Τελική Αναφορά

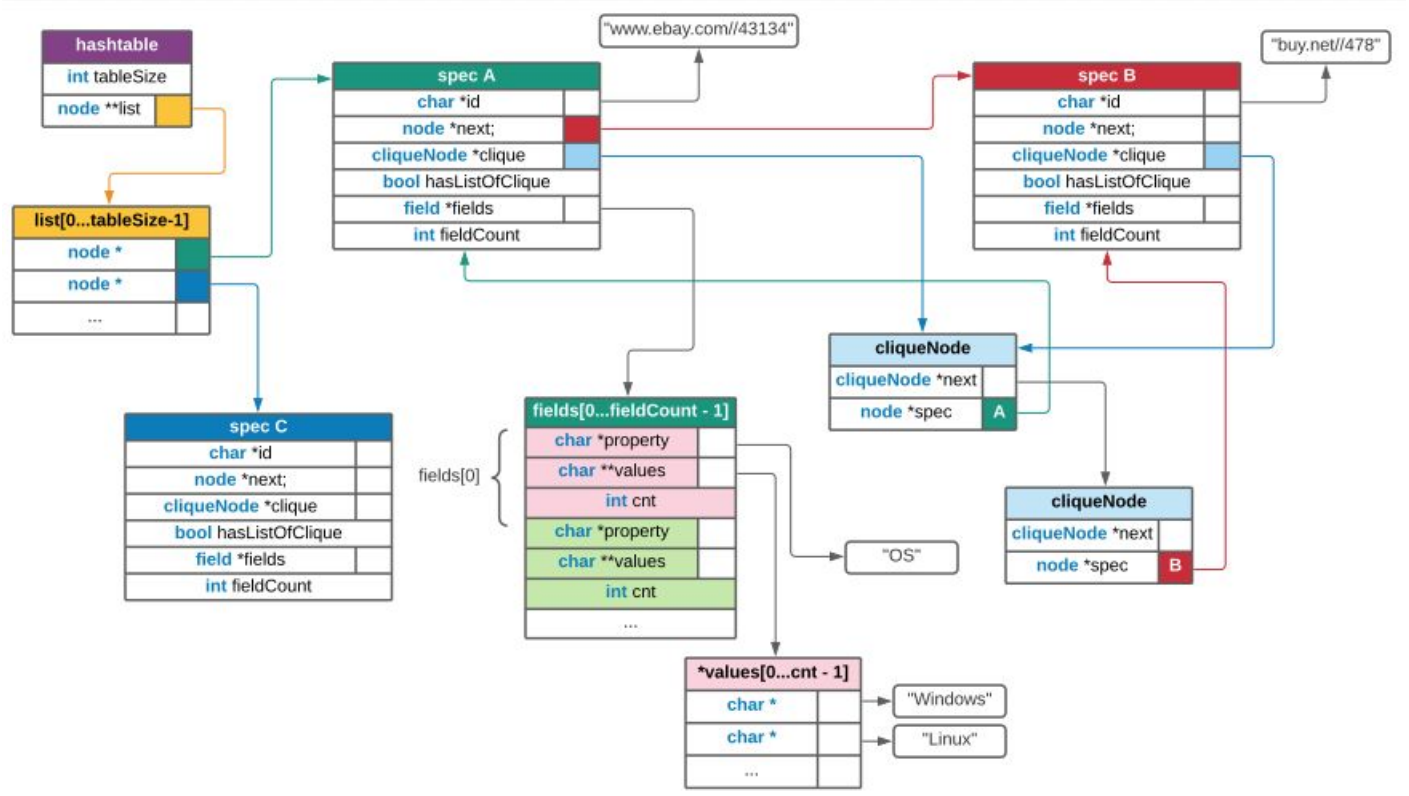
Project 2020-21

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)** καθώς γίνεται στην αρχή της λίστας.
- Η ουρά σε μορφή συνδεδεμένης λίστας έχει πολυπλοκότητα **O(1)**.
- Η αρνητική συσχέτιση υλοποιείται με λίστες σε κάθε κλίκα όπου ένας κόμβος της κάθε λίστας δείχνει στην άλλη κλίκα της αρνητικής συσχέτισης. Είναι σχέση με δύο κατευθύνσεις δηλαδή. Κάθε κλίκα επίσης περιέχει τους δείκτες προς τα specs τα οποία ανήκουν σε αυτή τη κλίκα σε μορφή λίστας.
    - Τις σχέσεις μεταξύ των specs επιλέξαμε να τις υλοποιήσουμε με λίστα καθώς δεν έχουμε λειτουργικότητα που απαιτεί να αναζητούμε μια συγκεκριμένη σχέση ενώ άλλες εναλλακτικές δεν παρουσιάζουν καλύτερη πολυπλοκότητα από αυτή της λίστας για τις λειτουργίες που θέλουμε.
- Η δομή του hashtable έχει χρησιμοποιηθεί ευρέως στην εργασίας μας καθώς όπως προαναφέραμε ευνοεί λειτουργίες αναζήτησης είτε πρόκειται για προϊόντα, είτε για ιδιότητες αυτών των προϊόντων ή ακόμη και για τη λειτουργία του preprocessing.
- Για το validation χρησιμοποιήθηκε πάλι λίστα καθώς θέλαμε απλά να καταγράψουμε τις σχέσεις που δημιουργούν conflicts και δε θα κάνουμε καμία αναζήτηση πάνω σε αυτήν παρά μόνο θα τη διατρέξουμε για την επίλυση των conflicts.

