

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	3
Priprema podataka	4
YouTube algoritam i očekivanja od analize	4
Analiza podataka	5
Analiza popularnih klipova	5
Analiza dužine popularnosti	5
Vreme do početka popularnosti	7
Vreme objave i popularnost istog dana	8
Broj klipova sa naslovom u velikim slovima	8
Analiza kategorija	9
Najpopularnije kategorije	9
Dužina popularnosti po kategoriji	10
Poređenje zemalja	11
Procenat istih popularnih klipova između zemalja	11
Poređenje broja pregleda, ocena i komentara	14
Popularnost kategorija po zemljama	16
Klasterovanje	21
Klasifikacija	23
KNN algoritam	23
Naivni Bajesov algoritam	25
Mašinsko učenje RProp algoritmom	26
Pravila pridruživanja	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](#) 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.

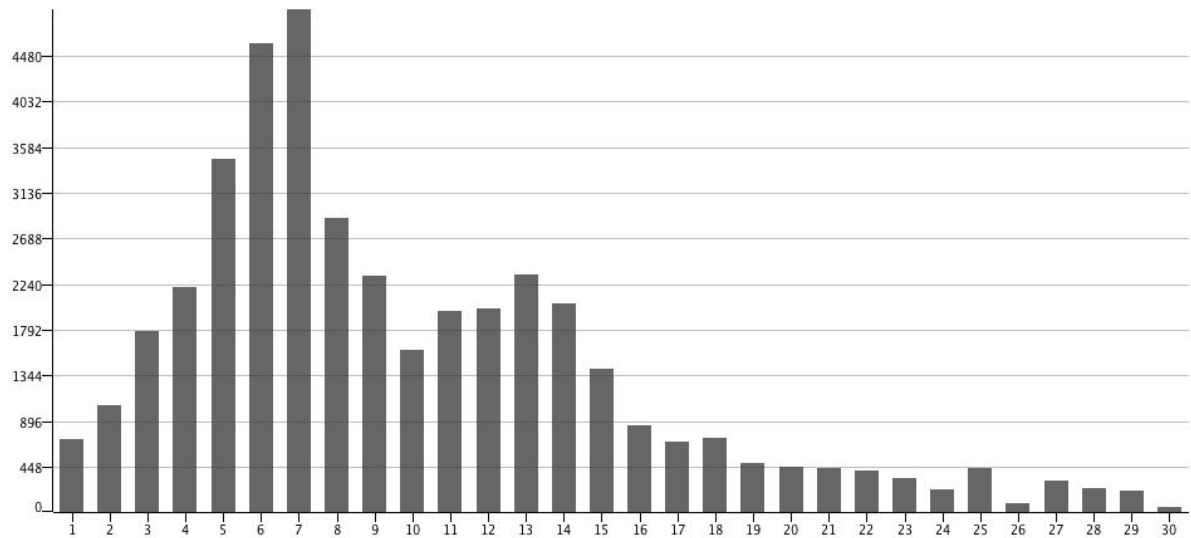
Google takođe tvrdi da algoritam uzima u obzir tematiku klipa, veličinu ciljne publike, kvalitet i slično, ali [brojni članci](#) koji govore o nesposobnosti platforme da suzbije netačne “lažne vesti” i sličan nekvalitetan sadržaj govore suprotno, pa ove tvrdnje (koje je ujedno teško izmeriti i proveriti) nećemo uzeti u obzir u analizi.

Analiza podataka

Analiza popularnih klipova

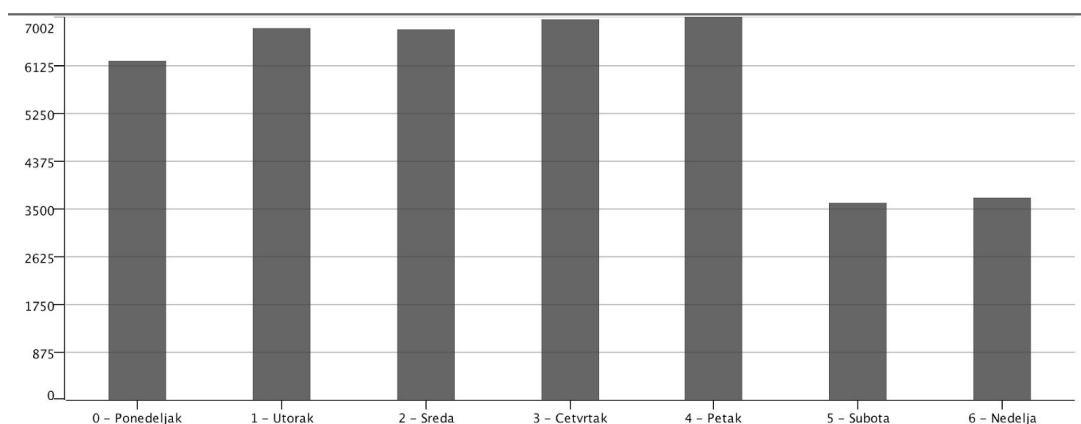
Analiza dužine popularnosti

Za početak analiziramo dužinu popularnosti klipova u danima:



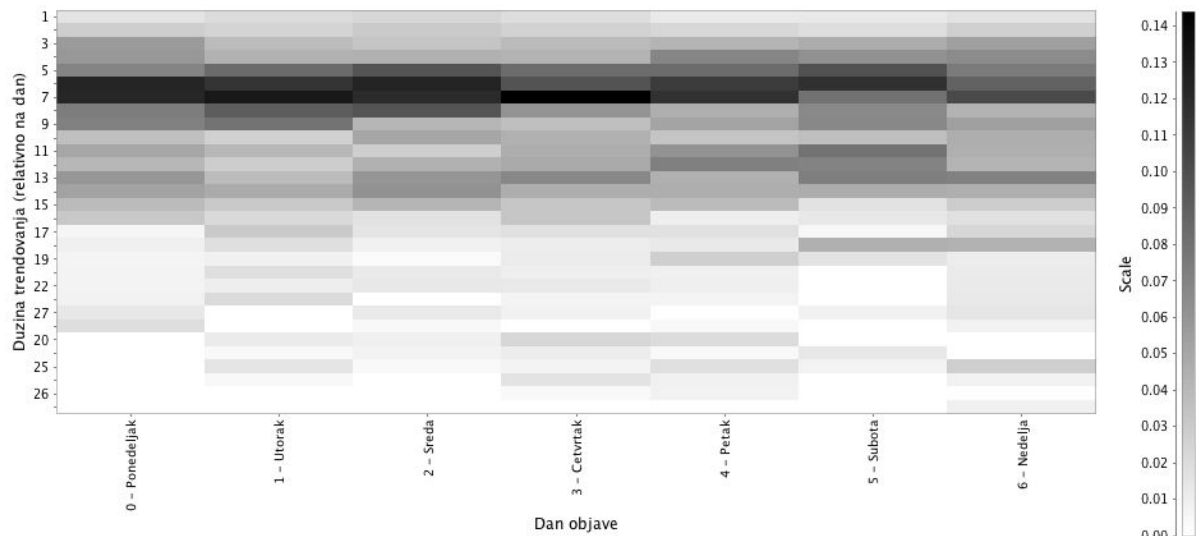
Odmah primećujemo da nijedan klip nije popularan duže od 30 dana. Može se pretpostaviti da algoritam to ne dozvoljava. Međutim pri analizi drugih zemalja vidimo da neki klipovi u Velikoj Britaniji su popularni i do 38 dana, dok su u Nemačkoj, Francuskoj i Kanadi klipovi popularni mnogo kraće. Nameće 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.



