Διαχείριση Μεγάλων Δεδομένων
2η Προγραμματιστική Εργασία

Διδάσκουσα:
Π. Ραυτοπούλου

Παράδοση μέχρι: **Κυριακή 20/02/2022 ώρα 23.59**
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 (ομάδες 2 ατόμων).
2. Αφού έχετε ολοκληρώσει την εργασία που θέλετε να παραδώσετε την υποβάλετε στο eclass στο υποσύστημα «Εργασίες φοιτητών». 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 που σας ζητούνται 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 των παραγόντων που επηρεάζουν πόσο δημοφιλές θα είναι ένα βίντεο YouTube

Το YouTube[1] είναι μια συνεχώς ανερχόμενη διαδικτυακή πλατφόρμα πολυμεσικού περιεχομένου, η οποία, κατά το διάστημα 2020-2021, αναδείχτηκε ως η δεύτερη πιο δημοφιλής μηχανή αναζήτησης (μετά το Google). Η πλατφόρμα ιδρύθηκε στις 14 Φεβρουαρίου του 2005, εξαγοράστηκε από την Google έναντι $1.65 δισεκατομμυρίων τον Οκτώβρη του 2006, διατηρεί περισσότερους από ένα δισεκατομμύριο μηνιαίους χρήστες που παρακολουθούν συνολικά περισσότερες από ένα δισεκατομμύριο ώρες βίντεο κάθε μέρα.

Μετά την αγορά από τη Google, η πλατφόρμα άλλαξε το επιχειρηματικό της μοντέλο αλλά και την πολιτική των προσφερόμενων υπηρεσιών. Το YouTube δεν παράγει πλέον έσοδα μόνο από διαφημίσεις, αλλά προσφέρει και περιεχόμενο επί πληρωμή, όπως για παράδειγμα ταινίες. Επιπλέον, το YouTube επεκτάθηκε ώστε να υποστηρίζει εφαρμογές για κινητά, δικτυακή τηλεόραση, καθώς και τη δυνατότητα σύνδεσης με άλλες υπηρεσίες. Οι κατηγορίες βίντεο που συναντά κανείς στο YouTube περιλαμβάνουν μουσικά βίντεο, βίντεο κλιπ, ειδήσεις, ταινίες μικρού μήκους, ταινίες μεγάλου μήκους, ντοκιμαντέρ, ηχογραφήσεις, τρέιλερ ταινιών, και άλλα. Το μεγαλύτερο μέρος του περιεχομένου δημιουργείται από χρήστες (αλλιώς YouTubers ☺), και μπορεί να περιλαμβάνει και συνεργασίες μεταξύ χρηστών και εταιρικών χορηγών. Από το 2015, γνωστές εταιρείες που παράγουν πολυμεσικό περιεχόμενο, όπως η Disney, η ViacomCBS και η WarnerMedia, έχουν δημιουργήσει τα δικά τους κανάλια στο YouTube για να έχουν απήχηση σε μεγαλύτερο κοινό.

Το μεγάλο ενδιαφέρον του κοινού προς τη συγκεκριμένη πλατφόρμα οφείλεται στη ψηφιοποίηση στην οποία υπόκειται τα τελευταία χρόνια ο τομέας της ψυχαγωγίας, καθώς και στην αυξανόμενη ζήτηση των χρηστών για οπτικοακουστικό υλικό. Το YouTube προσφέρεται να καλύψει τις παραπάνω ανάγκες, δίνοντας τη δυνατότητα στους χρήστες να κοινοποιούν εύκολα βίντεο, μικρής ή μεγάλης διάρκειας, αλλά και να αλληλεπιδρούν με το υλικό άλλων δημιουργών. Έτσι, αν κι ο αρχικός σκοπός του YouTube ήταν η δημιουργία ενός μέσου για την παρακολούθηση βίντεο, η τάση του κοινού προς την ανάπτυξη διαδικτυακών κοινωνικών σχέσεων συνέβαλλε στη μετεξέλιξη της πλατφόρμας σε ένα κοινωνικό δίκτυο, με χαρακτήρα παρόμοιο με αυτό των υπόλοιπων

