UNIVERSITY OF THESSALY
SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL
& COMPUTER ENGINEERING

Προχωρημένη Κατανεμημένη Υπολογιστική
Χειμερινό Εξάμηνο 2019-2020
Δημήτριος Κατσαρός

Σειρά προβλημάτων: 1η: **ΑΤΟΜΙΚΕΣ ΕΡΓΑΣΙΕΣ**

Ημέρα ανακοίνωσης: Monday, November 04, 2019
Προθεσμία παράδοσης: Κυριακή, Δεκέμβριος 01, 2019

Πρόβλημα-01 [1 μονάδα]
Στο αρχείο bible+shakes.nopunc.gz (όλα τα έργα του Shakespeare, και η Βίβλος) http://inf-server.inf.uth.gr/courses/CE623/adc_fall17/assignments/bible+shakes.nopunc.gz να βρείτε πόσες φορές εμφανίζεται κάθε "λέξη" και κάθε ζεύγος λέξεων. Ως "λέξη" νοείται κάθε σύνολο χαρακτήρων που διαχωρίζεται από την προηγούμενη και επόμενή της με κάποιον διαχωριστή (delimiter), δηλαδή space(s), tab, … Αγνοήστε πεζά-κεφαλαία γράμματα.

Πρόβλημα-02 [1 μονάδα]
Στην σελίδα http://ita.ee.lbl.gov/html/contrib/EPA-HTTP.html θα βρείτε όλες τις ημερήσεις HTTP αιτήσεις στον EPA WWW server που βρίσκεται (βρισκόταν) στο Research Triangle Park, NC. Διαβάστε στο Διαδίκυτο για το format στο οποίο είναι καταγεγραμμένα. Να φτιάξετε Map-Reduce για να βρείτε ποιο directory παρήγαγε τα περισσότερα hits.

Πρόβλημα-03 [2 μονάδες]
Στο αρχείο http://inf-server.inf.uth.gr/courses/CE623/adc_fall17/assignments/wikipedia.txt να βρείτε τις 25 πιο συχνές λέξεις που αποτελούνται από περισσότερους από 4 χαρακτήρες, σε διάταξη φθίνουσας συχνότητας. Αγνοήστε πεζά-κεφαλαία, καθώς και όποιες stop-words εμφανίζονται εδώ http://xpo6.com/list-of-english-stop-words/.
Έξοδος (ενδεικτική):
education 10546
states 8867

Πρόβλημα-04 [1 μονάδα]
Το αρχείο http://inf-server.inf.uth.gr/courses/CE623/adc_fall17/assignments/Largest-Urban-areas-world-with-density.xlsx περιέχει τις μεγαλύτερες πόλεις του κόσμου και προέρχεται από μια σχεσιακή βάση δεδομένων από τον πίνακα City. Αφού το αποθηκεύσετε ως αρχείο txt, να γράψετε κώδικα Map-Reduce για να εκτελέστε την λειτουργία που παρακάτω περιγράφεται ως SQL statement:
SELECT City, Population
FROM City
WHERE population > 12000000

Πρόβλημα-05 [3 μονάδες]
Στο αρχείο ecoli.fa.gz
http://inf-server.inf.uth.gr/courses/CE623/adc_fall17/assignments/ecoli.fa.gz
περιέχεται το πλήρες γονιδίωμα του E.coli (με την πρώτη γραμμή να είναι πληροφοριακή).

Θα πρέπει να υλοποιηθεί ο μετρητής 3-μερών. Εάν το γονιδίωμα ήταν μόνο το παρακάτω:
ACACACAGT
τότε η έξοδος του προγράμματος θα ήταν:
ACA 3
CAC 2
CAG 1
AGT 1
επειδή δεν υπάρχουν κενά που να διακρίνουν τις λέξεις στο γονιδίωμα, θα μετρήσουμε επικαλυπτέμενα 3-μερή αντί για διακριτές λέξεις. Η έξοδος να είναι ταξονομημένη με φθίνοντα μετρητή εμφάνισης.

Πρόβλημα-06 [1 μονάδα]
Εξηγήστε την εφαρμογή του Θεωρήματος CAP στο σύστημα δέσμευσης θέσεων μιας αεροπορικής εταιρείας.

Πρόβλημα-07 [1 μονάδα]
Στο Amazon Dynamo, όπως χρησιμοποιείται από το Amazon Storefront, υπάρχει ένα key που περιέχει ταυτόχρονα το "session" και "object", και ένα μια serialized value (τιμή) που αναπαριστά ένα καλάθι αγορών (shopping cart).

- Η μεταβλητή "session" προσδιορίζει (identifies) τον χρήστη και τον browser.
- Η μεταβλητή "object" ιχνηλατεί (tracks) αντικείμενα μέσα σε ένα shopping cart του χρήστη.
- Η (serialized) value αναπαριστά ένα cart.

Υποθέστε ότι κάποιος χρήστης ανοίγει δυο instances του ίδιου shopping cart σε δυο tabs του ίδιου browser, και προχωρά σε ενημέρωση του cart και από τα δυο tabs διαγράφοντας όμως διαφορετικό item (δηλαδή 'αγορασμένο' προϊόν) σε κάθε tab.

- Θεωρείτε ότι κάποιος αλγόριθμος συνέπειας π.χ., vector clocks θα αναγνώριζε πρόβλημα consistency του cart, δηλαδή ότι υπάρχουν δυο versions (με διαφορετικά περιεχόμενα) του συγκεκριμένου shopping cart; Γιατί;
  - Εάν πιστεύτε πως 'ΝΑΙ', τότε στο συγκεκριμένο πραγματικό εμπορικό σύστημα
    - Ποιος πιστεύετε ότι πρέπει να την επιλύσει; Το σύστημα (amazon storefront) ή ο χρήστης; Γιατί;
    - Ποια πιστεύετε ότι θα είναι η λύση, δηλαδή πώς θα επιλεγεί ποια version του shopping cart θα κρατηθεί;

Χρηστικές πληροφορίες:
Η προθεσμία παράδοσης είναι αυστηρή. Είναι δυνατή η παροχή παράτασης (μέχρι 2 ημέρες), αλλά μόνο αφού δώσει ο διδάσκων την έγκρισή του και αυτή η παράταση στοιχίζει 10% ποινή στον τελικό βαθμό της συγκεκριμένης Σειράς Προβλημάτων. Η παράδοση γίνεται με email (dkatsar@inf.uth.gr) του αρχείου λύσεων σε μορφή pdf (typeset). Το subject του μηνύματος πρέπει να είναι: CE623-Problem set 01: AEM

Ερμηνεία συμβόλων:

Δεν απαιτεί την χρήση υπολογιστή ή/και την ανάπτυξη κώδικα.

Απαιτεί την χρήση του Web για ανεύρεση πληροφοριών ή διεξαγωγή πειράματος.

Απαιτεί την ανάπτυξη κώδικα MapReduce στο Hadoop. Προτιμητέα γλώσσα προγραμματισμού η Java. Το παραδοτέο θα περιέχει:

- Τον ψευδοκώδικα υλοποίησης (για όσα ζεύγη map-reduce απαιτούνται).
- Ένα παράδειγμα εκτέλεσης (με μικρή είσοδο) που θα παρουσιάζει τις φάσεις εκτέλεσης των map και reduce που σχεδιάσετε.
- Τον πραγματικό πηγαίο κώδικα (π.χ., Java) υλοποίησης