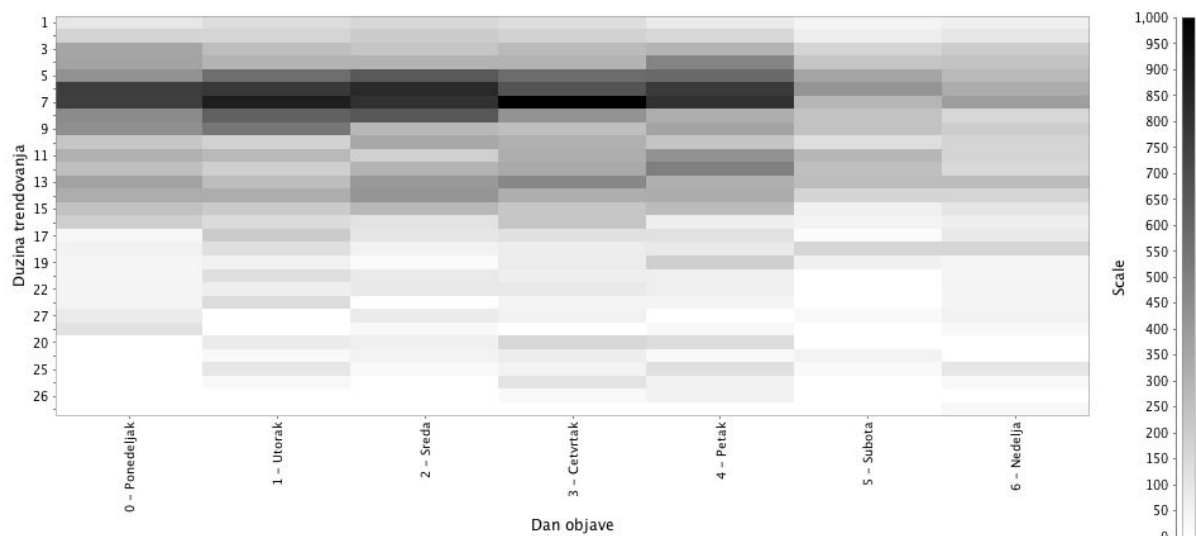
Primećujemo da je broj klipova objavljenih subotom i nedeljom skoro duplo manji od bilo kog drugog dana. Ovo potvrđuje preporuku [mnogih članaka](#) koji tvrde da su subota i nedelja najnepovoljniji dani za objavu.

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

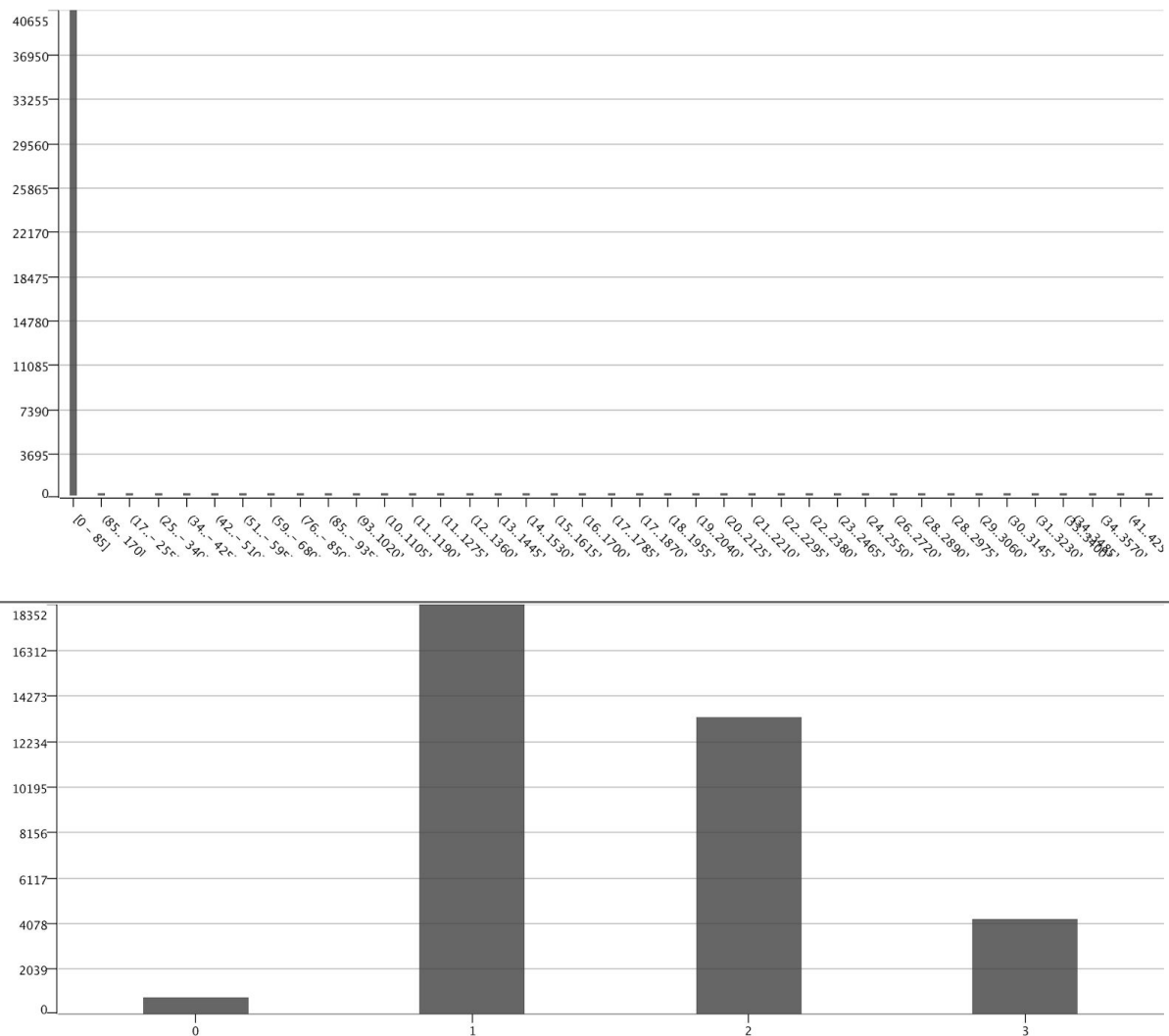


(Grafik je normalizovan tako da prikazuje gustnu relativnu na broj klipova objavljenih tog dana)

