Διαχείριση Μεγάλων Δεδομένων
Project 2

Στούππος Σάββας Α.Μ. 2022201800191 dit18191@go.uop.gr
Μότση Πολυξένη Α.Μ. 2022201800125 dit18125@go.uop.gr
Τσάκαλου Αναστασία Α.Μ. 2022201900226 dit19226@go.uop.gr

Ιανουάριος 2023

Περιεχόμενα

1 Αποθήκευση δεδομένων στο MongoDB

Καθένας από μας για τα ερωτήματά του, δημιούργησε ένα Database μέσω του Compass. Έπειτα για κάθε ξεχωριστό αρχείο, φτιάξαμε μία **συλλογή** και μέσω του Compass τα εισάγαμε στη συλλογή αυτή (αφού τα έχουμε μετατρέψει όλα σε **JSON format**). Τα αρχεία του τύπου country_youtube_trending_data.json μεταφορτώνονταν σωστά και χωρίς σφάλματα. Όμως, τα αρχεία του τύπου country_category_id.json, χρειάστηκαν προεπεξεργασία, καθώς το compass δεν αναγνώριζε τα arrays. Γι΄ αυτό κλείσαμε όλο το περιεχόμενο κάθε αρχείου μέσα σε

`'[...]'`.

1.1 Μετατροπή των δεδομένων σε JSON format

Τη μετατρτοπή των δεδομένων από **CSV** σε **JSON** την κάναμε μέσω του **Compass**. Δημιουργήσαμε προσωρινές συλλογές, έπειτα κάναμε import τα δεδομένα ως **CSV** και τέλος τα κάναμε export ως **JSON**.

2 Ανάλυση των δεδομένων

2.1 Τι ξέρουμε για τις δημοσιεύσεις του πολύ δημοφιλούς καναλιού Saturday Night Live·

Για το ερώτημα 2.1 αυτό που ζήταγε να κάνουμε είναι για κάθε βίντεο του καναλιού Saturday Night Live να εμφανίσουμε τον τίτλο του βίντεο, την ημερομηνία δημοσίευσης, τις προβολές, τα likes και τα dislikes, σε αύξουσα ταξινόμηση με βάση την ημερομηνία δημοσίευσης.

2.1.1 Οδηγίες εκτέλεσης ερωτήματος 2.1

Το πρώτο μας βήμα είναι να πάρουμε όλα τα βίντεο που ανήκουν για το συγκεκριμένο κανάλι που αναζητούμε.

```
$match:
{
    channelTitle: "Saturday Night Live"
}
```

Έπειτα, με την εντολή project αποφασίζουμε ποιά συγκεκριμένα πεδία θέλουμε να εμφανίσουμε μόνο, κρατήσαμε αυτά που ζητάγατε αλλά και τα id των βίντεο, τον τίτλο του καναλιού και τα trending dates διότι θα μας χρειαστούν στην συνέχεια.

```
$project:
{
    title: 1,
    video_id: 1,
    publishedAt: 1,
    channelTitle: 1,
    likes: 1,
    dislikes: 1,
    view_count: 1,
    trending_date: 1
}
```

Υπάρχει ένα θεματάκι όμως, στα αρχεία, τα ίδια βίντεο μπορεί να εμφανιστούν πάνω από μία φορές, και αυτό διότι μπορεί να έχει γίνει δημοφιλές πολλές φορές. Γι΄ αυτό τον λόγο πρέπει να διαλέξουμε μόνο ένα από όλα αυτά. Για να πραγματοποιηθεί αυτό θα πρέπει, με βάση το trending date, να επιλέξουμε το πιο πρόσφατο. Αυτό μπορεί να γίνει ομαδοποιώντας τα βίντεο και στην συνέχεια να επιλέξουμε το πρώτο στην σειρά. Ωστόσο, αν δεν κάνουμε μία ταξινόμηση προηγουμένως τότε θα επιλέξει το πρώτο έτσι όπως ήταν στο αρχείο και δε θα ήταν σωστό. Γι΄ αυτό τον λόγο πριν την εντολή group κάνουμε μία ταξινόμηση με βάση το trending date σε φθίνουσα σειρά. Έχουμε κάνει ταυτόχρονα και την ημερομηνία δημοσίευσης σε αύξουσα αλλά αυτό δεν επηρεάζει τα αποτελέσματα. Το βάλαμε για δική μας διευκόλυνση έτσι ώστε να βλέπουμε αν όντως πήραμε τα σωστά δεδομένα.

```
$sort:
{
   publishedAt: 1,
```

```
    trending_date: -1,
  }
```

Συνεπώς, αφού έγινε η ταξινόμηση τώρα ήρθε η ώρα να τα ομαδοποιήσουμε. Η ομαδοποίηση θα γίνει με βάση τα id των βίντεο και στην συνέχεια για κάθε ομάδα θα πάρουμε τα πρώτα δεδομένα στην λίστα. Εκτός από αυτό, για την ημερομηνία δημοσίευσης αποφασίσαμε να αφαίρεσουμε την ώρα και να κρατήσουμε μόνο την ημερομηνία.

```
$group:
{
   _id: "$video_id",
   publishedAt: {
     $first: {
       $substr: ["$publishedAt", 0, 10],
     },
   },
   title: {
     $first: "$title",
   },
   trending_date: {
     $first: "$trending_date",
   },
   channelTitle: {
     $first: "$channelTitle",
   },
   likes: {
     $first: "$likes",
   },
   dislikes: {
     $first: "$dislikes",
   },
   view_count: {
     $first: "$view_count",
   },
 }
```

Επειδή μετά την ομαδοποίηση η ταξινόμηση χάλασε, την ξαναβάλαμε.

```
$sort: {
   publishedAt: 1,
 }
```

Τέλος, αφαιρούμε τα πεδία που δεν χρειαζόμαστε πια, και έχουμε τα αποτελέσματα μας.

```
$unset: ["trending_date", "_id"],
```

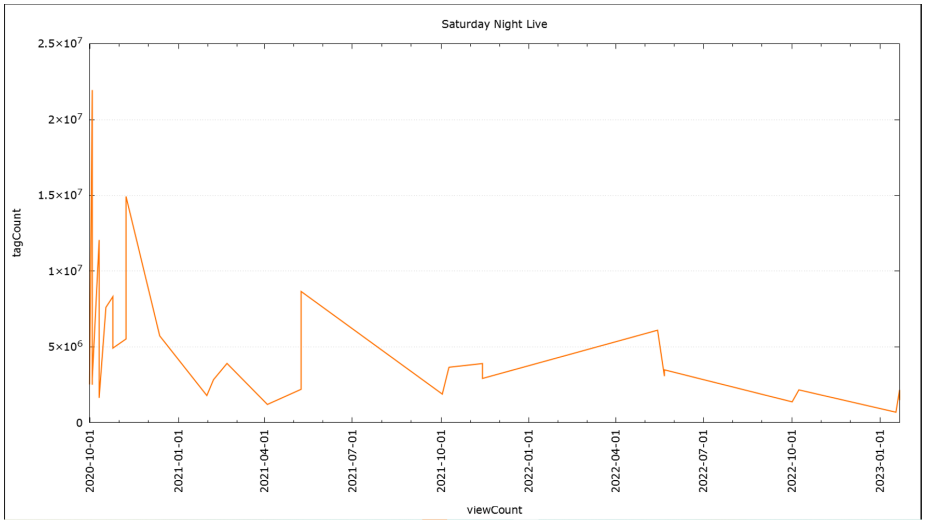
2.1.2 Αποτελέσματα ερωτήματος 2.1

Παρακάτω βλέπουμε τα 20 πρώτα αποτελέσματα που βρήκαμε.

	publishedAt	title	channelTitle	likes	dislikes	view_count
1						
2	2020-10-01	Jim Carrey and Maya	Saturday Night Live	42891	2650	2526816
3	2020-10-04	First Debate Cold Op	Saturday Night Live	437775	57125	21954683
4	2020-10-04	Superspreader Event	Saturday Night Live	43426	1370	2504202
5	2020-10-04	Weekend Update: Tr	Saturday Night Live	40076	2449	2916986
6	2020-10-11	VP Fly Debate Cold	Saturday Night Live	188194	31682	12070479
7	2020-10-11	New Normal - SNL	Saturday Night Live	29237	1326	1646765
8	2020-10-18	Dueling Town Halls (	Saturday Night Live	100317	11249	7610686
9	2020-10-25	Final Debate Cold O	Saturday Night Live	136664	20003	8321084
10	2020-10-25	Adele Monologue - S	Saturday Night Live	65916	1621	5002750
11	2020-10-25	The Bachelor - SNL	Saturday Night Live	93511	1247	4929699
12	2020-11-08	Weekend Update: Ru	Saturday Night Live	83187	4679	5530282
13	2020-11-08	Biden Victory Cold C	Saturday Night Live	275185	29394	14927666
14	2020-12-13	Rap Roundtable - SN	Saturday Night Live	214936	4382	5727960
15	2021-01-31	Twins - SNL	Saturday Night Live	38236	870	1794947
16	2021-02-07	Super Bowl Pre-gam	Saturday Night Live	35177	2714	2845426
17	2021-02-21	Britney Spears Cold	Saturday Night Live	53080	5830	3912795
18	2021-04-04	Salt Bae - SNL	Saturday Night Live	22567	1144	1207224
19	2021-05-09	Weekend Update: Fl	Saturday Night Live	47815	2361	2193521
20	2021-05-09	Wario - SNL	Saturday Night Live	103978	36296	4951441

Σχήμα 1: Τα 20 πρώτα αποτελέσματα

2.1.3 Γραφικές απεικονίσεις ερωτήματος 2.1



Σχήμα 2: Αποτελέσματα

2.1.4 Σχόλια για ερώτημα 2.1

Από το παραπάνω γράφημα παρατηρούμε ότι τα βίντεο του καναλιού Saturday Live Night είχαν μεγάλη απήχηση τον Φθινόπωρο-Χειμώνα του 2020 και ξανά μία απήχηση τον Φθινόπωρο-Χειμώνα του 2021. Αυτό θα μπορούσε να εξηγηθεί ίσως από το γεγονός ότι τότε πλησιάζουν και τα Χριστούγεννα και πολλοί διάσημοι καλλιτέχνες ή άνθρωποι να εμφανίζονται στο κανάλι αυτό. Μία άλλη εξήγηση θα μπορούσε να είναι ίσως επειδή τα βίντεο αυτά θα περιέχουν εορταστικό περιεχόμενο και αυτό να αρέσει στους θεατές. Τέλος, ίσως επειδή στις εποχές αυτές κάνει κρύο και ο κόσμος προτιμά να μείνει ζεστά στο σπίτι, να έχει περισσότερο χρόνο και να απολαμβάνει να βλέπει να βίντεο του καναλιού αυτού.