[1] https://www.youtube.com

μέσων κοινωνικής δικτύωσης. Το YouTube έχει στις μέρες μας πρωτοφανή κοινωνικό αντίκτυπο, επηρεάζοντας την κουλτούρα εκατομμυρίων χρηστών, τις τάσεις στο Διαδίκτυο, και δημιουργώντας πολυεκατομμυριούχους-διάσημους Youtubers. Παρά την ανάπτυξη και την επιτυχία του, το YouTube έχει επίσης επικριθεί ευρέως ως ένας ιστότοπος που διευκολύνει τη διάδοση παραπληροφόρησης, για ζητήματα πνευματικών δικαιωμάτων και παραβιάσεων του απορρήτου των χρηστών του, επειδή επιτρέπει τη λογοκρισία, καθώς επίσης διότι θεωρείται ότι θέτει σε κίνδυνο την ασφάλεια των παιδιών.

Το YouTube συνδυάζει διάφορες παραμέτρους[2] (π.χ., αριθμό προβολών, ταχύτητα δημιουργίας προβολών, ηλικία του βίντεο, κ.α.) για να καταλήξει κάθε μέρα σε μια λίστα με τα πιο δημοφιλή βίντεο (trending videos[3]) σε κάθε χώρα. Η λίστα αυτή αντικατοπτρίζει το περιεχόμενο της πλατφόρμας και τις προτιμήσεις των χρηστών της. Προσέξτε ότι το βίντεο με τον υψηλότερο αριθμό προβολών μια δεδομένη ημέρα ενδέχεται να μην είναι το #1 στις τάσεις και να εμφανίζεται στη λίστα κάτω από βίντεο με λιγότερες προβολές.

Είστε έτοιμοι να εξερευνήσετε τον κόσμο του YouTube; ☺

Σκοπός της εργασίας

Σκοπός αυτής της εργασίας είναι να διαχειριστείτε τα δεδομένα αρκετών μηνών που αφορούν στα καθημερινώς ανερχόμενα βίντεο της πλατφόρμας YouTube. Η αποθήκευση και η διαχείριση των δεδομένων θα πρέπει να γίνει μέσω μιας NoSQL βάσης δεδομένων, και συγκεκριμένα του **MongoDB**[4] document store. Αρχικά θα πρέπει να εγκαταστήσετε το MongoDB στον υπολογιστή σας, έπειτα να αποθηκεύσετε τα δεδομένα σας στο MongoDB, και τέλος να κάνετε μια μελέτη/ανάλυση των δεδομένων αυτών.

Εκτός από τις διαφάνειες του μαθήματος, στις οποίες συζητήσαμε για το MongoDB, πληροφορίες για την ακριβή λειτουργία του μπορείτε να βρείτε και στους ακόλουθους συνδέσμους (είναι διαθέσιμοι και στο eclass του μαθήματος):

[1] https://www.mongodb.com/try/download/
[2] https://www.mongodb.com/developer-tools
[3] https://docs.mongodb.com/manual/
[4] https://docs.mongodb.com/guides/

Η εργασία θα εκπονηθεί από ομάδες των **2 ατόμων** και μπορεί να γίνει σε οποιοδήποτε λειτουργικό σύστημα.

1. Αποθήκευση δεδομένων στο MongoDB [25 μονάδες]

Τα δεδομένα θα τα βρείτε μέσω του Kaggle, το οποίο είναι ένα τεράστιο αποθετήριο δημοσιευμένων και δωρεάν διαθέσιμων συνόλων δεδομένων. Για να τα κατεβάσετε στον υπολογιστή σας, επισκεφείτε τη σελίδα https://www.kaggle.com/datasnaek/youtube-new/data και πατήστε Download (539MB) πάνω δεξιά. Με αυτό τον τρόπο θα αποκτήσετε πρόσβαση στο πλήρες σύνολο δεδομένων, το οποίο περιλαμβάνει εκείνα τα βίντεο που αναδείχθηκαν ως τα πιο δημοφιλή για το διάστημα από τις 14 Νοεμβρίου 2017 ως και τις 5 Μαρτίου 2018· τα δεδομένα βρίσκονται σε 20 αρχεία, δέκα εκ των οποίων σε JSON μορφή και τα υπόλοιπα σε CSV.