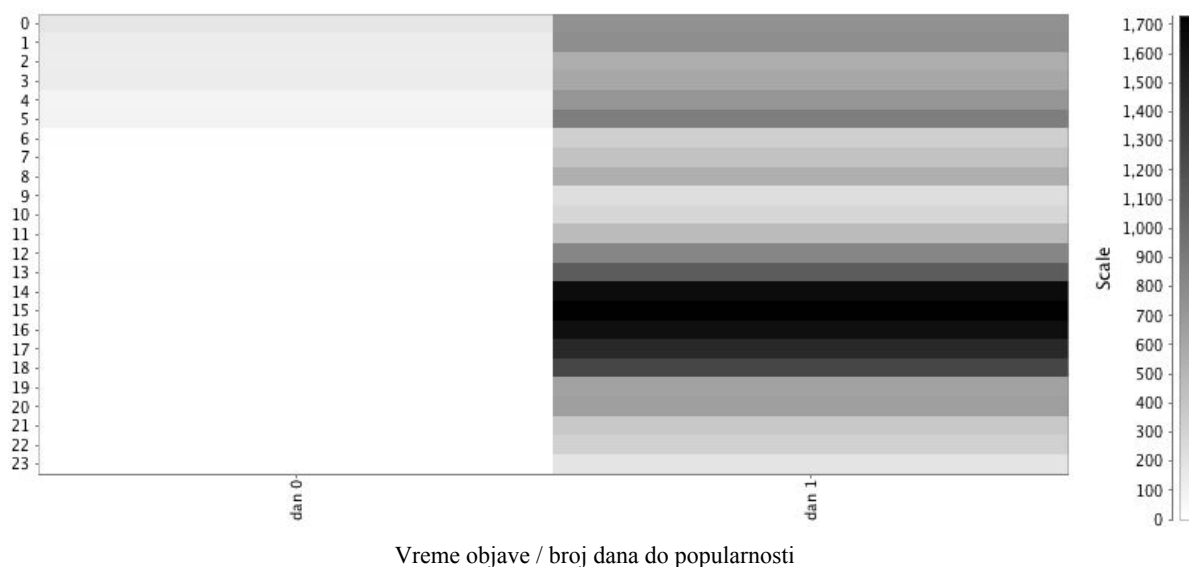
Ovo ipak nije značajan podatak kada uzmemo u obzir apsolutan broj klipova:



Vreme do početka popularnosti

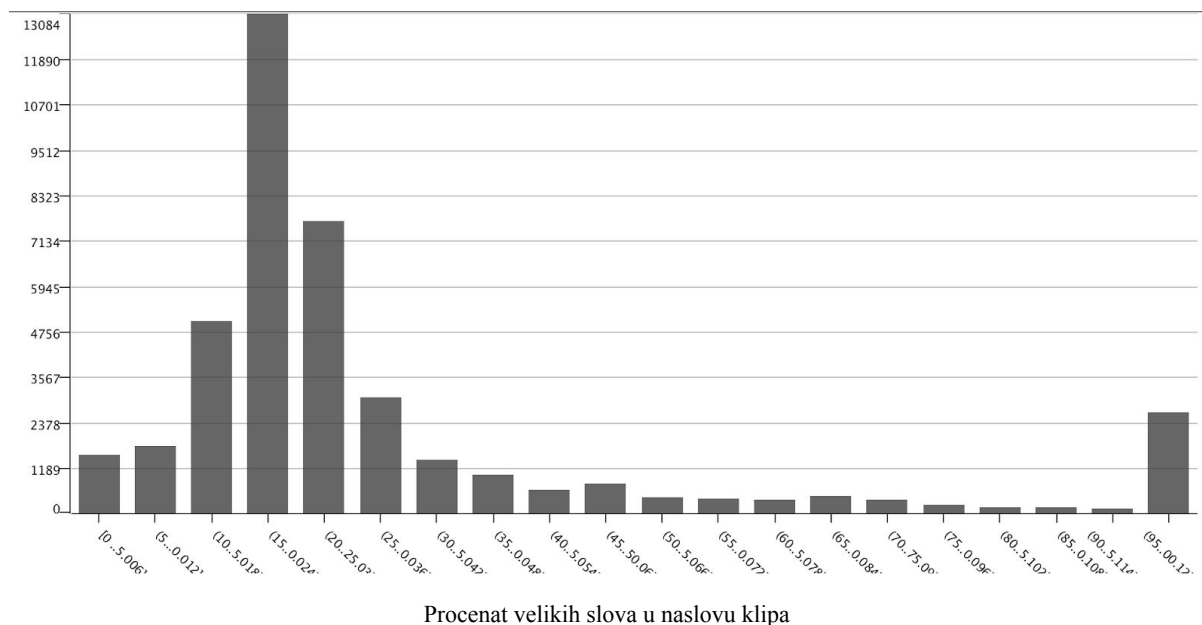


Vreme objave i popularnost istog dana



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

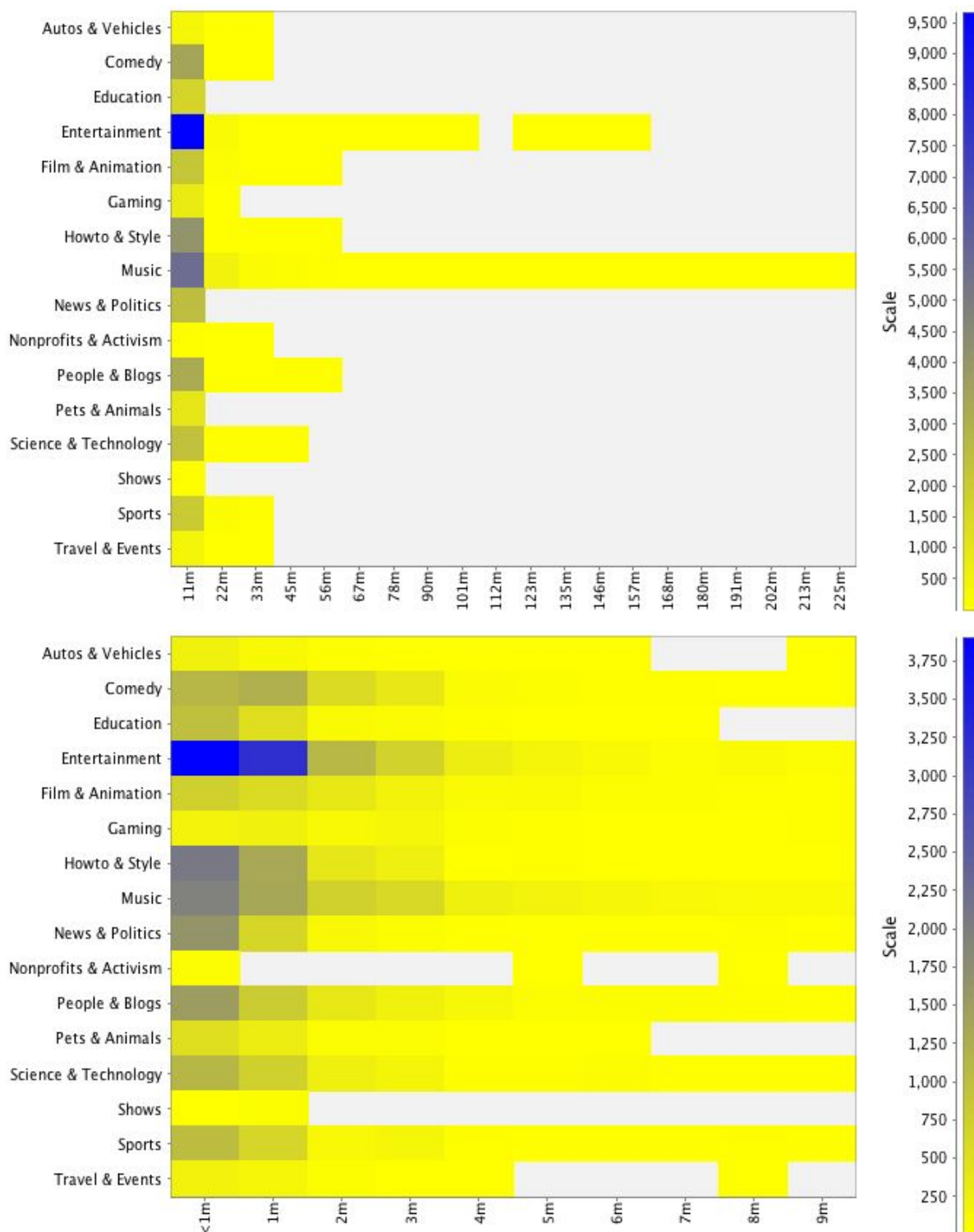
Broj klipova sa naslovom u velikim slovima



6.4% klipova je imalo naslov sa isključivo velikim slovima.