2.2 Πόσες ετικέτες χρησιμοποιούνται συνήθως στις δημοσιεύσεις των βίντεο·

2.2.1 Οδηγίες εκτέλεσης ερωτήματος 2.2

Αρχικά, πρέπει να πάρουμε τα πεδία που θα μας χρειαστούν, μέσω ενός σταδίου **project**:

```
$project:
  {
    _id: 1,
    title: 1,
    video_id: 1,
    trending_date: 1,
    tags: {
      $split: ["$tags", "|"],
    },
    view_count: 1,
  }
```

Επιλέγουμε τον τίτλο του βίντεο, το **id** του βίντεο, την **ημερομηνία** που έγινε trending, τον **αριθμό προβολών** και τέλος τα **tags** που είχε αποθηκεύοντάς τα σε έναν **πίνακα**.

Έπειτα, ταξινομούμε τα αποτελέσματα κατα φθίνουσα σειρά, με βάση την ημερομηνία μέσω ενός σταδίου **sort**:

```
$sort:
  {
    trending_date: -1,
  }
```

Ακόμη, τα ομαδοποιούμε με βάση το **video_id** και παίρνουμε τον τίτλο, το πλήθος προβολών και το πλήθος των **tags** με την εντολή **size**.

```
$group:
  {
    _id: "$video_id",
    trending_date: {
      $first: "$trending_date",
    },
    title: {
      $first: "$title",
    },
    view_count: {
      $first: {
        $toInt: "$view_count",
      },
    },
    tag_count: {
      $first: {
```

```
            $size: "$tags",
          },
        },
      }
```

Τέλος, με τα επόμενα στάδια **project, unset, sort** εμφανίζουμε τα δεδομένα ακριβώς όπως τα ζητάει η εκφώνηση.

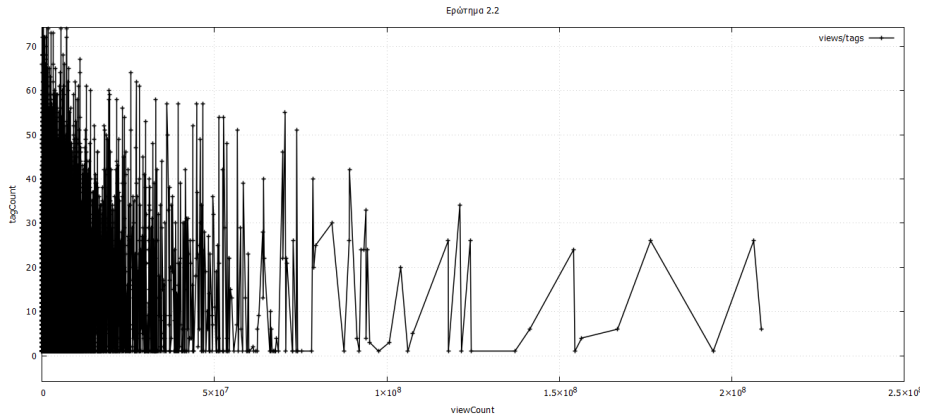
```
    $project:
      {
        video_id: "$_id",
        view_count: 1,
        tag_count: 1,
      },
  },
  {
    $unset:
      ["trending_date", "title", "_id"],
  },
  {
    $sort:
      /**
       * Provide any number of field/order pairs.
       */
      {
        view_count: -1,
      },
  },
```

2.2.2 Αποτελέσματα ερωτήματος 2.2

	A	B	C
1	view_count	tag_count	video_id
2	208581468	6	gdZLi9oWNZg
3	206344829	26	gQlMMD8auMs
4	194625542	1	qF0N19MgI3Q
5	176467113	26	vRXZj0DzXIA
6	166895681	6	WMweEpGlu_U
7	156482499	4	CuklIb9d3fI
8	154513998	1	ia6fRSeK8I0
9	154134590	24	awkkyBH2zEo
10	141428767	6	-5q5mZbe3V8
11	137068663	1	0e3GPea1Tyg
12	124476794	1	v4YjoKim3HA
13	124180499	26	dyRsYk0LyA8
14	121633557	1	SN0WiRJGBXc
15	121159003	34	8dJyRm2jJ-U
16	117832430	1	hdmx71UjBXs
17	117791538	26	POe9SOEKotk
18	107534237	5	T6n7lZirQkA
19	106089141	1	Fw7fbKoK3e8
20	104017073	20	U3ASj1L6_sY
21	100694053	3	46SbB0lHplE
22	97611742	1	LrJYKxyrMwg
23	95023322	3	pFtsvQuefXQ
24	94356729	24	CKZvWhCqx1s

Σχήμα 3: Αποτελέσματα

2.2.3 Γραφικές απεικονίσεις ερωτήματος 2.2



Σχήμα 4: Γραφική Απεικόνιση

2.2.4 Σχόλια για ερώτημα 2.2

Λαμβάνοντας υπόψη τα αποτελέσματα καθώς και τη γραφική απεικόνιση, καταλήγουμε στο συμπέρασμα ότι τα βίντεο που χρησιμοποιούν τα περισσότερα tags, καταλήγουν να μην επιτυγχάνουν τόσο όσο άλλα βίντεο με πιό λογικό αριθμό tags.

2.3 Πώς φέρονται οι vloggers και οι χρήστες ανά περιοχή·

Για το ερώτημα 2.3 αυτό που ζήταγε να κάνουμε είναι να βρούμε τον μέσο όρο πλήθος ετικετών ανά περιοχή και τον μέσο όρο προβολών των βίντεο ανά περιοχή. Ζήταγε επίσης να εμφανίσουμε στα αποτελέσματα τον κωδικό περιοχής, τον μέσο όρο των ετικετών και τον μέσο όρο των προβολών.

2.3.1 Οδηγίες εκτέλεσης ερωτήματος 2.3

Αρχικά, πρέπει οπωσδήποτε να βρούμε το πλήθος των συνολικών βίντεο ανά περιοχή χωρίς τα διπλότυπα τους βίντεο. Συνεπώς, για κάθε περιοχή θα πρέπει να βρούμε ξεχωριστά το πλήθος και να το προσθέσουμε σε ένα άλλο collection. Γι΄ αυτό, τα παρακάτω stages που θα δείξουμε θα είναι ίδια για όλες τις 11 περιοχές, με την μόνη διαφορά να αλλάζει ο κωδικός περιοχής.
Το πρώτο βήμα που πρέπει να κάνουμε είναι να ομαδοποιήσουμε τις εγγραφές με βάση των κωδικό των βίντεο για να αφαιρέσουμε τα διπλότυπα. Μας ενδιαφέρει μόνο η αφαίρεση τους.

```
$group:
{
  _id: "$video_id",
}
```

Έπειτα, θα ξανα κάνουμε πάλι ομαδοποίηση αλλά αυτή την φορά με null για να πάρουμε όλα τα βίντεο. Έτσι, αφού θα έχουμε μόνο μία ομαδοποίηση ταυτόχρονα θα μετράμε και το πλήθος τους, εμφανίζοντας ως αποτέλεσμα το πλήθος όλων των βίντεο της περιοχής.

```
$group:
{
  _id: null,
  count: {
    $sum: 1,
  },
}
```

Η προτελευταία εντολή μας είναι να προσθέσουμε των κωδικό της περιοχής.

```
$project:
{
  _id: "CA",
  count: 1,
}
```

Τέλος, με την εντολή merge θα βάλουμε τα αποτελέσματα μας σε ένα άλλο collection με όνομα sum.

```
$merge:
{
```

```
    into: "sum",
}
```

Έτσι, κάνοντας τα παραπάνω για κάθε ξεχωριστή περιοχή, στο τέλος θα έχουμε το εξής collection.

sum

	_id String	count Int32
1	"RU"	99294
2	"BR"	27530
3	"CA"	35516
4	"DE"	39093
5	"FR"	37467
6	"GB"	32798
7	"IN"	52520
8	"JP"	23186
9	"KR"	20315
10	"MX"	23787
11	"US"	32942

Τώρα, για να βρούμε τους μέσους όρους των ετικετών και των προβολών θα κάνουμε τα εξής stages για όλες τις περιοχές ξεχωριστά με την διαφορά ότι θα αλλάξουμε τους κωδικούς περιοχής και το πλήθος των βίντεο καθώς η κάθε περιοχή, όπως είδαμε και παραπάνω έχει διαφορετικό πλήθος των βίντεο. Έτσι, ξεκινάμε με το πρώτο βήμα μας που είναι η εμφάνιση συγκεκριμένων πεδίων που θα μας φανούν χρήσιμα αλλά και την ξεχώριση των ετικετών από ένα αλφαρηθμητικό σε πίνακα.

```
$project:
{
  video_id: 1,
  title: 1,
  publishedAt: 1,
  trending_date: 1,
  channelTitle: 1,
```

```
  tags: {
    $split: ["$tags", "|"],
  },
  view_count: 1,
}
```