Πιο αναλυτικά, τα δεδομένα αφορούν στις εξής δέκα περιοχές: Η.Π.Α. (US), Μεγάλη Βρετανία (GB), Γερμανία (DE), Καναδά (CA), Γαλλία (FR), Ρωσία (RU), Μεξικό (MX), Νότια Κορέα (KR), Ιαπωνία (JP), και Ινδία (IN), και περιλαμβάνουν έως και 200 καταγεγραμμένα δημοφιλή βίντεο ανά ημέρα ανά περιοχή (λείπουν τα στοιχεία για τις 10 και 11 Ιανουαρίου 2018). Τα δεδομένα κάθε περιοχής βρίσκονται σε ξεχωριστό αρχείο (XXvideos.csv, όπου XX η περιοχή), και περιλαμβάνουν τα παρακάτω χαρακτηριστικά/πεδία (attributes/fields):

- video_id - ο μοναδικός κωδικός του βίντεο,
- trending_date - η ημερομηνία που το βίντεο βρέθηκε στη λίστα δημοφιλών βίντεο (στη μορφή YY.DD.MM),
- title - ο τίτλος του βίντεο,
- channel_title - ο τίτλος του καναλιού το οποίο δημοσίευσε το βίντεο,
- category_id - ο κωδικός της κατηγορίας στην οποία ανήκει το βίντεο,
- publish_time - η ημερομηνία δημοσίευσης του βίντεο (σε ISO 8601 μορφή),
- tags - οι ετικέτες που έχουν χρησιμοποιηθεί στο βίντεο,
- views - ο αριθμός προβολών του,
- likes - ο αριθμός των "μου αρέσει" που έχει λάβει το βίντεο,

[2] https://support.google.com/youtube/answer/7239739?hl=en
[3] https://www.youtube.com/feed/trending
[4] https://www.mongodb.com

- dislikes - ο αριθμός των "δεν μου αρέσει" που έχει λάβει το βίντεο,
- comment_count - ο αριθμός σχολίων που έχει λάβει το βίντεο,
- thumbnail_link - ένας σύνδεσμο στο βίντεο,
- comments_disabled - FALSE αν επιτρέπεται να σχολιάζουν οι χρήστες, αλλιώς TRUE,
- ratings_disabled - FALSE αν επιτρέπεται να βαθμολογούν οι χρήστες, αλλιώς TRUE,
- video_error_or_removed - FALSE αν δεν ισχύει, αλλιώς TRUE, και
- description - η περιγραφή του βίντεο.

Σημειώστε ότι οι κατηγορίες των βίντεο διαφέρουν μεταξύ των περιοχών. Για να ανακτήσετε τις κατηγορίες για ένα συγκεκριμένο βίντεο, θα πρέπει να δείτε το συσχετισμένο JSON αρχείο. Προσέξτε επίσης, ότι στα αρχεία με τα δεδομένα των βίντεο δεν υπάρχει κάποιο πεδίο που να δηλώνει την περιοχή προέλευσης μιας εγγραφής.

Έτσι για παράδειγμα, τα δεδομένα που προέρχονται από τη Μεγάλη Βρετανία (GB) περιλαμβάνουν 38.916 εγγραφές (αρχείο GBvideos.csv) και μπορεί να ανήκουν σε 31 διαφορετικές κατηγορίες (αρχείο GB_category_id.json)· παραδείγματα κατηγοριών είναι "Film & Animation", "Autos & Vehicles", "Music", κοκ.

