Διαχείριση Μεγάλων Δεδομένων
1η Προγραμματιστική Εργασία

Διδάσκουσα:
Π. Ραυτοπούλου

Παράδοση μέχρι: **Δευτέρα 20/12/2021 ώρα 23.59**
Προσωπική εξέταση: στο τέλος του εξαμήνου
(η ακριβής ημερομηνία θα ανακοινωθεί έγκαιρα)

ΣΗΜΑΝΤΙΚΕΣ ΣΗΜΕΙΩΣΕΙΣ:

1. Σε όλα τα αρχεία που θα παραδώσετε θα πρέπει **ΟΠΩΣΔΗΠΟΤΕ** να βάλετε τα ονόματα, τους Α.Μ., και τα username/email των μελών της ομάδας (κατά προτίμηση ομάδες 2 ατόμων).
2. Αφού έχετε ολοκληρώσει την εργασία που θέλετε να παραδώσετε, την υποβάλετε στο eclass στο υποσύστημα «Εργασίες» (Assignments). Προσοχή: μόνο 1 άτομο από την ομάδα χρειάζεται να παραδώσει την εργασία μέσω του e-class! Η υποβολή πρέπει να γίνει **ΠΡΙΝ** την καταληκτική ημερομηνία παράδοσης. Παραδίδετε όλα τα αρχεία μαζί σε ένα συμπιεσμένο αρχείο (το οποίο θα φέρει τα ονόματα της ομάδας π.χ., RaftopoulouPapadopoulos.zip).
3. Περιπτώσεις αντιγραφής θα μηδενίζονται κι οι εμπλεκόμενοι δε θα έχουν δικαίωμα παράδοσης άλλων εργασιών.
4. Η ημερομηνία παράδοσης είναι αυστηρή, και η παράδοση γίνεται μόνο μέσω του eclass και όχι με email στη διδάσκουσα. Ασκήσεις που παραδίδονται μετά τη λήξη της προθεσμίας δε γίνονται δεκτές.

Ανάλυση της προσωπικότητα των πελατών - Ποιοι είναι οι ιδανικοί πελάτες;

Η Αμερικανική ψυχολογική εταιρία ορίζει την προσωπικότητα ως «ατομικές διαφορές στα χαρακτηριστικά πρότυπα σκέψης, συναισθημάτων, και συμπεριφοράς».[1] Για τους επαγγελματίες του μάρκετινγκ, τους υπεύθυνους επικοινωνίας, ακόμη και τις υπηρεσίες δημόσιας υγείας (που θέλουν για παράδειγμα να προωθήσουν πιο υγιεινές συμπεριφορές σε μεγάλους πληθυσμούς, όπως δίαιτα, διατροφή, άσκηση, διακοπή του καπνίσματος), η ανάλυση της προσωπικότητας βοηθά να ταιριάξουν πρακτικές με συγκεκριμένα προφίλ ατόμων και να προβλέψουν συμπεριφορές με βάση τα χαρακτηριστικά της προσωπικότητάς τους.[2] Για παράδειγμα, κανένας έμπορος δεν θέλει να παρουσιάσει ένα μήνυμα που είναι άσχετο με (ή ίσως και να θίγει) τα θέλω του πελάτη του. Η ανάλυση της προσωπικότητας προσφέρει λοιπόν, την ευκαιρία να κατανοήσουμε τους ανθρώπους και να τους δώσουμε το μήνυμα, τη διαφήμιση ή το περιεχόμενο με έναν τρόπο που είναι πιο πιθανό να έχει απήχηση σε εκείνους.

