Κ23α - Ανάπτυξη Λογισμικού Για Πληροφοριακά Συστήματα

Χειμερινό Εξάμηνο 2020 – 2021

Καθηγητής Ι. Ιωαννίδης

Άσκηση 1 – Παράδοση: Δευτέρα 16 Νοεμβρίου 2020

Στα sites ηλεκτρονικού εμπορίου, δεν υπάρχουν ακριβή πεδία για την καταχώρηση των προϊόντων. Συνεπώς, τα αναφερόμενα χαρακτηριστικά ίδιων προϊόντων διαφέρουν τόσο στα πεδία (keys) που καταγράφονται (αριθμός πεδίων, όνομα πεδίου), όσο και στις τιμές των πεδίων αυτών. Στόχος της εργασίας είναι να κατασκευαστούν οι δομές δεδομένων και οι κατάλληλοι αλγόριθμοι που να διευκολύνουν την αναγνώριση ίδιων προϊόντων που αντιστοιχούν σε διαφορετικές καταχωρήσεις. Στη βιβλιογραφία απαντάται ως αναγνώριση ταυτότητας (entity resolution) ή αποσαφήνιση (disambiguation).

Περιγραφή Προβλήματος

Θα σας δοθεί ένα σύνολο δεδομένων (dataset) που περιέχει περίπου 30000 προδιαγραφές προϊόντων (spec) σε μορφή JSON. Κάθε spec περιέχει μία λίστα από ζεύγη <όνομα_ιδιότητας, τιμή_ιδιότητας> που έχουν εξαχθεί από διαφορετική ιστοσελίδα και έχουν συλλεχθεί από 24 διαφορετικά web sites. Το σύνολο δεδομένων αυτό αναφέρεται ως dataset X.

- Κάθε spec αποθηκεύεται ως ξεχωριστό αρχείο και τα αρχεία οργανώνονται σε καταλόγους (directories), όπου κάθε κατάλογος αντιστοιχεί σε διαφορετική πηγή δεδομένων (π.χ. www.alibaba.com, www.ebay.com)
- Όλα τα specs αναφέρονται σε φωτογραφικές μηχανές και περιλαμβάνουν πληροφορίες για το μοντέλο της φωτογραφικής μηχανής (π.χ. Canon EOS 5D Mark II) και, πιθανώς, εξαρτήματα/αξεσουάρ (π.χ. σετ φακών, τσάντα, τρίποδα). Τα αξεσουάρ δεν συνεισφέρουν στην αναγνώριση του προϊόντος. Για παράδειγμα μία μηχανή Canon EOS 5D Mark II που πωλείται μαζί με μία τσάντα μεταφοράς θεωρείται το ίδιο προϊόν με μία Canon EOS 5D Mark II που πωλείται μόνη της.

Παράδειγμα εγγραφής σε μορφή JSON:

```
{
   "<page title>": "Samsung Smart WB50F Digital Camera White Price in India with Offers & Full Specifications | PriceDekho.com",
  "brand": "Samsung",
  "dimension": "101 x 68 x 27.1 mm",
  "display": "LCD 3 Inches",
  "pixels": "Optical Sensor Resolution (in MegaPixel)\n16.2 MP"
  "battery": "Li-Ion"
}
```

Σημειώστε ότι ενώ η ιδιότητα ‘page title’ θα είναι πάντα παρούσα, όλα τα υπόλοιπα ονόματα ιδιοτήτων είναι προαιρετικά και μεταβαλλόμενα, ακόμη και εντός της ίδιας διαδικτυακής πηγής. Σημειώστε επίσης ότι δύο ιδιότητες με το ίδιο όνομα (ομώνυμα) μπορεί να έχουν διαφορετική σημασία (π.χ. η ιδιότητα battery μπορεί να αναφέρεται σε τύπο μπαταρίας, όπως “AAA”, ή συστατικά μπαταρίας, όπως “Li-Ion”). Ακόμη, δύο ιδιότητες με την ίδια σημασία μπορεί να έχουν διαφορετικό όνομα (συνώνυμα), (π.χ. “resolution” και “pixels).

Σας παρέχεται επίσης και ένα σημειωμένο dataset σε μορφή CSV, που περιέχει 3 στήλες “left_spec_id”, “right_spec_id” και “label”. Το σύνολο δεδομένων αυτό αναφέρεται ως dataset W.