Έπειτα, θα κάνουμε ταξινόμηση

```
 $sort:
{
  publishedAt: 1,
  trending_date: -1,
}
```

Επειδή μετά την ομαδοποίηση η ταξινόμηση χάλασε, την ξαναβάλαμε.

```
 $sort: {
    publishedAt: 1,
  }
```

Τέλος, αφαιρούμε τα πεδία που δεν χρειαζόμαστε πια, και έχουμε τα αποτελέσματα μας.

```
 $unset: ["trending_date", "_id"],
```

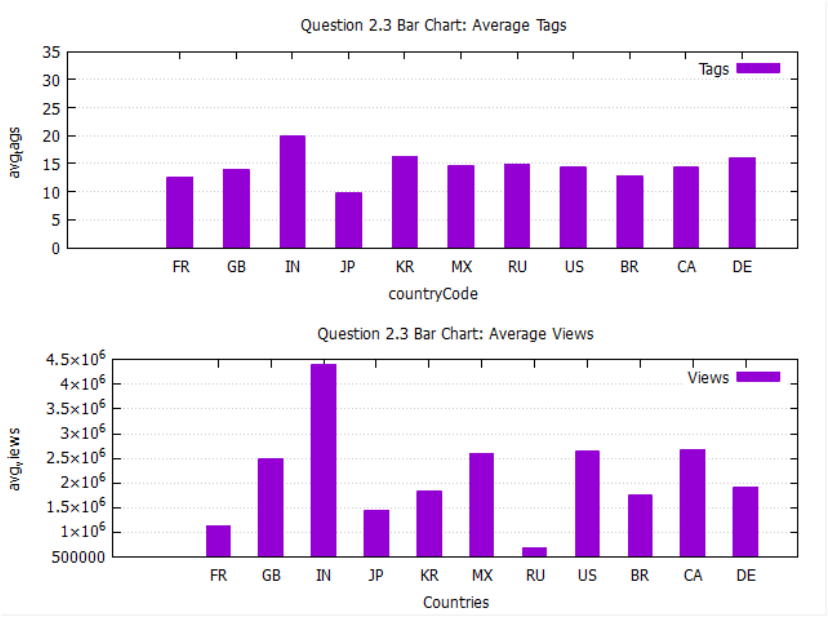
2.3.2 Αποτελέσματα ερωτήματος 2.3

1	avg_tags	avg_views	countryCode
2	12.56404302	1130088.616	FR
3	14.02899567	2488449.486	GB
4	19.85825355	4407015.015	IN
5	9.744371604	1444529.459	JP
6	16.31656412	1825054.683	KR
7	14.57808887	2604893.668	MX
8	14.90557335	686841.2749	RU
9	14.31112258	2638957.114	US
10	12.80087178	1746685.122	BR
11	14.34570334	2677568.287	CA
12	16.01539918	1909150.608	DE

Σχήμα 5: Τα αποτελέσματα του ερωτήματος 2.3

2.3.3 Γραφικές απεικονίσεις ερωτήματος 2.3

Οι γραφικές απεικονίσεις του ερωτήματος μας. Η πρώτη εικόνα περιέχει τους μέσους όρους των ετικετών και η δεύτερη εικόνα τους μέσους όρους των προβολών.



Σχήμα 6: Γραφική απεικόνιση του ερωτήματος 2.3

2.3.4 Σχόλια για ερώτημα 2.3

Από την παραπάνω γραφική απεικόνιση αυτό που παρατηρούμε είναι ότι σε γενικές γραμμές όλες οι χώρες χρησιμοποιούν ίδιο ποσοστό ετικετών, με την Ινδία να έχει λίγα αρκετά παραπάνω. Όσο αφορά την επιρροή των προβολών θα μπορούσαμε να πούμε και ναι και όχι. Για την Ινδία παραδείγματος χάριν, χρησιμοποιεί πολλές περισσότερες ετικέτες και έχει πράγματι μεγάλο πλήθος προβολών. Ωστόσο όμως, η Γαλλία που χρησιμοποιεί και αυτή αρκετές ετικέτες δεν έχει πολύ χαμηλές προβολές, το ίδιο επίσης παρατηρείτε και στην Ρωσία η οποία είναι σε τελείως ακραία περίπτωση. Παρόλα αυτά όμως, η Ινδία είναι μία χώρα με τεράστιο πλυθησμό, συνεπώς αυτό ίσως να δικαιολογεί τα αποτελέσματα αυτά με τους μέσους όρους των προβολών και να καταλήγουμε στο συμπέρασμα ότι το πλήθος των προβολών επηρεάζει ελάχιστα.

2.4 Ποιες είναι οι πιο δημοφιλείς ετικέτες στα ανερχόμενα βίντεο·

2.4.1 Οδηγίες εκτέλεσης ερωτήματος 2.4

Αρχικά, πρέπει να πάρουμε τα πεδία που θα μας χρειαστούν, μέσω ενός σταδίου **project**:

```
$project:
  {
    video_id: 1,
    tags: {
      $split: ["$tags", "|"],
    },
  },
```

Έπειτα, επειδή έχουμε αποθηκεύσει τα **tags** του κάθε βίντεο σε πίνακα, πρέπει να τα ¨ξετυλίξουμε' για να μπορέσουμε να τα επεξεργαστούμε ένα ένα. Αυτό το επιτυγχάνουμε μέσω ενός σταδίου **unwind**:

```
$unwind:
  {
    path: "$tags",
    preserveNullAndEmptyArrays: false,
  },
```

Τέλος, ομαδοποιούμε τα documents κατά **tags** και με μερικά επιπλέον στάδια, καταλήγουμε στα αποτελέσματα που ζητούνται:

```
 $group:
      {
        _id: "$tags",
        video_count: {
          $sum: 1,
        },
      },
  },
  {
    $sort:
      {
        video_count: -1,
      },
  },
  {
    $project:
      /**
       * specifications: The fields to
       *   include or exclude.
       */
      {
```

```
      tag: "$_id",
      video_count: 1,
    },
},
{
  $unset:
    "_id",
},
```

2.4.2 Αποτελέσματα ερωτήματος 2.4

Τρέχοντας τα παραπάνω pipelines για τις δύο διαφορετικές περιοχές GB και US έχουμε τα παρακάτω αποτελέσματα:

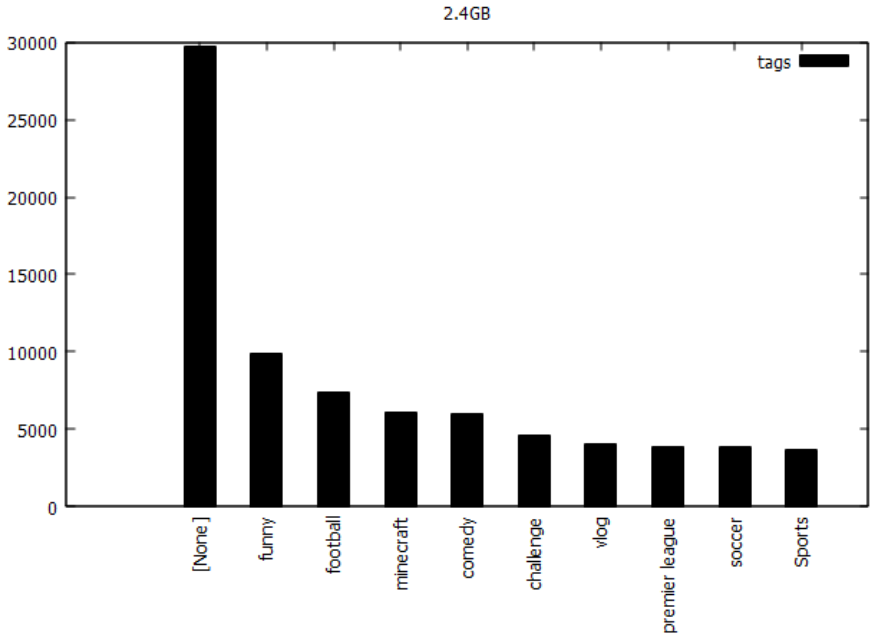
	A	B
1	video_count	tag
2	29669	[None]
3	9830	funny
4	7321	football
5	6067	minecraft
6	5964	comedy
7	4578	challenge
8	4070	vlog
9	3863	premier league
10	3804	soccer
11	3617	Sports
12	3597	highlights

Σχήμα 7: Πρώτα 10 αποτελέσματα GB

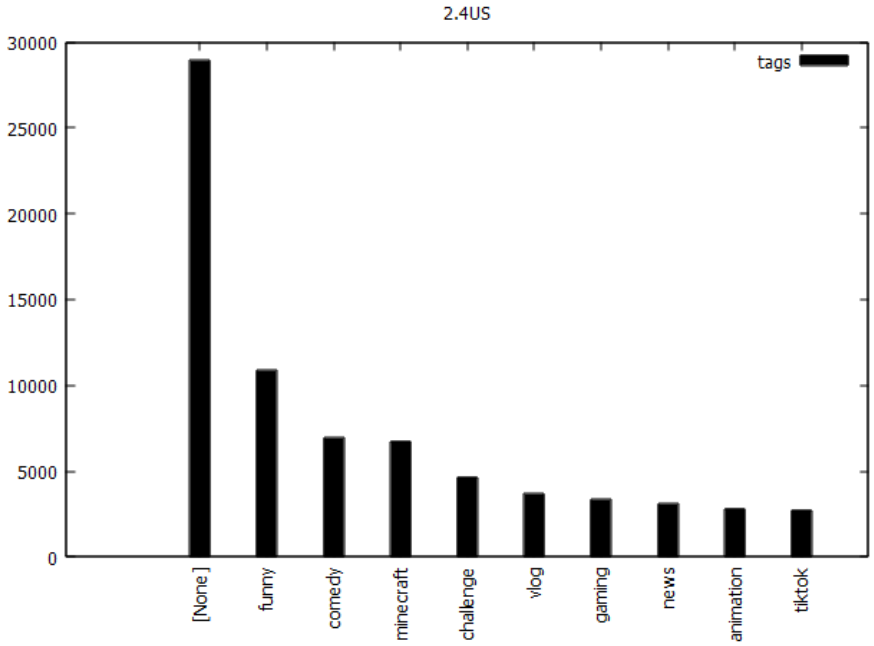
	A	B
1	video_count	tag
2	28925	[None]
3	10887	funny
4	6941	comedy
5	6720	minecraft
6	4655	challenge
7	3697	vlog
8	3357	gaming
9	3121	news
10	2798	animation
11	2708	tiktok
12	2664	highlights