Αλήθεια εσείς τι τύπος είστε; Γιατί δεν κάνετε ένα quiz: https://www.scienceofpeople.com/personality/

Τα τελευταία χρόνια, οι ερευνητές στον τομέα της ανάλυσης προσωπικοτήτων, αντί να βασίζονται μόνο στις απαντήσεις των ανθρώπων σε ερωτηματολόγια, άρχισαν να χρησιμοποιούν τα ψηφιακά αποτυπώματα των ανθρώπων (π.χ., τις αγορές τους σε ένα ηλεκτρονικό κατάστημα, τα likes στο Facebook, τα tweets, το ιστορικό περιήγησης, κ.α.) για να βγάλουν συμπεράσματα σχετικά με την προσωπικότητά τους. Με βάση μεγάλα σύνολα δεδομένων που περιέχουν τόσο τις απαντήσεις των ανθρώπων σε παραδοσιακά ψυχομετρικά ερωτηματολόγια όσο και τις πληροφορίες που συλλέγονται στα ψηφιακά προφίλ τους, οι ερευνητές μπόρεσαν να εντοπίσουν εμπειρικές σχέσεις μεταξύ συγκεκριμένων ψηφιακών αποτυπωμάτων και συγκεκριμένων ψυχολογικών χαρακτηριστικών.[3] Για παράδειγμα, προέκυψαν ορισμένοι συσχετισμοί μεταξύ της αρέσκειας ενός

[1] https://www.apa.org/topics/personality
[2] https://towardsdatascience.com/a-data-science-approach-to-personality-models-d3c7e18a377
[3] https://www.researchgate.net/publication/273391924_Tracking_the_Digital_Footprints_of_Personality

συγκεκριμένου είδους μουσικής ή φαγητού και συγκεκριμένων χαρακτηριστικών προσωπικότητας. Και φυσικά, όσο περισσότερα είναι τα διαθέσιμα προς ανάλυση δεδομένα, τόσο πιο ακριβής είναι κι η τελική αξιολόγηση.

Το μάρκετινγκ με βάση την προσωπικότητα είναι μόνο μια πτυχή αυτής της σχετικά νέας, ταχέως αναδυόμενης προσέγγισης για την κατανόηση των ανθρώπων. Μπορούμε τώρα να περάσουμε από τυχαίες παρατηρήσεις στην αποκωδικοποίηση του τι πραγματικά κινεί τα άτομα και τι τα προκαλεί να δεθούν με έναν προϊόν ή να συμμετέχουν σε μια ενέργεια. Η ανάλυση προσωπικότητας πελατών βοηθά μια επιχείρηση να τροποποιήσει για παράδειγμα τα προϊόντα της με βάση τους πελάτες-στόχους της, ή αντί να δαπανά χρήματα για την προώθηση ενός νέου προϊόντος σε κάθε πελάτη στη βάση δεδομένων της, η εταιρία μπορεί να αναλύσει ποιοι πελάτες (ή ποια ομάδα πελατών) είναι πιο πιθανό να αγοράσουν το προϊόν και στη συνέχεια να διαθέσουν ή να προτείνουν το προϊόν μόνο σε αυτούς τους συγκεκριμένους πελάτες (ή ομάδα πελατών). Ο τρόπος βέβαια, με τον οποίο θα το κάνουμε αυτό θα καθορίσει εάν το μάρκετινγκ με βάση την προσωπικότητα χρησιμοποιείται για επικοινωνία με ενσυναίσθηση και θετικά αποτελέσματα (και για τους πελάτες) ή για χειραγώγηση και εκμετάλλευση.[4]

Σκοπός της εργασίας

Σκοπός αυτής της εργασίας είναι να μελετήσετε ένα σύνολο δεδομένων που αφορά στους πελάτες μιας αλυσίδας super market, η οποία διαθέτει φυσικά καταστήματα αλλά έχει και ψηφιακή παρουσία. Τα δεδομένα που σας δίνονται είναι δωρεάν και δημοσίως διαθέσιμα.

Η εργασία θα εκπονηθεί κατά προτίμηση από ομάδες των **2 ατόμων**, θα υλοποιηθεί σε **Java**, και μπορεί να γίνει σε οποιοδήποτε λειτουργικό σύστημα (συστήνεται το GNU/Linux).

Για τις ανάγκες της εργασίας, θα κατεβάσετε και θα εγκαταστήσετε το Hadoop στο μηχάνημά σας, κατά προτίμηση σε **Single-Node Local (Standalone) Mode**. Το mode αυτό προορίζεται για debugging και έχει ευκολότερη διαδικασία εγκατάστασης, οπότε το συστήνω μιας και θα κάνετε την υλοποίηση στο μηχάνημά σας κι όχι κάποιον cluster. Προσοχή, θα πρέπει να κατεβάσετε και να εγκαταστήσετε την έκδοση του Hadoop που ταιριάζει με το setup του μηχανήματός σας!