Δείτε ένα μικρό υποσύνολο των δεδομένων (για λόγους παρουσίασης παραλείπονται πεδία):

video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes
Jw1Y-zhQURU	17.14.11	John Lewis Christmas Ad 2017 - #MozTheMonster	John Lewis	26	2017-11-10T07:38:29.000Z	christmas|"john lewis christmas"|"john lewis"|"christmas ad"|"mozthemonster"|"christmas 2017"|"christmas ad 2017"|"john lewis christmas advert"|"moz"	7224515	55681	10247
3s1rvMFUweQ	17.14.11	Taylor Swift: …Ready for It? (Live) - SNL	Saturday Night Live	24	2017-11-12T06:24:44.000Z	SNL|"Saturday Night Live"|"SNL Season 43"|"Episode 1730"|"Tiffany Haddish"|"Taylor Swift"|"Taylor Swift Ready for It"|"s43"|"s43e5"|"episode 5"|"live"|"new york"|"comedy"|"sketch"|"funny"|"hilarious"|"late night"|"host"|"music"|"guest"|"laugh"|"impersonation"|"actor"|"improv"|"musician"|"comedian"|"actress"|"If Loving You Is Wrong"|"Oprah Winfrey"|"OWN"|"Girls Trip"|"The Carmichael Show"|"Keanu"|"Reputation"|"Look What You Made Me Do"|"ready for it?"	1053632	25561	2294
n1WpP7iowLc	17.14.11	Eminem - Walk On Water (Audio) ft. Beyoncé	Eminem VEVO	10	2017-11-10T17:00:03.000Z	Eminem|"Walk"|"On"|"Water"|"Aftermath/Shady/Interscope"|"Rap"	17158579	787420	43420
PUTEiSjKwJU	17.14.11	Goals from Salford City vs Class of 92 and Friends at The Peninsula Stadium!	Salford City Football Club	17	2017-11-13T02:30:38.000Z	Salford City FC|"Salford City"|"Salford"|"Class of 92"|"University of Salford"|"Salford Uni"|"Non League"|"National League"|"National League North"	27833	193	12

Μελετήστε τα δεδομένα σας, δείτε προσεκτικά τις τιμές για κάθε χαρακτηριστικό, παρατηρήστε ότι κάποια πεδία (π.χ., tags) μπορεί έχουν πολλαπλές τιμές, ή ενδέχεται να υπάρχουν εγγραφές που έχουν κενά κάποια από τα πεδία τους, προσέξτε επίσης ότι οι trending και publish ημερομηνίες έχουν διαφορετική μορφή, κοκ. Σε μια σχεσιακή βάση δεδομένων, θα έπρεπε να αποφασίσουμε ποιο είναι το schema των δεδομένων μας και πώς θα αντιμετωπίσουμε τέτοιου είδους διαφοροποιήσεις στις τιμές των χαρακτηριστικών. Στα πλαίσια αυτής της εργασίας, δε χρειαζόμαστε schema καθώς θα δουλέψουμε με μια NoSQL βάση δεδομένων, και συγκεκριμένα με το MongoDB document store όπου τα δεδομένα αποθηκεύονται ως JSON έγγραφα.

JavaScript Object Notation (JSON) format

Το JSON είναι ένα format περιγραφής κι ανταλλαγής δεδομένων, το οποίο είναι εύκολα κατανοητό από τους ανθρώπους, αλλά κυρίως προσφέρει έναν εύκολο τρόπο για αποθήκευση, ανάκτηση και ανάλυση δεδομένων στα μηχανήματα. Με λίγα λόγια, το JSON είναι μια μορφή κειμένου που είναι εντελώς ανεξάρτητη από τη γλώσσα, υποστηρίζει ενσωματωμένα πεδία (επομένως σχετιζόμενα δεδομένα ή λίστες δεδομένων) και χρησιμοποιεί συμβάσεις που είναι γνωστές στους προγραμματιστές· αυτές οι ιδιότητες καθιστούν το JSON ιδανικό για αποθήκευση και ανταλλαγή δεδομένων. Εκτός από τις διαφάνειες του μαθήματος, περισσότερα για το JSON και για την μοντελοποίηση των δεδομένων στο MongoBD μπορείτε να βρείτε και στους παρακάτω συνδέσμους (είναι διαθέσιμοι και στο eclass του μαθήματος):