- Το “spec_id” είναι μία μοναδική ταυτότητα για ένα spec και αποτελείται από το σχετικό μονοπάτι (path) του αρχείου spec. Σημειώστε ότι αντί για τον χαρακτήρα “/”, το spec_id χρησιμοποιεί τον ειδικό χαρακτήρα “//” και ότι παραλείπεται η επέκταση “.json”. Για παράδειγμα, το spec_id “www.ebay.com//1000” αναφέρεται στο αρχείο 1000.json, το οποίο βρίσκεται στον κατάλογο www.ebay.com. Όλα τα spec_id στο σημειωμένω dataset W αναφέρονται σε χαρακτηριστικά προϊόντων που περιλαμβάνονται στο dataset X.
- Κάθε γραμμή του σημειωμένου dataset αναπαριστά ένα ζεύγος spec. Αν η τιμή της στοιχείου label είναι 1 σημαίνει ότι το αριστερό και το δεξί spec αναφέρονται στο ίδιο προϊόν (δηλαδή ταιριάζουν). Η τιμή του στοιχείου label 0 σημαίνει ότι τα δύο specs αναφέρονται σε διαφορετικό προϊόν (δηλαδή ότι δεν ταιριάζουν).

Παράδειγμα γραμμής στο σημειωμένο dataset:

```
left_spec_id, right_spec_id, label
www.ebay.com//1, www.ebay.com//2, 1
www.ebay.com//3, buy.net//10, 0
```

Σημειώστε ότι μπορεί να υπάρχουν ζεύγη specs που να ταιριάζουν ακόμη και εντός της ίδιας διαδικτυακής πηγής και ότι ισχύει η μεταβατική ιδιότητα (αν δηλαδή το A ταιριάζει με το B και το B ταιριάζει με το C, τότε το A ταιριάζει με το C).

Με βάση αυτά τα δεδομένα, θα βρεθούν όλα τα ζεύγη specs στο dataset X που ταιριάζουν, δηλαδή αναφέρονται στο ίδιο προϊόν. Η έξοδος θα αποθηκεύεται σε ένα αρχείο CSV με δύο στήλες, “left_spec_id” και “right_spec_id”. Κάθε γραμμή περιέχει μόνο τα δύο IDs, χωρισμένα με “,”.

Παράδειγμα αρχείο εξόδου:

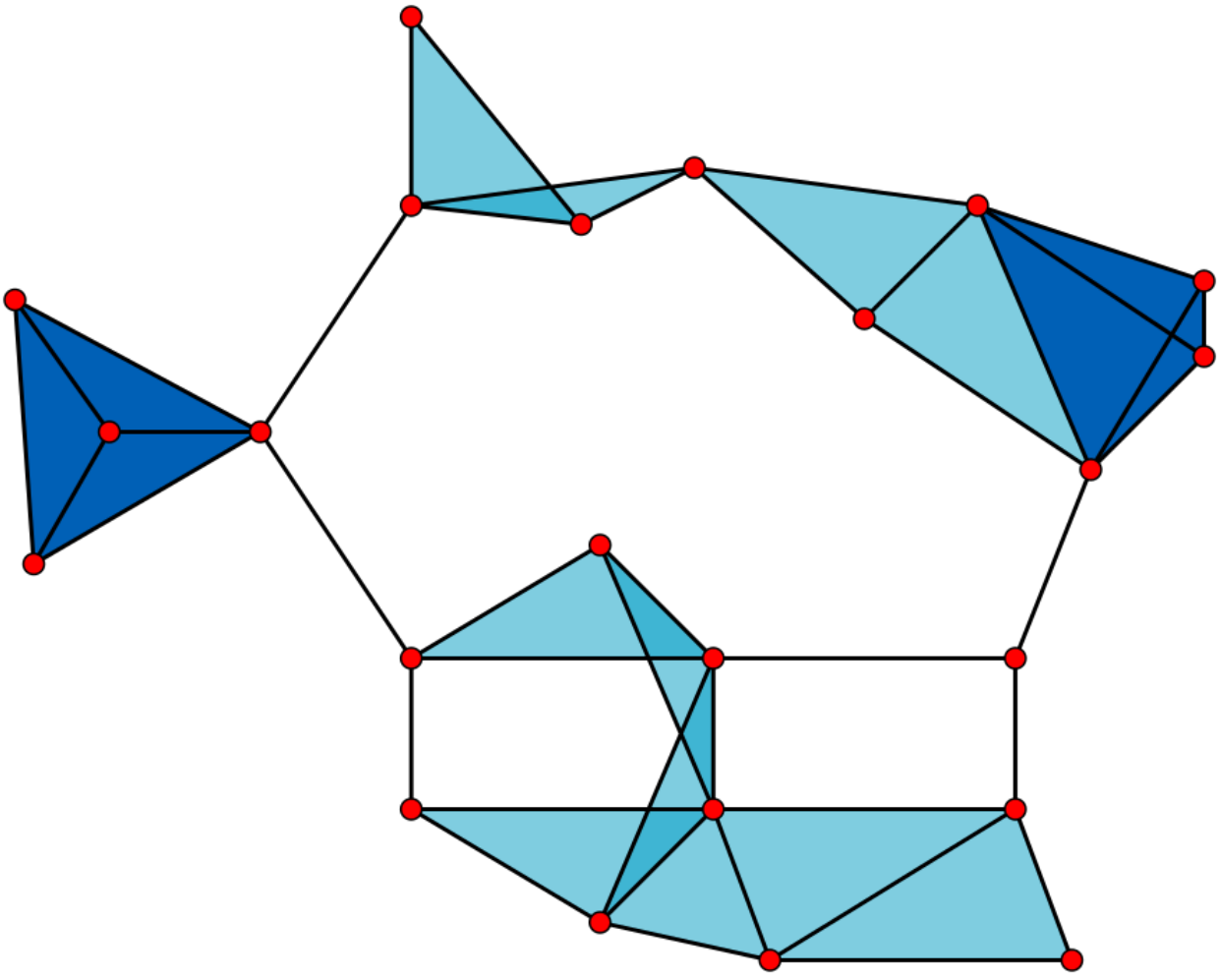
```
left_spec_id, right_spec_id
www.ebay.com//10, www.ebay.com//20
www.ebay.com//10, buy.net//100
```

