Analiza skupa podataka o popularnim YouTube video klipovima

Seminarski rad u okviru kursa

Istraživanje podataka

Matematički fakultet

Stefan Mirić 333/2015

Nikola Vuković 71/2015

Septembar 2018.

[**Uvod**](#_uqyhk3qs9ztu) **3**

[**Priprema podataka**](#_7xz5cgd5db1k) **4**

[**YouTube algoritam i očekivanja od analize**](#_55xq92yfknzc) **4**

[**Analiza podataka**](#_ayrokampksm) **5**

[Analiza popularnih klipova](#_8z0lkl77xcws) 5

[Analiza dužine popularnosti](#_ka5gn6ojrik1) 5

[Vreme do početka popularnosti](#_nbsve7tsebtf) 7

[Vreme objave i popularnost istog dana](#_aakjwiyqjb6h) 8

[Broj klipova sa naslovom u velikim slovima](#_fy36gkx5rrev) 8

[Analiza kategorija](#_jbt5lu9l3l5l) 9

[Najpopularnije kategorije](#_tbt91tvi0rd6) 9

[Dužina popularnosti po kategoriji](#_jcip08m2g7at) 10

[Poređenje zemalja](#_kr6nxeqgw8bu) 11

[Procenat istih popularnih klipova između zemalja](#_auh00nw3w2o1) 11

[Poređenje broja pregleda, ocena i komentara](#_dcfdii9i95ls) 14

[Popularnost kategorija po zemljama](#_ddqn8jo5pkd9) 16

[**Klasterovanje**](#_5unvyakxdxry) **21**

[**Klasifikacija**](#_785zwb5p58d) **23**

[KNN algoritam](#_ry67soeli9zi) 23

[Naivni Bajesov algoritam](#_6o5t8i6gzyfq) 25

[Mašinsko učenje RProp algoritmom](#_5ev0unay1qhb) 26

[**Pravila pridruživanja**](#_6bfuhigsdc4q) **27**

# Uvod

U radu će biti analiziran skup podataka “Trending YouTube Video Statistics” (<https://www.kaggle.com/datasnaek/youtube-new>) koji prikazuje 200 najpopularnijih klipova po YouTube listi popularnih klipova za svaki dan u periodu 14. 11. 2017. - 14. 6. 2018. Uzima se u obzir da se neki klipovi pojavljuju više puta uz različite datume jer su bili popularni više dana.

Set zadrži podsetove podataka za Kanadu, Nemačku, Francusku, Veliku Britaniju, i SAD. Svi podsetovi su identičnog formata i opsega. U prvom delu analize ćemo se, zarad jednostavnosti, ograničiti na analizu podataka za SAD. Prikazani koraci analize se preslikavaju na druge podsetove iz seta.

Skup sadrži sledeće kolone:

* video\_id: (niska) identifikator video klipa sa samog sajta
* trending\_date: (niska) dan na koji je video klip bio popularan (ako je bio popularan više dana biće više redova sa istim identifikatorom i različitim datumom)
* title: (niska) naziv video klipa
* channel\_title: (niska) naziv korisnika koji je objavio video klip
* category\_id: (ceo broj) identifikator kategorije klipa (informacije o kategorijama, kao što je naziv kategorije, se nalaze u odvojenom fajlu za svaku zemlju)
* publish\_time: (niska) datum i vreme objave
* tags: (niz niski odvojenih karakterom “||”) labele kojima je korisnik obeležio svoj klip
* views: (ceo broj) broj pregleda u trenutku sačinjavanja skupa
* likes: (ceo broj) broj korisnika koji su označili da im se klip sviđa
* dislikes: (ceo broj) broj korisnika koji su označili da im se klip ne sviđa
* comment\_count: (ceo broj) broj komentara na klipu
* thumbnail\_link: (niska) http link na sličici koja predstavlja klip na sajtu
* comments\_disabled: (*boolean*) označava da li su komentari isključeni
* ratings\_disabled: (*boolean*) označava da li je zabranjeno označavanje sviđanja
* video\_error\_or\_removed: (*boolean*) označava da li je video klip uklonjen ili se desila neka greška pri unosu u skup podataka
* description: (niska) opis video klipa sa sajta

# 

# Priprema podataka

Za pripremu podataka je korišćena Python skripta (prepareData.py) uz Pandas biblioteku. U setu nije bilo redova kojima fale atributi (neki redovi imaju prazan opis ali to ne predstavlja problem).

U pripremi su dodati atributi:

* category: (niska) naziv kategorije
* days\_to\_trend: (ceo broj) broj dana između objave i početka popularnosti
* days\_trending: (ceo broj) broj dana tokom kojih je video klip bio popularan
* publish\_weekday: (ceo broj) dan u nedelji kada je objavljen klip

Uključena je Python skripta korišćenja za pripremu podataka.

U pripremi su izbačeni atributi: thumbnail\_link, description.

# YouTube algoritam i očekivanja od analize

Google [tvrdi](https://support.google.com/youtube/answer/7239739?hl=en) da algoritam koji bira popularne klipove uzima u obzir broj pregleda, brzinu rasta broja pregleda, i starost klipa. Detalji o algoritmu nisu poznati. Pretpostavimo da se u obzir uzimaju i druge vrednosti kao broj korisnika pretplaćenih na kanal korisnika koji je objavio klip, broj korisnika koji su označili da im se svideo klip i broj komentara. Očekujemo slično ponašanje između klipova.

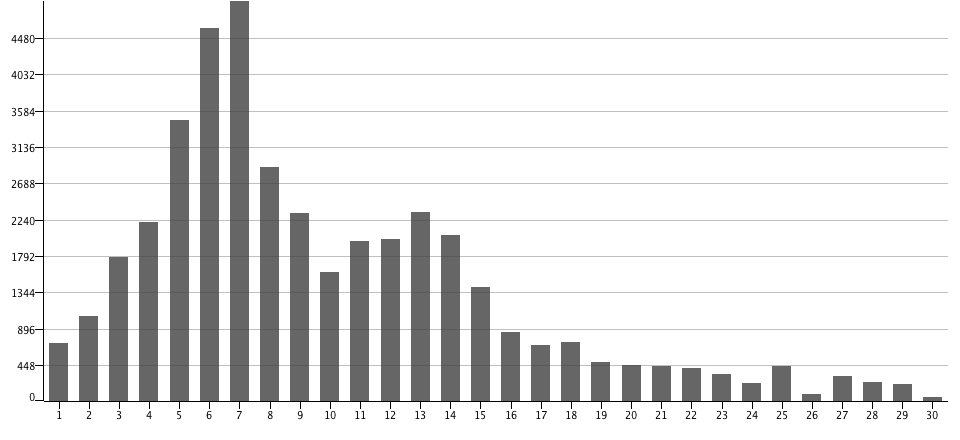
Google takođe tvrdi da algoritam uzima u obzir tematiku klipa, veličinu ciljne publike, kvalitet i slično, ali [brojni članci](https://www.google.com/search?q=youtube+trending+algorithm&tbm=nws) koji govore o nesposobnosti platforme da suzbije netačne “lažne vesti” i sličan nekvalitetan sadržaj govore suprotno, pa ove tvrdnje (koje je ujedno teško izmeriti i proveriti) nećemo uzeti u obzir u analizi.

# Analiza podataka

## Analiza popularnih klipova

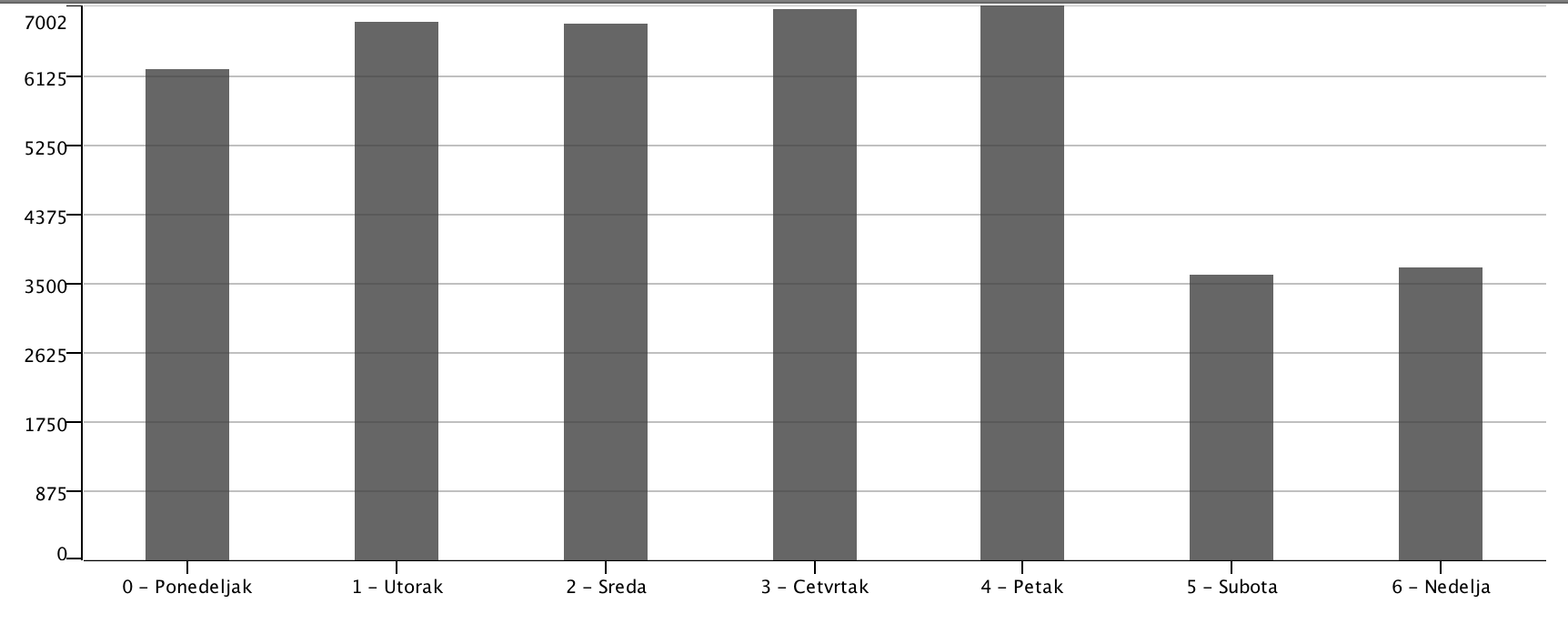
### Analiza dužine popularnosti

Za početak analiziraćemo dužinu popularnosti klipova u danima:



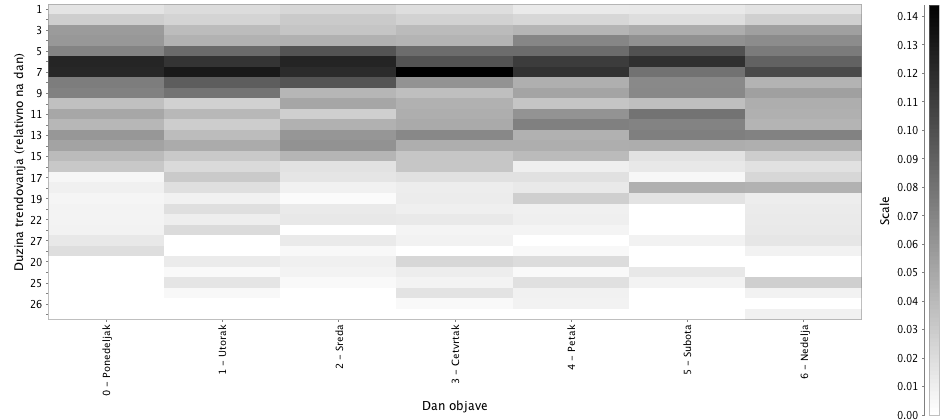
Odmah primećujemo da nijedan klip nije popularan duže od 30 dana. Može se pretpostaviti da algoritam to ne dozvoljava. Međutim pri analizi drugih zemalja vidimo da neki klipovi u Velikoj Britaniji su popularni i do 38 dana, dok su u Nemačkoj, Francuskoj i Kanadi klipovi popularni mnogo kraće. Nameće se pitanje da li se algoritam razlikuje od zemlje do zemlje ili je razlika u navikama gledalaca. Više o razlikama između zemalja kasnije.

Dalje primećujemo da su klipovi najčešće popularni 5 do 8 dana, ali takođe primećujemo da imamo dva lokalna maksimuma, jedan na nedelju dana a drugi na 13 dana - dan manje od dve nedelje. Sledi pitanje da li je dan u nedelji povezan sa dužinom popularnosti.



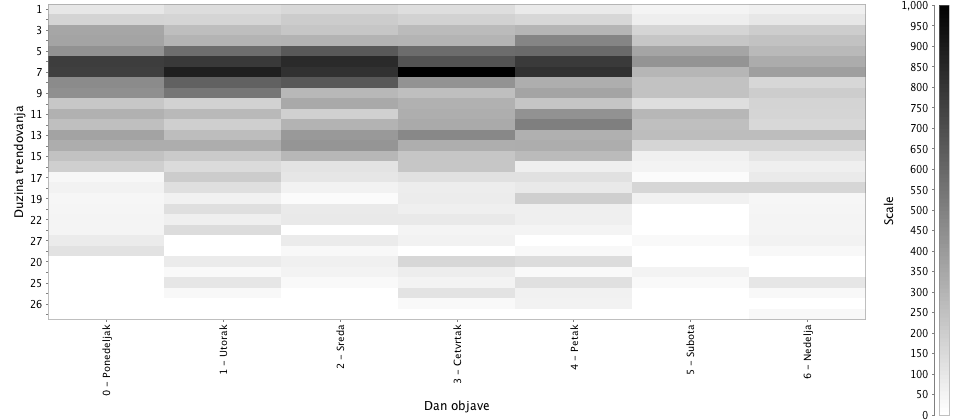
Primećujemo da je broj klipova objavljenih subotom i nedeljom skoro duplo manji od bilo kog drugog dana. Ovo potvrđuje preporuku [mnogih članaka](https://www.tubefilter.com/2015/01/12/best-days-times-to-post-youtube-videos-yearly-calendar/) koji tvrde da su subota i nedelja najnepovoljniji dani za objavu.