Εκτός από τις διαφάνειες του μαθήματος, στις οποίες συζητήσαμε για το Hadoop και για το MapReduce, επιπλέον πληροφορίες για την ακριβή λειτουργία τους μπορείτε να βρείτε στα Έγγραφα και στους Συνδέσμους (όλα διαθέσιμα στο eclass του μαθήματος).

[4] https://pubmed.ncbi.nlm.nih.gov/26348336/

Λειτουργικότητα του συστήματος

Σας δίνεται το αρχείο personality_analysis.csv που αποτελείται από 2.241 εγγραφές. Τα δεδομένα αφορούν στην 3/ετία 2019-2021. Κάθε εγγραφή σε αυτό το αρχείο αντιστοιχεί σε έναν πελάτη της εταιρίας και για κάθε εγγραφή έχουν αποθηκευτεί τα εξής 27 χαρακτηριστικά (attributes):

Ατομικά στοιχεία

1. ID: Μοναδικός κωδικός πελάτη
2. Year_Birth: Έτος γέννησης
3. Education: Μορφωτικό επίπεδο
4. Marital_Status: Οικογενειακή κατάσταση
5. Income: Ετήσιο οικογενειακό εισόδημα
6. Kidhome: Συνολικό πλήθος παιδιών
7. Teenhome: Αριθμός εφήβων
8. Dt_Customer: Ημερομηνία απόκτησης κωδικού στην εταιρία
9. Recency: Αριθμός ημερών από την τελευταία αγορά
10. Complain: 1 αν ο πελάτης έχει υποβάλει παράπονο τα τελευταία 2 χρόνια, 0 διαφορετικά

Στοιχεία αγορών

11. MntWines: Χρηματικό ποσό που έχει ξοδέψει σε κρασί τα τελευταία 2 χρόνια
12. MntFruits: Χρηματικό ποσό που έχει ξοδέψει σε φρούτα τα τελευταία 2 χρόνια
13. MntMeatProducts: Χρηματικό ποσό που έχει ξοδέψει σε κρέας τα τελευταία 2 χρόνια
14. MntFishProducts: Χρηματικό ποσό που έχει ξοδέψει σε ψάρι τα τελευταία 2 χρόνια
15. MntSweetProducts: Χρηματικό ποσό που έχει ξοδέψει σε γλυκά τα τελευταία 2 χρόνια
16. MntGoldProds: Χρηματικό ποσό που έχει ξοδέψει σε προϊόντα καλλωπισμού τα τελευταία 2 χρόνια

Στοιχεία προσφορών

17. NumDealsPurchases: Πλήθος αγορών που έγιναν με έκπτωση
18. AcceptedCmp1: 1 αν ο πελάτης δέχθηκε την προσφορά της 1ης προωθητικής ενέργειας, διαφορετικά 0
19. AcceptedCmp2: 1 αν ο πελάτης δέχθηκε την προσφορά της 2ης προωθητικής ενέργειας, διαφορετικά 0
20. AcceptedCmp3: 1 αν ο πελάτης δέχθηκε την προσφορά της 3ης προωθητικής ενέργειας, διαφορετικά 0
21. AcceptedCmp4: 1 αν ο πελάτης δέχθηκε την προσφορά της 4ης προωθητικής ενέργειας, διαφορετικά 0
22. AcceptedCmp5: 1 αν ο πελάτης δέχθηκε την προσφορά της 5ης προωθητικής ενέργειας, διαφορετικά 0
23. Response: 1 αν ο πελάτης δέχθηκε την προσφορά της πιο πρόσφατης προωθητικής ενέργειας, διαφορετικά 0