[7] https://docs.mongodb.com/manual/core/data-modeling-introduction/
[8] https://www.json.org/

Μετατροπή των δεδομένων σε JSON format

Στα πλαίσια αυτής της εργασίας, θα πρέπει να μετατρέψετε **σε JSON format** όσα από τα δεδομένα σας είναι σε CSV. Δείτε για παράδειγμα την εγγραφή:

```
{
  "video_id": "Jw1Y-zhQURU",
  "trending_date": "2017-11-14",
  "title": "John Lewis Christmas Ad 2017 - #MozTheMonster",
  "channel_title": "John Lewis",
  …
  "tags": ["Christmas","john lewis christmas","john lewis","christmas ad","mozthemonster","christmas 2017",
 "christmas ad 2017","john lewis christmas advert","moz"],
  …
 }
```

Σημειώστε ότι η παραπάνω εγγραφή είναι μόνο ένα παράδειγμα, το οποίο δε δείχνει όλη τη δυναμική του JSON format. Σε αυτό το παράδειγμα, οι ετικέτες είναι αποθηκευμένες σε πίνακα, αλλά υπάρχουν και άλλες επιλογές/δομές που μπορούν να υποστηρίξουν πολλαπλές τιμές σε κάποιο πεδίο.

Για την μετατροπή των δεδομένων σε JSON μπορείτε να χρησιμοποιήσετε κάποιον από τους csv σε json μετατροπείς που είναι διαθέσιμοι online (αρκεί να αναφέρετε την πηγή σας στην αναφορά) ή να φτιάξετε κάποιον δικό σας (μια καλή ιδέα θα ήταν με χρήση Shell Script[5]) ή να το κάνετε μέσω των εργαλείων του MongoDB. Σε κάθε περίπτωση σκεφτείτε προσεκτικά πώς θα διαχωρίσετε τα δεδομένα που προέρχονται από διαφορετικές περιοχές, πώς θα διαχειριστείτε τις πολλαπλές τιμές σε ένα πεδίο, καθώς επίσης και τον τρόπο που θα εισάγετε τις ημερομηνίες. Θα πρέπει να εξηγήστε τις επιλογές και τους χειρισμούς που κάνατε στην αναφορά σας.

Εγκατάσταση του MongoDB document store

Για τις ανάγκες της εργασίας, θα κατεβάσετε και θα εγκαταστήσετε τη **δωρεάν έκδοση** του **MongoDB** (Products -> Community Edition) στο μηχάνημά σας, επιλέγοντας είτε το Community είτε το Enterprise Server[6]. Η τελευταία σταθερή έκδοση είναι η **V5.0.5**, αλλά μπορείτε να χρησιμοποιήσετε και την παλαιότερη V4.4, προσέξτε μόνο να κατεβάσετε το κατάλληλο αρχείο ανάλογα με το λειτουργικό σας σύστημα. Για την εγκατάσταση δείτε και τις αναλυτικές οδηγίες από εδώ: https://docs.mongodb.com/manual/installation/