Primećujemo da klipovi objavljeni subotom imaju malo veću verovatnoću da budu popularni oko dve nedelje:



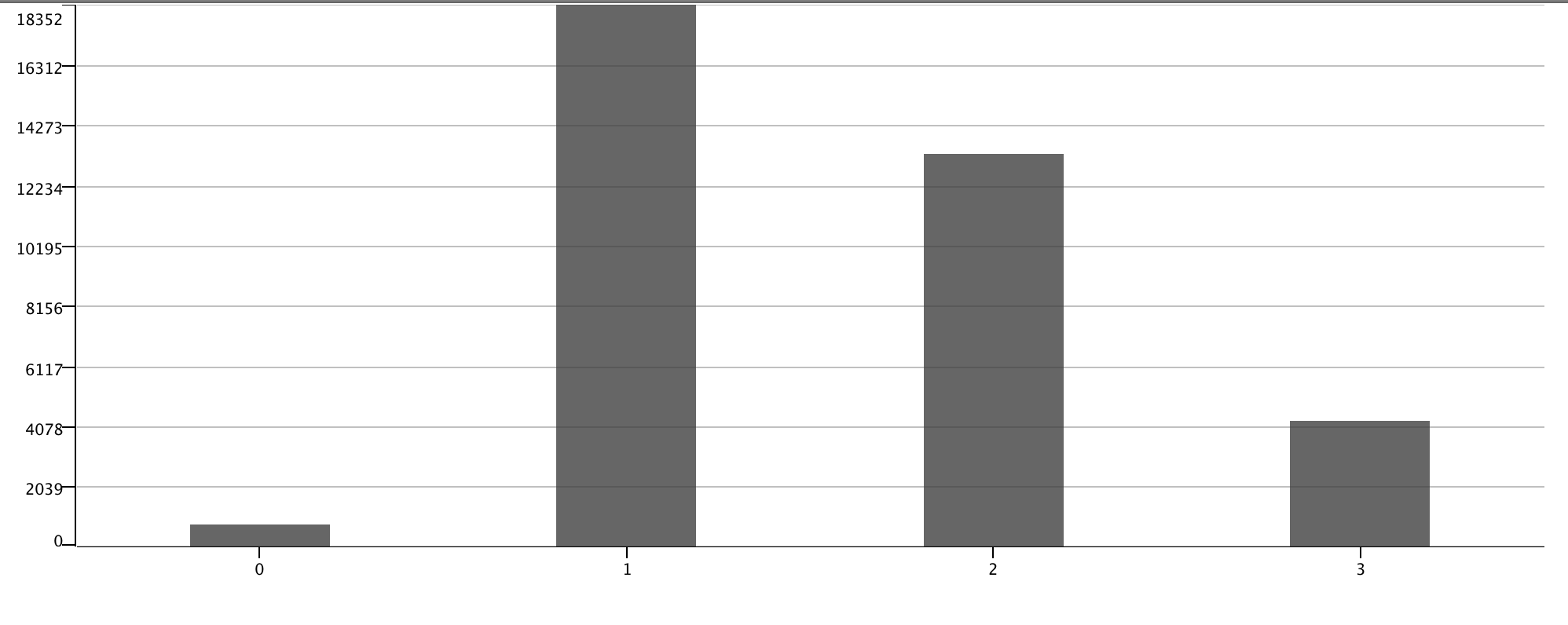
(Grafik je normalizovan tako da prikazuje gustnu relativnu na broj klipova objavljenih tog dana)

Ovo ipak nije značajan podatak kada uzmemo u obzir apsolutan broj klipova:



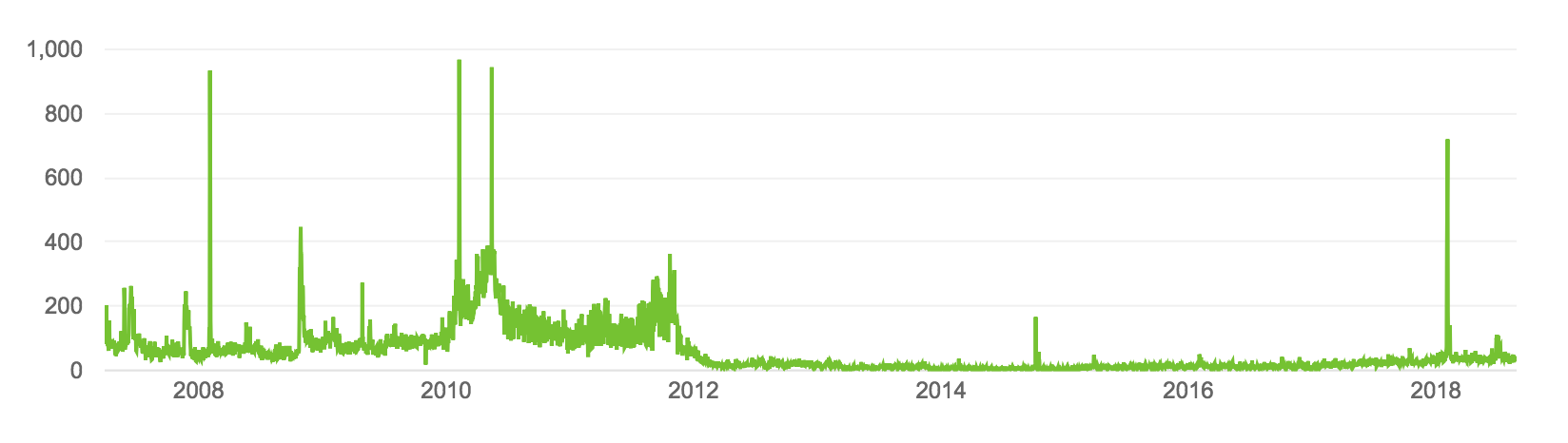
### Vreme do početka popularnosti

### 



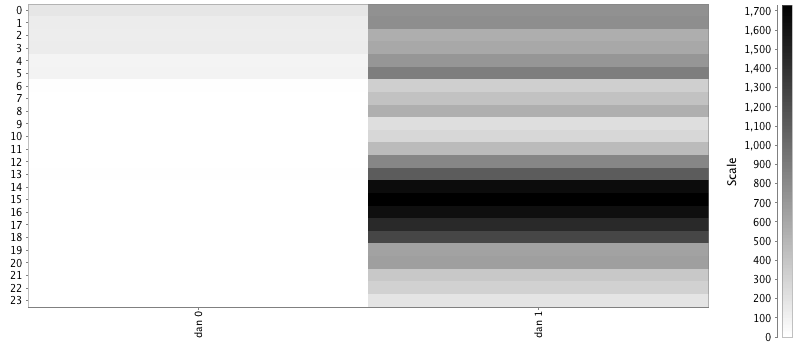
Broj popularnih klipova / broj dana do početka popularnosti

89.3% klipova počnu da budu popularniu roku od 3 dana nakon objave. Postoje slučajevi kada klipovi počnu da budu popularni mnogo vremena nakon svoje objave. Klip sa najviše vremena između objave i popularnosti je [reklama za pivo](https://www.youtube.com/watch?v=MJO3FmmFuh4) koja je bila popularna u februaru 2018, skoro 12 godina nakon objave 2006.



Pregledi reklame kroz vreme

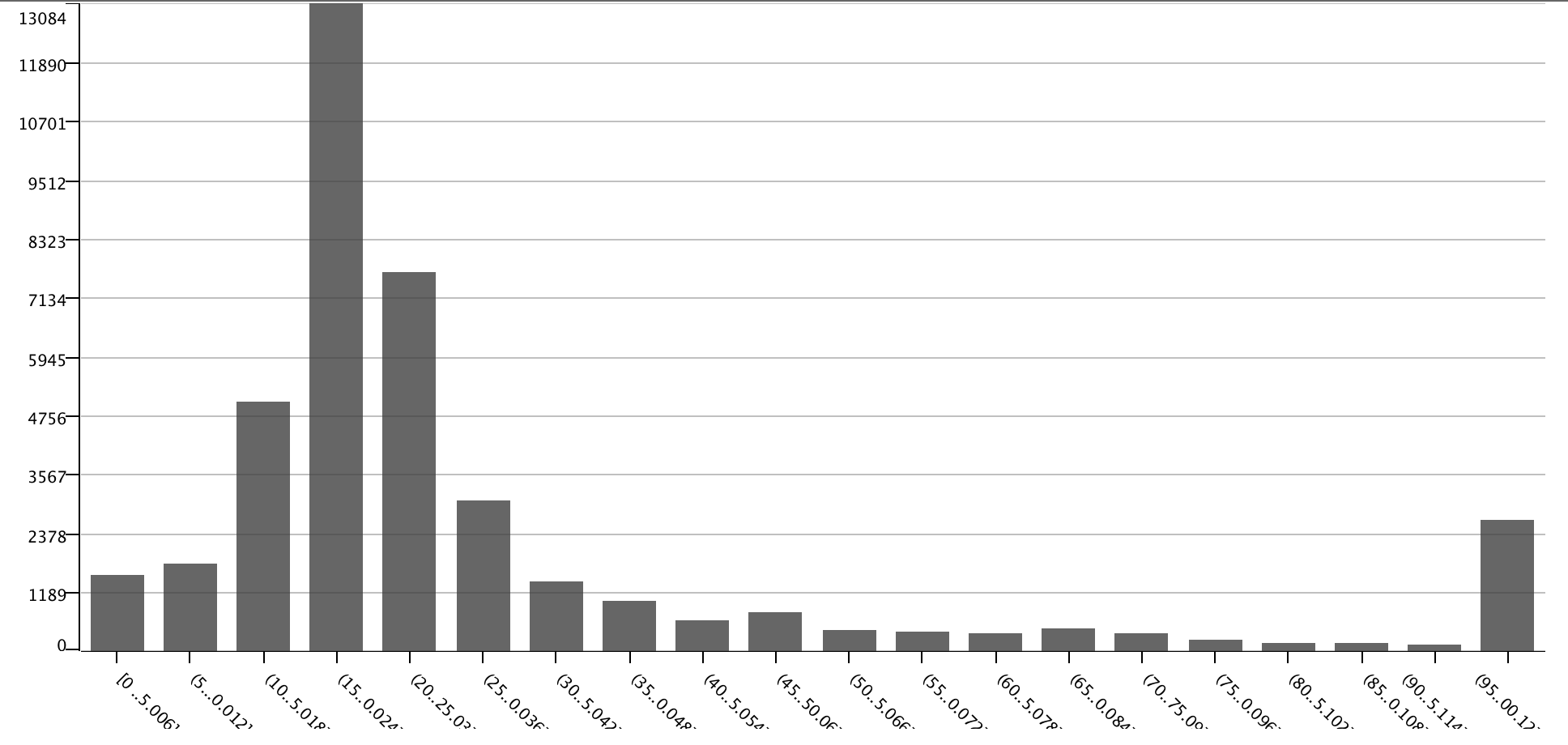
### Vreme objave i popularnost istog dana



Vreme objave / broj dana do popularnosti

Vidimo da klipovi ponekad uspeju da postanu popularni istog dana kada su objavljeni, ali samo ako su objavljeni do 6 ujutru.

### Broj klipova sa naslovom u velikim slovima

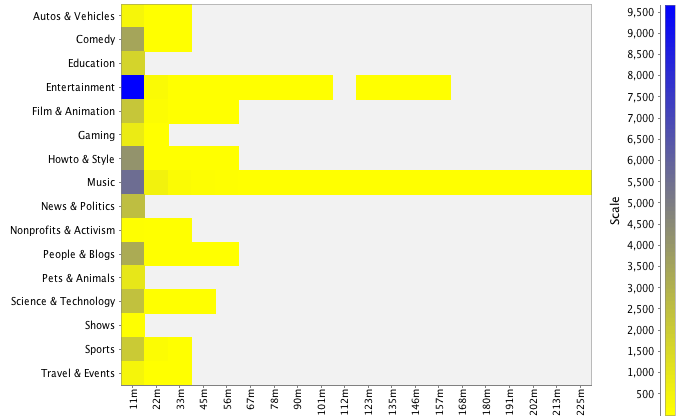


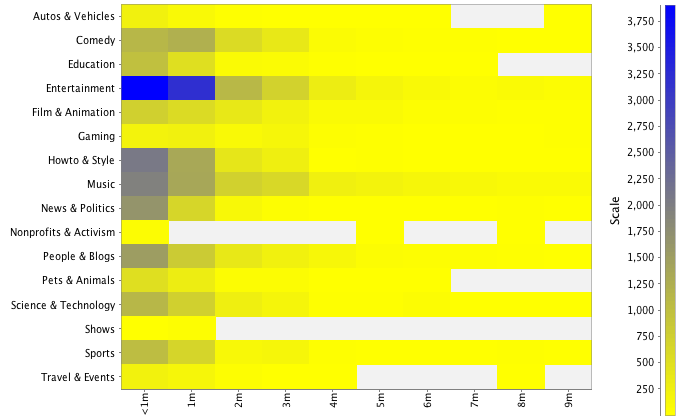
Procenat velikih slova u naslovu klipa

6.4% klipova je imalo naslov sa isključivo velikim slovima.