Σχήμα 8: Πρώτα 10 αποτελέσματα US

2.4.3 Γραφικές απεικονίσεις ερωτήματος 2.4



Σχήμα 9: Γραφική Απεικόνιση GB



Σχήμα 10: Γραφική Απεικόνιση US

2.4.4 Σχόλια για ερώτημα 2.4

Καθώς το πλήθος των tags ήταν αρκετά μεγάλο όσο για τα αποτελέσματα, τόσο για τα γραφήματα, τα συμπεράσματα αποπνέουν απο τα 10 πρώτα αποτελέσματα για κάθε χώρα. Παρατηρήσαμε λοιπόν, ότι και στις δύο χώρες τα βίντεο με τις περισσότερες προβολές ήταν αυτά που δεν είχαν κάποιο tag. Έπειτα παρατηρήθηκε ότι τα βίντεο με αστείο ή γενικότερα ψυχαγωγικό περιεχόμενο ήταν και τα πιο δημοφιλή. Τέλος, βίντεο με αθλητικό περιεχόμενο ήταν πιο δημοφιλή στη Βρετανία παρά στην Αμερική.

2.5 Τι αντίκτυπο έχει στο κοινό η απενεργοποίηση των σχολίων·

Το ερώτημα αυτό ζητούσε να βρούμε τι αντίκτυπο έχει στο κοινό η απενεργοποίηση των σχολίων. Οπότε, έπρεπε να ελέγξουμε αν τα βίντεο με τα απενεργοποιημένα σχόλια είχαν σημαντική διαφορά στις προβολές, likes και dislikes σε σχέση με τα βίντεο με τα ενεργοποιημένα σχόλια.

2.5.1 Οδηγίες εκτέλεσης ερωτήματος 2.5

Αρχικά, το 1ο στάδιο είναι το match, στο οποίο βρίσκουμε τις εγγραφές οι οποίες έχουν απενεργοποιημένα σχόλια, και το 2ο στάδιο είναι το sort, που ταξινομούμε σε αύξουσα σειρά τα βίντεο ως προς την ημερομηνία δημοσίευσης τους, και σε φθίνουσα σειρά ως προς την ημέρα που μπήκαν στα trending βίντεο.

```
{
  $match:
    {
      comments_disabled: "True",
    },
},
{
  $sort:
    {
      publishedAt: 1,
      trending_date: -1,
    },
},
```

Στη συνέχεια, στο 3ο στάδιο, κάνουμε project, δηλαδή εμφανίζουμε τα πεδία που λέει η εκφώνηση. Επίσης, στο 4ο στάδιο κάνουμε group, ομαδοποιούμε δηλαδή τα βίντεο ως προς το videoid τους και κρατάμε μόνο τη πρώτη φορά που εμφανίζονται για να φύγουν τα διπλότυπα.

```
{
  $project:
    {
      title: 1,
      video_id: 1,
      likes: 1,
      dislikes: 1,
      view_count: 1,
      comments_disabled: 1,
    },
},
{
  $group: {
```

```
      _id: "$video_id",
      title: {
        $first: "$title",
      },
      view_count: {
        $first: "$view_count",
      },
      likes: {
        $first: "$likes",
      },
      dislikes: {
        $first: "$dislikes",
      },
    },
  },
```

Εν συνεχεία, στο 5ο στάδιο έχουμε το sort, που κάνουμε ταξινόμηση κατά αύξουσα σειρά ως προς την ημερομηνία δημοσίευσης των βίντεο και κατά φθίνουσα σειρά ως προς την ημερομηνία που τα βίντεο ήταν στα τρενδινγ βίντεο. Είναι προαιρετικό το στάδιο αλλά βοηθάει στην πιο εύληπτη παρουσίαση των εγγραφών μας. Έπειτα, στο 6ο στάδιο κάνουμε group και βρίσκουμε το άθροισμα των προβολών όλων των βίντεο με απενεργοποιημένα σχόλια.

```
  {
    $sort:
      {
        publishedAt: 1,
        trending_date: -1,
      },
  },
  {
    $group:
      {
        _id: null,
        view_count_sum: {
          $sum: {
            $toInt: "$view_count",
          },
        },
        likes_sum: {
          $sum: {
            $toInt: "$likes",
          },
        },
        dislikes_sum: {
          $sum: {
            $toInt: "$dislikes",
          },
```

```
      },
    },
},
```

Ακόμα, στο 7ο στάδιο, βρίσκουμε τον Μ.Ο. των views διαιρώντας το άθροισμα που βρήκαμε στο 6ο στάδιο, διά το πλήθος των βίντεο με απενεργοποιημένα σχόλια, το οποίο είναι 741. Επίσης, προσθέσαμε το CountryCode: "GB" αλλά και το commentsDisabled: "True" ούτως ώστε μετά, στην ενοποίηση των μέσων όρων, να ξεχωρίζουμε ποιος Μ.Ο. είναι για ποιο query. Εδώ συγκεκριμένα ψάχνουμε για τα βίντεο με απενεργοποιημένα σχόλια, ποιος είναι ο μέσος όρος προβολών τους.

```
{
  $project:
    {
      countryCode: "GB",
      commentsDisabled: "TRUE",
      avg_views: {
        $round: [
          {
            $divide: ["$view_count_sum", 741],
          },
          2,
        ],
      },
      avg_likes: {
        $round: [
          {
            $divide: ["$likes_sum", 741],
          },
          2,
        ],
      },
      avg_dislikes: {
        $round: [
          {
            $divide: ["$dislikes_sum", 741],
          },
          2,
        ],
      },
    },
},
{
  $unset:
    "_id",
},
{
```

```
    $merge:
      {
        into: "Average_Results",
        //let: 'specification(s)',
        //whenMatched: 'string',
        //whenNotMatched: 'string'
      },
  },
```

Στο 8ο στάδιο αφαιρούμε το πεδίο id γιατί δεν χρειάζεται να το εμφανίσουμε στα αποτελέσματά μας. Τέλος, στο 9ο στάδιο, κάνουμε merge για να εξάγουμε το αποτέλεσμα, δηλαδή την 1 εγγραφή που παρήχθηκε στο 8ο στάδιο, σε ένα νέο collection όπου θα έχουμε τους συνολικά έξι μέσους όρους. Οπότε εδώ εξάγεται το πρώτο αποτέλεσμα στο collection.

```
        $unset:{
             "_id"
        }
        $merge: {
            into: "Average_results,
            on: "_id"
       }
```

Αυτό που θα αλλάξει για την εξαγωγή του M.O. views, likes, dislikes για τα βίντεο με τα ενεργοποιημένα σχόλια, είναι ότι στο comments disabled θα το θέσουμε ίσο με false, κι επίσης θα διαιρέσουμε δια το πλήθος των βίντεο με ενεργοποιημένα σχόλια, που ο αριθμός είναι πλέον 31719.

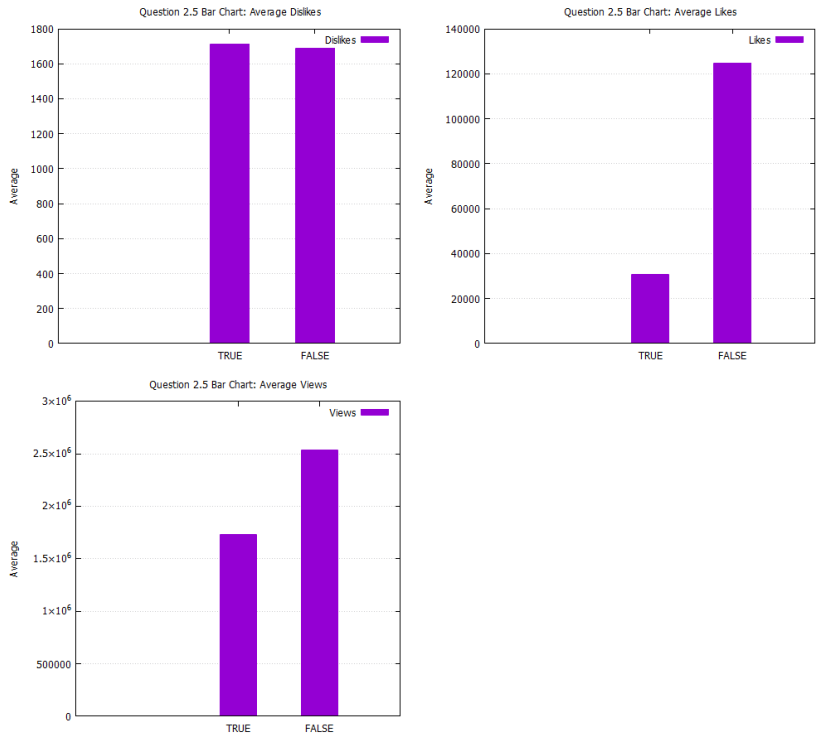
2.5.2 Αποτελέσματα ερωτήματος 2.5

Παρακάτω βλέπουμε τα 2 αποτελέσματα που έχει το collection από το merge των 2 διαφορετικών pipelines.