Εισαγωγή των δεδομένων στο MongoDB

Στη συνέχεια, αφού εγκαταστήσετε τη MongoDB στον υπολογιστή σας, θα πρέπει να εισάγετε τα δεδομένα σας στη βάση. Δείτε παρακάτω ότι η ανάλυση που καλείστε να κάνετε αφορά σε υποσύνολο των δεδομένων, οπότε μπορείτε να εισάγετε στη βάση σας μόνο εκείνα που θα χρειαστείτε. Η εισαγωγή των δεδομένων μπορεί να γίνει μέσω του **mongo Shell**[7] (διανέμεται μαζί με το MongoDB). Αντί αυτού, μια καλή ιδέα θα ήταν να χρησιμοποιήσετε το εργαλείο **mongoimport**[8] για μαζική εισαγωγή δεδομένων στο document store. Αυτό το εργαλείο δε διανέμεται μαζί με το MongoDB, οπότε, αν θέλετε να το χρησιμοποιήσετε, θα πρέπει να το

[5] https://www.shellscript.sh
[6] Οποιαδήποτε από τις δυο κάνει για τις ανάγκες της εργασίας, ως developer έχω μια προτίμηση στην Enterprise server, αλλά εσείς εγκαταστήστε όποια από τις δυο επιθυμείτε.
[7] https://docs.mongodb.com/manual/introduction/
[8] https://docs.mongodb.com/database-tools/

κατεβάσετε και να το εγκαταστήσετε ανεξάρτητα. Για την εγκατάστασή του μπορείτε να ακολουθήσετε τις αναλυτικές οδηγίες από εδώ: https://www.mongodb.com/try/download/database-tools

Στη συνέχεια, ανακτήστε τα έγγραφα από τη συλλογή σας, δείτε κι ελέγξτε τα αποτελέσματά σας, σιγουρευτείτε ότι όλα πήγαν καλά ως εδώ.

2. Ανάλυση των δεδομένων

Τι μπορείτε να κάνετε με αυτά τα δεδομένα; Θα αρχίσετε απαντώντας ερωτήσεις και στη συνέχεια, δεδομένων των απαντήσεων, θα κάνετε μια μελέτη του περιεχομένου της πλατφόρμας YouTube.

Για κάθε ένα από τα παρακάτω ερωτήματα θα εξάγετε όλες τις απαντήσεις σε αρχείο αποτελεσμάτων, το οποίο και θα ανεβάσετε στο eclass μαζί με τα υπόλοιπα παραδοτέα της άσκησης. Επίσης, για κάθε ένα από τα ερωτήματα θα συμπεριλάβετε στη γραπτή αναφορά σας το ερώτημα που συντάξατε, τις 20 πρώτες εγγραφές των αποτελεσμάτων, τις γραφικές αναπαραστάσεις που σας ζητούνται, καθώς επίσης και τα σχόλιά σας. Αν οι απαντήσεις για κάποιο ερώτημα είναι λιγότερες από 20, τότε θα τις συμπεριλάβετε όλες στην αναφορά σας.

Η διατύπωση των ερωτημάτων σας πάνω από τα αποθηκευμένα δεδομένα (json documents) μπορεί να γίνει μέσω του **mongo Shell** και κάνοντας χρήση της mongoDB γλώσσας ερωτήσεων[9,10] (mongoQL). Αντί αυτού, μια καλή ιδέα θα ήταν να χρησιμοποιήσετε **MongoDB Compass**[11] που ουσιαστικά είναι το GUI για το MongoDB, οπότε και δεν χρειάζεται να αποκτήσετε κάποια ιδιαίτερη εξοικείωση με τη σύνταξη ερωτημάτων σε mongoQL. Αν θέλετε να χρησιμοποιήσετε το compass θα πρέπει να το κατεβάσετε και να το εγκαταστήσετε ανεξάρτητα. Για την εγκατάσταση του μπορείτε να ακολουθήσετε τις αναλυτικές οδηγίες από εδώ: https://docs.mongodb.com/compass/master/install

Για τις τις γραφικές απεικονίσεις θα μπορούσατε να χρησιμοποιήσετε **Gnuplot**[12], **robomongo**[13], **Grafana**[14], **Visual Studio Code**[15], ή και κάποιο άλλο εργαλείο που εσείς θα επιλέξετε σε συνεννόηση με τη διδάσκουσα. Η επιλογή του εργαλείου απεικόνισης (μαζί με λεπτομέρειες που αξίζει να σημειωθούν) θα πρέπει να συμπεριληφθούν στην αναφορά.