## Analiza kategorija

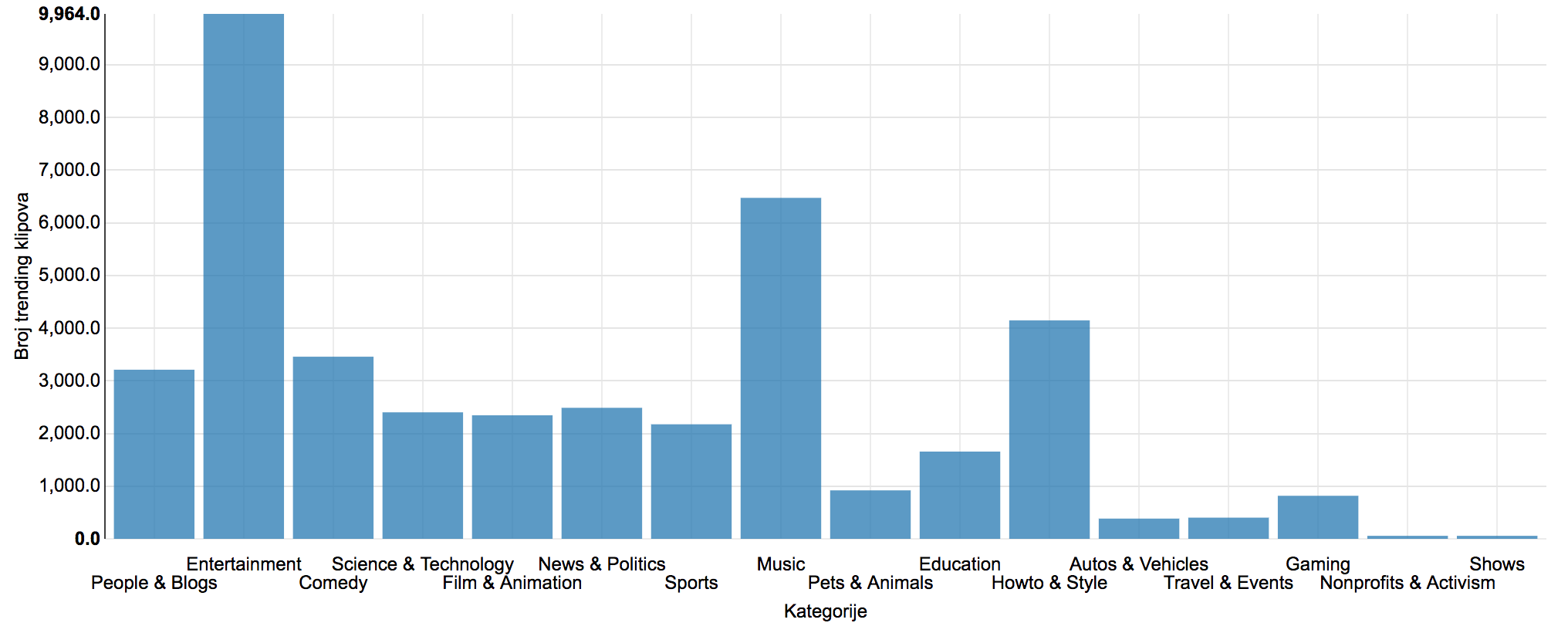
### Najpopularnije kategorije



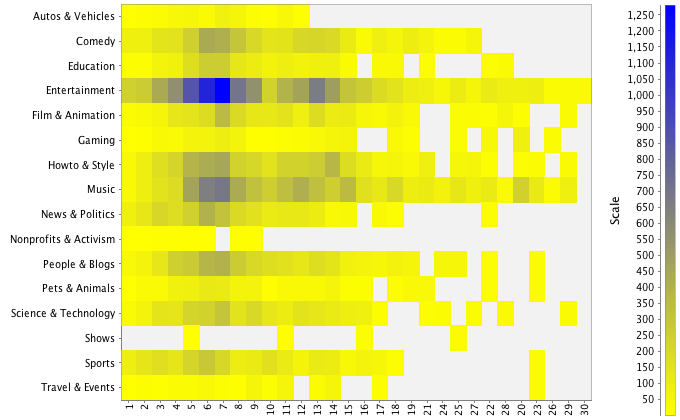


Kategorije po gustini broja pregleda - USA

Vidimo da su “Music” i “Entertainment” kategorije popularne značajno više od drugih. Primećujemo i razliku između njih - “Music” kategorija je jedina sa klipovima sa više od 150 miliona pregleda, ali “Entertainment” kategorija ima više popularnih klipova:



### Dužina popularnosti po kategoriji



Primećujemo da, očekivano, klipovi iz popularnijih kategorija su duže popularni u proseku.

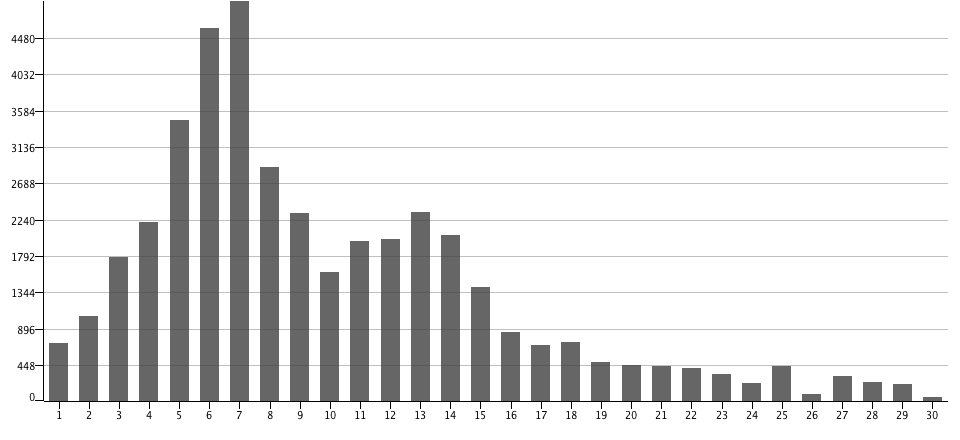
## Poređenje zemalja

### Procenat istih popularnih klipova između zemalja

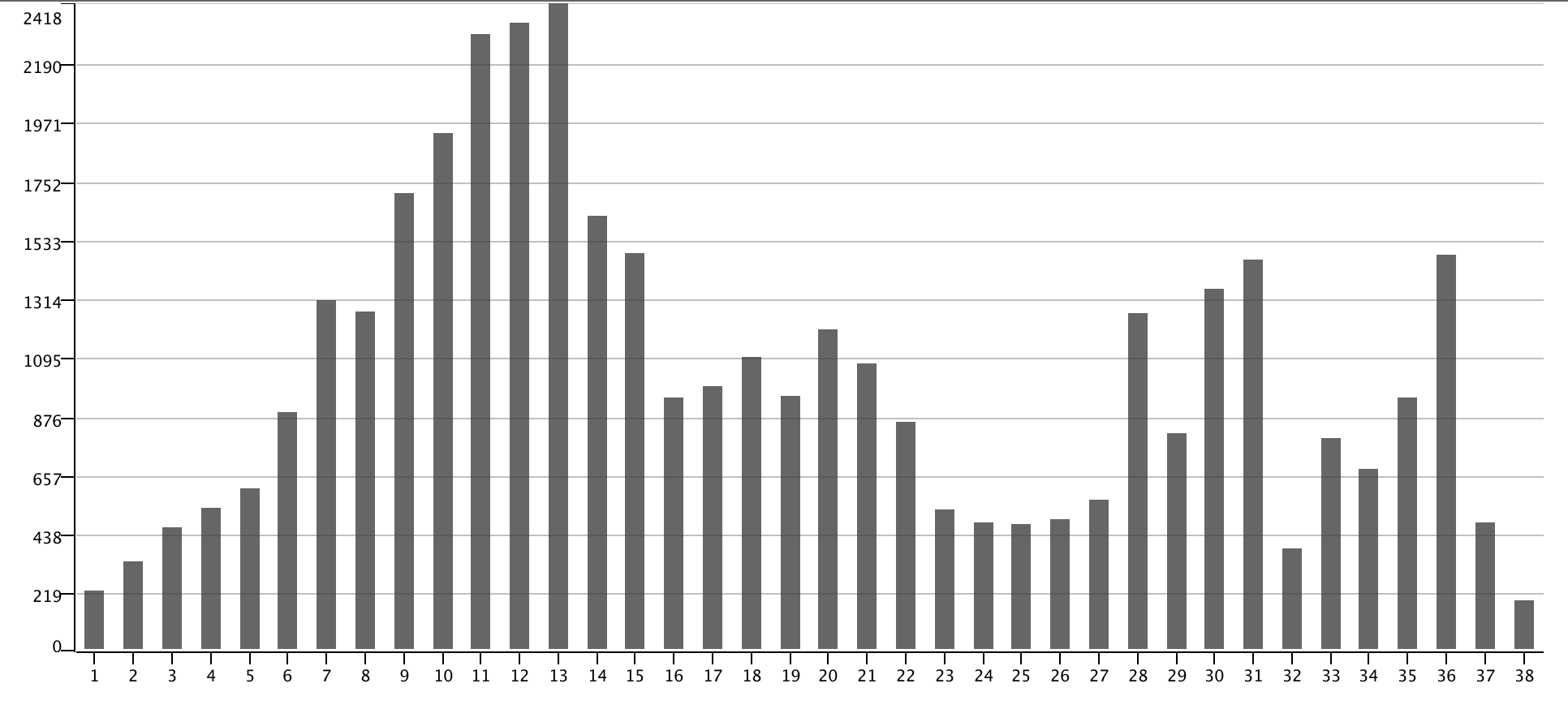
### 

Procenat popularnih klipova u drugoj zemlji

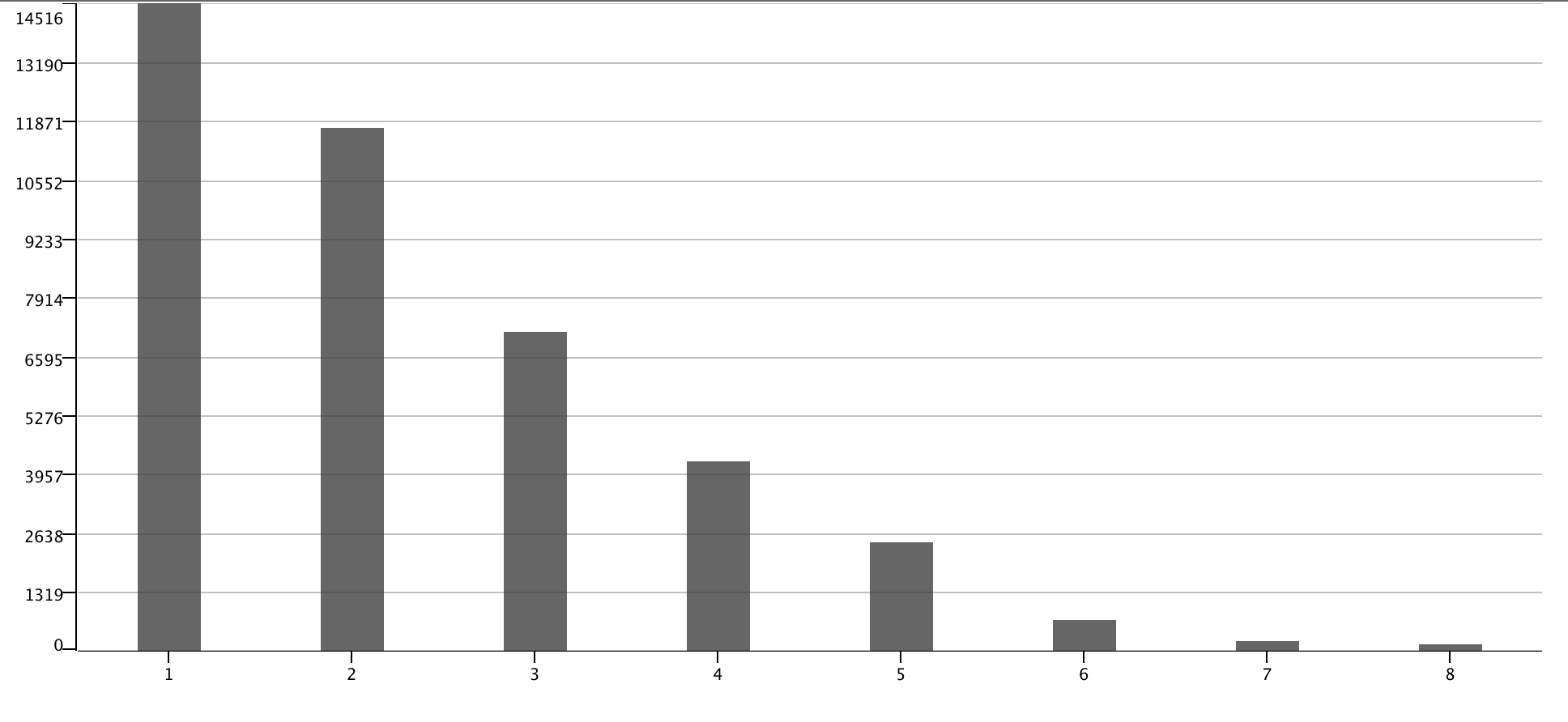
Najveću sličnost vidimo između SAD, Kanade i Velike Britanije. Ovo se donekle očekuje jer države u istom govornom području. Primećujemo takođe da procenat sličnosti nije simetričan, iz razloga što se dužina popularnosti drastično razlikuje između SAD, Velike Britanije, i ostalih država:



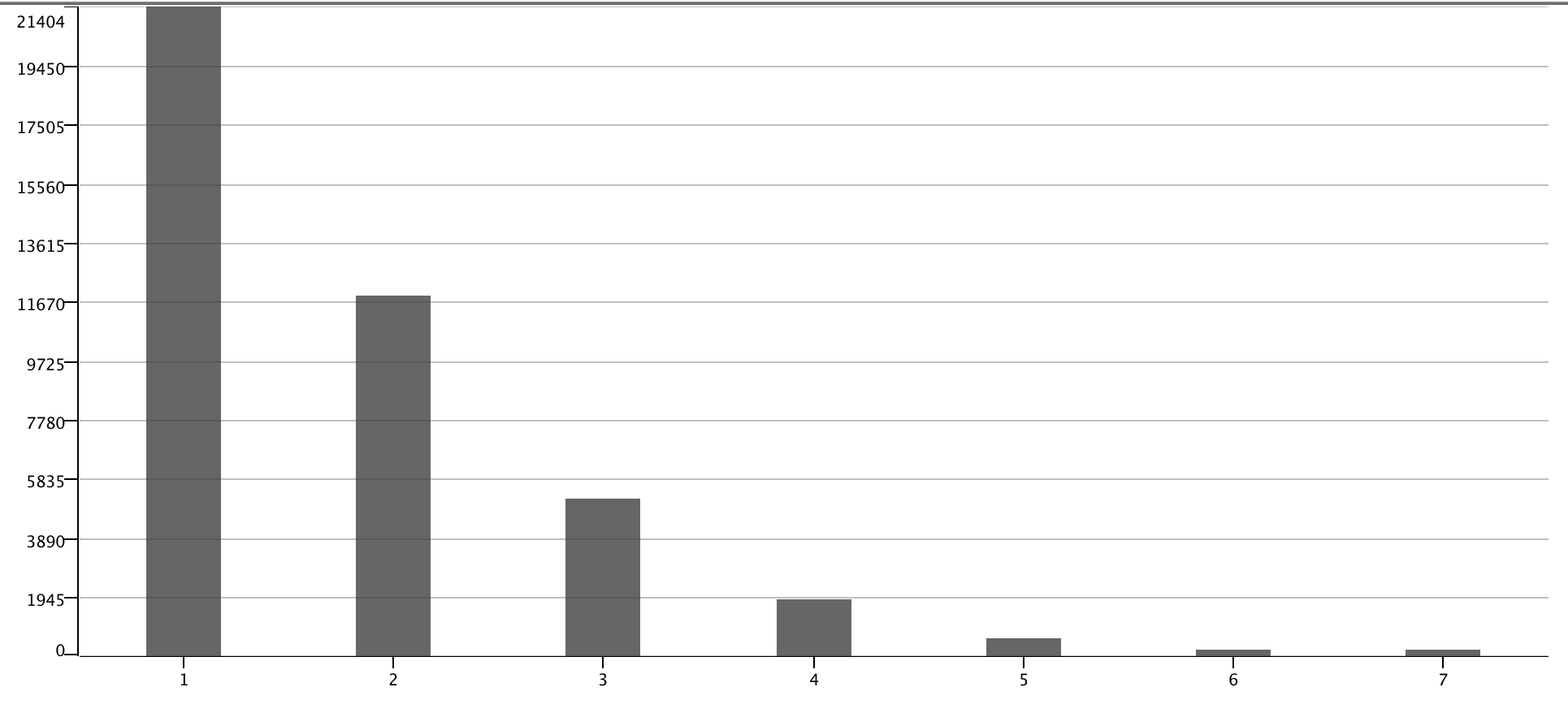
Dužina popularnosti u SAD



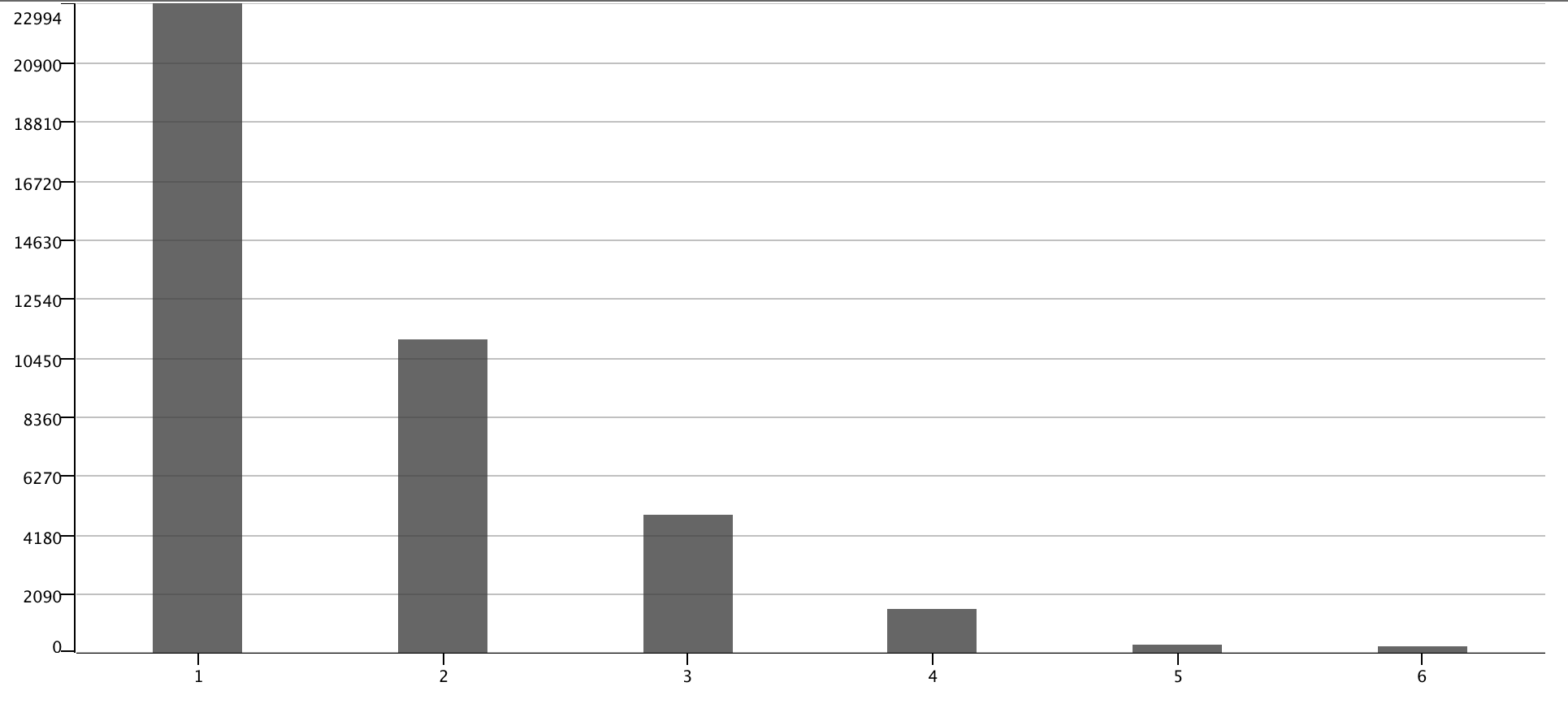
Dužina popularnosti u Velikoj Britaniji



Dužina popularnosti u Kanadi



Dužina popularnosti u Nemačkoj

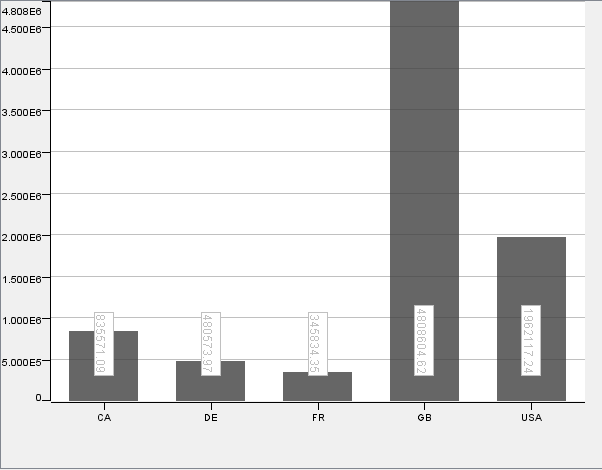


Dužina popularnosti u Francuskoj

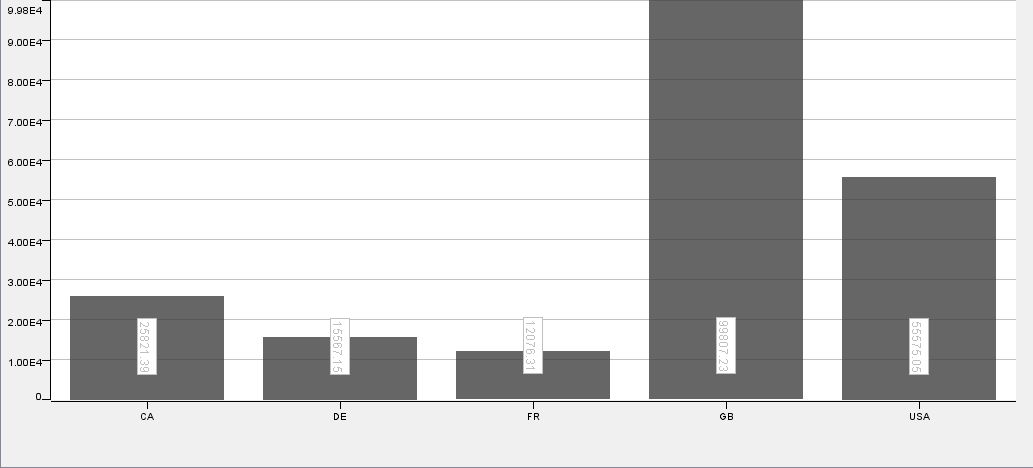
Primećuje se da su klipovi popularni mnogo duže u SAD i Velikoj Britaniji nego u ostalim državama. Iz ovog razloga broj jedinstvenih popularnih klipova je različit između država pa samim tim i asimetričnu tabelu sličnosti.

|  |  |  |
| --- | --- | --- |
| Država | Ukupno redova | Jedinstvenih klipova |
| Kanada | 40881 | 24427 |
| Nemačka | 40840 | 29627 |
| Francuska | 40724 | 30581 |
| Velika Britanija | 38916 | 3272 |
| SAD | 40949 | 6351 |

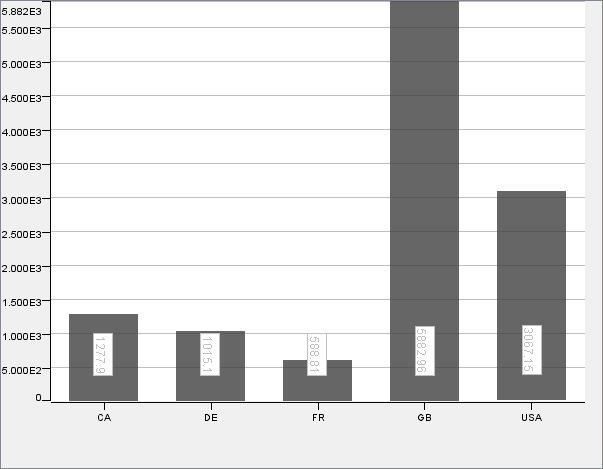
### Poređenje broja pregleda, ocena i komentara



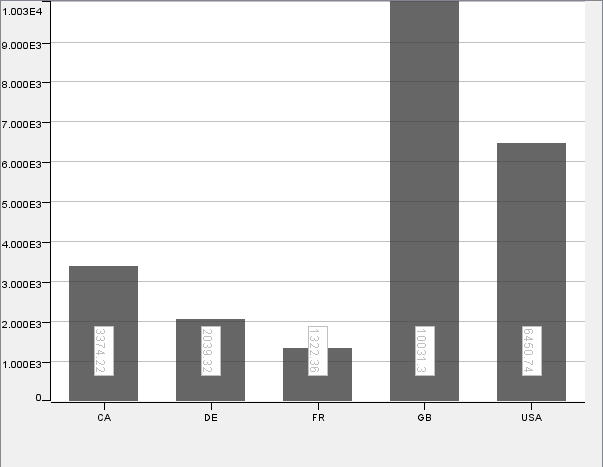
Prosečan broj pregleda klipa po državi



Prosečan broj pozitivnih ocena (“sviđa mi se”) na klipu po državi



Prosečan broj negativnih ocena (“Ne sviđa mi se”) na klipu po državi

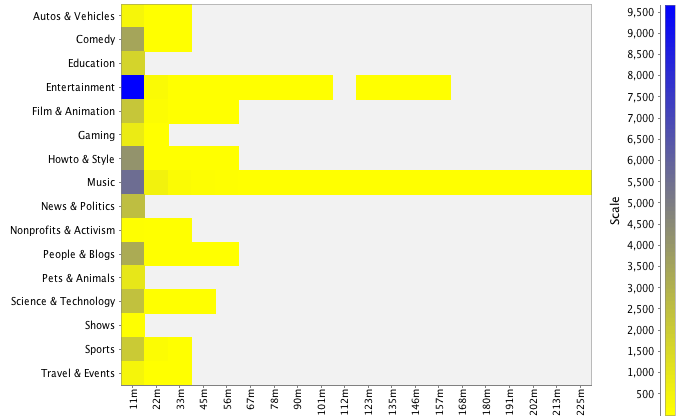
**

Prosečan broj komentara na klipu po državi

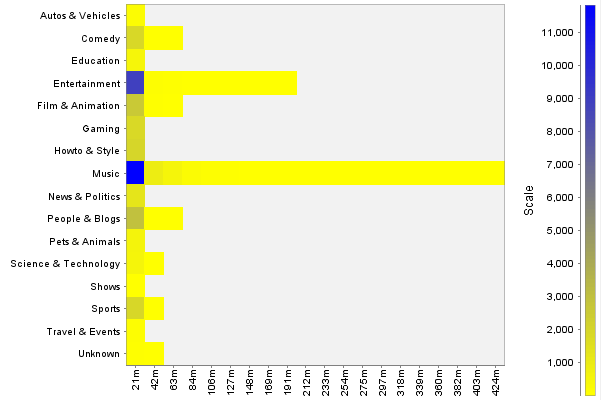
Primećujemo da je svaki grafik skoro identičnog oblika i da se značajno razlikuju samo brojevi između njih. Naime kako smo već pokazali da video najduže popularanu Velikoj Britaniji pa zatim u Americi, vidimo da to značajno utiče na statistike tog video klipa - ukoliko video klip je duže popularan imaće tendenciju da ima više pregleda, ocena i komentara.