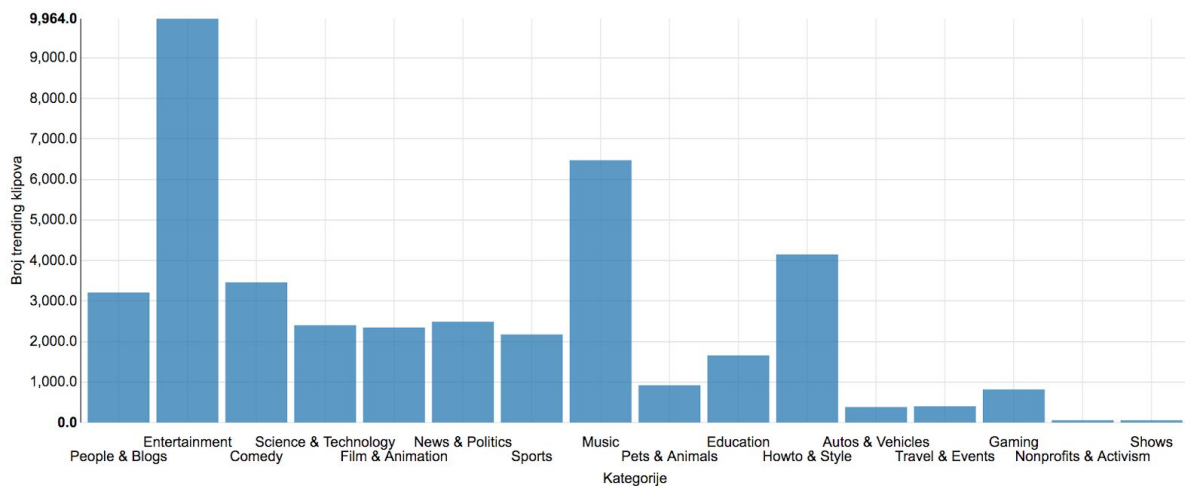
Analiza kategorija

Najpopularnije kategorije

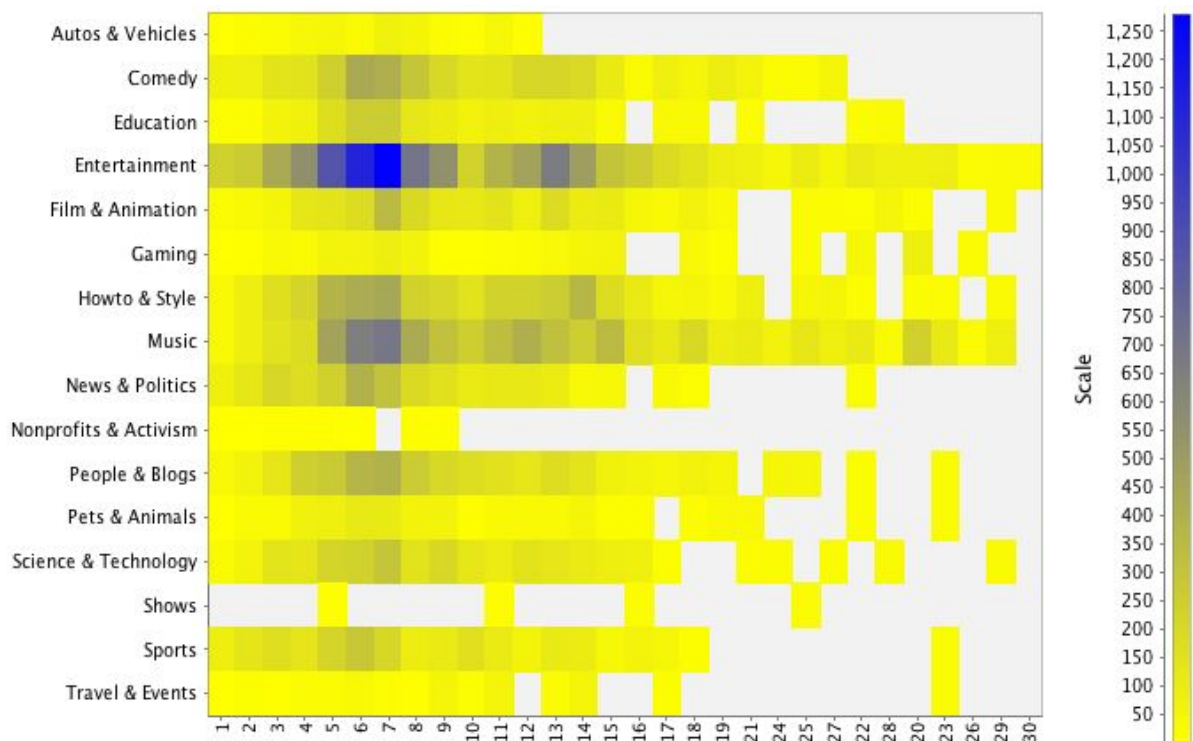


Kategorije po gustini broja pregleda - USA

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.