2.1 Τι ξέρουμε για τις δημοσιεύσεις του πολύ δημοφιλούς καναλιού Saturday Night Live[16]; [10 μονάδες]

Για την περιοχή GB, βρείτε τις δημοσιεύσεις τους καναλιού «Saturday Night Live». Για κάθε εγγραφή στο αρχείο αποτελεσμάτων, εμφανίστε τον τίτλο του βίντεο, τον αριθμό των προβολών του, καθώς και τους αριθμούς των «μου αρέσει» και «δεν μου αρέσει» που έχει λάβει το βίντεο, ταξινομημένα ως προς το πλήθος προβολών σε φθίνουσα σειρά. Σχολιάστε τα αποτελέσματά σας.

2.2 Πόσες ετικέτες χρησιμοποιούνται συνήθως στις δημοσιεύσεις των βίντεο; [15 μονάδες]

Για την περιοχή GB, βρείτε το πλήθος των ετικετών που χρησιμοποιούνται ανά βίντεο. Για κάθε εγγραφή στο αρχείο αποτελεσμάτων, εμφανίστε τον κωδικό του βίντεο, το πλήθος των ετικετών, και τον αριθμό των προβολών του βίντεο, ταξινομημένα ως προς το πλήθος προβολών σε φθίνουσα σειρά. Κάντε μια γραφική απεικόνιση (scatter plot) των αποτελεσμάτων που πήρατε βάζοντας στον x άξονα το πλήθος των προβολών και στο y άξονα το πλήθος των ετικετών. Τι συμπεράσματα βγάζετε; Σχολιάστε τα αποτελέσματά σας. Έχουν τα βίντεο με τις περισσότερες ετικέτες το μεγαλύτερο αριθμό προβολών; Βοηθά η χρήση πολλών ετικετών ένα βίντεο να έχει περισσότερες προβολές; Διακρίνετε εξαιρέσεις (outliers) στα αποτελέσματα; κοκ.

2.3 Ποιες είναι οι πιο δημοφιλείς ετικέτες στα ανερχόμενα βίντεο; [15 μονάδες]

Για τις περιοχές GB και USA, μετρήστε για κάθε ετικέτα τον αριθμό των βίντεο όπου αυτή εμφανίζεται. Θα έχετε 2 αρχεία αποτελεσμάτων, ένα για κάθε περιοχή, και για κάθε εγγραφή στα αρχεία αποτελεσμάτων εμφανίστε την ετικέτα και το πλήθος των βίντεο που χρησιμοποιούν τη συγκεκριμένη ετικέτα. Τα αποτελέσματά σας θα πρέπει να είναι ταξινομημένα σε φθίνουσα σειρά ως προς το πλήθος των βίντεο. Φτιάξτε για κάθε περιοχή bar chart με τα αποτελέσματα που πήρατε. Σχολιάστε τα αποτελέσματά σας.

---

[9] https://docs.mongodb.com/manual/tutorial/query-documents/
[10] https://docs.mongodb.com/manual/reference/operator/aggregation/index.html
[11] https://docs.mongodb.com/compass/
[12] http://www.gnuplot.info
[13] https://robomongo.org
[14] https://grafana.com
[15] https://code.visualstudio.com
[16] https://www.youtube.com/snl

2.4 Τι αντίκτυπο έχει στο κοινό η απενεργοποίηση των σχολίων; [15 μονάδες]

Για την περιοχή GB, βρείτε τους μέσους όρους των προβολών, των «μου αρέσει» και των «δε μου αρέσει» για τα βίντεο που έχουν απενεργοποιημένα τα σχόλια (comments_disabled=TRUE). Πράξτε αντίστοιχα για τα βίντεο που επιτρέπουν στους χρήστες να τους αφήνουν σχόλια (comments_disabled=FALSE). Φτιάξτε bar charts με τα αποτελέσματα που πήρατε. Σχολιάστε τα αποτελέσματά σας. Τα βίντεο που δεν επιτρέπουν τα σχόλια έχουν τελικά περισσότερες αντιπάθειες; Μήπως αποτρέπουν τους χρήστες να τα προβάλουν; Ή μήπως η απενεργοποίηση των σχολίων είναι ένας καλός τρόπος να αποφύγουν της αρνητική προσοχή; κοκ.