### Popularnost kategorija po zemljama

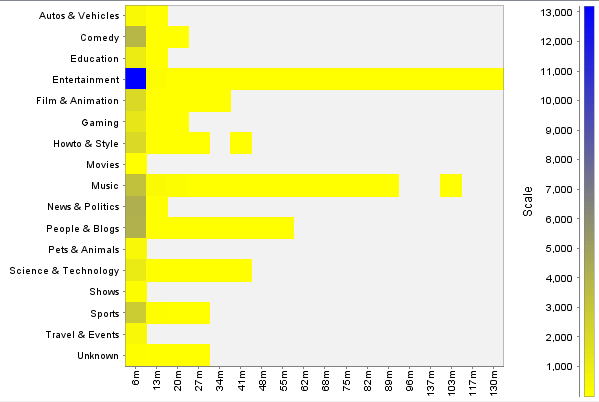
Primećuje se ponovo da su grafici sličnog oblika između zemalja, i da se razlikuju po broju pregleda - zbog prosečne dužine popularnosti klipa.



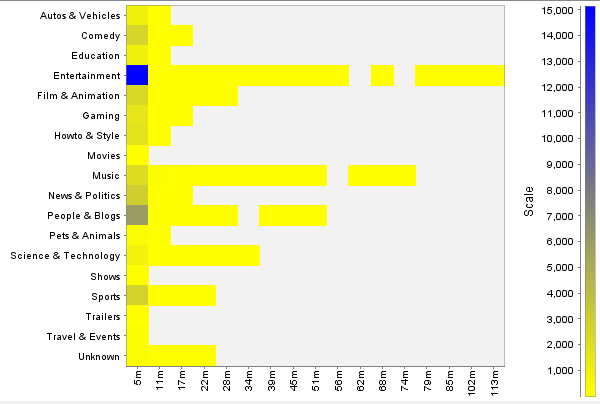
Broj pregleda po kategoriji u SAD



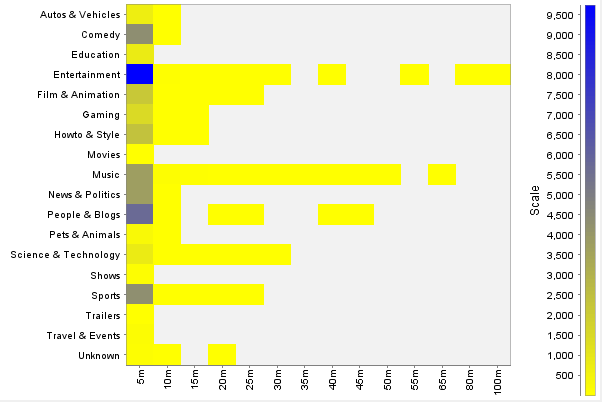
Broj pregleda po kategoriji u Velikoj Britaniji



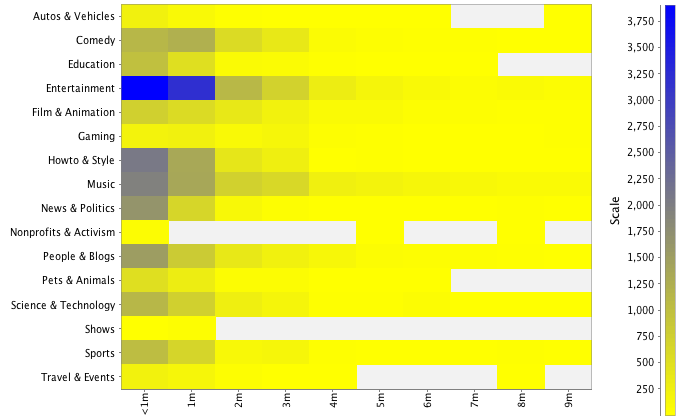
Broj pregleda po kategoriji u Kanadi



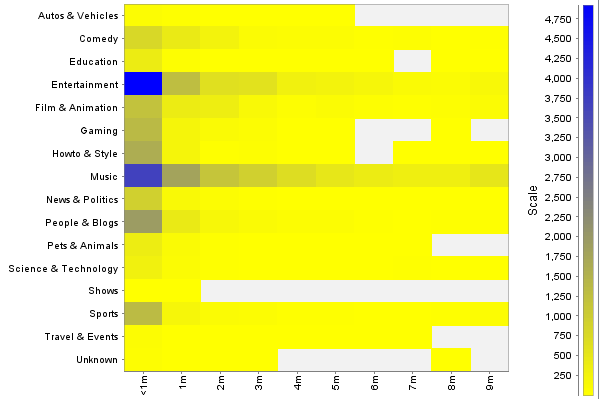
Broj pregleda po kategoriji u Nemačkoj



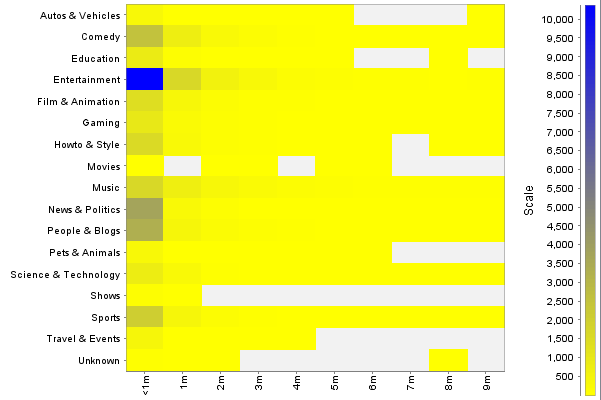
Broj pregleda po kategoriji u Francuskoj



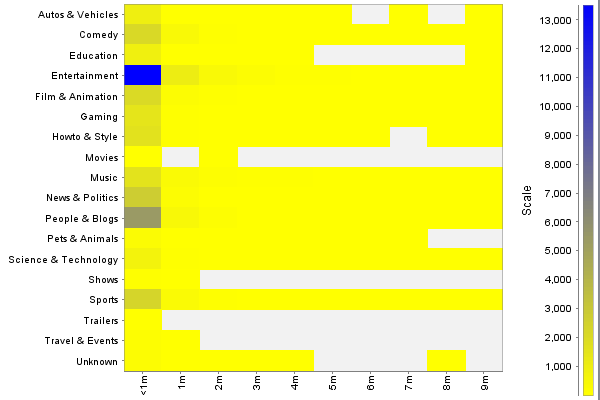
Broj pregleda po kategoriji u SAD do 10 mil.



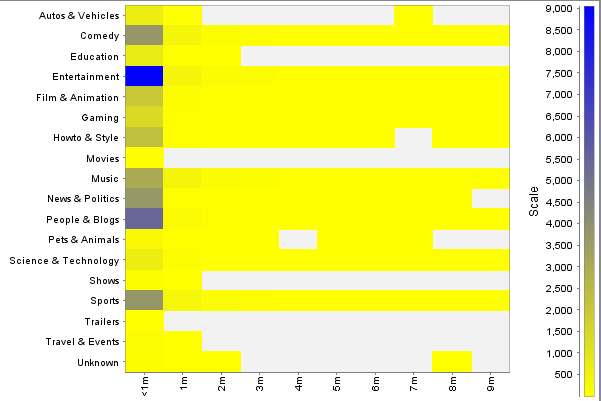
Broj pregleda po kategoriji u Velikoj Britaniji do 10mil.



Broj pregleda po kategoriji u Kanadi do 10mil.



Broj pregleda po kategoriji u Nemačkoj do 10mil.

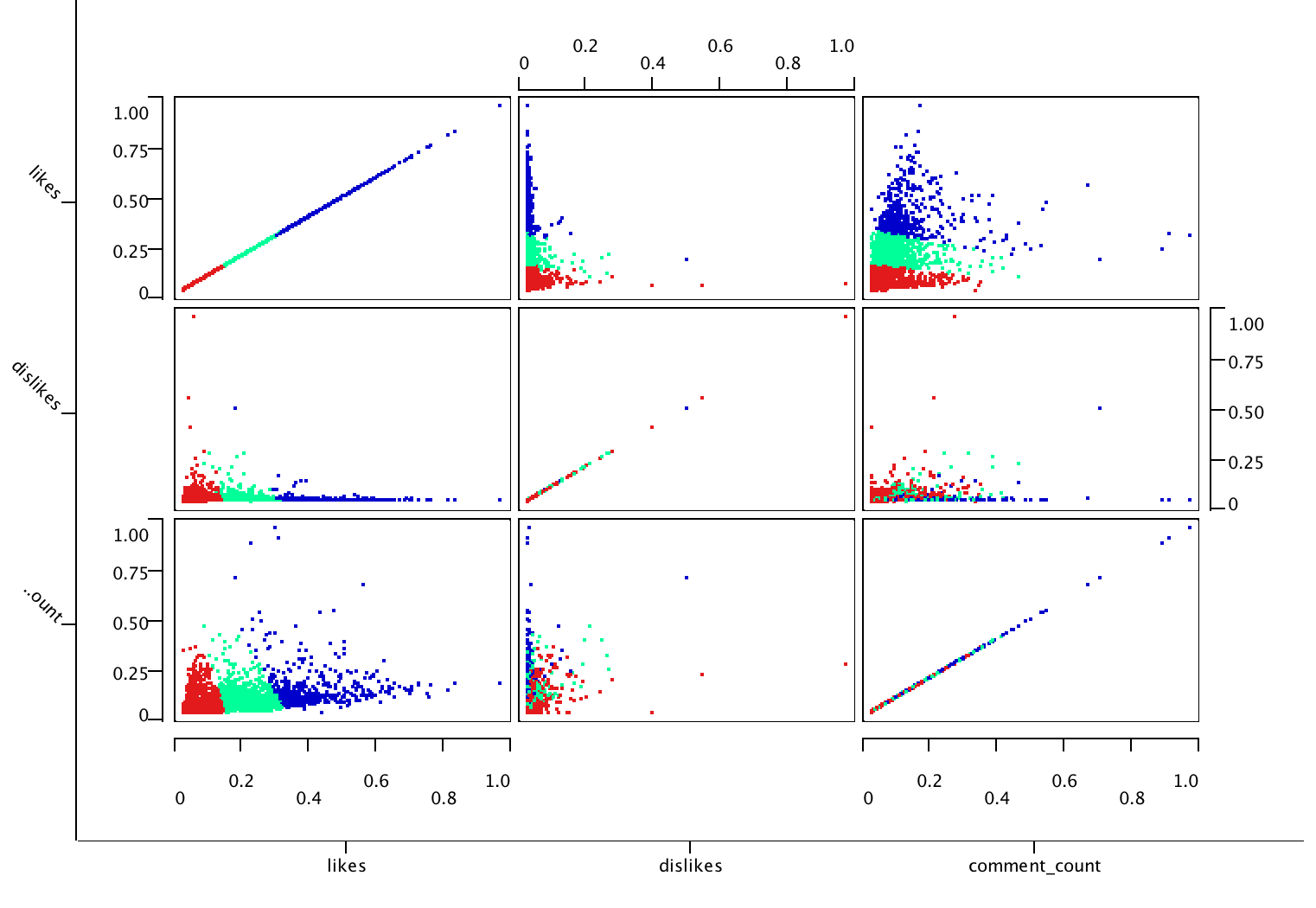


Broj pregleda po kategoriji u Francuskoj do 10mil.

# Klasterovanje

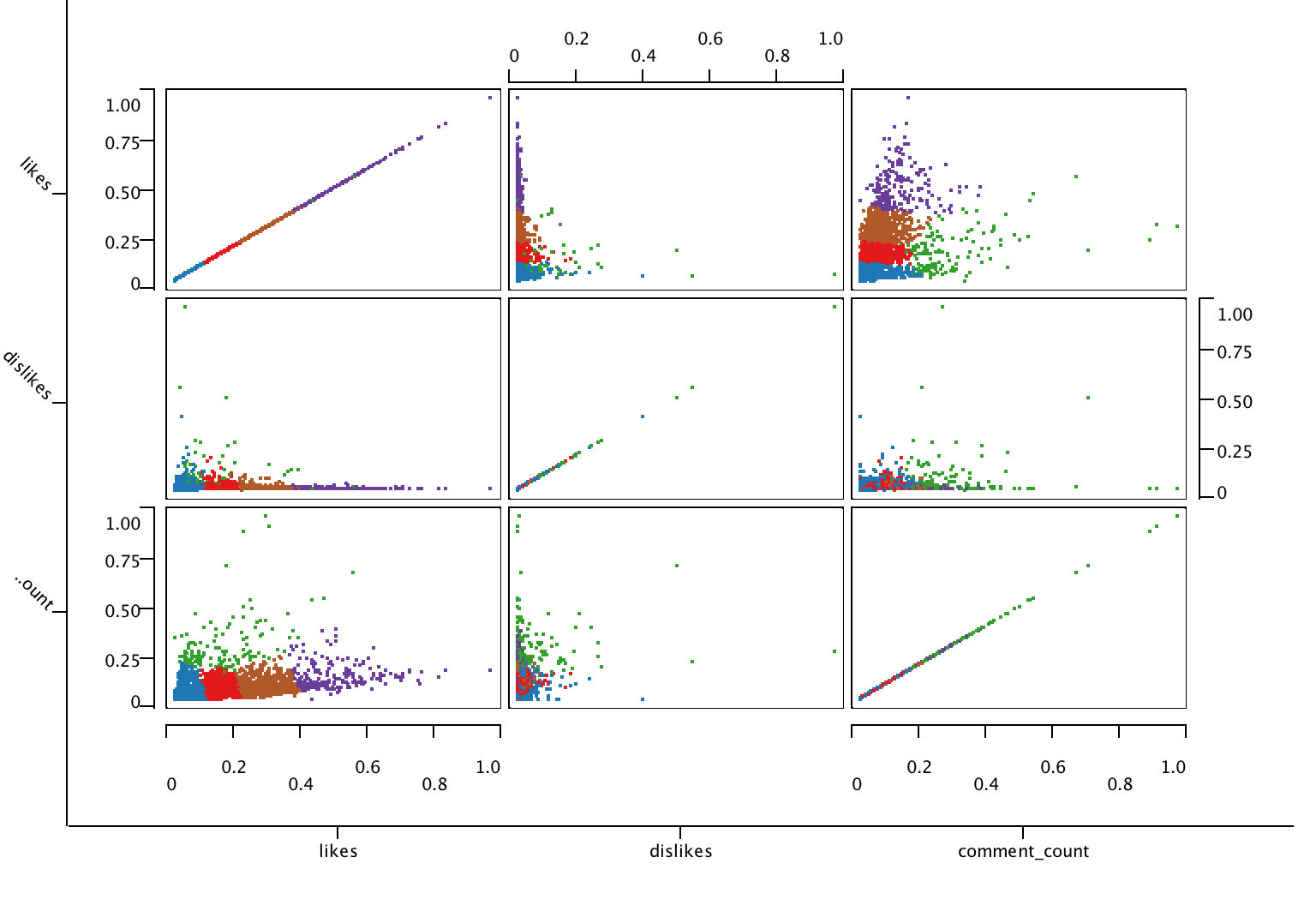
Rađeno je klasterovanje klipova K-Means algoritmom uzimajući u obzir broj komentara, pozivitnih i negativnih ocena klipa. Brojevi se uzimaju relativno na broj pregleda tako da ocenjujemo prosečnu reakciju gledalaca klipa.