Προδιαγραφές του μηχανήματος δοκιμών :

Operating system: ubuntu (64-bit)
Ram: 8gb
Storage: 10gb

Δοκιμές

1ο Μέρος Εργασίας

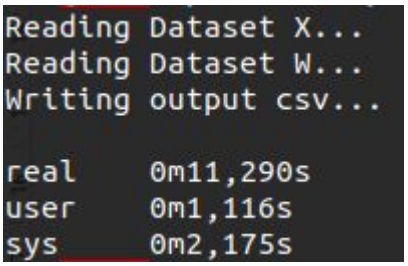
Το μεγαλύτερο μέρος του χρόνου εκτέλεσης καταλαμβάνεται από το διάβασμα των Datasets, με κυριότερο το **Dataset X**:

Εντολή εκτέλεσης:

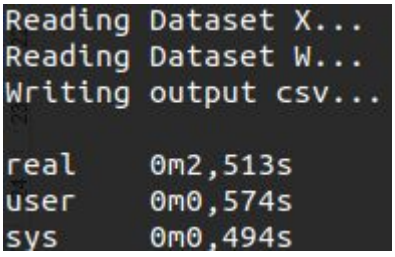
```
time ./specs -x ../Datasets/2013_camera_specs -w
../Datasets/sigmod_large_labelled_dataset.csv -o out.csv
```

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).

2ο Μέρος Εργασίας

Στην δεύτερη εργασία ζητήθηκε να βρεθεί για κάθε πιθανό ζεύγος από προϊόντα το εάν σχετίζονται (1) ή όχι (0). Η υλοποίηση της εργασίας έγινε σειριακά.

Τα αποτελέσματα για την εκτέλεση όλου του προγράμματος ,δηλαδή με τη δημιουργία κλικών και αρνητικών κλικών, ήταν:

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

```
[Reading Dataset X]
\Distinct words: 29328
[Preprocessing specs]
Distinct words after trim: 5222
[Reading Dataset W]
[Partitioning Derived Dataset W']
Total lines in expanded dataset: 341929
Training:     205157 lines
Validation:    68385 lines
Test:          68387 lines
Training model...
Validation set... Prediction accuracy: 83.89% (57366 hits)
Testing set... Prediction accuracy: 71.98% (49223 hits)
[Writing output csv (cliques)]
```

```
Command being timed: "./specs -x Datasets/2013_camera_specs -w Datasets/sigmod_large_labelled_dataset.csv"
User time (seconds): 12.56
System time (seconds): 1.99
Percent of CPU this job got: 93%
Elapsed (wall clock) time (h:mm:ss or m:ss): 0:15.54
Average shared text size (kbytes): 0
Average unshared data size (kbytes): 0
Average stack size (kbytes): 0
Average total size (kbytes): 0
Maximum resident set size (kbytes): 1943708
Average resident set size (kbytes): 0
```

Χρόνοι μόνο για το preprocessing, training:

```
Command being timed: "./specs -x Datasets/2013_camera_specs -w Datasets/sigmod_large_labelled_dataset.csv"
User time (seconds): 8.56
System time (seconds): 1.56
Percent of CPU this job got: 56%
Elapsed (wall clock) time (h:mm:ss or m:ss): 0:17.86
Average shared text size (kbytes): 0
Average unshared data size (kbytes): 0
Average stack size (kbytes): 0
Average total size (kbytes): 0
Maximum resident set size (kbytes): 1943688
```

3ο Μέρος εργασίας

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

Για την εκτέλεση όλου του προγράμματος:

Για 2 threads:

```
Testing set... Prediction accuracy: 69.78% (47720 hits)
        Command being timed: "./specs -x Datasets/2013_camera_specs -w Datasets/sigmod_large_labelled_dataset.csv"
        User time (seconds): 287.41
        System time (seconds): 1.67
        Percent of CPU this job got: 100%
        Elapsed (wall clock) time (h:mm:ss or m:ss): 4:47.85
        Average shared text size (kbytes): 0
        Average unshared data size (kbytes): 0
        Average stack size (kbytes): 0
        Average total size (kbytes): 0
        Maximum resident set size (kbytes): 2008936
```