Υλοποίηση Δομών Δεδομένων

Κλίκες Γράφων

Πλήρης γράφος n κορυφών είναι ο γράφος του οποίου κάθε κορυφή είναι παρακείμενη σε όλες τις άλλες.

Κλίκα (clique) σε ένα γράφο ονομάζεται ένας πλήρης υπογράφος του. Μπορεί να υπάρχουν περισσότερες από μία κλίκες σε ένα γράφο.



Εικόνα από wikipedia, David Eppstein - Own work, Public Domain,

Στην παραπάνω εικόνα εμφανίζονται

- 23 × κλίκες 1 κορυφής (οι κορυφές),

- 42 × 2 κλίκες 2 κορυφών (οι ακμές)
- 19 × 3 κλίκες 3 κορυφών (γαλάζια και μπλε τριγωνα)
- 2 × 4 κλίκες 4 κορυφών (σκούρες μπλε περιοχές).

Χρήση κλίκας για την εφαρμογή

Στην περίπτωσή μας, θεωρούμε ότι κάθε spec είναι κορυφή του γράφου. Οι ακμές του γράφου έχουν βάρη και συγκεκριμένα την τιμή +1, αν τα specs ταιριάζουν και -1 αν τα specs δεν ταιριάζουν. Μας ενδιαφέρουν οι κλίκες που σχηματίζονται από ακμές με βάρος +1. Όλες οι κορυφές που ανήκουν σε μία κλίκα, ταιριάζουν.

Η υλοποίηση που θα πραγματοποιηθεί θα χρησιμοποιήσει την ιδιότητα αυτή και θα δημιουργήσει μόνο τις κλίκες των γράφων και όχι τους πλήρεις γράφους. Θα εκμεταλλευτούμε τη μεταβατική κλειστότητα (transitive closure) για να εισάγουμε νέες κορυφές στις κλίκες σε περίπτωση θετικής γειτνίασης, αλλά και για να αποκλείσουμε κάποιες άλλες στην περίπτωση αρνητικής γειτνίασης.

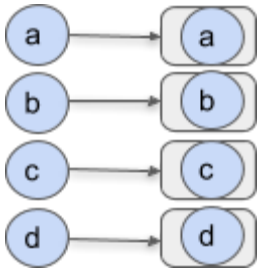
Υλοποίηση - Αλγόριθμος

Δεν θα χρειαστεί να υλοποιηθεί μία πλήρης δομή γράφου για την εργασία, παρά μόνο η δομή της κλίκας. Σε αυτή τη δομή, θα υπάρχει μία 1-1 αντιστοιχία μεταξύ των κορυφών του γράφου και των specs των προϊόντων.