Ako tražimo podelu u 3 klastera vidimo da algoritam deli klipove skoro isključivo po broju lajkova, što nam nije preterano zanimljivo.



K-Means klasterovanje u 3 klastera

Klasterovanje postaje zanimljivije pri podeli na 5 klastera. Vidimo da algoritam i dalje prvenstveno klasterjue po broju lajkova ali sada kreira i jedan klaster za klipove sa natprosečnim brojem komentara. U početku smo očekivali da algoritam odvoji klipove sa značajnim brojem negativnih ocena ali nije iz prostog razloga što je takvih klipova malo.

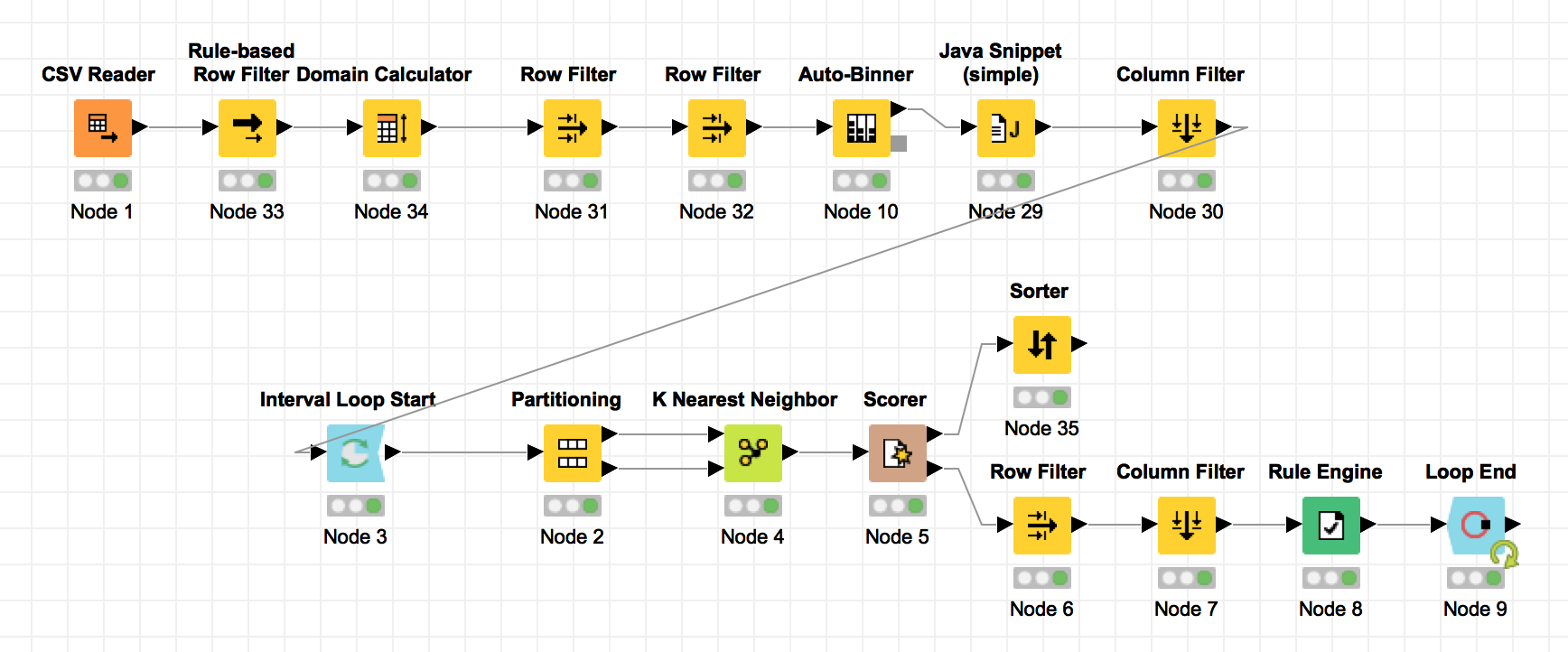


K-Means klasterovanje u 5 klastera

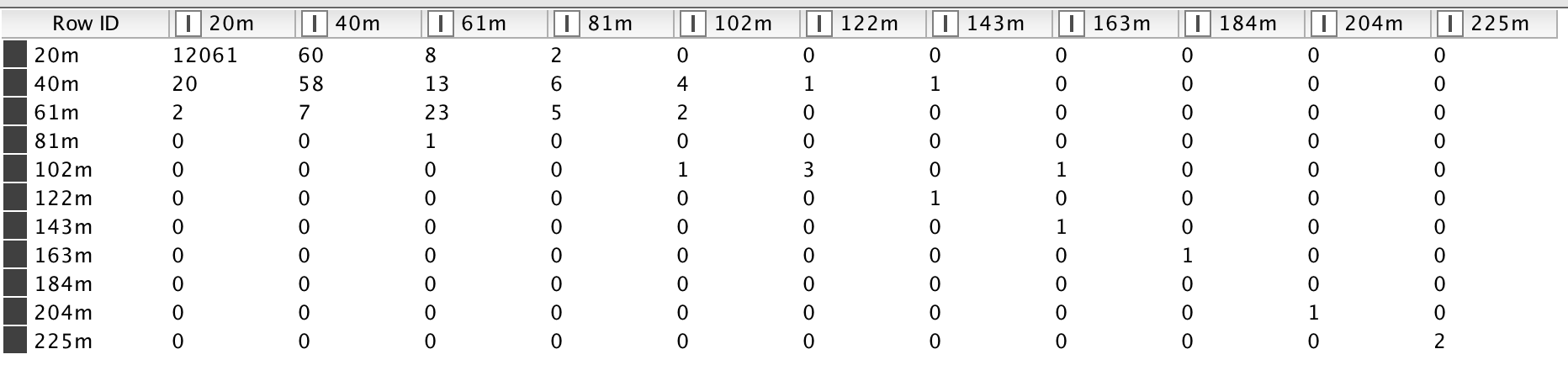
# Klasifikacija

Rađena je klasifikacija klipova po broju komentara i ocena. Cilj je bio da se pretpostavi okviran broj pregleda po tim podacima.

## KNN algoritam

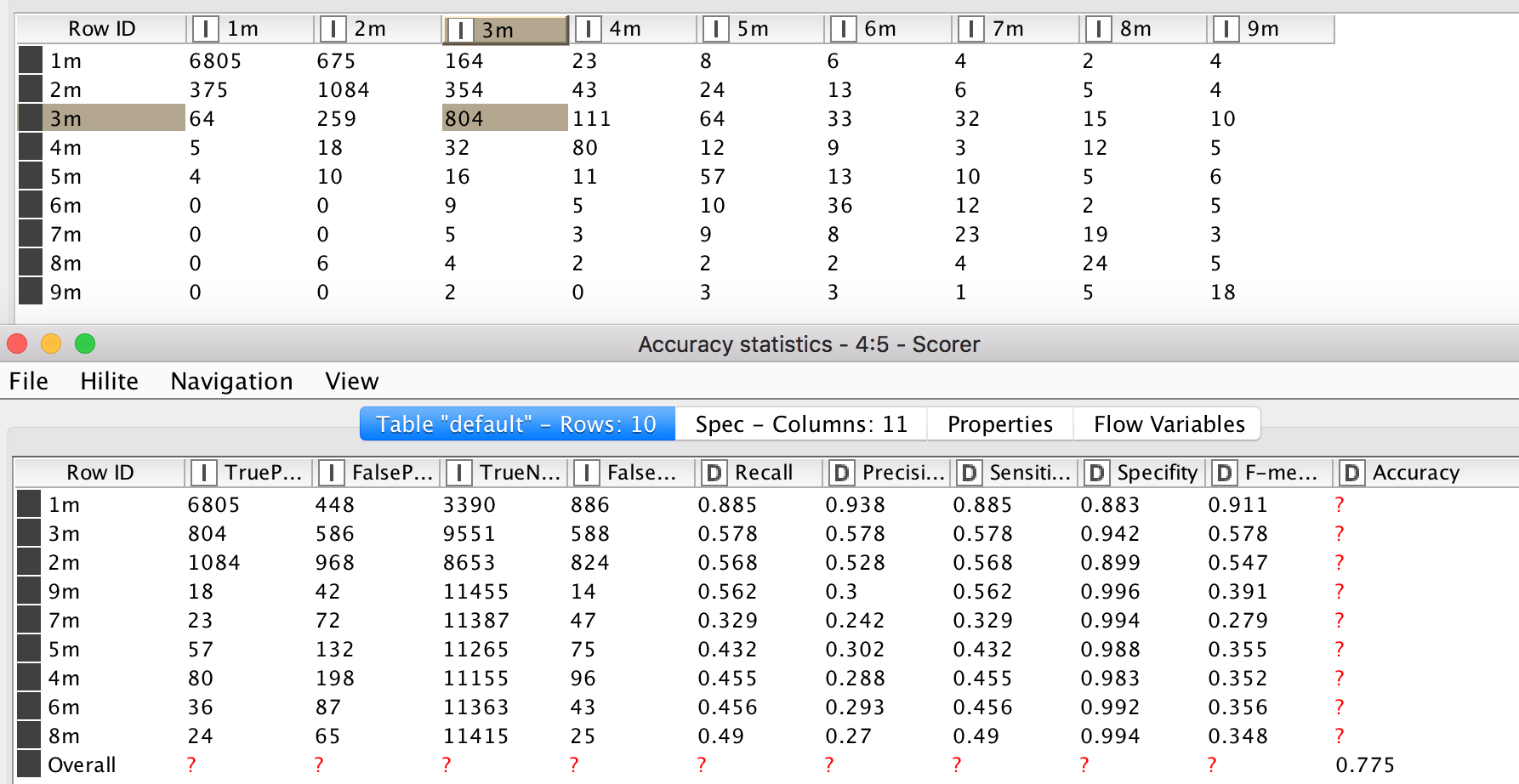


Prvo smo podelili klipove u 10 kategorija po broju pregleda i istrenirali KNN klasifikator. Primetili smo odlične rezultate od 99% tačnosti. Međutim pri analizi rezultata primetili smo da veliki broj klipova pripada kategoriji sa do 20 miliona pregleda. Dakle ne postoji raznolikost po podacima i klasifikator skoro uvek zaključuje da je klip sa do 20 miliona pregleda. Takođe toliko neprecizna informacija da klip ima ispod 20 miliona pregleda nam nije korisna.



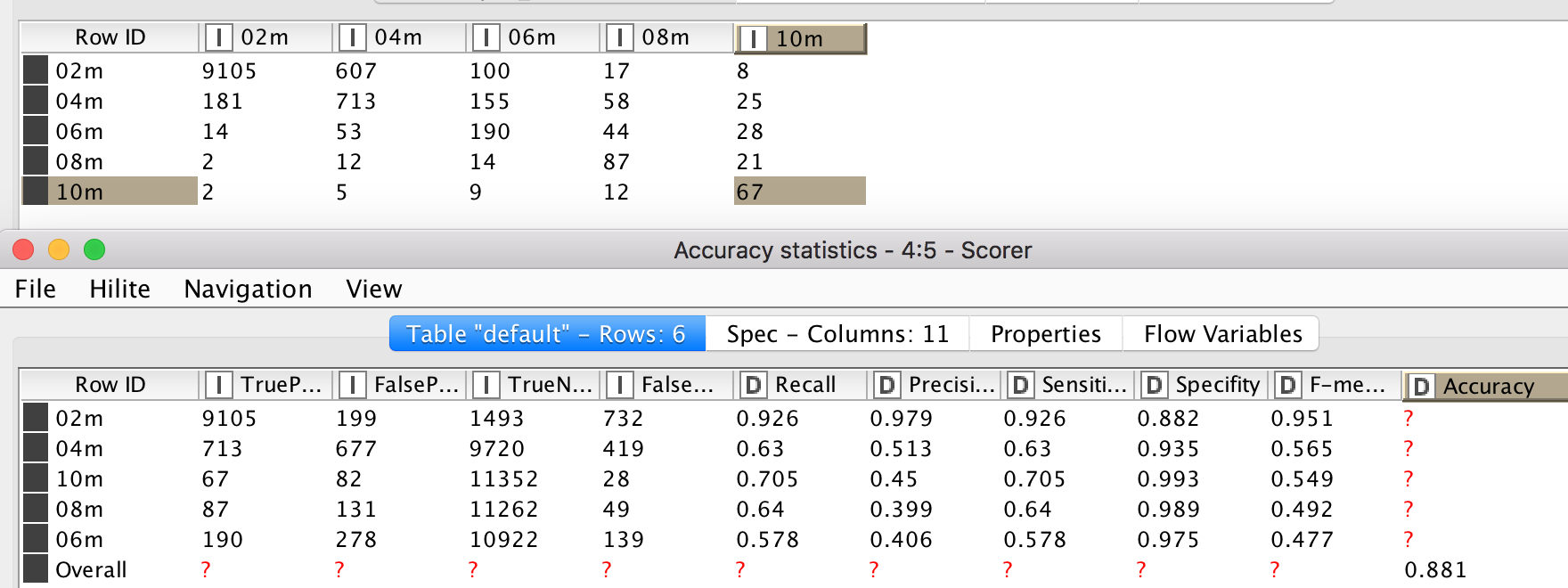
Matrica konfuzije za KNN sa svim klipovima

Sledeći korak je bio da se ograničimo na klipove sa do 10 miliona pregleda. Odradili smo istu analizu i odmah videli smanjenje preciznosti na 78% (zavisno od K). Ipak, primećujemo da matrica konfuzije ima najveće vrednosti blizu dijagonale, odnosno greška je često za jednu ili dve kategorije (npr. klipovi sa 2 miliona pregleda se klasifikuju kao da imaju 1 ili 3 miliona).



