Ονομα: Θεοδώρα
Επώνυμο: Αναστασίου
A.M : 1115201400236
Ανάπτυξη Λογισμικού για Αλγοριθμικά Προβλήματα : Εργασία 3

Παρατηρήσεις:

1)Το πρόγραμμα έχει υλοποιηθεί σε c++ .

2)Στον φάκελο περιέχεται αρχείο Makefile το οποίο μεταγλωτίζει το πρόγραμμα.

3)Εκτελείται με τις εξής εντολές εκτελείται με τις εξής εντολές :

make
./recommendation –d <input file> -o <output file> inputnum -validate

//τα υπόλοιπα αρχεία είναι στον φάκελο και αρχικοποιουνται στην main.
οπου inputnum = 0 τρεχει και τις 4 μεθοδους
= 1 τρεχει A1
= 2 τρεχει A2
= 3 τρεχει B1
= 4 τρεχει B2

4)Το έχω τρεξει με valgrind για απαλοιφή errors/leaks

6) Κατα την εκτέλεση του προγράμματος δημιουργείται αρχείο με τα δεδομένα και τους υπολογισμούς χρόνων που ζητήθηκαν.

Έχω υλοποιήσει τις εξής δομές/αρχεία :

- (απο 1η+2η εργασια διαμορφωμένες κατάλληλα για να δέχονται δεδομένα τύπου double) DataSet.cpp/DataSet.h , Hash_Εκτελείται με τις εξής εντολές :uclidean.cpp/Hash_Εκτελείται με τις εξής εντολές :uclidean.h,
Hash_Cosine.cpp/Hash_Cosine.h, HF_Bucket.cpp/HF_Bucket.h, Lsh.cpp/Lsh.h,
HyperCube.cpp/Hypercube.h,Cluster.cpp,Cluster.h,Combinations.cpp,
Combinations.h,,notfun.cpp,notfun.h.

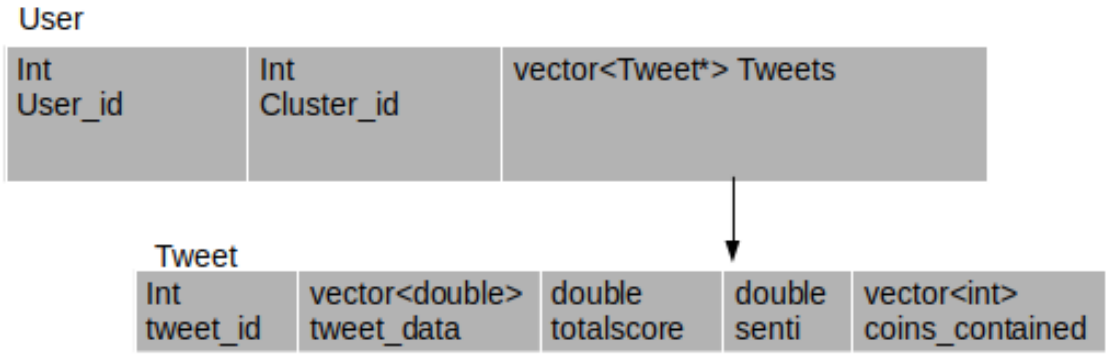
Για το τρίτο μέρος της εργασίας χρειάστηκε να υλοποιήσω :

Δομές: Feelings,User,Tweet οι οποίες θα επεξηγηθούν πιο κάτω σχηματικά.
Αρχεια: rec.cpp/rec.h το οποίο περιέχει τις συναρτήσεις A1,A2,B1,B2
functions.cpp/functions.h το οποίο περιέχει όλες τις συναρτήσεις που υλοποιήθηκαν για να ικανοποιηθούν οι απαιτήσεις της εργασίας.