Συγκεκριμένα κάθε spec θα αντιστοιχιστεί με ένα id (έστω a, b, c,..). Κάθε id θα δείχνει κάθε στιγμή σε ένα σύνολο από άλλα id με τα οποία ταιριάζει.

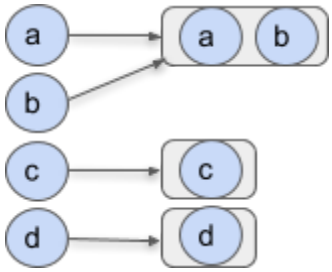
Επομένως τα βήματα είναι ως εξής:

Ανάγνωση του dataset X, σχηματισμός των nodes με όλα τα πεδία που περιγράφονται στα specs και σχηματισμός της αρχικής κατάστασης των κλικών στις οποίες ανήκουν.

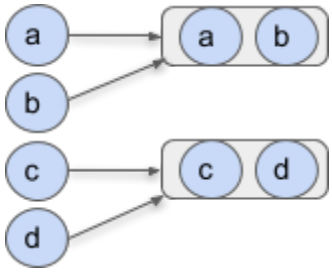


Στην αρχική κατάσταση τα specs ταιριάζουν μόνο με τον εαυτό τους.

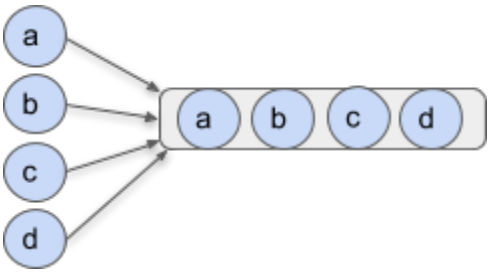
Στη συνέχεια, διαβάζεται το dataset W. Οι γραμμές που δηλώνουν ότι δεν ταιριάζουν τα specs, δηλαδή αυτές που στην τελευταία στήλη έχουν την τιμή 0, θα αγνοηθούν για αυτό το τμήμα της εργασίας. Για κάθε γραμμή που δηλώνει ότι 2 specs ταιριάζουν ακολουθείται η εξής διαδικασία:
Έστω ότι υπάρχει η γραμμή (a, b, 1). Τότε το σύνολο των specs με τα οποία ταιριάζει τόσο το a, όσο και το b θα είναι το {a, b}, δηλαδή θα έχουμε τα εξής:



δηλαδή, τόσο ο a, όσο και ο b θα δείχνουν στο κοινό σύνολο κόμβων με τους οποίους ταιριάζουν.
Αν η επόμενη γραμμή είναι η (c, d, 1), δηλαδή η δήλωση ότι ο c και ο d ταιριάζουν, τότε η επόμενη κατάσταση θα είναι η εξής:



Σε περίπτωση που η επόμενη γραμμή είναι η (a, c, 1), τότε αναλογικά θα επέλθει η εξής κατάσταση:



Έξοδος εφαρμογής

Όπως αναφέρθηκε και προηγουμένως, η έξοδος της εφαρμογής θα είναι όλα τα ζεύγη specs στο dataset X που ταιριάζουν, δηλαδή αναφέρονται στο ίδιο προϊόν.

Παράδοση εργασίας

Η εργασία είναι ομαδική, **2 ή 3 ατόμων**.
Προθεσμία παράδοσης: 16/11/2020
Γλώσσα υλοποίησης: C / C++ χωρίς χρήση stl.
Περιβάλλον υλοποίησης: Linux (gcc > 5.4+).

Παραδοτέα: Η παράδοση της εργασίας θα γίνει με βάση το τελευταιο commit πριν την προθεσμία υποβολής στο git repository σας. **Η χρήση git είναι υποχρεωτική**.
Επιπλέον, εκτός από τον πηγαίο κώδικα, θα παραδώσετε μια σύντομη αναφορά, με τις σχεδιαστικές σας επιλογές καθώς και να εφαρμόσετε ελέγχους ως προς την ορθότητα του λογισμικού με τη χρήση ανάλογων βιβλιοθηκών (Software testing).