Για 4 threads:

```
esting set... Prediction accuracy: 69.78% (47720 hits)
        Command being timed: "./specs -x Datasets/2013_camera_specs -w Datasets/sigmod_large_labelled_dataset.csv"
        User time (seconds): 273.33
        System time (seconds): 2.07
        Percent of CPU this job got: 98%
        Elapsed (wall clock) time (h:mm:ss or m:ss): 4:40.83
        Average shared text size (kbytes): 0
        Average unshared data size (kbytes): 0
        Average stack size (kbytes): 0
        Average total size (kbytes): 0
        Maximum resident set size (kbytes): 2008864
```

Για 8 threads:

```
Testing set... Prediction accuracy: 69.78% (47720 hits)
        Command being timed: "./specs -x Datasets/2013_camera_specs -w Datasets/sigmod_large_labelled_dataset.csv"
        User time (seconds): 268.39
        System time (seconds): 1.57
        Percent of CPU this job got: 100%
        Elapsed (wall clock) time (h:mm:ss or m:ss): 4:28.40
        Average shared text size (kbytes): 0
        Average unshared data size (kbytes): 0
        Average stack size (kbytes): 0
        Average total size (kbytes): 0
        Maximum resident set size (kbytes): 2009396
```

Για 16 threads:

```
Testing set... Prediction accuracy: 69.78% (47720 hits)
        Command being timed: "./specs -x Datasets/2013_camera_specs -w Datasets/sigmod_large_labelled_dataset.csv"
        User time (seconds): 269.55
        System time (seconds): 2.35
        Percent of CPU this job got: 100%
        Elapsed (wall clock) time (h:mm:ss or m:ss): 4:31.51
        Average shared text size (kbytes): 0
        Average unshared data size (kbytes): 0
        Average stack size (kbytes): 0
        Average total size (kbytes): 0
        Maximum resident set size (kbytes): 2008528
```

Χρόνοι μόνο για preprocessing, training:

Για 1 thread:

```
[Run sets (1 threads)]
Training model...
        Command being timed: "./specs -x Datasets/2013_camera_specs -w Datasets/sigmod_large_labelled_dataset.csv"
        User time (seconds): 11.35
        System time (seconds): 1.41
        Percent of CPU this job got: 63%
        Elapsed (wall clock) time (h:mm:ss or m:ss): 0:20.05
```

Για 2 threads:

```
[Run sets (2 threads)]
Training model...
        Command being timed: "./specs -x Datasets/2013_camera_specs -w Datasets/sigmod_large_labelled_dataset.csv"
        User time (seconds): 11.51
        System time (seconds): 1.77
        Percent of CPU this job got: 72%
        Elapsed (wall clock) time (h:mm:ss or m:ss): 0:18.31
```

Για 4 threads:

```
[Run sets (4 threads)]
Training model...
        Command being timed: "./specs -x Datasets/2013_camera_specs -w Datasets/sigmod_large_labelled_dataset.csv"
        User time (seconds): 15.48
        System time (seconds): 3.61
        Percent of CPU this job got: 76%
        Elapsed (wall clock) time (h:mm:ss or m:ss): 0:24.88
```

Για 8 threads:

```
[Run sets (8 threads)]
Training model...
        Command being timed: "./specs -x Datasets/2013_camera_specs -w Datasets/sigmod_large_labelled_dataset.csv"
        User time (seconds): 12.43
        System time (seconds): 2.58
        Percent of CPU this job got: 74%
        Elapsed (wall clock) time (h:mm:ss or m:ss): 0:20.21
```

Για 16 threads:

```
Run sets (16 threads)]
raining model...
        Command being timed: "./specs -x Datasets/2013_camera_specs -w Datasets/sigmod_large_labelled_dataset.csv"
        User time (seconds): 11.72
        System time (seconds): 1.30
        Percent of CPU this job got: 76%
        Elapsed (wall clock) time (h:mm:ss or m:ss): 0:16.95
```

Παρατηρήσεις:

- Το δεύτερο μέρος της εργασίας έχει μεγαλύτερο prediction accuracy (71.98%), καθώς το τρίτο μέρος της εργασίας,το οποίο περιλαμβάνει threads και επίλυση των conflicts, έχει prediction accuracy 69.78% .Η διαφορά μεταξύ τους είναι 2.2% (για το large dataset).
- Στο training τμήμα της 3ης εργασίας παρατηρείται μεγαλύτερο elapsed time,user time και system time στα 4 threads.
- Στη 3η εργασία ο elapsed time μειώνεται όσο αυξάνονται τα threads (2-8 threads). Στα 16 threads παρατηρούμε μειωση elapsed time συγκριτικά με τα 2,4 threads.