Τόπος αγορών

24. NumWebPurchases: Πλήθος αγορών που έγιναν μέσω της ιστοσελίδας της εταιρίας
25. NumCatalogPurchases: Πλήθος αγορών που έγιναν τηλεφωνικά
26. NumStorePurchases: Πλήθος αγορών που έγιναν σε φυσικό κατάστημα
27. NumWebVisitsMonth: Πλήθος επισκέψεων στην ιστοσελίδα της εταιρίας τον τελευταίο μήνα

Δείτε παρακάτω ένα παράδειγμα 3 εγγραφών:

5524, 1957, Graduation, Single, 58138, 0, 0, 04-09-2012, 58, 635, 88, 546, 172, 88, 88, 3, 8, 10, 4,7, 0, 0, 0, 0, 0, 0, 1
4141, 1965, Graduation, Together, 71613, 0, 0, 21-08-2013, 26, 426, 49, 127, 111, 21, 42, 1, 8, 2, 10, 4, 0, 0, 0, 0, 0, 0, 0
8180, 1952, Master, Divorced, 59354, 1, 1, 15-11-2013, 53, 233, 2, 53, 3, 5, 14, 3, 6, 1, 5, 6, 0, 0, 0, 0, 0, 0, 0

Για λόγους παρουσίασης, στο παραπάνω παράδειγμα έχουμε εισάγει κόμματα ανάμεσα στα χαρακτηριστικά κάθε εγγραφής. Σημειώστε ότι η σειρά με την οποία εμφανίζονται τα χαρακτηριστικά σε αυτές τις εγγραφές αντιστοιχεί στο αρχείο των δεδομένων κι όχι στην αρίθμηση της λίστας που έχω δώσει παραπάνω. Δείτε με προσοχή στο αρχείο των δεδομένων τη σειρά εμφάνισης των χαρακτηριστικών.

Μελέτη των δεδομένων [10%]

Μελετήστε τα δεδομένα και τα χαρακτηριστικά των εγγραφών που εμφανίζονται στο αρχείο personality_analysis.csv. Γράψτε στην αναφορά σας πόσα/ποια από τα χαρακτηριστικά είναι numeric και πόσα/ποια είναι nominal. Στη συνέχεια, στην αναφορά σας καθορίστε/υπολογίστε τις τιμές ηλικιών και τα ποσά εισοδημάτων που θα θεωρούσατε ως εξαιρέσεις (outliers) και αιτιολογήστε την απάντησή σας.

Προ-επεξεργαστείτε ή καθαρίστε τα δεδομένα σας (για όσες εγγραφές ή χαρακτηριστικά το κρίνετε σκόπιμο) και εξηγήσετε στην αναφορά το σκεπτικό σας. Οποιαδήποτε προ-επεξεργασία κάνετε στα δεδομένα σας δε θα πρέπει να γίνει με φίλτρα του excel (ή αντίστοιχου προγράμματος), αλλά θα πρέπει να χρησιμοποιήσετε κώδικα (π.χ., script) δικό σας ή τρίτων (αρκεί να δίνετε την πηγή στη γραπτή αναφορά σας), τον οποίο θα συμπεριλάβετε στα αρχεία που θα ανεβάσετε στο eclass.

Έπειτα, χρησιμοποιείστε τα (επεξεργασμένα) δεδομένα σας για να κάνετε τους παρακάτω υπολογισμούς και

καταγράψτε σε αντίστοιχα αρχεία εξόδου τα αποτελέσματα που θα πάρετε. Σημειώστε ότι οι εγγραφές που εμφανίζονται στα παραδείγματα που ακολουθούν είναι τυχαίες και δεν αντιστοιχούν σε πραγματικούς υπολογισμούς.

Εκπαιδευτικό υπόβαθρο πελατών [15%]

(α) Γράψτε τον **ψευδοκώδικα** για τη λειτουργία Map και για τη λειτουργία Reduce και βρείτε το πλήθος των πελατών ανά βαθμίδα εκπαίδευσης. Δώστε ένα **παράδειγμα** και μια **σχηματική εκτέλεση** που να εξηγεί τη λογική πίσω από το ψευδοκώδικά σας.