1	avg_dislikes	avg_likes	avg_views	commentsDisabled	countryCode
2	1712.65	30648.04	1729656.43	TRUE	GB
3	1690.66	124776.82	2534043.19	FALSE	GB

Σχήμα 11: Αποτελέσματα

2.5.3 Γραφικές απεικονίσεις ερωτήματος 2.5



Σχήμα 12: Γραφική Απεικόνιση

2.5.4 Σχόλια για ερώτημα 2.5

Λαμβάνοντας υπόψη τα αποτελέσματα καθώς και τη γραφική απεικόνιση, καταλήγουμε στο συμπέρασμα ότι τα βίντεο που έχουν ενεργοποιημένα τα σχόλια, έχουν σημαντικά παραπάνω απήχηση, και σε likes, και λίγη παραπάνω σε views σε σχέση με τα βίντεο που έχουν απενεργοποιημένα τα σχόλια. Ωστόσο, στα dislikes δεν υπάρχει καμία διαφορά.

2.6 Ποιες ήταν οι πιο δημοφιλείς ημερομηνίες για δημοσίευση βίντεο·

Το ερώτημα ζητούσε να βρούμε ποιες είναι οι πιο δημοφιλείς ημερομηνίες για δημοσίευση βίντεο εντός διαστήματος 4 μηνών. Θα χρειαστεί να μετρήσουμε το πλήθος των βίντεο που έγιναν trending, κάθε μέρα, και έπειτα να τα ταξινομήσουμε από την παλαιότερη ημερομηνία έως την νεότερη.

2.6.1 Οδηγίες εκτέλεσης ερωτήματος 2.6

Πρώτα, στο 1ο στάδιο κάνουμε ταξινόμηση στην ημερομηνία δημοσίευσης του βίντεο και στην ημέρα που ήταν στα trending βίντεο. Έπειτα, στο 2ο στάδιο κάνουμε project τα πεδία που θέλουμε να δείξουμε. Ακόμα, στο 3ο στάδιο προσθέτουμε κάποια πεδία που θα χρειαστούμε για να υπολογίσουμε το πλήθος των βίντεο που δημοσιεύτηκαν κάθε μέρα, για 4 μήνες πάνω κάτω (από 1 Σεπτεμβρίου έως μέσα Ιανουαρίου, εφόσον η εκφώνηση έγραφε μέχρι ¨σήμερα¨, και η εκφώνηση ανέβηκε μέσα Ιανουαρίου περίπου).

```
[
  {
    $sort:
      {
        publishedAt: 1,
        trending_date: -1,
      },
  },
  {
    $project:
      {
        video_id: 1,
        title: 1,
        view_count: 1,
        trending_date: 1,
        channelTitle: 1,
        publishedAt: 1,
        likes: 1,
        dislikes: 1,
      },
  },
  {
    $addFields:
      {
        year: {
          $substr: ["$publishedAt", 0, 4],
        },
        month: {
          $substr: ["$publishedAt", 5, 2],
        },
        day: {
          $substr: ["$publishedAt", 8, 2],
```

```
      },
      date: {
        $substr: ["$publishedAt", 0, 10],
      },
    },
},
```

Μετά, στο 4ο στάδιο κάνουμε match τα πεδία που δημιουργήσαμε προηγουμένως, δηλαδή το έτος, μήνα και μέρα, τα οποία πρέπει να ταιριάζουν με το χρονικό διάστημα που ορίσαμε, δηλαδή Σεπτέμβριος 2022 μέχρι Ιανουάριος 2023. Στη συνέχεια, στο 5ο στάδιο κάνουμε group με id να είναι το video id για να απαλοίψουμε τις διπλοεγγραφές. Στο 6ο στάδιο, κάνουμε ταξινόμηση ως προς την ημερομηνία.

```
{
  $match:
    {
      $or: [
        {
          month: {
            $in: ["09", "10", "11", "12"],
          },
          year: "2022",
        },
        {
          month: {
            $in: ["01"],
          },
          year: "2023",
        },
      ],
    },
},
{
  $group:
    {
      _id: "$video_id",
      title: {
        $first: "$title",
      },
      date: {
        $first: "$date",
      },
      year: {
        $first: "$year",
      },
      month: {
        $first: "$month",
      },
```

```
      day: {
        $first: "$day",
      },
      publishedAt: {
        $first: "$publishedAt",
      },
      trending_date: {
        $first: "$trending_date",
      },
      view_count: {
        $first: "$view_count",
      },
      channelTitle: {
        $first: "$channelTitle",
      },
      likes: {
        $first: "$likes",
      },
      dislikes: {
        $first: "$dislikes",
      },
    },
},
{
  $sort:
    {
      publishedAt: 1,
      trending_date: -1,
    },
},
```

Προχωρώντας, στο 7ο στάδιο κάνουμε group με id το date και μετράμε το άθροισμα των βίντεο που ανέβηκαν την κάθε μέρα. Τέλος, στο 8ο στάδιο κάνουμε ταξινόμηση ως προς το date καθώς το ζητάει η εκφώνηση να είναι κατά αύξουσα σειρά της ημερομηνίας.

```
{
  $group:
    {
      _id: "$date",
      date: {
        $first: "$date",
      },
      count: {
        $sum: 1,
      },
```

```
        },
    },
    {
      $sort:
        {
          date: 1,
        },
    },
]
```

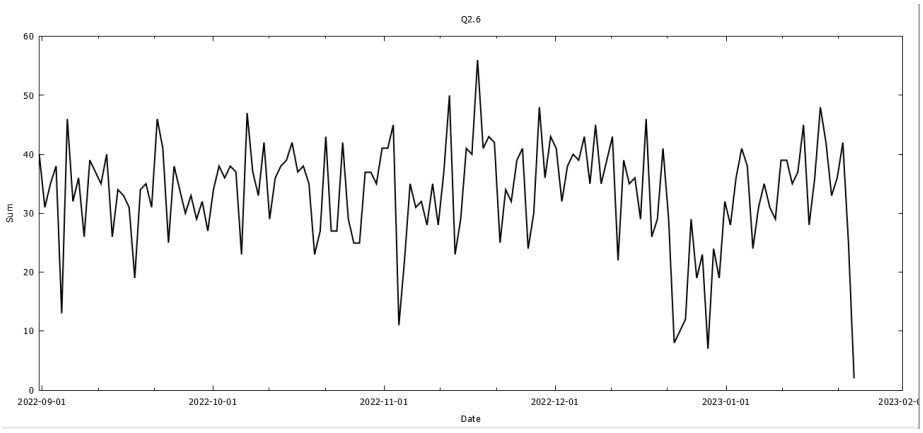
2.6.2 Αποτελέσματα ερωτήματος 2.6

Παρακάτω βλέπουμε τα 20 αποτελέσματα που παράγονται μετά το pipeline.

	A
1	_id,date,count
2	2022-09-01,2022-09-01,40
3	2022-09-02,2022-09-02,31
4	2022-09-03,2022-09-03,35
5	2022-09-04,2022-09-04,38
6	2022-09-05,2022-09-05,13
7	2022-09-06,2022-09-06,46
8	2022-09-07,2022-09-07,32
9	2022-09-08,2022-09-08,36
10	2022-09-09,2022-09-09,26
11	2022-09-10,2022-09-10,39
12	2022-09-11,2022-09-11,37
13	2022-09-12,2022-09-12,35
14	2022-09-13,2022-09-13,40
15	2022-09-14,2022-09-14,26
16	2022-09-15,2022-09-15,34
17	2022-09-16,2022-09-16,33
18	2022-09-17,2022-09-17,31
19	2022-09-18,2022-09-18,19
20	2022-09-19,2022-09-19,34

Σχήμα 13: Αποτελέσματα

2.6.3 Γραφικές απεικονίσεις ερωτήματος 2.6



Σχήμα 14: Αποτελέσματα

2.6.4 Σχόλια για ερώτημα 2.6

Απο τα παραπάνω παρατηρούμε ότι τα περισσότερα βίντεο έχουν δημοσιευτεί το 2022 στον μήνα Δεκέμβριο.Επιπρόσθετα θα μπορούσαμε να πούμε ότι δεν συνάδουν με κάποια εορτή ή κάτι γιατί υπάρχει μία ¨σταθερότητα¨. Αυτό που έχει όμως ενδιαφέρον είναι ότι στις μέρες που είναι κοντά στην αλλαγή του χρόνου υπάρχουν πολλά λιγότερα βίντεο και αυτό ίσως να δικαιολογείται από το γεγονός ότι τότε οι vloggers περνάνε χρόνο με τις οικογένειες τους λόγω Χριστουγέννων και αλλάγη της νέας χρονιάς και γι΄ αυτό δεν έχουν ανεβεί πολλά βίντεο.

2.7 Bonus

Για το bonus αποφασίσαμε να κάνουμε την πρώτη ιδέα που αναφέρατε στην εκφώνηση η οποία είναι να βρούμε και την κατηγορία των βίντεο. Αυτό που κάναμε όμως εν τέλει, είναι να βρούμε όλες την κατηγορία που έχει ανά βίντεο για όλες τις 4 περιοχές της Ασίας (γεωγραφικά), δηλαδή για τις περιοχές RU, IN, KR και JP. Έτσι, βγάλαμε 3 αποτελέσματα, το πλήθος των βίντεο ανά κατηγορία, το πλήθος των βίντεο ανά χώρα που χρησιμοποιούν την κατηγορία και φυσικά τον μέσο όρο των προβολών των βίντεο που έχουν την εξής κατηγορία. Με αυτά τα αποτελέσματα θα σχολιάσουμε ερωτήματα όπως, ποια κατηγορία είναι η πιο δημοφιλής και ποιά η λογότερη δημοφιλής, ποια χώρα έχει τις περισσότερες προβολές σε κάποια κατηγορία και γιατί κτλ.

