Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Πληροφορικής
Μάθημα: Προγραμματισμός Ηλεκτρονικών Υπολογιστών με C++
Ακαδημαϊκό έτος: 2011-12
Διδάσκων: Ι. Ανδρουτσόπουλος

1η Εργασία

Δίνεται μια συλλογή μηνυμάτων ηλεκτρονικού ταχυδρομείου. Η συλλογή αποτελείται από ανεπιθύμητα διαφημιστικά μηνύματα (ΑΔΜ, spam) και μηνύματα που διακινήθηκαν μέσω της λίστας ηλεκτρονικού ταχυδρομείου Linguist (http://www.linguistlist.org/). Οι συντονιστές της Linguist διενεργούν ελέγχους που εγγυώνται ότι δεν διακινούνται ΑΔΜ μέσω της λίστας. Τα μηνύματα που δεν είναι ΑΔΜ, στην περίπτωσή μας τα μηνύματα που προέρχονται από τη Linguist, θα λέγονται «επιθυμητά» (ham). Τα ΑΔΜ της συλλογής βρίσκονται στο συμπιεσμένο φάκελο αρχείων **spam.tar.gz** και τα επιθυμητά στο φάκελο **ling.tar.gz**. Οι φάκελοι αυτοί περιέχουν ένα μήνυμα ανά αρχείο. Για τη διευκόλυνσή σας, δίνονται επίσης τα αρχεία **ling_filenames.txt** και **spam_filenames.txt**, τα οποία περιέχουν τα ονόματα των αρχείων των δύο προηγουμένων φακέλων. Δίνεται επίσης το αρχείο **keywords.txt**, το οποίο περιέχει λέξεις-κλειδιά που είναι χαρακτηριστικές είτε για ΑΔΜ είτε για μηνύματα της λίστας Linguist. Το αρχείο αυτό περιέχει μια λέξη-κλειδί ανά γραμμή. Δεν γίνεται διαχωρισμός μεταξύ λέξεων-κλειδιών που είναι χαρακτηριστικές της μίας ή της άλλης κατηγορίας μηνυμάτων.

Ζητείται να γράψετε ένα πρόγραμμα σε C++, το οποίο να επεξεργάζεται τη συλλογή μηνυμάτων της εργασίας και να τυπώνει στο cout τα ακόλουθα:

```
< messagecollection messages = "n" >

< message file = "…" category = "spam" features = "m" >
< feature token = "k1" id = "i1" freq = "fi1" />
< feature token = "k2" id = "i2" freq = "fi2" />
…
< feature token = "km" id = "im" freq = "fim" />
</ message >

…

< message file = "…" category = "ham" features = "r" >
< feature token = "q1" id = "j1" freq = "fj1" />
< feature token = "q2" id = "j2" freq = "fj2" />
…
< feature token = "qr" id = "jr" freq = "fjr" />
</ message >

…

</ messagecollection >
```

Πρέπει να υπάρχουν συνολικά *n* ενότητες <message … > … </message>, μία για κάθε μήνυμα της συλλογής. Το πεδίο file κάθε ενότητας έχει ως τιμή (μέσα στα εισαγωγικά) το όνομα του αρχείου που περιέχει το αντίστοιχο μήνυμα, μαζί με το φάκελο-πατέρα του (αλλά όχι τους πιο πάνω φακέλους, π.χ. «file="spam/spmsga1.txt"» ή «file="ling/3-1msg1.txt"»). Πρέπει να τυπώνονται πρώτα οι ενότητες <message … > … </message> των ΑΔΜ (με τη σειρά που εμφανίζονται στο spam_filenames.txt) και στη συνέχεια οι ενότητες των επιθυμητών μηνυμάτων (με τη σειρά που εμφανίζονται στο ling_filenames.txt). Το category κάθε ενότητας δείχνει αν η ενότητα παριστάνει επιθυμητό μήνυμα ή μήνυμα ΑΔΤ. Το features δείχνει πόσες λέξεις-κλειδιά του keywords.txt περιέχει το αντίστοιχο μήνυμα. Αν μια λέξη-κλειδί εμφανίζεται πολλές φορές σε ένα μήνυμα, μετριέται μόνο μία φορά στο features.

Σε κάθε ενότητα <message … > … </message> πρέπει να υπάρχει μία υποενότητα <feature … /> για κάθε λέξη-κλειδί που εμφανίζεται στο αντίστοιχο μήνυμα. Αν μια λέξη-κλειδί εμφανίζεται πολλές φορές σε ένα μήνυμα, παράγεται μόνο μια υποενότητα <feature … /> για αυτή τη λέξη-κλειδί στην ενότητα <message … > … </message> του μηνύματος. Τα *k1*, *k2*, ..., *km* (αντίστοιχα *q1*, *q2*, …, *qr*) πρέπει να είναι οι λέξεις-κλειδιά που εμφανίζονται στο μήνυμα. Τα *i1*, *i2*, ..., *im* (αντίστοιχα *j1*, *j2*, …, *jr*) πρέπει να είναι οι αριθμοί των γραμμών του αρχείου keywords.txt που αντιστοιχούν στα *k1*, *k2*, ..., *km* αντίστοιχα (ή στα *q1*, *q2*, …, qr). Η αρίθμηση των γραμμών του keywords.txt να ξεκινάει από το 0. Τα *fi1*, *fi2*, ..., *fim* (αντίστοιχα *fj1*, *fj2*, …, *fjr*) είναι θετικοί ακέραιοι που δείχνουν πόσες φορές εμφανίζονται οι αντίστοιχες λέξεις-κλειδιά στο συγκεκριμένο μήνυμα. Σε κάθε ενότητα <message … > … </message>, οι υποενότητες <feature … /> πρέπει να είναι ταξινομημένες κατά αύξουσα σειρά των *i1*, *i2*, ..., *im* (αντίστοιχα *j1*, *j2*, …, *jr*).