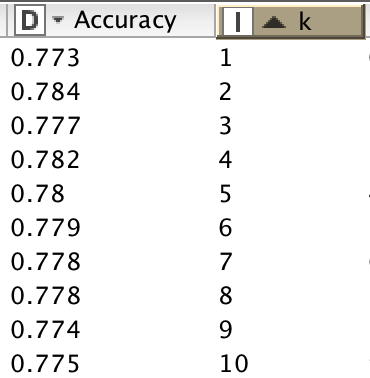
Matrica konfuzije i analiza preciznosti za KNN sa klipovima do 10 miliona pregleda

Podela podataka na 5 kategorija umesto 10 je davala bolje rezultate klasifikacije - u proseku 88% preciznosti (zavisno od K), ali ipak takva klasifikacija nije previše korisna jer pogađa opseg pregleda sa preciznošću od 2 miliona.

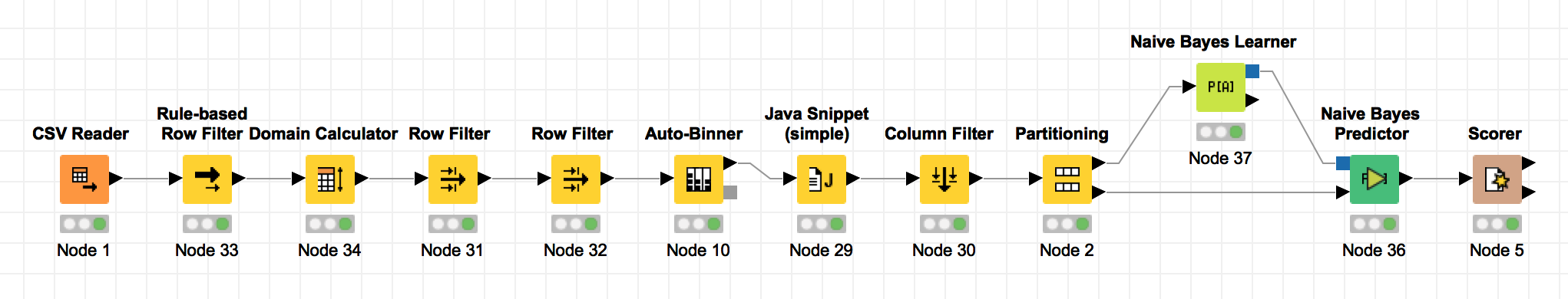


Matrica konfuzije i analiza preciznosti za KNN sa klipovima do 10 miliona pregleda u 5 kategorija

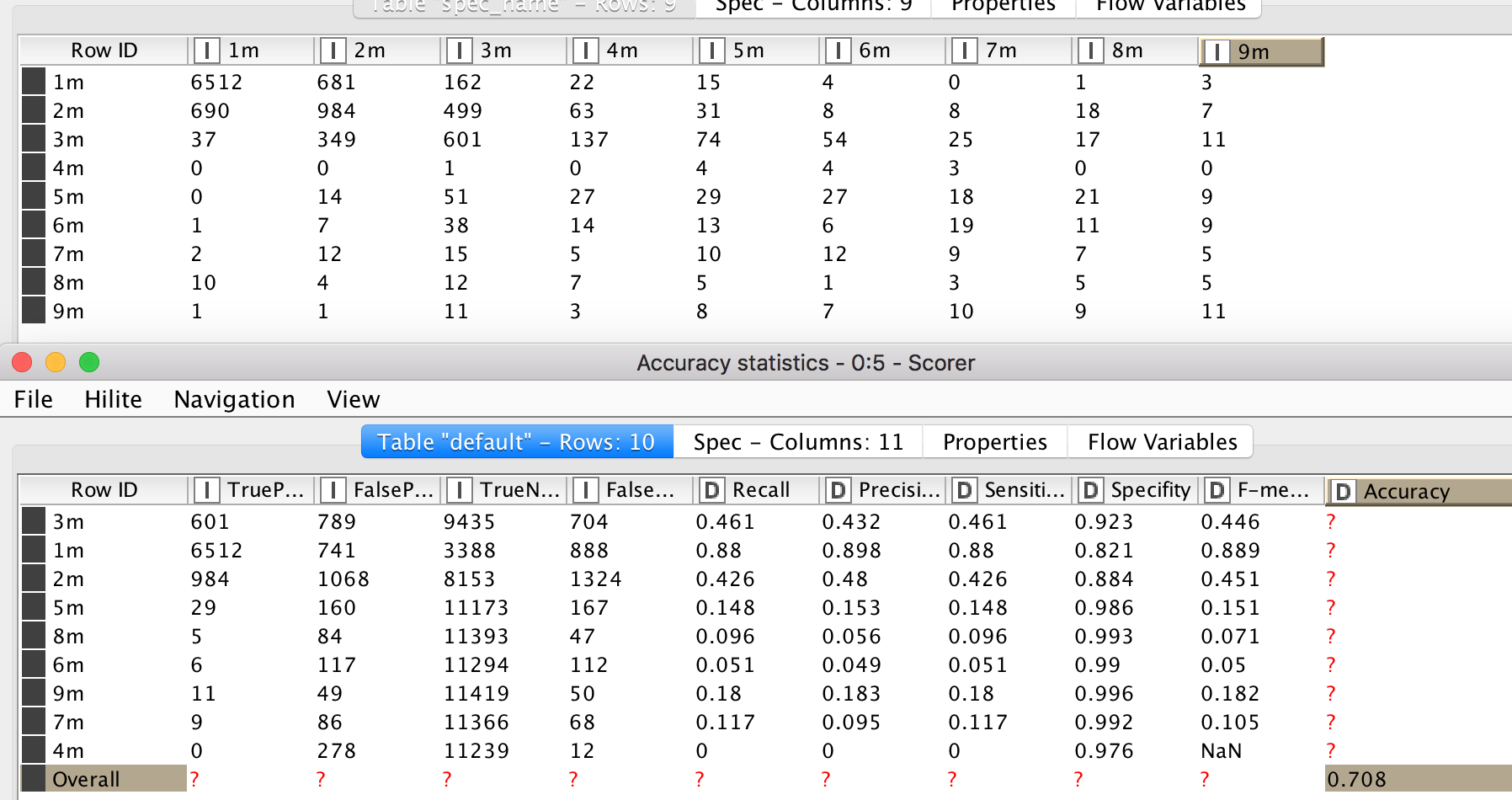
Isprobane su vrednosti za K od 1 do 10. Nije bilo većih razlika između preciznosti - do 2%. Najbolje se pokazali klasifikatori koji su uzimali u obzir 2 i 4 najbliža elementa.



## Naivni Bajesov algoritam



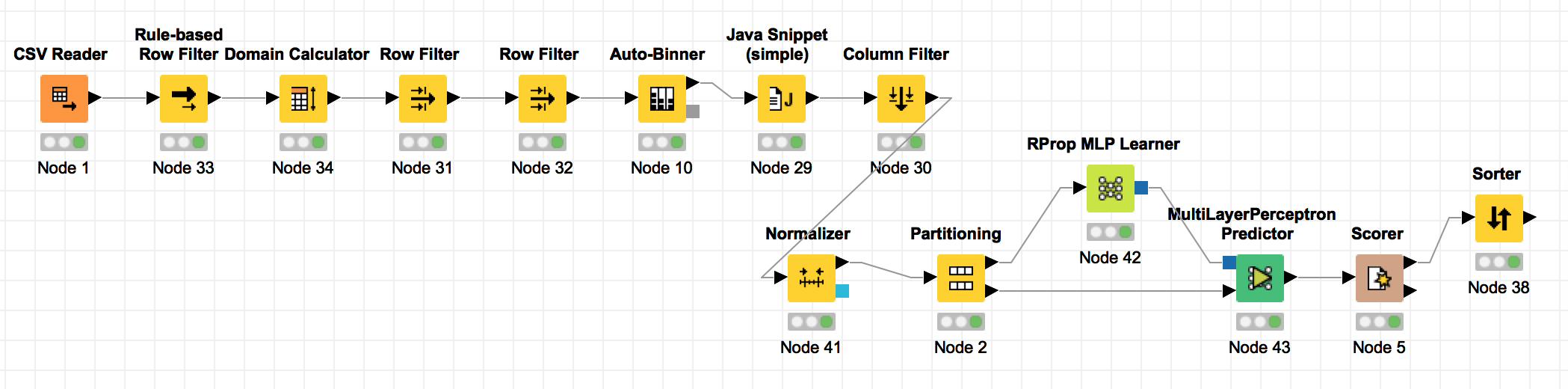
Primenjena je ista podela na 10 kategorija kao za KNN. Naivni Bajesov algoritam se ponašao značajno loše od KNN algoritma. Osim preciznosti od oko 70%, primetili smo da nijedan klip u kategoriji od 4 miliona nije tačno klasifikovan. Kada smo detaljnije analizirali statistiku klasifikatora primetili smo preprilagođavanje kategoriji klipova sa manje od 1 milion pregleda, a kako ti klipovi čine oko 60% trening skupa dobijamo relativno visoku prosečnu preciznost. Ovaj klasifikator dakle nije koristan.



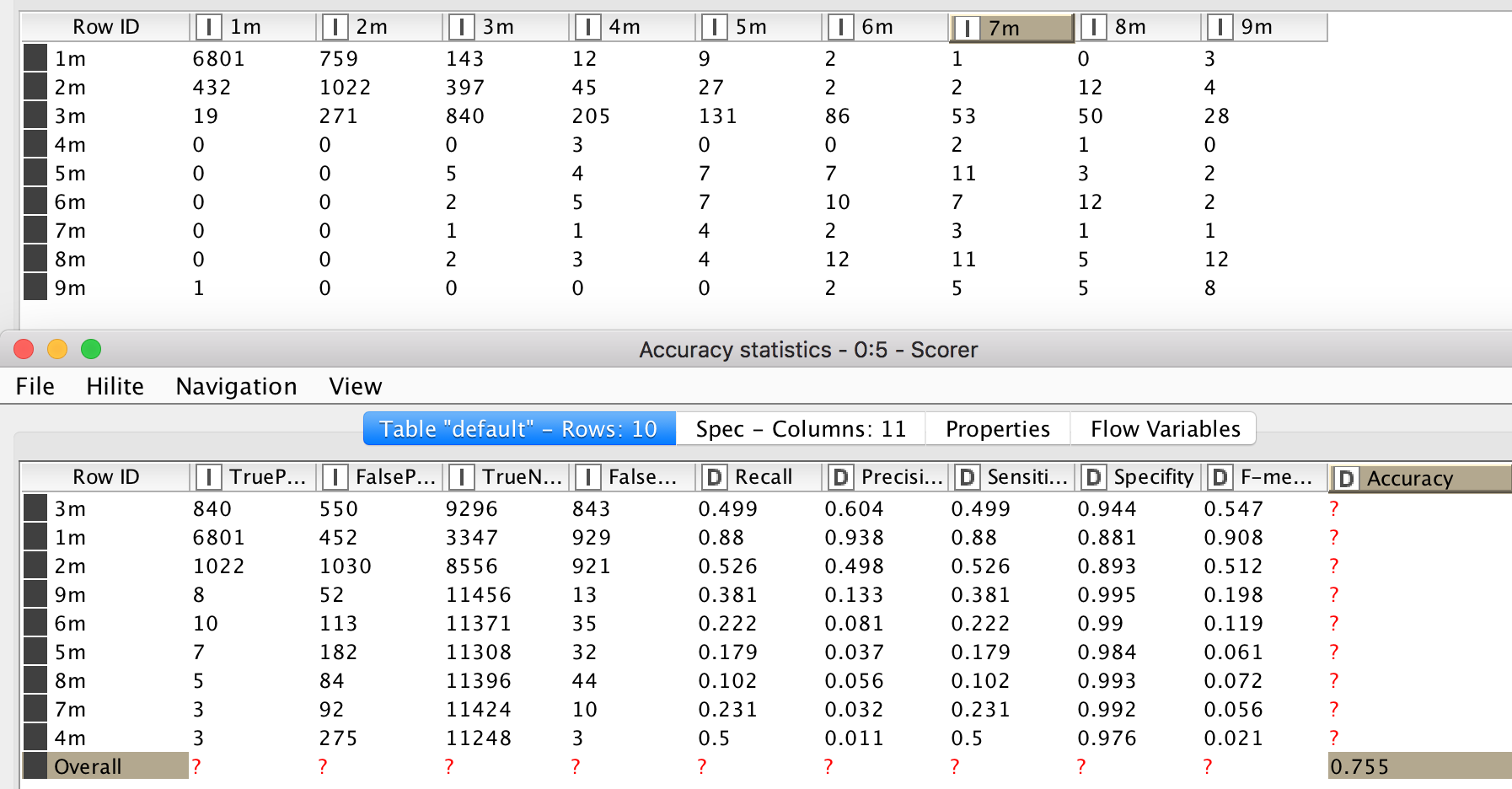
Matrica konfuzije i analiza preciznosti za Naivni Bajesov algoritam

sa klipovima do 10 miliona pregleda u 10 kategorija

## Mašinsko učenje RProp algoritmom



Uz ustanovljenu podelu na 10 kategorija RProp algoritam daje naizgled solidne rezultate od 75% preciznosti. Ipak opet smo primetili preprilagođavanje kategoriji klipova sa do 1 milion pregleda pa ni ovaj klasifikator nije koristan. Menjanje maksimalnog broja iteracija nije značajno pomoglo preciznosti.