2.7.1 Οδηγίες εκτέλεσης ερωτήματος 2.7 για το πλήθος των κατηγοριών

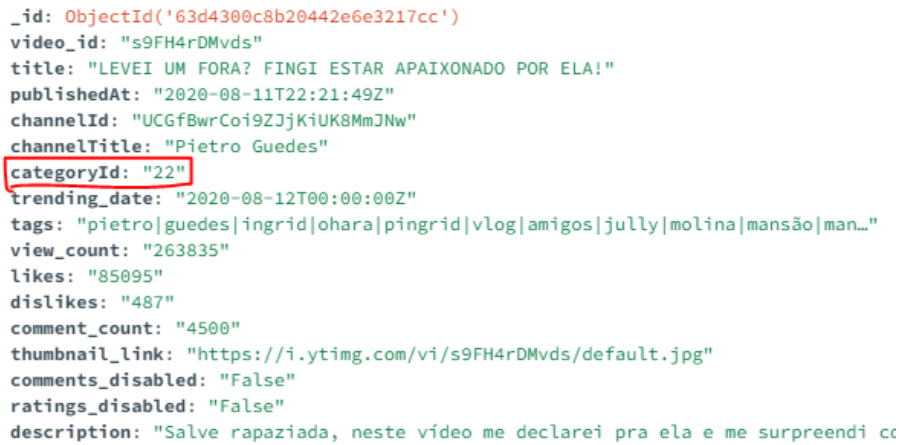
Αρχικά, αποφασίσαμε να αφαιρέσουμε τα διπλότυπα διότι ο σκοπός μας είναι απλά να εμφανίσουμε τις κατηγορίες των βίντεο. Συνεπώς ξεκινάμε κάνοντας αρχικά ταξινόμηση και έπειτα ομαδοποίηση με βάση των κωδικό των βίντεο. Φυσικά, επιλέγουμε με βάση την πιο πρόσφαση ημερομηνία δημοφιλότητας όπως έχουμε κάνει και στα προηγούμενα ερωτήματα.

```
$sort:
{
  title: 1,
  trending_date: -1,
},

$group:
{
  _id: "$video_id",
  video_id: {
    $first: "$video_id",
  },
  title: {
    $first: "$title",
  },
  categoryId: {
    $first: "$categoryId",
  },
  view_count: {
    $first: "$view_count",
  },
}
```

Πριν όμως προχωρήσουμε, πάμε να δούμε από το collection που θα φτιάξουμε τις εντολές τι πληροφορία μας δίνει για τις κατηγορίες. Όπως βλέπουμε παρακάτω, μας δίνει τον κωδικό της κατηγορίας που είναι αριθμός.

Έχοντας αυτό υπόψιν μας, πάμε να δούμε πως είναι το collection με τις κατηγορίες. Παρακάτω, παρατηρούμε υπάρχει μόνο 1 γραμμή όπως βλέπουμε στο Σχήμα 10.



Σχήμα 15: Κωδικός κατηγορίας

Σχήμα 16: Κατηγορίες περιοχής

Ωστόσο, άμα ανοίξουμε τον πίνακα items, όπως γίνεται στο Σχημα 11, θα δούμε ότι όλες οι κατηγορίες βρίσκονται εκεί. Ουσιασικά αυτό που παρατηρούμε είναι ότι ο πίνακας items περιέχει μια σειρά απο αντικείμενα πινάκων όπου το κάθε αντικείμενο ξεχωριστά έχει κάποιες πληροφορίες, όπως τον κωδικό της κατηγορίας που μας ενδιαφρει και από ένα άλλο αντικείμενο πίνακας με όνομα snippet που περιέχει άλλες πληροφορίες, όπως ο τίτλος της κατηγορίας που εν τέλει αυτό θα τραβήξουμε από αυτό το collection.

Συνεπώς, με βάση τα παραπάνω, αυτό που πρέπει να κάνουμε είναι με την εντολή lookup να κάνουμε join από το collection που έχουμε τις κατηγορίες. Έτσι, ξεκινάμε γράφοντας το όνομα του collection και έπειτα, με την εντολή let λέμε ότι από το τρέχων collection που είμαστε κράτα τον κωδικό της κατηγορίας, έτσι ώστε να μπορέσουμε να την χρησιμοποίουμε στην επόμενη εντολή, στην pipeline. Στην εντολή αυτή λοιπόν, μπορούμε να κάνουμε κανονικά ότι έντολες έχουμε κάνει και στα άλλα ερωτήματα, πχ. match, project, sort κτλ., έτσι ώστε να φτάσουμε στο επιθημητό αποτέλεσμα. Επειδή όπως είδαμε και στις προηγούμενες εικόνες, ο πίνακας items είναι γεμάτος από οβθεςτ, έπρεπε να κάνουμε υνωινδ για να τα διώξουμε. Αφού κάνουμε αυτό, εκτελούμε την εντολή match και θα το δεχτούμε

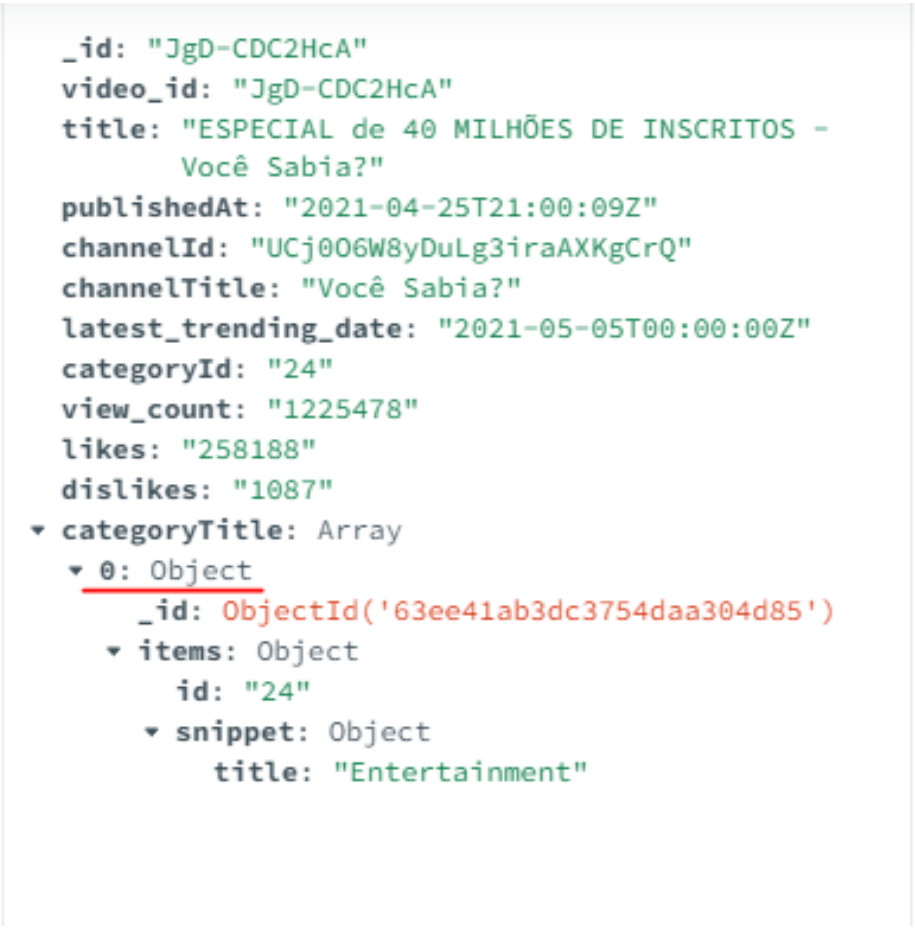
Σχήμα 17: κατηγορίες περιοχής

με την προυπόθεση ότι ο κωδικός που βρίσκεται μέσα στον πίνακα items είναι ίδιος με τον κωδικό που έχουμε κρατήσει με το let. Τέλος, απλά εμφανίζουμε αυτά που χρειαζόμαστε και καταλήγουμε να το εμφανίσουμε στο τρέχων collection μας ως categoryTitle.

```
$lookup:
{
  from: "C_BR",
  let: {
    cid: "$categoryId",
  },
  pipeline: [
    {
      $unwind: {
        path: "$items",
      },
    },
    {
      $match: {
        $expr: {
          $eq: ["$items.id", "$$cid"],
        },
      },
    },
    {
      $project: {
```

```
            "items.id": 1,
            "items.snippet.title": 1,
          },
        },
      ],
      as: "categoryTitle",
    }
```

Ωστόσο, επειδή ο πίνακας μας το εμφάνισε κανονικά με το αντικείμενο του, όπως ήταν δηλαδή αρχικά, οπως βλέπουμε στην εικόνα παρακάτω.



έπρεπε να κάνουμε unwind και μετά να πάρουμε μόνο τον τίτλο και να το εμφανίσουμε.

```
      $unwind:
      {
        path: "$categoryTitle",
      },
      $project:
      {
```

```
  _id: 1,
  video_id: 1,
  title: 1,
  categoryId: 1,
  view_count: 1,
  categoryTitle:
    "$categoryTitle.items.snippet.title",
}
```

2.7.2 Έυρεση πλήθους κατηγοριών

Τώρα, για το πλήθος των κατηγοριών αποφασίσαμε να βρούμε 2 πράγματα, το πλήθος των κατηγοριών από όλα τα βίντεο των 4 περιοχών αλλά και το πλήθος των κατηγοριών ανά χώρα. Συνεπώς, για να πραγματοποιηθεί αυτό θα χρειαστεί να εκτελέσουμε τις παρακάτω εντολές, εν συνέχεια του τέλους από την προηγούμενη ενότητα, δηλαδή από την εντολή project.

