε σχήμα κωδικοποίησης που θα χρησιμοποιήσετε να αναπτύξετε και τον απαραίτητο αποκωδικοποιητή και να ελέγξετε ότι αποκωδικοποιείται σωστά το κωδικοποιημένο αρχείο.

Στην εργασία σας θα παραδώσετε μία αναφορά μεγέθους από 3 - 5 σελίδες που να περιέχει συνοπτική περιγραφή των τεχνικών που θα χρησιμοποιήσετε, η δομή ενδεχόμενων αρχείων εξόδου των κωδικοποιητών, τα πειράματα και τα συμπεράσματα σας σε σχέση με τα ερωτήματα που τίθενται στην εκφώνηση. Για τη σύνταξη του κειμένου σας να ακολουθήσετε το πρότυπο που βρίσκεται στη δικτυακή θέση : http://www.ieee.org/documents/MSW_A4_format.doc

Μαζί με την αναφορά θα παραδώσετε και τον κώδικα που αναπτύξατε και πιθανό εκτελέσιμο που μπορεί να έχει προκύψει. (Η ανάπτυξη μπορεί να γίνει σε οποιοδήποτε προγραμματιστικό περιβάλλον επιθυμείτε).

Η εργασία πρέπει να παραδοθεί σε ηλεκτρονική μορφή στο e-mail: nsgouros@outlook.com έως 21/4/2015 και ώρα 23.59.59.

Σε περίπτωση που υποβληθεί μετά την παραπάνω καταληκτική ημερομηνία ο τελικός βαθμός της εργασίας (ΤΒΕ) θα προκύψει συναρτήσει του βαθμού που θα λάβει (ΒΕ) και των ωρών (Ω) που μεσολάβησαν μεταξύ της παραπάνω ημερομηνίας και της ημερομηνίας παράδοσης όπως καθορίζεται από την ώρα παραλαβής του e-mail σύμφωνα με τον παρακάτω τύπο:

$$TBE = BE \cdot 2^{-45 \cdot 10^{-4} \cdot \Omega}$$

Καλή επιτυχία

26/3/2015
Ν. Σγούρος