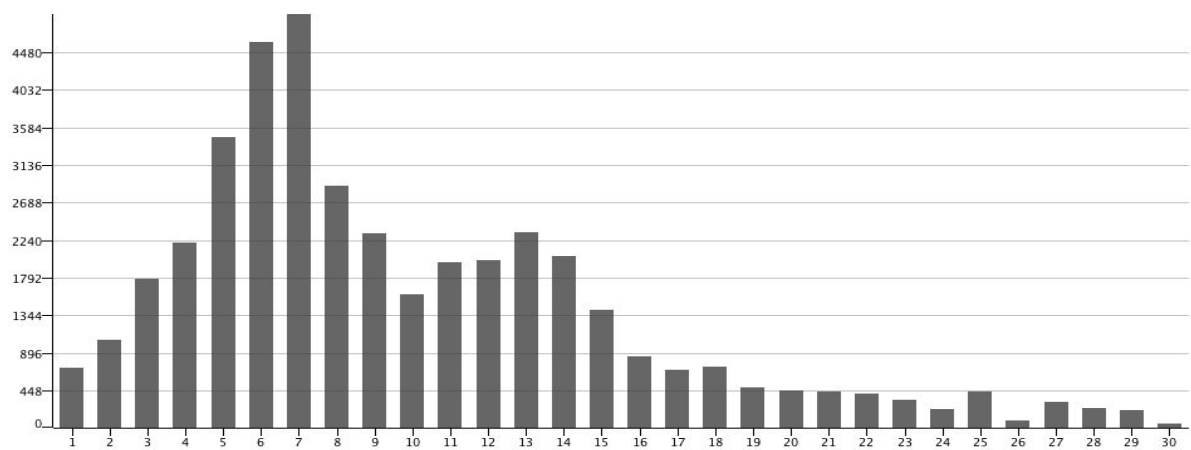
Poređenje zemalja

Procenat istih popularnih klipova između zemalja

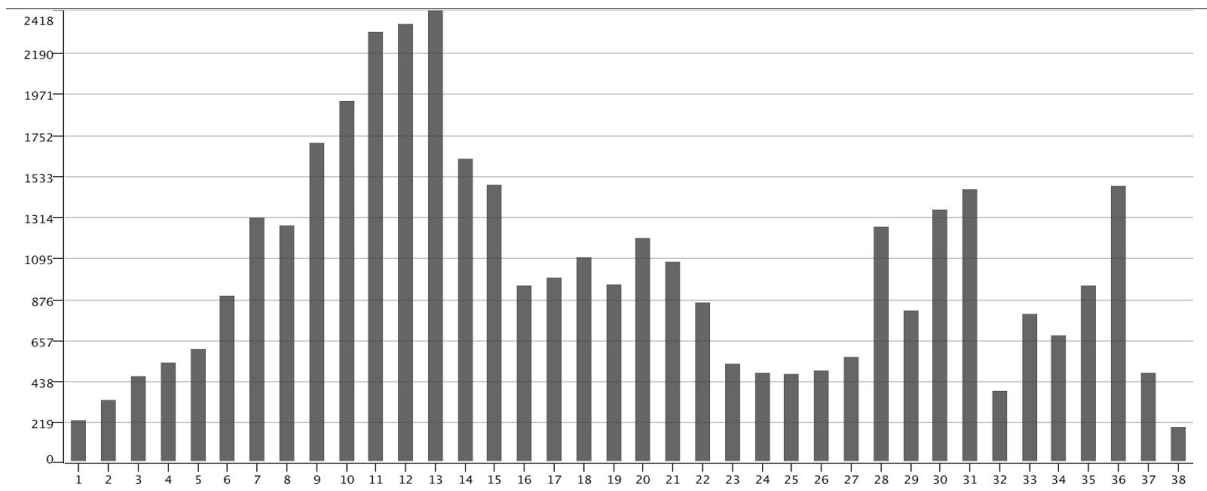
	CA	DE	FR	GB	US
CA		21%	17%	3%	9%
DE	17%		13%	1%	3%
FR	13%	13%		1%	2%
GB	29%	16%	13%		31%
US	35%	17%	10%	15%	

Procenat popularnih klipova u drugoj zemlji

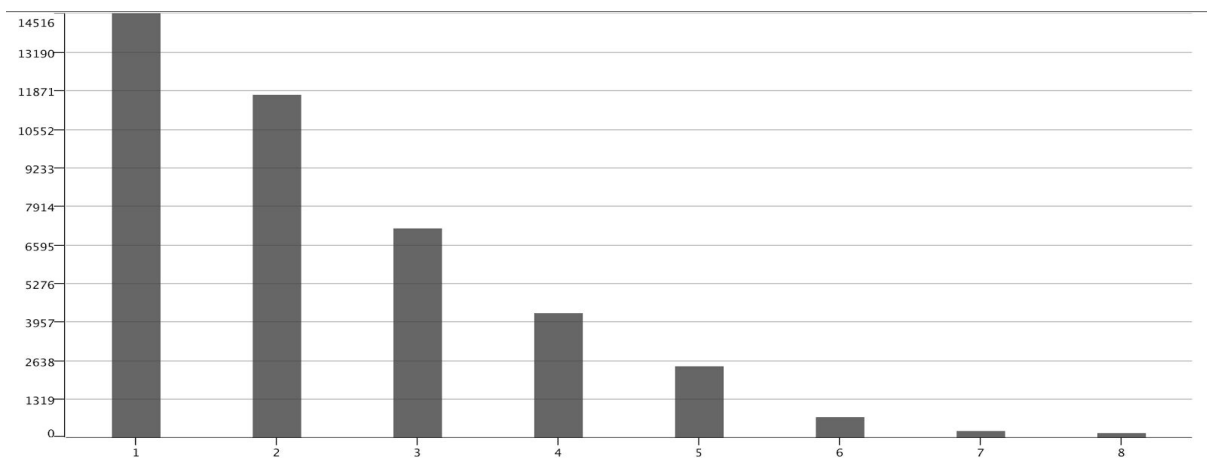
Najveću sličnost vidimo između SAD, Kanade i Velike Britanije. Ovo se donekle očekuje jer države u istom govornom području. Primećujemo takođe da procenat sličnosti nije simetričan, iz razloga što se dužina popularnosti drastično razlikuje između SAD, Velike Britanije, i ostalih država:



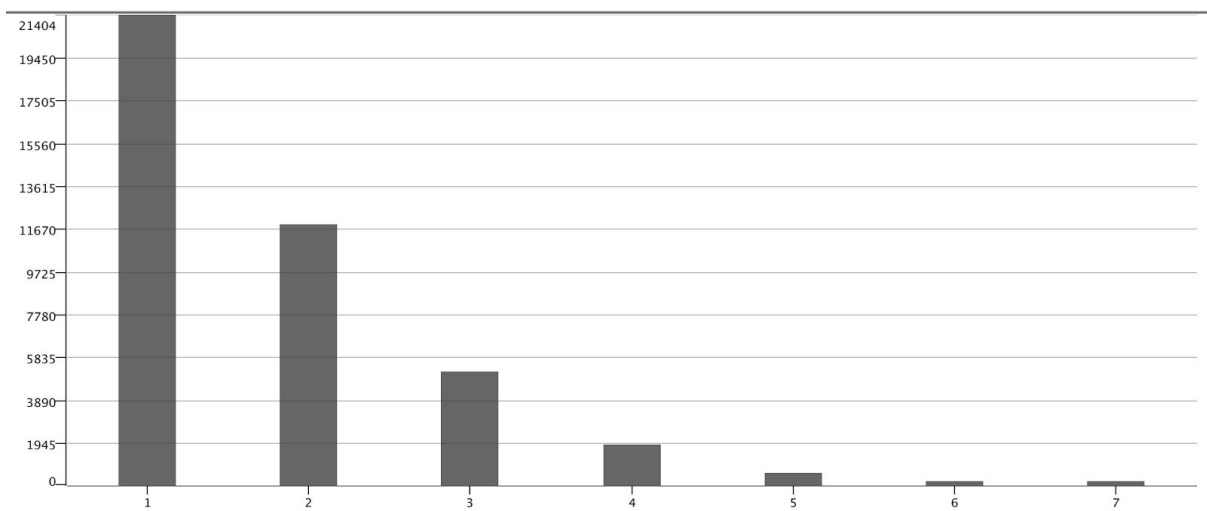
Dužina popularnosti u SAD



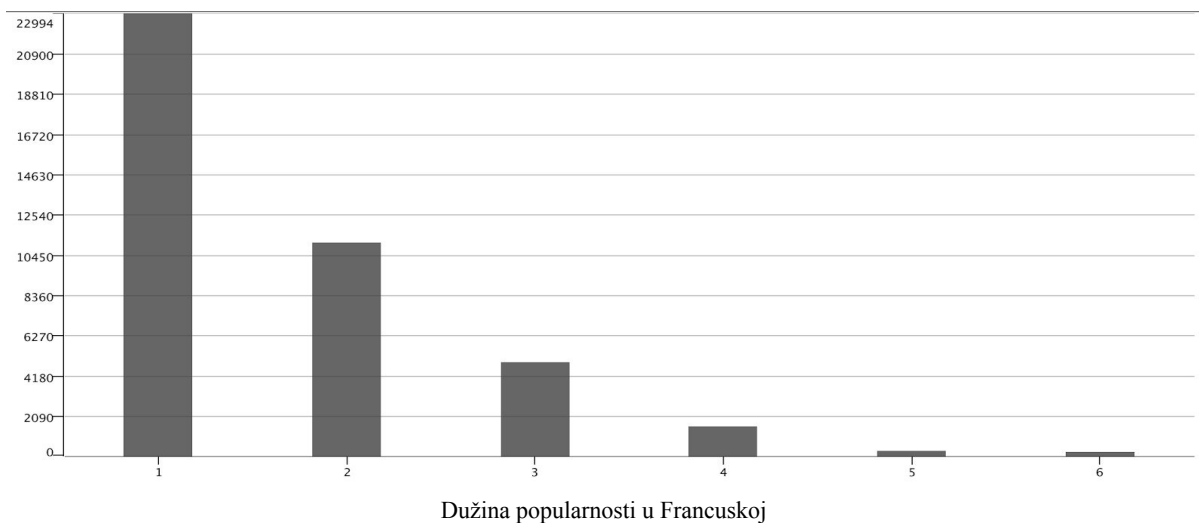
Dužina popularnosti u Velikoj Britaniji



Dužina popularnosti u Kanadi



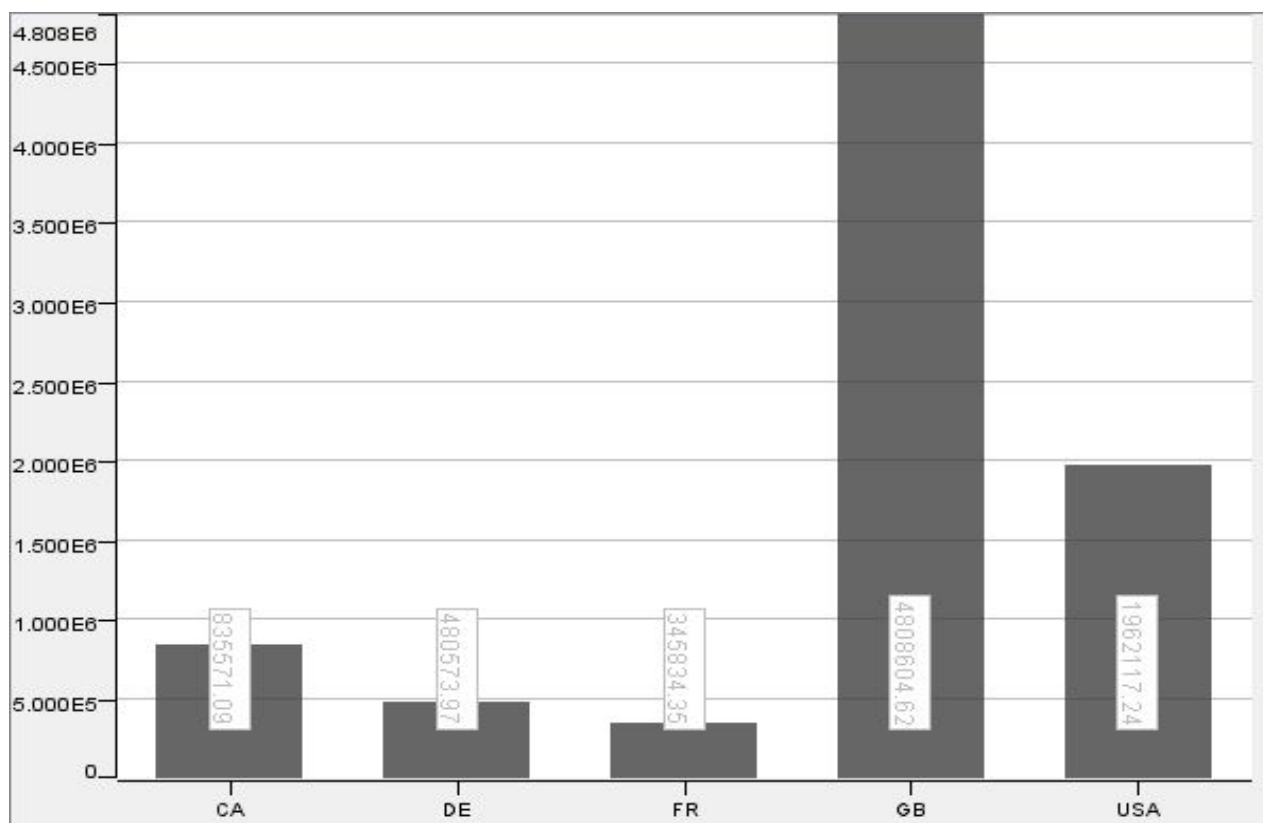
Dužina popularnosti u Nemačkoj



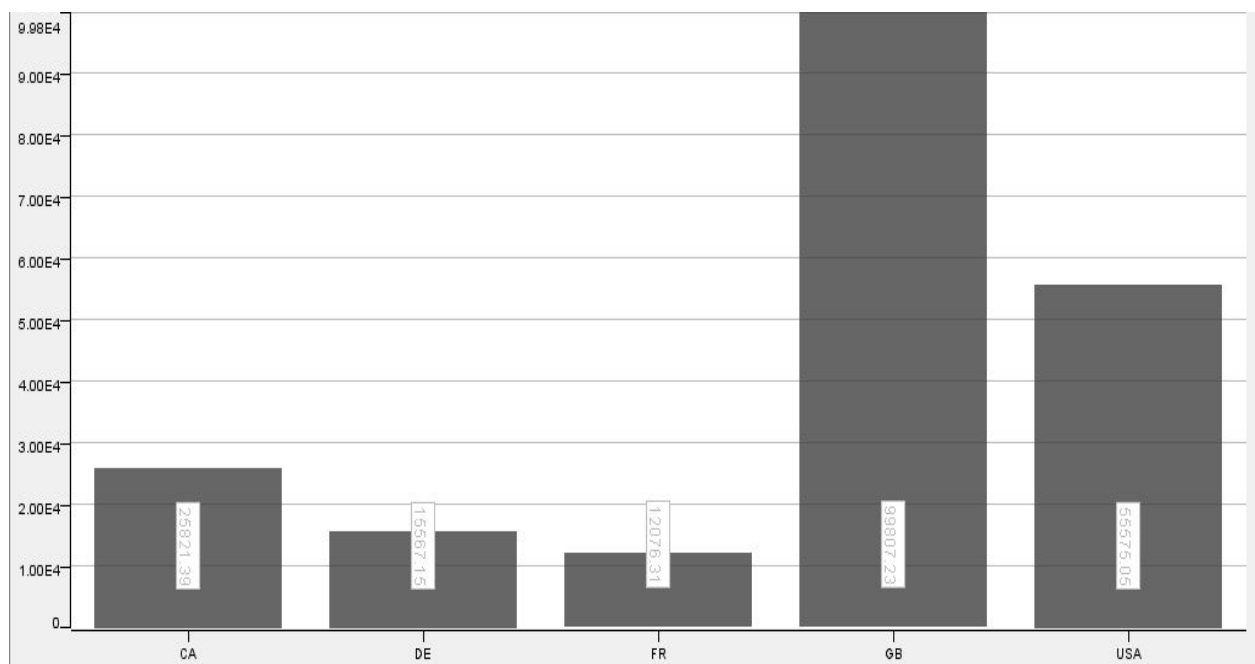
Primećuje se da su klipovi popularni mnogo duže u SAD i Velikoj Britaniji nego u ostalim državama. Iz ovog razloga broj jedinstvenih popularnih klipova je različit između država pa samim tim i asimetričnu tabelu sličnosti.

