

## 3η Προγραμματιστική Εργασία

### Υλοποίηση αλγορίθμων υπόδειξης κρυπτονομισμάτων (Recommendation)

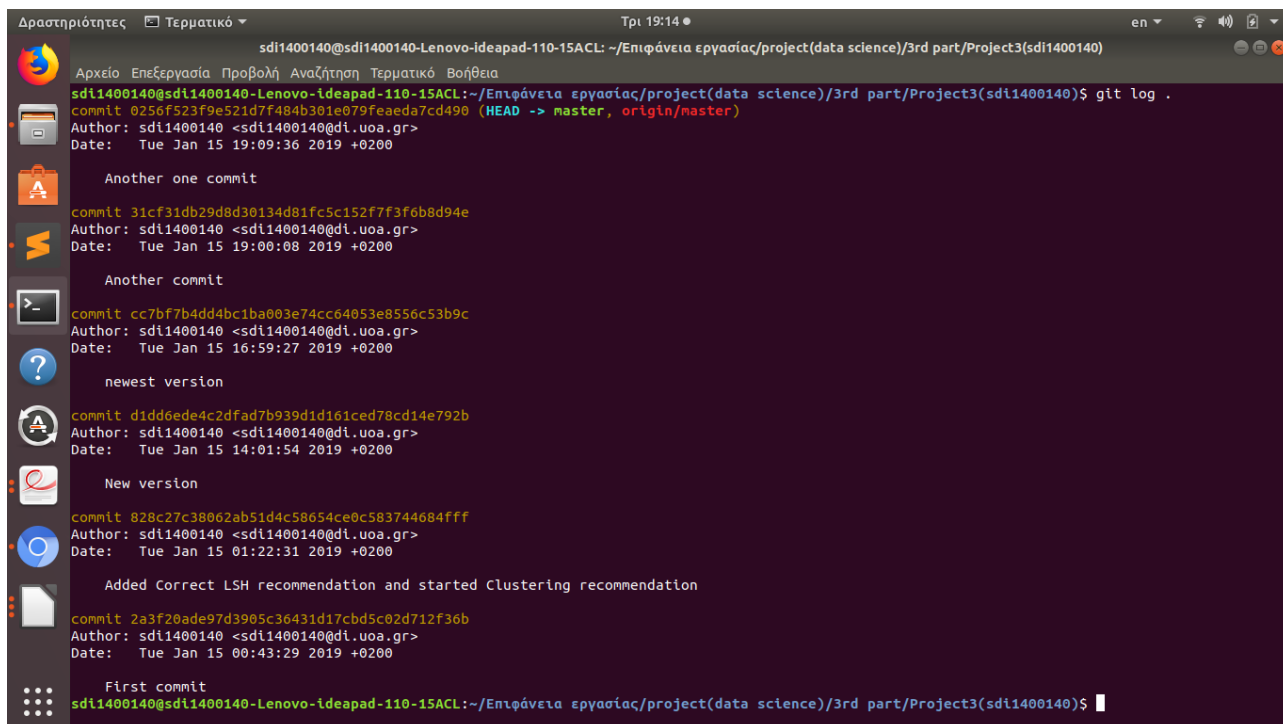
**ΟΝΟΜΑΤΕΠΩΝΥΜΟ:** Παπαγεωργίου Βασίλειος-Νικηφόρος  
**ΑΜ:** 1115201400140

#### Γενικές Πληροφορίες:

Η εργασία μου υλοποιεί το ζητούμενο και είναι χωρισμένο στα παρακάτω αρχεία .cpp , .h: το recommendation.cpp(που αποτελεί την main στην ουσία), τα lsh.h και lsh.cpp(βιβλιοθήκη lsh απο την πρώτη εργασία) και τα cluster.h και cluster.cpp(βιβλιοθήκη clustering απο την δεύτερη εργασία)).Επίσης, υπάρχει το αρχείο επικεφαλίδας structures.h με τον ορισμό των κλάσεων, τα datasets που μας έχετε δώσει και το αρχείο: structures.cpp , που περιέχει τις υλοποιήσεις των διαφόρων δευτερεύουσων και κύριων **ΔΙΚΩΝ ΜΟΥ** δομών(hashtable, lists καθώς επέλεξα να μην κάνω χρήση STL-εκτός από κάποια unordered map για τα λεξικά-διότι ήθελα να έχω πλήρη επίγνωση της κάθε δομής κι όχι να χρησιμοποιήσω κάτι έτοιμο) και αντίστοιχων συναρτήσεων της εργασίας(των ζητούμενων μεθόδων).Επίσης στο φακελο έχω φτιάξει και το αντίστοιχο Makefile και συνεπώς με την εντολή make μεταγλωττίζεται το αντίστοιχο πρόγραμμα και παράγεται το εκτελέσιμο recommendation!Συνεπώς, τα προγράμματα μου εκτελούνται με την εντολή ./recommendation, ακολουθούμενη από τις ζητούμενες παραμέτρους(πχ ***./recommendation -l vader\_lexicon.csv -cl coins\_queries.csv -d tweets\_dataset\_small.csv -o output -o2 output2 -o3 output3***).Στην υλοποίηση μου, έχω υλοποιήσει τις ζητούμενες μεθόδους Recommendation(LSH, Clustering-Random Init, Lloyds Assign, PAM Update).Ακόμη, έχω υλοποιήσει τη μορφή εκτύπωσης στο αρχείο output, ΑΚΡΙΒΩΣ, όπως ζητείται ,εκτυπώνοντας τις ονομασίες των προτεινόμενων κρυπτονομισμάτων για κάθε χρήστη(την 5η στήλη αν υπάρχει, αλλιώς την πρώτη).Επίσης έχω κάνει **version control , με χρήση git**(παρατίθεται screenshot παρακάτω).