Επεξήγηση Προγράμματος:

Διαβάζοντας το αρχείο ινπουτ αρχικά φτιάχνω τα λεξικά που θα μου χρειαστούν αργότερα

1) λεξικό με λέξεις και σκορ
2) λεξικό με coins <id coin,namecoin> //ανάλογα με το αν υπάρχει η 5η στήλη η όχι
3)λεξικό με ολα τα ονόματα των coins και id coin
4) στη συνέχεια φτιάχνω τις δομές που φαίνοντε στο πάρακάτω σχήμα .
Κάθε χρήστης έχει ένα πίνακα(α1.1) με tweets στα οποία βρίσκω ποιά coins αναφέρονται ούτως ώστε να φτιάξω τον πίκανα (σχήμα α1.2).



Σχήμα αβ.1

A

	Sentiment Of coin1	Sentiment Of coin2			Sentiment Of coinK
User1 feelings_withInf	0,3421	0.241	inf	inf	inf
User1 feelings	0.3421	0.241	Μεσος όρος των γνωστων	Μεσος όρος των γνωστων	Μεσος όρος των γνωστων

(σχήμα α.1)

Το σχήμα(α.1) απεικονίζει τους πραγματικούς users , η πρώτη γραμμή του πίνακα δίχνει τον vector συναισθημάτων που αντιστοιχεί σε User[s1,s2,...sk] που με βοηθά να βρίσκω ποια coins έχουν βαθμολογιθεί και πια όχι.
Η δεύτερη γραμμή του πίνακα δίχνει τον συμπληρωμένο πίνακα στον οποίο αυτά που δέν είχαν βαθμολογιθεί αντικαθιστούνται με τον μέσο όρο των υπολοίπων.

A1

1) Χρησιμοποιώντας τον πάραπάνω πίνακα τον μετατρέπω σε δομή DataSet ούτως ώστε να μπορώ να χρησιμοποιήσω την LSH της 1ης εργασίας , η οποία και δεχόταν DataSet* σαν όρισμα.
2) Χρησιμοποιώ τη LSH Cosine της 1ης εργασίας για να κατανέμω τους πραγματικούς users και τα sentiments για κάθε coin σε πινακες κατακερματισμού L και buckets.
3) Αφού φτιάξω τους πίνακες κατακερματισμού μου, δίνω σαν query πλέον ένα ένα τους πραγματικούς users ουτως ώστε να βρώ σε ποιό απο όλα τα bucket θα άνηκε.
4)Αφού έχω το bucket στο οποίο θα άνηκε ο συγκεκριμένος πραγματικός μου user ,παίρνω ολα τα στοιχεία που περιέχει το bucket αυτό.
5)Αφού έχω όλα τα στοιχεία που περιέχει το bucket ψάχνω να βρώ τους P κοντινότερους, υπολογίζοντας τις αποστάσεις απο τον user μου, αποθηκεύω όλα τα distances σε vector και στην συνέχεια εξάγω τις P μικρότερες.
6)Ετσι αφού βρώ τις P μικρότερες αποστάσεις βρίσκω και τα P κοντινότερα αντικείμενα στον user μου.
7)Βάση των P_nearest πάω και κάνω τους υπολογισμούς με τον τύπο για τις προβλέψεις ratings των άγνωστων coin.(με cosine similarity)

$$R(u,i)= R(u)+ z* \Sigma sim(u,υ)*(R(υ,i)-R(υ))$$

όπου το R(u) ο μέσος όρος των γνωστών sentiments του target μας,
το z είναι 1/αθροισμα similarity των P κοντινότερων.
R(υ,i) είναι το sentiment του συγκεκριμένου γειτονα για το συγκεκριμένο coin
R(υ) είναι το average του συγκεκριμένου γείτονα για όλα τα coins που έχει βαθμολογίσει.

8) Αφού έχω υπολογίσει τα Ratings των coins που ήταν άγνωστα για τον κάθε User πάω και βρίσκω τα top5 δηλαδή αυτά που έχουν το μεγαλυτερο rating.