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

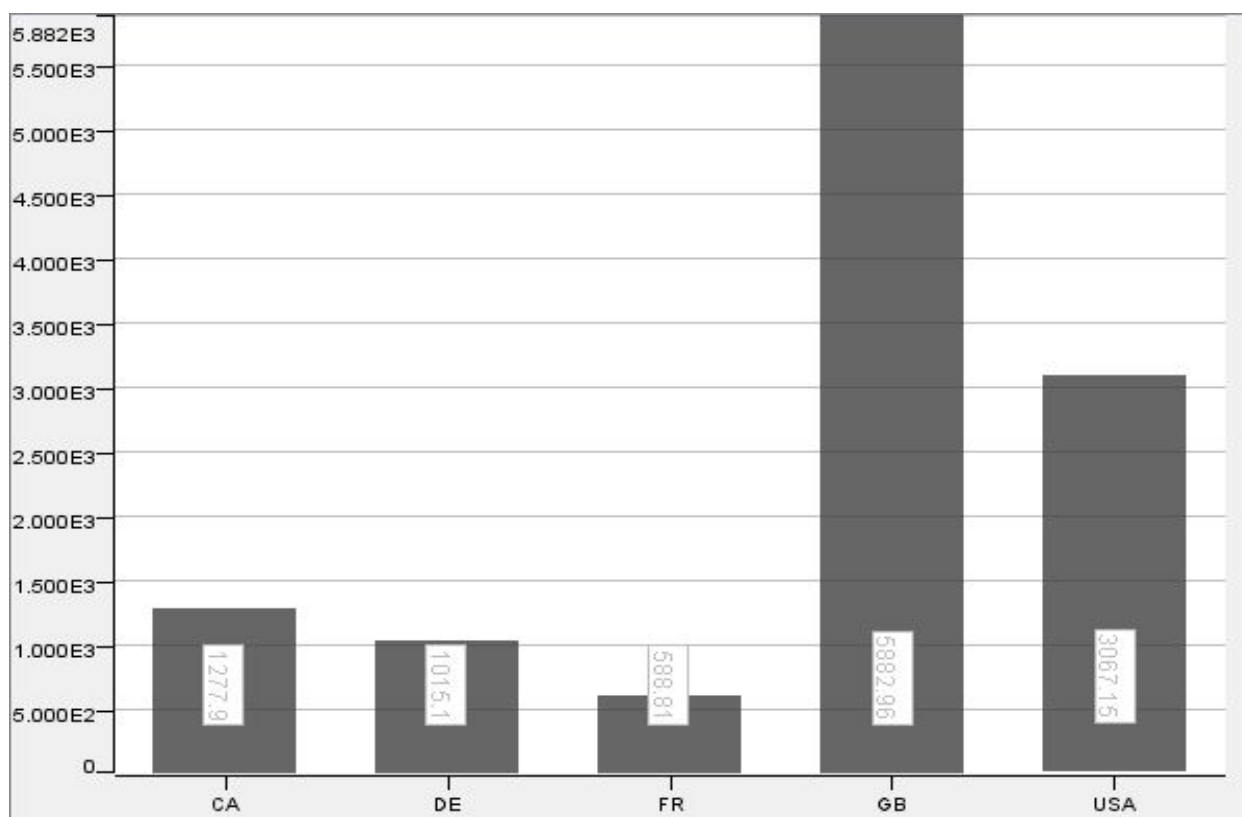
Poređenje broja pregleda, ocena i komentara