2.5 Ποιες ήταν οι πιο δημοφιλείς ημερομηνίες για δημοσίευση βίντεο; [20 μονάδες]

Για την περιοχή GB, βρείτε το πλήθος των βίντεο που δημοσιεύτηκαν ανά ημέρα για το τελευταίο 3/μηνο, δηλαδή κατά το διάστημα από 5 Δεκεμβρίου 2017 ως και 5 Μαρτίου 2018. Για κάθε εγγραφή στο αρχείο αποτελεσμάτων, εμφανίστε την ημερομηνία δημοσίευσης (μόνο έτος, μήνα, και μέρα) και το αντίστοιχο πλήθος των βίντεο, ταξινομημένα ως προς την ημερομηνία από την παλαιότερη προς την νεότερη. Κάντε μια γραφική απεικόνιση (scatter plot) των αποτελεσμάτων που πήρατε βάζοντας στον x άξονα τις ημερομηνίες και στο y άξονα το πλήθος των βίντεο. Τι συμπεράσματα βγάζετε; Σχολιάστε τα αποτελέσματά σας. Σε ποιες ημέρες δημοσιεύτηκαν τα περισσότερα βίντεο; Συνάδουν οι περισσότερες δημοσιεύσεις με κάποια γιορτή ή κάποιο εποχικό θέμα; Παρατηρείτε οι δημοσιεύσεις να ακολουθούν κάποιο μοτίβο; κοκ.

Bonus υλοποίηση (έως 20 μονάδες)

Μπορείτε να πάρετε bonus μέχρι 20% απαντώντας και σχολιάζοντας πιο σύνθετα ερωτήματα. Για παράδειγμα, βρίσκοντας έναν τρόπο να εμφανίζετε και τον τίτλο της κατηγορίας στην οποία ανήκει ένα βίντεο ή υπολογίζοντας το χρονικό διάστημα (σε πλήθος ημερών) που περνάει από τη μέρα που θα δημοσιευτεί ένα βίντεο μέχρι να τη μέρα που θα βρεθεί στη λίστα με τα πιο δημοφιλή βίντεο του YouTube, ή κάποιο άλλο που εσείς θα σκεφτείτε σε συνεννόηση με τη διδάσκουσα.

Παραδοτέα και βαθμολόγηση

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

Τα παραδοτέα της εργασίας είναι:

- τα scripts που ενδεχομένως χρησιμοποιήσατε
- τα αρχεία με τα αποτελέσματα για κάθε ένα από τα παραπάνω ερωτήματα
- μία γραπτή αναφορά που θα περιέχει:
    - τις ενέργειες σας (μαζί με κατάλληλη αιτιολόγηση) σε σχέση με την μετατροπή των δεδομένων σας από csv σε json,
    - τα εργαλεία που εγκαταστήσατε και χρησιμοποιήσατε (μαζί με λεπτομέρειες που αξίζει να σημειωθούν),
    - τα ερωτήματα που υποβάλατε στη βάση σας,
    - οδηγίες για την εκτέλεση των ερωτημάτων,
    - τις 20 πρώτες εγγραφές των αποτελεσμάτων για κάθε ένα από τα ερωτήματα,
    - τις γραφικές απεικονίσεις των αποτελεσμάτων όπου ζητούνται, και
    - τα σχόλιά σας σε σχέση με τα αποτελέσματα για κάθε ένα από τα ερωτήματα.

Καλή επιτυχία!