A2

1) Χρησιμοποιώντας τον πάραπάνω πίνακα τον μετατρέπω σε δομή DataSet ούτως ώστε να μπορώ να χρησιμοποιήσω το clustering της 2ης εργασίας , η οποία και δεχόταν DataSet* σαν όρισμα.
2)Χρησιμοποιώ τον συνδιασμό Random – Lloyds – Kmeans για να κάνω το clustering των User μου.
3) Αφού κάνω το Clustering ψάχνω σε ποιό cluster ανήκει ανατρέχοντας ένα ένα τα clusters μέχρι να βρώ το συγκεκριμένο user βάση του user_id του.
4) αφού έχω το cluster στο οποίο ανήκει ψάχνω να βρώ τους πιό κοντινούς του με τον ίδιο τρόπο του A1
5)Αφού έχω τους P κοντινότερους με euclidean similarity ψάχνω να βρω τα ratings των αγνώστων και κρατάω τα top2.

B

	Sentiment Of coin1	Sentiment Of coin2	...	...	Sentiment Of coinK
Cluster1 feelings_withInf virtualusers	0,3421	0.241	inf	inf	inf
Cluster1 feelings_virtual users	0.3421	0.241	Μεσος όρος των γνωστων	Μεσος όρος των γνωστων	Μεσος όρος των γνωστων

(σχήμα β.1)

Το σχήμα(β.1) απεικονίζει τους εικονικούς users που βγαίνουν απο το clustering του dataset της 2ης εργασίας , η πρώτη γραμμή του πίνακα δίχνει τον vector συναισθημάτων που αντιστοιχεί σε cluster[s1,s2,...sk] που με βοηθά να βρίσκω ποια coins έχουν βαθμολογιθεί και πια όχι. Για να φτιαχτεί έπρεπε να αντρέχω να βρίσκω αντιστοιχα id των tweet στο dataset της 3ης εργασίας ούτως ώστε να βρίσκω ποια coins ηταν γνωστά και να προσθέτω στην αντίστοιχη θέση το sentiment.
Η δεύτερη γραμμή του πίνακα δίχνει τον συμπληρωμένο πίνακα στον οποίο αυτά που δέν είχαν βαθμολογιθεί αντικαθιστούνται με τον μέσο όρο των υπολοίπων.

B1

1) 1) Χρησιμοποιώντας τον πάραπάνω πίνακα τον μετατρέπω σε δομή DataSet ούτως ώστε να μπορώ να χρησιμοποιήσω την LSH της 1ης εργασίας , η οποία και δεχόταν DataSet* σαν όρισμα.
2) Χρησιμοποιώ τη LSH Cosine της 1ης εργασίας για να κατανέμω τους εικονικούς users και τα sentiments για κάθε coin σε πινακες κατακερματισμού L και buckets.
3) Αφού φτιάξω τους πίνακες κατακερματισμού μου, δίνω σαν query πλέον ένα ένα τους πραγματικούς users ουτως ώστε να βρώ σε ποιό απο όλα τα bucket θα άνηκε.
Ακολουθώ ακριβώς τα ίδια βήματα όπως και στο Α1.

B2

1) Χρησιμοποιώντας τον πάραπάνω πίνακα τον μετατρέπω σε δομή DataSet ούτως ώστε να μπορώ να χρησιμοποιήσω το clustering της 2ης εργασίας , η οποία και δεχόταν DataSet* σαν όρισμα.
2)Χρησιμοποιώ τον συνδιασμό Random – Lloyds – Kmeans για να κάνω το clustering των User μου.
3) Αφού κάνω το Clustering ψάχνω σε ποιό cluster ανήκει ο κάθε πραγματικός user υπολογίζοντας την απόσταση απο το συγκεκριμένο κεντροειδές κάθε φορά.
Όταν βρώ σε ποιο cluster θα άνηκε ο πραγματικός μου user κάνω ακριβώς τα ίδια βήματα που έκανα και στο Α2 για υπολογισμό των ratings.

Validation:

Το κομμάτι αυτό προσπάθησα να το υλοποιήσω αλλα λόγο περιορισμένου χρόνου, μεγάλου χρόνου εκτέλεσης και ενός λάθους στον διωρισμό των users στα validation set τα νομίσματα βαθμολογούνταν ακριβώς με τις ίδιες τιμές που είχαν και προηγουμένος, με αποτέλεσμα ο μέσος όρος των MAE να είναι ίσος με 0.