Το μέγεθος του dataset(πλήθος tweets) , καθώς και το μέγεθος του coins\_lexicon, προσδιορίζονται απ τις αντίστοιχες σταθερές COIN\_SIZE, SIZE(5000 to small, 161972 to big) με #define.Επίσης έχω μία σταθερά USERTYPE, που καθορίζει αν θα γίνει η διαδικασία με εικονικούς ή

με κανονικούς χρήστες.



The screenshot shows a terminal window with a dark background. The title bar at the top indicates the user is 'sdi1400140' on a 'Lenovo-ideapad-110-15ACL' machine, working in the directory '~/Επιφάνεια εργασίας/project(data science)/3rd part/Project3(sdi1400140)'. The terminal displays the output of the 'git log' command, showing a series of commits from the 'First commit' to the 'newest version'. Each commit entry includes a commit hash, the author's name and email, and the date. The commits are as follows:

- commit 0256f523f9e521d7f484b301e079feada7cd490 (HEAD -> master, origin/master)  
Author: sdi1400140 <sdi1400140@di.uoa.gr>  
Date: Tue Jan 15 19:09:36 2019 +0200
- Another one commit  
commit 31cf31db29d8d30134d81fc5c152f7f3f6b8d94e  
Author: sdi1400140 <sdi1400140@di.uoa.gr>  
Date: Tue Jan 15 19:00:08 2019 +0200
- Another commit  
commit cc7bf7b4dd4bc1ba003e74cc64053e8556c53b9c  
Author: sdi1400140 <sdi1400140@di.uoa.gr>  
Date: Tue Jan 15 16:59:27 2019 +0200
- newest version  
commit d1dd6ede4c2dfad7b939d1d161ced78cd14e792b  
Author: sdi1400140 <sdi1400140@di.uoa.gr>  
Date: Tue Jan 15 14:01:54 2019 +0200
- New version  
commit 828c27c38062ab51d4c58654ce0c583744684fff  
Author: sdi1400140 <sdi1400140@di.uoa.gr>  
Date: Tue Jan 15 01:22:31 2019 +0200
- Added Correct LSH recommendation and started Clustering recommendation  
commit 2a3f20ade97d3905c36431d17cbd5c02d712f36b  
Author: sdi1400140 <sdi1400140@di.uoa.gr>  
Date: Tue Jan 15 00:43:29 2019 +0200
- First commit  
commit 0256f523f9e521d7f484b301e079feada7cd490  
Author: sdi1400140 <sdi1400140@di.uoa.gr>  
Date: Tue Jan 15 19:09:36 2019 +0200

Για την αντιμετώπιση της κανονικοποίησης σε διανύσματα της μορφής  $[inf, a, a, inf]$  όπου με κανονικοποίηση πριν το hashing, γίνονται  $[0, 0, 0, 0]$  εφαρμόζω την κανονικοποίηση συνολικά META το hashing, σύμφωνα με τον δεύτερο τύπο των διαφανειών, ο οποίος για το rating λαμβάνει υπ' όψιν και τον μέσο όρο. Επίσης κατά την προεπεξεργασία, ΑΦΑΙΡΩ από το dataset διανύσματα που αποτελούνται μόνο από 0 και inf.

Για την διαδικασία με εικονικούς χρήστες (για εξοικονόμηση χρόνου), έχω ήδη τρέξει το clustering δημιουργώντας 100 clusters, και διαβάζω απλά από την τρίτη εργασία το output αρχείο της δεύτερης (αρχείο outputof2nd\_inputof3rd), ακολουθώντας στη συνέχεια τα απαραίτητα βήματα αντιστοίχισης των dataset, για την αντίστοιχη δημιουργία των χρηστών. Να σημειωθεί ότι για πιο γρήγορα αποτελέσματα-εκτελέσεις κάνω για τους εικονικούς χρήστες την αντιστοίχιση μεταξύ των small input των εργασιών 2 και 3.

Επίσης παραμετροποιώντας και δοκιμάζοντας διάφορες τιμές k του Clustering, για το small input, κατέληξα ότι για τους κανονικούς χρήστες ένα καλό πλήθος clusters είναι 150, ενώ για 100 κανονικούς ένα καλό πλήθος clusters είναι 20.

Τα paths των αρχείων, δίνονται μέσω της γραμμής εντολών, από τις αντίστοιχες παραμέτρους (χρησιμοποιώ και κάποια ενδιάμεσα αρχεία)!

**Ενδεικτικές εκτελέσεις(για κανονικούς χρήστες):**

***Για το tweets\_dataset\_small.csv:***

LSH execution time: 108.544s

Clustering execution time: 36.2243s

***Για το tweets\_dataset\_big.csv:***

LSH execution time: 2975.6s

Clustering execution time:1440.8s

**ΣΥΜΠΕΡΑΣΜΑ**

Συγκρίνοντας τα Clustering και LSH Recommendation, παρατηρούμε ότι το Clustering είναι αρκετά πιο γρήγορο απ το LSH για ίδιο πλήθος dataset(5000 tweets), άρα και ίδιο πλήθος προτάσεων.

Περαιτέρω λεπτομέρειες παρέχονται σε σχόλια στα αρχεία και οτιδήποτε προκύψει είμαι διαθέσιμος να το αιτιολογήσω στην προφορική εξέταση!