(β) Υλοποιήστε στο Hadoop το σχεδιασμό που κάνατε παραπάνω. Το αρχείο εξόδου θα πρέπει να περιέχει τόσες γραμμές όσες και οι διαφορετικές βαθμίδες εκπαίδευσης που εμφανίζονται στα δεδομένα, οι βαθμίδες να είναι ταξινομημένες σε αλφαβητική σειρά, και σε κάθε γραμμή να εμφανίζονται (i) η βαθμίδα εκπαίδευσης και (ii) το πλήθος των πελατών που ανήκουν σε αυτήν τη βαθμίδα.

Δείτε παρακάτω ένα παράδειγμα:

```
Basic 54
Graduation 1127
...
```

Ανάδειξη των πελατών που είναι πιο πιθανό να αγοράσουν ένα νέο προϊόν κρασιού [25%]

(α) Γράψτε τον **ψευδοκώδικα** για τη λειτουργία Map και για τη λειτουργία Reduce και βρείτε τους πελάτες που τα τελευταία δύο έτη ξόδεψαν για κρασί 50% περισσότερα χρήματα από το μέσο όρο των πελατών της εταιρίας. Δώστε ένα **παράδειγμα** και μια **σχηματική εκτέλεση** που να εξηγεί τη λογική πίσω από το ψευδοκώδικά σας.

(β) Υλοποιήστε στο Hadoop το σχεδιασμό που κάνατε παραπάνω. Το αρχείο εξόδου θα πρέπει να περιέχει τους πελάτες ταξινομημένους σε φθίνουσα σειρά ως προς τα συνολικά χρήματα που ξόδεψαν για κρασί τα τελευταία δύο χρόνια και σε κάθε γραμμή να εμφανίζονται (i) η σειρά κατάταξης, (ii) ο μοναδικός κωδικός του πελάτη (ID), (iii) η ηλικία του, (iv) το επίπεδο εκπαίδευσής του (Education), (v) η οικογενειακή του κατάσταση (Marital_Status), (vi) το ετήσιο εισόδημά του (Income), και (vii) το συνολικό ποσό που ξόδεψε σε κρασί (MntWines). Σε περίπτωση που υπάρχουν παραπάνω από ένας πελάτες που ξόδεψαν το ίδιο ποσό, τότε πιο ψηλά στην κατάταξη θεωρείστε τον πελάτη με το μεγαλύτερο εισόδημα.

Δείτε παρακάτω ένα παράδειγμα:

```
1 7431 62 Graduation Together 87771 1492
2 5547 39 PhD Married 84169 1478
3 11088 50 PhD Together 78642 1396
...
```

Κατηγοριοποίηση των πελατών ανάλογα με την αγοραστική τους δυναμική [35%]

Ας θεωρήσουμε τις εξής 4 κατηγορίες πελατών.

1. Χρυσοί πελάτες - Gold: Οι πελάτες που έχουν έρθει πρόσφατα στην εταιρία (δηλ. έχουν αποκτήσει κωδικό το τελευταίο έτος), με υψηλά ετήσια οικογενειακά εισοδήματα (>69500), και με την τάση να ξοδεύουν πολλά χρήματα σε κάθε τους αγορά (50% περισσότερα από το μέσο όρο).
2. Αργυροί πελάτες - Silver: Οι παλιοί πελάτες της εταιρίας (δηλ. έχουν αποκτήσει κωδικό σε προηγούμενα έτη), με υψηλά ετήσια οικογενειακά εισοδήματα (>69500), και με την τάση να ξοδεύουν πολλά χρήματα σε κάθε τους αγορά (50% περισσότερα από το μέσο όρο).
3. Χάλκινοι πελάτες - Bronze: Οι πελάτες που έχουν έρθει πρόσφατα στην εταιρία (δηλ. έχουν αποκτήσει κωδικό το τελευταίο έτος), με ετήσια οικογενειακά εισοδήματα κάτω από τον μέσο όρο, και με την τάση να ξοδεύουν λίγα χρήματα σε κάθε τους αγορά (25% του μέσου όρου).
4. Χάρτινοι πελάτες - Paper: Οι παλιοί πελάτες της εταιρίας (δηλ. έχουν αποκτήσει κωδικό σε προηγούμενο έτος), με ετήσια οικογενειακά εισοδήματα κάτω από τον μέσο όρο, και με την τάση να ξοδεύουν λίγα χρήματα σε κάθε τους αγορά (25% του μέσου όρου).