Τα μηνύματα της συλλογής έχουν υποστεί προεπεξεργασία που αφαίρεσε όλα τα τμήματα των κεφαλίδων (header) τους εκτός από το θέμα (Subject), μετέτρεψε όλα τα κεφαλαία γράμματα του θέματος και του κυρίου μέρους σε πεζά και χώρισε τα σημεία στίξης και άλλους ειδικούς χαρακτήρες (π.χ. "$") από τις υπόλοιπες λέξεις με κενά. Για τους σκοπούς της εργασίας, λέξη θεωρείται κάθε ακολουθία μη-κενών χαρακτήρων που χωρίζεται από τους υπόλοιπους χαρακτήρες του μηνύματος με κενό ή αλλαγή γραμμής (π.χ. το κείμενο "*from $ 5 . 00 to $ 15*" περιέχει 8 λέξεις). Ένα μήνυμα θεωρείται ότι περιέχει μια λέξη-κλειδί, αν η λέξη-κλειδί εμφανίζεται ως αυτοτελής λέξη μέσα στο μήνυμα, με την ίδια ακριβώς μορφή που έχει στο keywords.txt.

Για να αποσυμπιέσετε τα spam.tar.gz και ling.tar.gz σε Windows, μπορείτε να χρησιμοποιήσετε ένα πρόγραμμα σαν το WinZip (http://www.winzip.com/). Σε Unix, χρησιμοποιήστε τις εντολές:

gunzip spam.tar.gz
tar xvf spam.tar

gunzip ling.tar.gz
tar xvf ling.tar

Φροντίστε το πρόγραμμά σας να περιέχει επαρκή σχόλια. Δομήστε το πρόγραμμά σας ιεραρχικά, χρησιμοποιώντας συναρτήσεις που να αντιστοιχούν σε υποεργασίες, υπο-υποεργασίες κλπ., ώστε να είναι εύκολα κατανοητός ο ρόλος κάθε συνάρτησης και οι συναρτήσεις να είναι σύντομες. Μπορείτε να υποθέσετε ότι όλα τα αρχεία έχουν την απαιτούμενη μορφή, χωρίς να περιλάβετε στο πρόγραμμά σας σχετικούς ελέγχους. Χρησιμοποιήστε την εντολή const για να ορίσετε σταθερές (π.χ. ονόματα αρχείων) που ενδεχομένως να χρειαστεί να αλλάξουν τιμές στο μέλλον.

Για να ανοίξετε ένα αρχείο του οποίου το όνομα είναι αποθηκευμένο σε ένα string *filename*, χρησιμοποιήστε την ακόλουθη σύνταξη που μετατρέπει το string στην παλιότερη μορφή πίνακα χαρακτήρων:

ifstream inFile(filename.c_str());
ofstream outFile(filename.c_str());

Προσοχή: Το πρόγραμμά σας θα ελεγχθεί σε μεγάλο βαθμό αυτόματα χρησιμοποιώντας διαφορετικά δεδομένα εισόδου. Φροντίστε η έξοδος του προγράμματός σας να συμφωνεί απόλυτα με τις παραπάνω προδιαγραφές.

Προσοχή: Διαβάστε οπωσδήποτε το έγγραφο «Γενικές πληροφορίες για τις εργασίες του μαθήματος» (αρχείο «cpp_assignments_general_info.pdf»), που βρίσκεται στα έγγραφα του μαθήματος στο e-class. Μεταξύ άλλων, περιλαμβάνει οδηγίες για τον τρόπο παράδοσης και εξέτασης των εργασιών.

Προσοχή: Το πρόγραμμά σας θα πρέπει να θεωρεί ότι κατά την εκτέλεσή του θα βρίσκονται στον ίδιο φάκελο με αυτό:

- Ένας υποφάκελος με όνομα «ling», με αρχεία της ίδιας μορφής με εκείνα του ling.tar.gz. Κατά τον έλεγχο του προγράμματός σας ενδέχεται, όμως, τα αρχεία αυτά να περιέχουν διαφορετικά μηνύματα και τα μηνύματα να είναι περισσότερα ή λιγότερα από εκείνα του ling.tar.gz.
- Ένας υποφάκελος με όνομα «spam», με αρχεία της ίδιας μορφής με εκείνα του spam.tar.gz. Κατά τον έλεγχο του προγράμματός σας ενδέχεται, όμως, τα αρχεία αυτά να περιέχουν διαφορετικά μηνύματα και τα μηνύματα να είναι περισσότερα ή λιγότερα από εκείνα του spam.tar.gz.
- Ένα αρχείο με όνομα «ling_filenames.txt», αντίστοιχο εκείνου που συνοδεύει την εκφώνηση, που θα περιέχει τα ονόματα των αρχείων του υποφακέλου «ling».
- Ένα αρχείο με όνομα «spam_filenames.txt», αντίστοιχο εκείνου που συνοδεύει την εκφώνηση, που θα περιέχει τα ονόματα των αρχείων του υποφακέλου «spam».
- Ένα αρχείο με όνομα «keywords.txt», αντίστοιχο εκείνου που συνοδεύει την εκφώνηση. Κατά τον έλεγχο της εργασίας σας, όμως, ενδέχεται το αρχείο να περιέχει άλλες λέξεις-κλειδιά.