Τελική Αναφορά

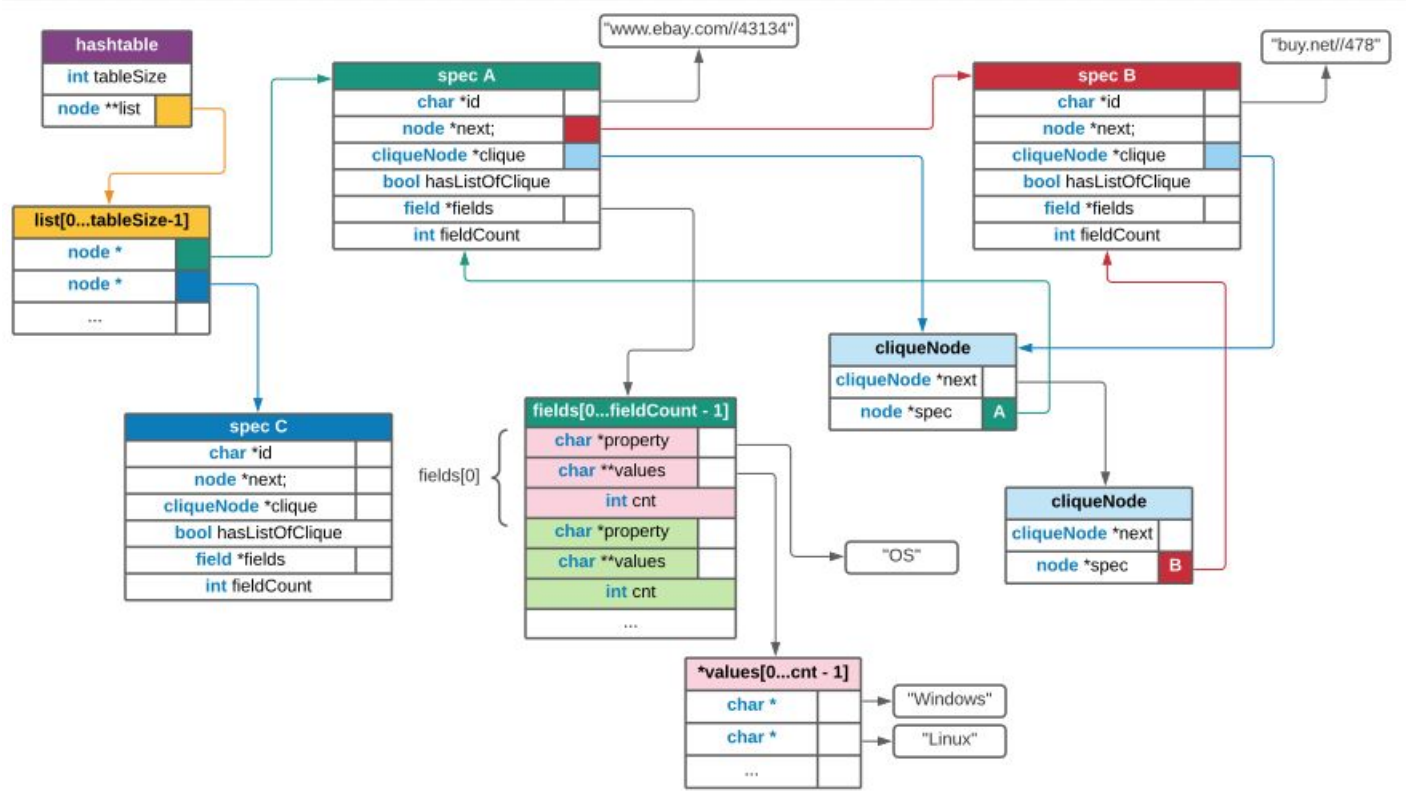
Project 2020-21

Μέλη ομάδας	
Κατσούλη Ιωάννα	1115201400067
Κονόμη Μαρίνα	1115201700054
Παπαδάκος Λεωνίδας-Παναγιώτης	1115201700117

Σύνοψη

Η ζητούμενη εφαρμογή διαχειρίζεται πραγματικές εγγραφές προϊόντων από ηλεκτρονικά καταστήματα, σε μορφή json. Υλοποιείται σε 3 στάδια:

1. Αποθήκευση των **Datasets X** (οι εγγραφές προϊόντων - **spec**) και **W** (οι ομοιότητες/διαφορές μεταξύ τους) σε εσωτερικές δομές και εξαγωγή συμπερασμάτων σχετικά με την ομοιότητα των προϊόντων (**κλίκες**)



2. Εισαγωγή μηχανικής μάθησης (μοντέλο logistic regression) με σκοπό την εύρεση ομοίων προϊόντων, όχι μόνο από το **Dataset W**, αλλά και με βάση τις ιδιότητές τους.

3. Παραλληλοποίηση της εκμάθησης του μοντέλου, με υλοποίηση mini-batch.

Παραδοχές

- Χρησιμοποιούμε τη βιβλιοθήκη acutest για unit testing.
- Επιλέξαμε να χρησιμοποιήσουμε **hashtable** που οδηγεί σε **λίστες** εγγραφών για τα προϊόντα-specs. Η χρήση αυτής της δομής είναι επιθυμητή λόγω της ανάγκης για άμεση αναζήτηση στις καταχωρήσεις, με βάση το **id** (π.χ. **www.ebay.com//46243**), την οποία το hashtable παρέχει με πολυπλοκότητα **O(1)**.

  Οι λίστες έχουν πολυπλοκότητα αναζήτησης **O(n),** ενώ η εισαγωγή γίνεται γρήγορα σε χρόνο **O(1)**.
-

Παρατηρήσεις - Δοκιμές

Χρονισμός

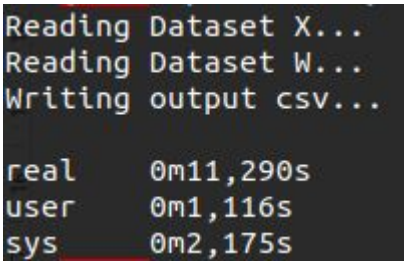
Το μεγαλύτερο μέρος του χρόνου εκτέλεσης καταλαμβάνεται από το διάβασμα των Datasets, με κυριότερο το **Dataset X**:

Εντολή εκτέλεσης:

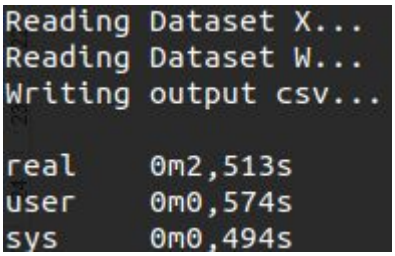
```
time ./specs -x ../Datasets/2013_camera_specs -w
../Datasets/sigmod_large_labelled_dataset.csv -o out.csv
```

Εδώ φαίνεται η εκτέλεση του πρώτου μέρους του προγράμματος (`git checkout part-1`)

1η εκτέλεση:



2η εκτέλεση λίγα δευτερόλεπτα μετά:



Η διαφορά στο χρόνο εκτέλεσης προκύπτει από το caching των αρχείων στη RAM. Συγκεκριμένα, στη 2η εκτέλεση, μεγαλύτερο μέρος του διαβάσματος των specs από το **Dataset X** μπορεί να γίνει χωρίς να απευθυνθεί το σύστημα στο δίσκο (“ακριβή” διαδικασία από την οπτική γωνία του επεξεργαστή).