Το πρώτο βήμα που πρέπει να κάνουμε είναι να ομαδοποιήσουμε τις κατηγορίες και να μετρήσουμε το πλήθος των βίντεο που ανήκουν στην εξής κατηγορία αλλά και το πλήθος των προβολών όλων αυτών των βίντεο που ανήκουν στην εξής κατηγορία.

```
 $group:
{
  _id: "$categoryTitle",
  sum: {
    $sum: 1,
  },
  views: {
    $sum: {
      $toInt: "$view_count",
    },
  },
}
```

Έπειτα, επειδή έχουμε σκοπό τα δεδομένα αυτά να τα βάλουμε μαζεμένα σε ένα collection θα πρέπει να το διαμορφώσουμε για κάθε περιοχή ξεχωριστά. Επειδή εδώ έχουμε την Ινδία, θα εμφανίσουμε την κατηγορία φυσικά και το πλήθος της κατηγορίας για την συγκεκριμένη χώρα. Για τα υπόλοιπα, θα προσθέσουμε την τιμή 0. Τέλος, θα αφαιρέσουμε το id διότι έχουμε βάλει ήδη το όνομα της κατηγορίας στο πεδίο "Category".

```
 $project:
{
  Category: "$_id",
  IN: "$sum",
  JP: {
    $toInt: "0",
  },
```

```
    KR: {
      $toInt: "0",
    },
    RU: {
      $toInt: "0",
    },
  }
```

Τελευταίο βήμα από αυτό το collection είναι η μεταφορά των δεδομένων μας στο καινούργιο collection. Φυσικά αυτό θα γίνει με την εντολή merge.

```
  $merge:
     {
       into: "bonus",
     }
```

Ωστόσο, τα δεδομένα μας στο καινούργιο collection, όπως βλέπουμε παρακάτω,

bonus

	_id ObjectId	Category String	IN Int32	JP Int32	KR Int32	RU Int32
1	ObjectId('63f1389c93e510fecc8…	"People & Blogs"	8558	0	0	0
2	ObjectId('63f1389c93e510fecc8…	"Music"	5652	0	0	0
3	ObjectId('63f1389c93e510fecc8…	"Film & Animation"	1024	0	0	0
4	ObjectId('63f1389c93e510fecc8…	"Science & Technology"	1895	0	0	0
5	ObjectId('63f1389c93e510fecc8…	"Education"	1432	0	0	0
6	ObjectId('63f1389c93e510fecc8…	"Sports"	1587	0	0	0
7	ObjectId('63f1389c93e510fecc8…	"Howto & Style"	1722	0	0	0
8	ObjectId('63f1389c93e510fecc8…	"Travel & Events"	367	0	0	0
9	ObjectId('63f1389c93e510fecc8…	"Gaming"	3047	0	0	0
10	ObjectId('63f1389c93e510fecc8…	"News & Politics"	1731	0	0	0
11	ObjectId('63f1389c93e510fecc8…	"Comedy"	3556	0	0	0
12	ObjectId('63f1389c93e510fecc8…	"Entertainment"	21587	0	0	0
13	ObjectId('63f1389c93e510fecc8…	"Autos & Vehicles"	317	0	0	0
14	ObjectId('63f1389c93e510fecc8…	"Pets & Animals"	24	0	0	0
15	ObjectId('63f1391093e510fecc8…	"Education"	0	123	0	0
16	ObjectId('63f1391093e510fecc8…	"Entertainment"	0	7609	0	0
17	ObjectId('63f1391093e510fecc8…	"People & Blogs"	0	2949	0	0
18	ObjectId('63f1391093e510fecc8…	"Music"	0	2976	0	0
19	ObjectId('63f1391093e510fecc8…	"Film & Animation"	0	1118	0	0
20	ObjectId('63f1391093e510fecc8…	"Sports"	0	1961	0	0

Σχήμα 18: κατηγορίες περιοχής

δεν είναι τα αποτελέσματα που ζητάμε, καθώς εμείς αυτό που ζητάμε είναι το πλήθος των κατηγοριών για την συγκεκριμένη χώρα αλλά και το συνολικό πλήθος των κατηγοριών για όλες τις χώρες. Στα αποτελέσματα πάνω κάποιες τιμές είναι 0 και το συνολικό πλήθος δεν υπάρχει πουθένα. Δυστυχώς αυτό δεν μας βολεύει και γι΄ αυτό τον λόγο θα χρειαστεί να επεξεργαστούμε λίγο τα δεδομένα μας. Επόμενως, αυτό που θα κάνουμε είναι να ομαδοποιήσουμε τις κατηγορίες και με βάση την κατηγοριοποίηση να κάνουμε την πρόσθεση, έτσι ώστε να πάρουμε πρώτα το πλήθος των κατηγοριών ανά χώρα, δηλαδή αυτό 0+334+0+0. Έπειτα, με την επόμενη εντολή θα κάνουμε ξανά πρόσθεση όλες τις χώρες για την συγκεκριμένη κατηγορία με αποτέλεσμα να αποκτήσουμε το πλήθος όλων των περιοχών ανά κατηγορία. Φυσικά, στο τέλος θα τα ταξινομήσουμε με βάση την κατηγορία και θα αποθηκεύσουμε τα δεδομένα μας σε .csv αρχείο.

```
{$group:
    {
      _id: "$Category",
      IN: {
        $sum: "$IN",
      },
      JP: {
        $sum: "$JP",
      },
      KR: {
        $sum: "$KR",
      },
      RU: {
        $sum: "$RU",
      },
    },
},
{
  $project:
    {
      _id: 1,
      totalSUM: {
        $add: ["$IN", "$JP", "$KR", "$RU"],
      },
      IN: 1,
      JP: 1,
      KR: 1,
      RU: 1,
    },
},
{
  $sort:
    {
      _id: 1,
    },
},
```

2.7.3 Έυρεση μέσο όρο κατηγοριών

Για την εύρεση των μέσο όρων είναι περίπου η ίδια διαδικασία. Θα ομαδοποιήσουμε τις κατηγορίες και απλά θα κάνουμε πρόσθεση τις προβολές. Έπειτα, επειδή έχουμε σκοπό τα δεδομένα αυτά να τα βάλουμε μαζεμένα σε ένα collection ξανά θα πρέπει να το διαμορφώσουμε για κάθε περιοχή ξεχωριστά. Επειδή εδώ έχουμε την Ινδία, θα εμφανίσουμε την κατηγορία φυσικά, το πλήθος της κατηγορίας και οι προβολές των βίντεο για την συγκεκριμένη κατηγορία με ονόμα INs και INv αντίστοιχα. Για τα υπόλοιπα, θα προσθέσουμε την τιμή 0. Τέλος, θα αφαιρέσουμε το id διότι έχουμε βάλει ήδη το όνομα της κατηγορίας στο πεδίο "Category", όπως κάναμε και προηγουμένως. Τέλος, θα τα μεταφέρουμε στο νέο μας collection.

```
 {
  $group:
    {
      _id: "$categoryTitle",
      sum: {
        $sum: 1,
      },
      views: {
        $sum: {
          $toInt: "$view_count",
        },
      },
    },
},
{
  $project:
    {
      Category: "$_id",
      INs: "$sum",
      INv: "$views",
      JPs: {
        $toInt: "0",
      },
      JPv: {
        $toInt: "0",
      },
      KRs: {
        $toInt: "0",
      },
      KRv: {
        $toInt: "0",
      },
      RUs: {
        $toInt: "0",
      },
      RUv: {
        $toInt: "0",
      },
    },
},
{
  $unset:
    "_id",
},
{
  $merge:
    {
      into: "bonus3",
    },
}
```

Φυσικά, αυτό δεν αρκεί καθώς ακόμα δεν έχουμε βρει τον μέσο όρο και αντιμετωπίζουμε παρόμοιο πρόβλημα όπως και προηγουμένως. Συνεπώς, στο νέο collection θα κάνουμε τις εξής εντολές. Θα ομαδοποιήσουμε με βάση την κατηγορία, θα αθροίσουμε με βάση την χώρα και στην συνέχεια θα κάνουμε πρόσθεση όλα τα νούμερα από όλες τις χώρες και θα το χρησιμοποιήσουμε για κάθε χώρα ως παρανομαστή έτσι ώστε να έχουμε τον μέσο όρο που θέλουμε για κάθε χώρα. Η εντολή round κόβει τα δεκαδικά.

```
 [
{
  $group:
    {
      _id: "$Category",
      INs: {$sum: "$INs"},
      JPs: {$sum: "$JPs"},
      KRs: {$sum: "$KRs"},
      RUs: {$sum: "$RUs"},
      INv: {$sum: "$INv"},
      JPv: {$sum: "$JPv"},
      KRv: {$sum: "$KRv"},
      RUv: {$sum: "$RUv"},
    },
},
{
  $project:
    {
      _id: 1,
      INavg: {$round: [{$divide: ["$INv",
              {
                $add: ["$INs","$JPs","$KRs","$RUs",]
                ,},],},2,],},
      JPavg: {
        $round: [{$divide: ["$JPv",
              {
                $add: ["$INs","$JPs","$KRs","$RUs",
                ],},],},2,],},
      KRavg: {
        $round: [{$divide: ["$KRv",
              {
                $add: ["$INs","$JPs","$KRs","$RUs",
                ],},],},2,],},
      RUavg: {
        $round: [
          {
            $divide: ["$RUv",
              {
                $add: ["$INs","$JPs","$KRs","$RUs",
                ],},],},2,],},},
},
```

```
    {
      $sort:
        {
          _id: 1,
        },
    },
]
```