(α) Γράψτε τον **ψευδοκώδικα** για τη λειτουργία Map και για τη λειτουργία Reduce για να βρείτε τους χρυσούς και τους αργυρούς πελάτες της εταιρίας. Δώστε ένα **παράδειγμα** και μια **σχηματική εκτέλεση** που να εξηγεί τη λογική πίσω από το ψευδοκώδικά σας.

(β) Υλοποιήστε στο Hadoop το σχεδιασμό που κάνατε παραπάνω. Το αρχείο εξόδου θα πρέπει να περιέχει δυο γραμμές, μια για την κατηγορία Gold και μια για την κατηγορία Silver, και σε κάθε γραμμή να εμφανίζονται (i) η κατηγορία πελατών και (ii) οι μοναδικοί κωδικοί (ID) των πελατών που ανήκουν σε αυτήν την κατηγορία χωρισμένοι με κόμματα και ταξινομημένοι σε αύξουσα σειρά.

Δείτε παρακάτω ένα παράδειγμα:

```
Gold 5524, 10026, …
Silver 1002, 4141, 5879, …
```

Αναφορά [15%]

Η αναφορά σας θα πρέπει να έχει έκταση τουλάχιστον πέντε σελίδες (χωρίς το εξώφυλλο και τις αναφορές/παραπομπές) και θα πρέπει να περιέχει:

- τη μελέτη των δεδομένων που σας ζητήθηκε παραπάνω
- τις ενέργειες σας μαζί με κατάλληλη αιτιολόγηση σε σχέση με την (προ-)επεξεργασία των δεδομένων (π.χ., κάνατε κάποιου είδους καθαρισμό των δεδομένων, πώς διαχειριστήκατε τις ημερομηνίες, κοκ.),
- τον ψευδοκώδικα, μαζί με ένα παράδειγμα και μια σχηματική εκτέλεση για κάθε ένα από τα ερωτήματα,
- οδηγίες για την εκτέλεση του προγράμματος (user manual),
- λεπτομέρειες της υλοποίησης που αξίζει να σημειωθούν,
- για κάθε ένα από τα ερωτήματα, τουλάχιστον μια γραφική απεικόνιση (εσείς θα αποφασίσετε ποια) των αποτελεσμάτων και τα σχόλιά σας σε σχέση με τα αποτελέσματα (π.χ., τι αναμένατε και τι πήρατε ως αποτέλεσμα, κοκ.).

Μην συμπεριλάβετε στην αναφορά σας μέρος ή σύνολο του κώδικα, εκτός αν αφορά σε κάποια σύντομη επεξήγηση!

Bonus υλοποίηση

Σε συνεννόηση με τη διδάσκουσα μπορείτε να πάρετε μέχρι 10% bonus για επιπλέον λειτουργικότητα ή ερωτήματα που θα υλοποιήσετε. Για παράδειγμα, θα μπορούσατε να βρείτε τους πελάτες της εταιρίας που ανήκουν στις άλλες δυο κατηγορίες, να υλοποιήσετε τα παραπάνω ερωτήματα σε **Pig Latin**, κα.

Παραδοτέα και βαθμολόγηση

Πλήρης θεωρείται η εργασία η οποία **υλοποιεί σωστά τις βασικές απαιτήσεις** που περιγράφονται παραπάνω. Ασκήσεις που υλοποιούν μόνο ένα μέρος των βασικών απαιτήσεων λαμβάνουν και αντίστοιχο μέρος του βαθμού.

Τα παραδοτέα της εργασίας είναι:

- τυχόν scripts ή κώδικα τρίτων που χρησιμοποιήσατε,
- τα αρχεία πηγαίου κώδικα,
- τα εκτελέσιμα αρχεία,
- τα αρχεία με τα αποτελέσματα για κάθε ένα από τα παραπάνω ερωτήματα,
- η γραπτή αναφορά.

Καλή επιτυχία!