Matrica konfuzije i analiza preciznosti za RPRop algoritam

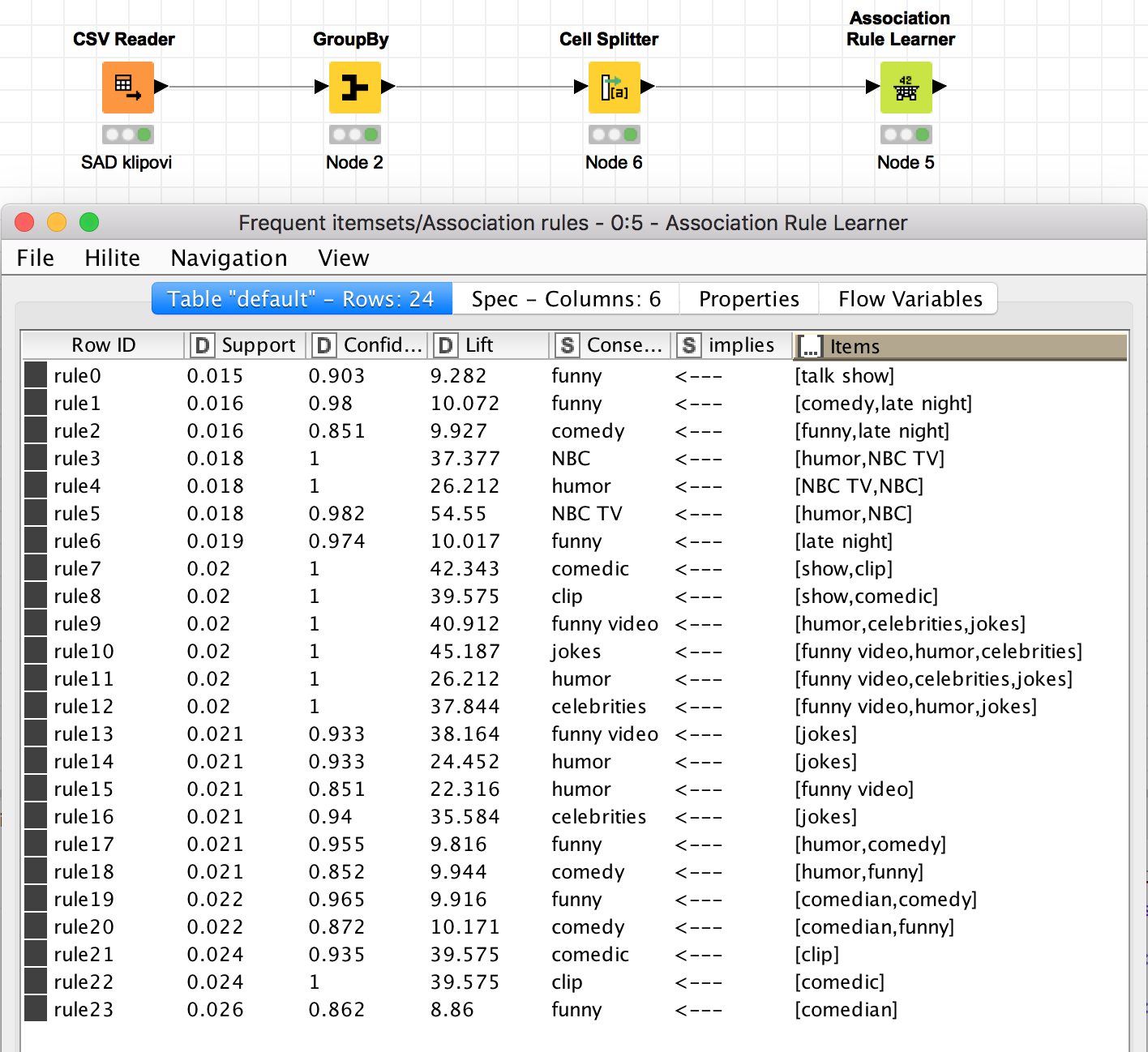
sa klipovima do 10 miliona pregleda u 10 kategorija

# 

# Pravila pridruživanja

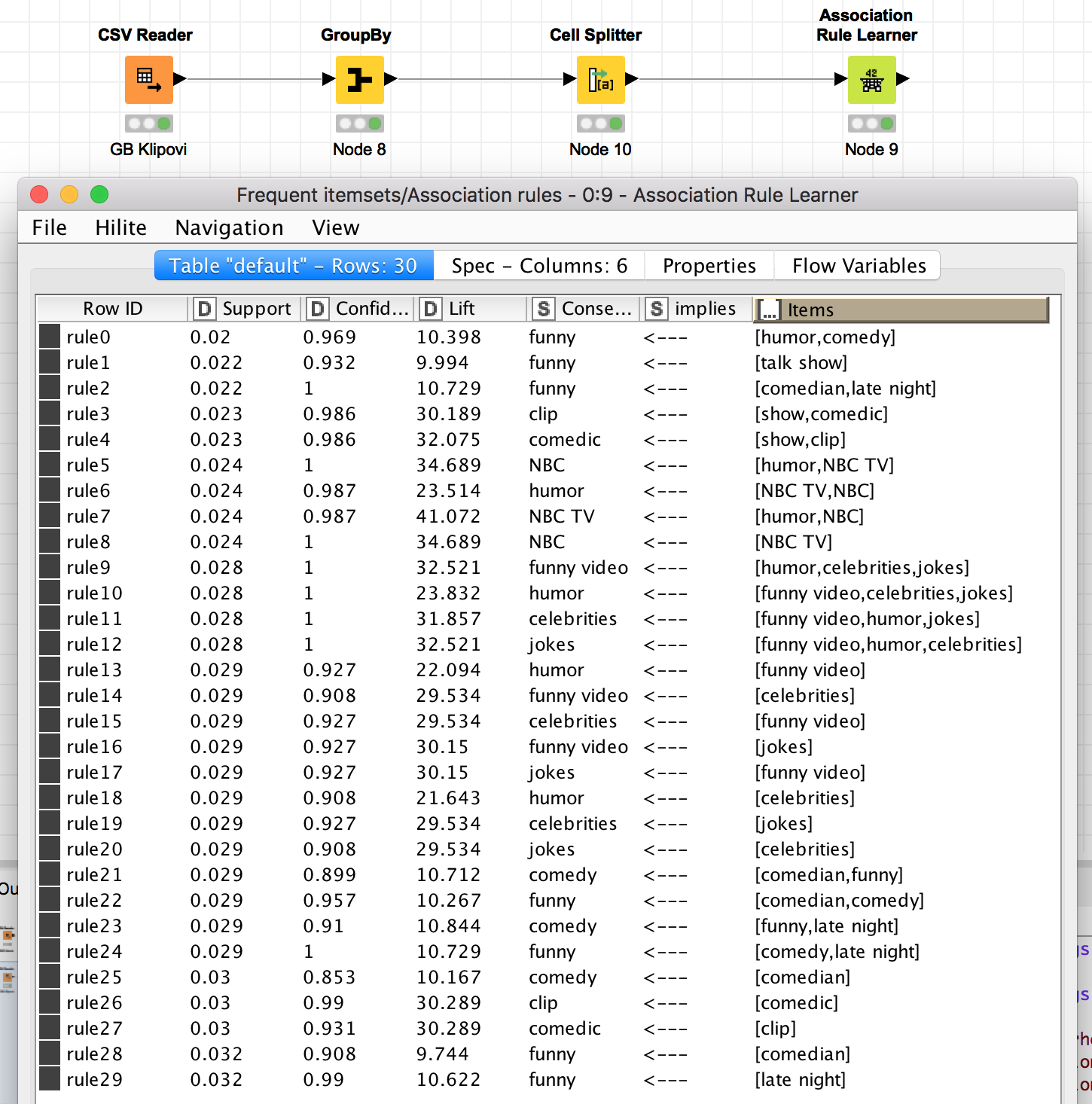
Svaki video klip ima listu oznaka (kolona “tags”) kojima je opisan od strane autora. Analiza je rađena za države u engleskom govornom području. Analizom je pronađena pronađena jaka veza među oznakama “humor”, “funny video”, “jokes” i sličnih oznaka koje označavaju smešne video klipove.

Pokušali smo da analiziramo sa što manjom podrškom a velikom sigurnošću kako bismo pronašli i ređe primećena ali sigurna pravila (za klipove sa tematikom koja je manje popularna). Analiza SAD skupa je rađena sa podrškom od 1.5% i sigurnošću od 0.85. Analizu nismo uspeli da odradimo sa manjom podrškom jer zbog velikog broja elemenata program ostajao bez memorije. Pri analizi podataka za Veliku Britaniju podršku smo morali da podignemo na 2%.

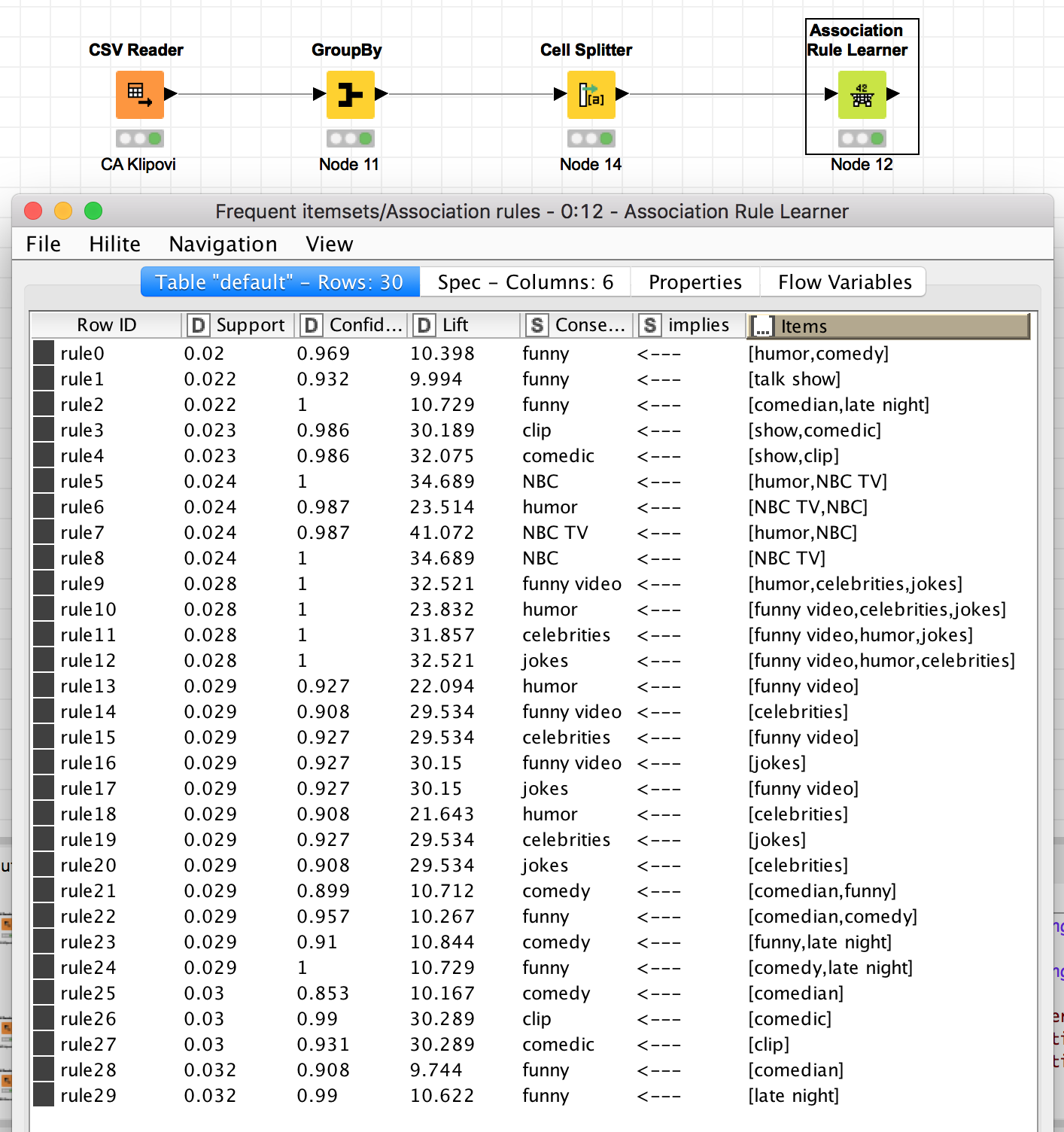


Pravila pridruživanja za klipove iz SAD skupa

(Prikazuje i proces analize u programu Knime)



Pravila pridruživanja za klipove iz GB skupa



Pravila pridruživanja za klipove iz CA skupa