2.7.4 Αποτελέσματα ερωτήματος 2.7

	_id	INavg	JPavg	KRavg	RUavg
2	Autos & Vehicles	71482.29	32836.1	34593.44	328336.59
3	Comedy	841552.21	150439.57	130147.39	414627.12
4	Education	685151.62	65772.45	132257.89	106362.25
5	Entertainment	1044005.82	162526.31	189914.09	315977.22
6	Film & Animation	372112.18	133748.58	86963.64	333629.24
7	Gaming	276494.08	118940.88	55100.01	293057.27
8	Howto & Style	460729.15	76584.33	112651.34	214480.54
9	Music	1859531.24	808630.03	855876.58	634307.28
10	News & Politics	92848.08	28260.37	71068.84	394009.86
11	People & Blogs	734490.98	73404.03	101916.51	254291.9
12	Pets & Animals	249860.03	228933.18	241672.42	270458.13
13	Science & Technolc	1156894.25	81201.03	123909.15	287827.06
14	Sports	300933.63	107327.6	154970.3	439919.48
15	Travel & Events	440287.25	40005.47	230951.94	181892.51

(α′) Μέσος Όρος Κατηγοριών

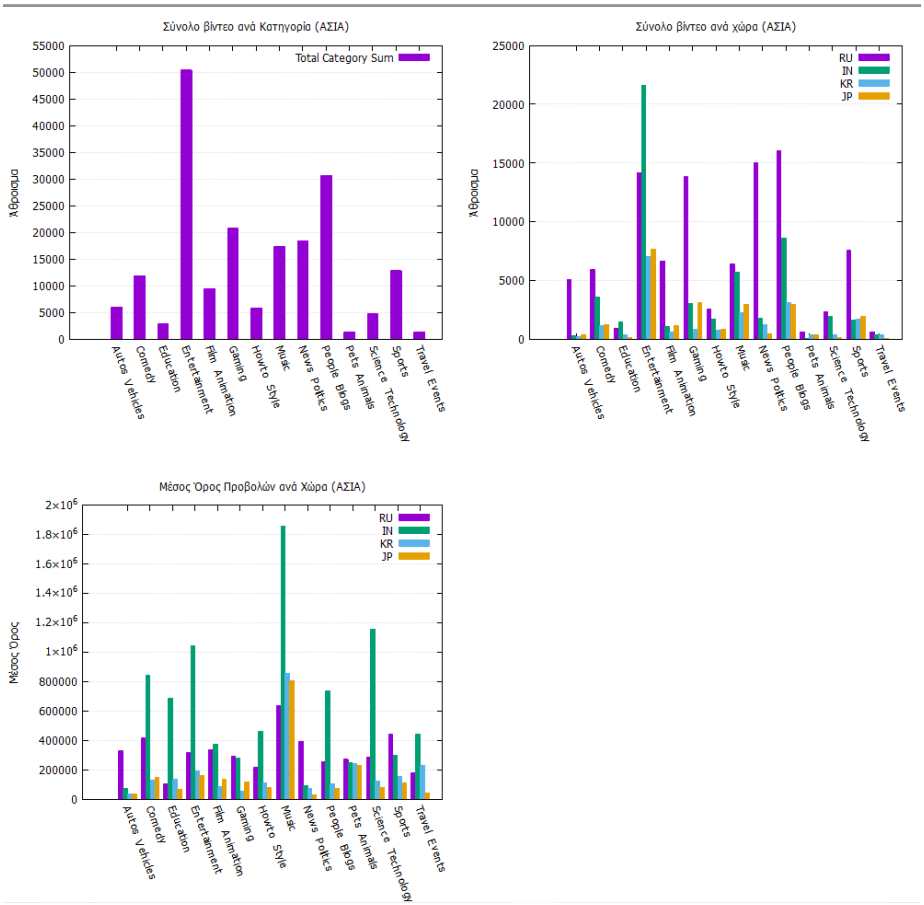
	_id	IN	JP	KR	RU	totalSUM
2	Autos & Vehicles	317	378	224	5052	5971
3	Comedy	3556	1250	1170	5909	11885
4	Education	1432	123	390	915	2860
5	Entertainment	21587	7609	7028	14144	50368
6	Film & Animation	1024	1118	616	6631	9389
7	Gaming	3047	3069	805	13803	20724
8	Howto & Style	1722	797	765	2520	5804
9	Music	5652	2976	2264	6361	17253
10	News & Politics	1731	462	1216	14997	18406
11	People & Blogs	8558	2949	3062	16041	30610
12	Pets & Animals	24	330	372	618	1344
13	Science & Technolc	1895	97	328	2354	4674
14	Sports	1587	1961	1714	7547	12809
15	Travel & Events	367	49	325	629	1370

(β′) Άθοισμα Κατηγοριών

Σχήμα 19: Αποτελέσματα ερωτήματος 2.7

2.7.5 Γραφικές απεικονίσεις ερωτήματος 2.7

Παρακάτω βλέπουμε την γραφική απεικόνιση του ερωτήματος 2.7. Θα σχολιάσουμε και τα 3 γραφήματα και θα επικεντρωθούμε σε πολλά ερωτήματα. Η πρώτη εικόνα περιέχει το σύνολο των κατηγοριών από όλες τις χώρες. Η δεύτερη εικόνα περιέχει το σύνολο των κατηγοριών αλλά ανά χώρα και η τελευταία εικόνα περιέχει τον μέσο όρο των κατηγοριών ανά χώρα. Για να βγει ο μέσος όρος, διαιρέσαμε με το πλήθος όλων των κατηγοριών.



Σχήμα 20: Γραφική Απεικόνιση Αποτελεσμάτων 2.7

2.7.6 Σχόλια για ερώτημα 2.7

Αρχικά, όπως παρατηρήσατε θα σχολιάσουμε την κατάσταση των κατηγοριών που χρησιμοποιούν οι χώρες της Ασίας. Από την πρώτη εικόνα παρατηρούμε ότι η πιο δημοφιλής κατηγορία βίντεο είναι το Entairtainment και η λιγότερη δημοφιλής η κατηγορία Pets Animals με μικρή διαφορά από την κατηγορία Travel Events. Επιπρόσθετα, από την δεύτερη εικόνα παρατηρούμε κάποια δεδομένα που ¨αποδεικνύουν' κάποια γεγονότα που θεωρούνται σαν δεδομένα στην καθημερινότητα μας. Για παράδειγμα, όσων αφορά με την Ρωσία παρατηρούμε ότι έχει πολύ υψηλό αριθμό βίντεο για την κατηγορία που αφορά τις Ειδήσεις και Πολιτική καθώς η

Ρωσία είναι μία πολυ δυνατή, ανεξάρτητη και μεγάλη χώρα όπου είναι λογικό, αφού έχει και μεγάλη επιρροή στον κόσμο να επικεντρώνονται οι άνθρωποι στην Ρωσία σε αυτήν την κατηγορία. Εκτός από αυτό, στον κόσμο του Gaming είναι δεδομένο ότι θα συναντήσεις Ρώσους να παίζουν αρκετά παιχνίδια, συνεπώς ήταν αναμενόμενο η Ρωσία να έχει μεγαλύτερο αριθμό βίντεο σε αυτήν την κατηγορία. Τέλος, κάτι που μπορούμε να παρατηρήσουμε από την τελευταία εικόνα είναι, για παράδειγμα για την κατηγορία Music η Ινδία να έχει πολύ υψηλό μέσο όρο. Αυτό είναι λογικό διότι η Ινδία φημίζεται για το Bollywood και γενικά έχουν πάρα πολλά τραγούδια που προέρχονται από τις σειρές και τις ταινίες τους. Επίσης, για την κατηγορία Music η Κορέα και η Ιαπωνία είναι σχεδόν ίσα και είναι λογικό διότι αυτές οι χώρες, ειδικά τον τελευταίο καιρό, φημίζονται πολύ για την K-Pop και J-Pop αντίστοιχα, και επειδή πολλοί Κορεάτες καλλιτέχνες συνηθίζουν να βγάζουν και Japanese Version των τραγουδιών τους ίσως αυτό να εξήγουσε και αυτά τα τόσο κοντινά αποτελέσματα. Ένα τελευταίο πράγμα που μπορούμε να σχολιάσουμε, είναι για την κατηγορία Science Technology που η Ινδία, ξανά, να ξεπερνάει σε τεράστιο βαθμό τις υπόλοιπες χώρες. Αυτό και πάλι μπορεί να δικαιολογηθεί καθώς, ειδικά στον τομέο της Πληροφορικής και Τεχνολογίας, οι Ινδοί να είναι οι 'σωτήρες' των φοιτητών και γενικά των ανθρώπων που εργάζονται σε αυτούς τους τομείς, και γι΄ αυτό τον λόγο βλέπουμε αυτά τα στατιστικά.

3 Εργαλεία που εγκαταστήσαμε

Τρέξαμε τα ερωτήματα στο λογισμικό MongoDB Compass. Ουσιαστικά, αυτό που έπρεπε να κάνουμε είναι να δημιουργήσουμε κάποια pipelines για κάθε ερώτημα, με κάποια stages που θα μας δίδανε τα επιθυμητά αποτελέσματα. Για κάθε pipeline θα εξηγήσαμε περιληπτικά βήμα βήμα τι γίνεται. Επίσης, για τις γραφικές απεικονίσης χρησιμοποιήσαμε το εργαλείο Gnuplot.