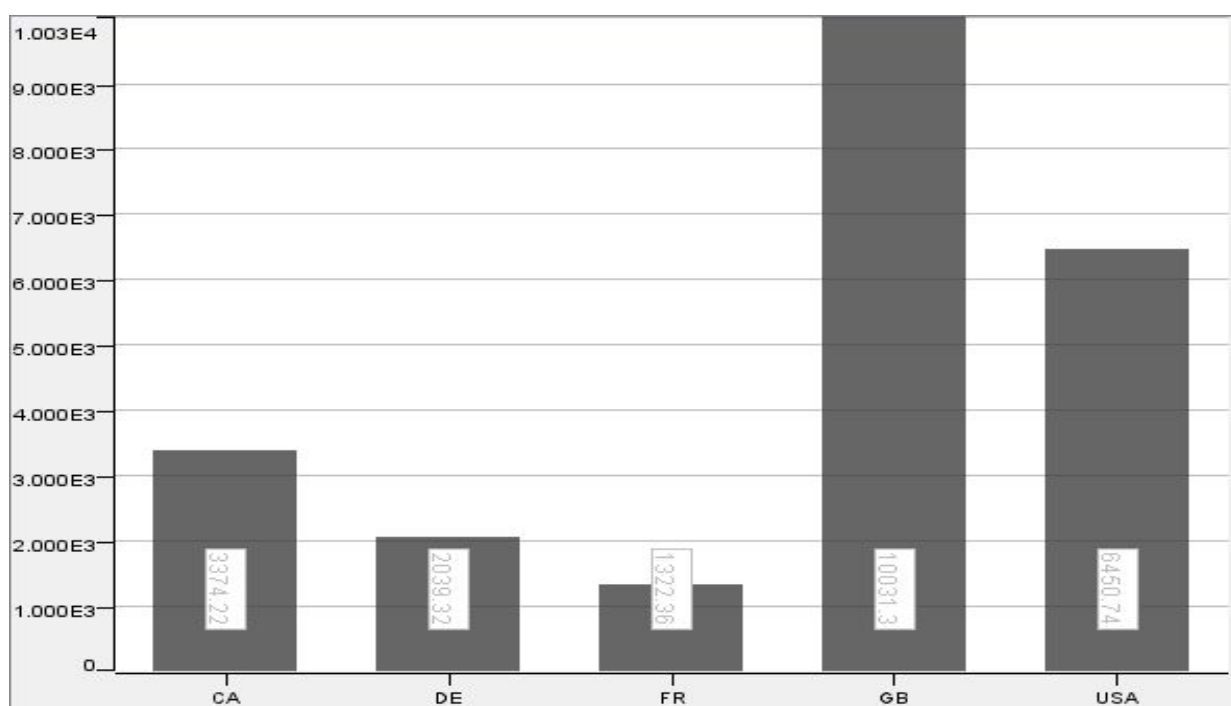
Prosečan broj pregleda klipa po državi



Prosečan broj pozitivnih ocena ("svidi mi se") na klipu po državi



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

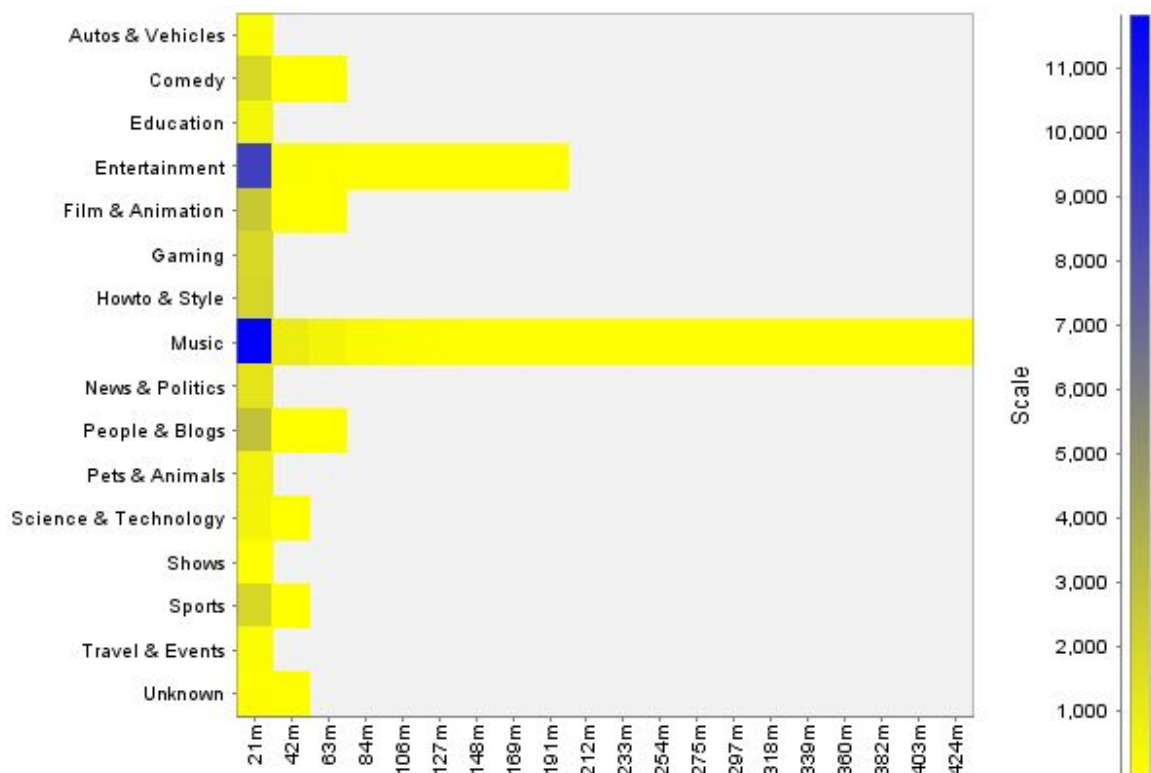
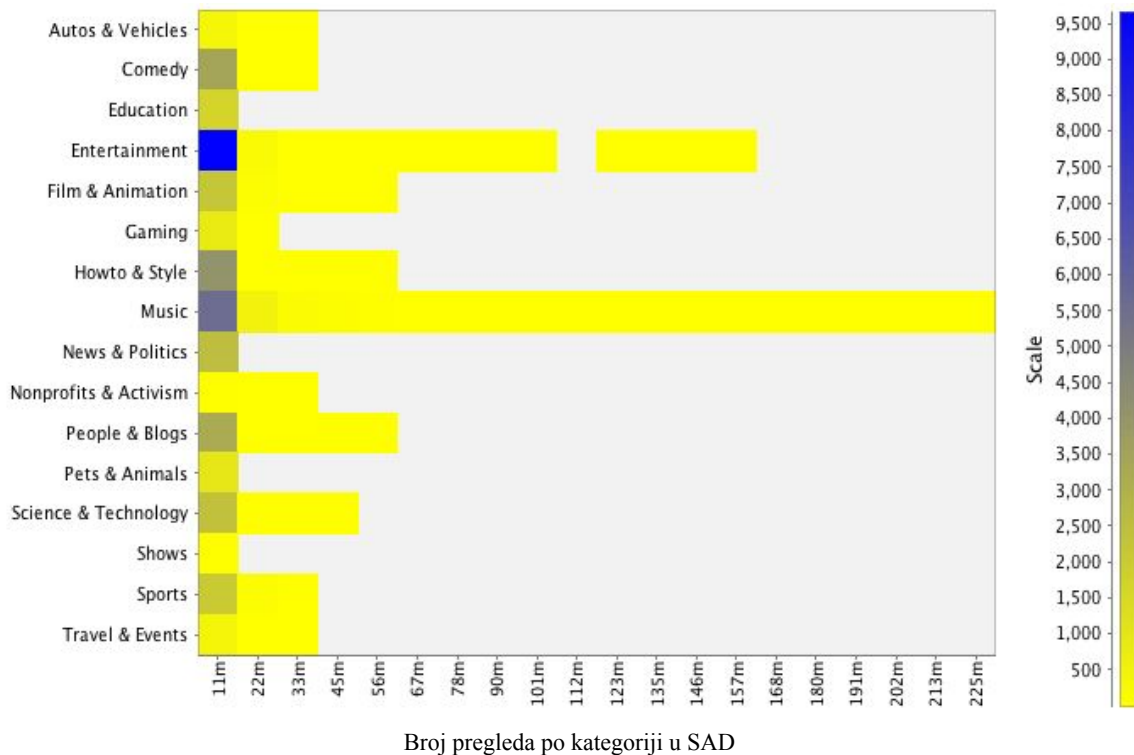


Prosečan broj komentara na klipu po državi

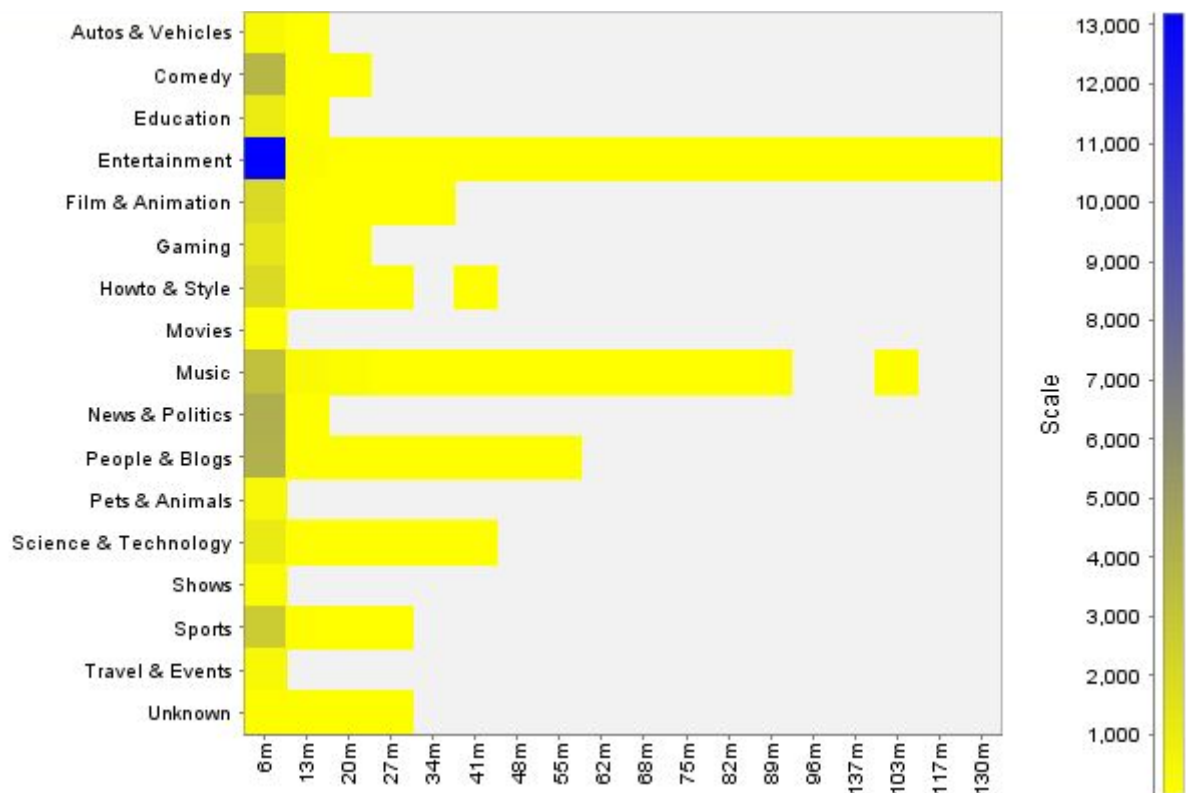
Primećujemo da je svaki grafik skoro identičnog oblika i da se značajno razlikuju samo brojevi između njih. Naime kako smo već pokazali da video najduže popularan u Velikoj Britaniji pa zatim u Americi, vidimo da to značajno utiče na statistike tog video klipa - ukoliko video klip je duže popularan imaće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

Popularnost kategorija po zemljama

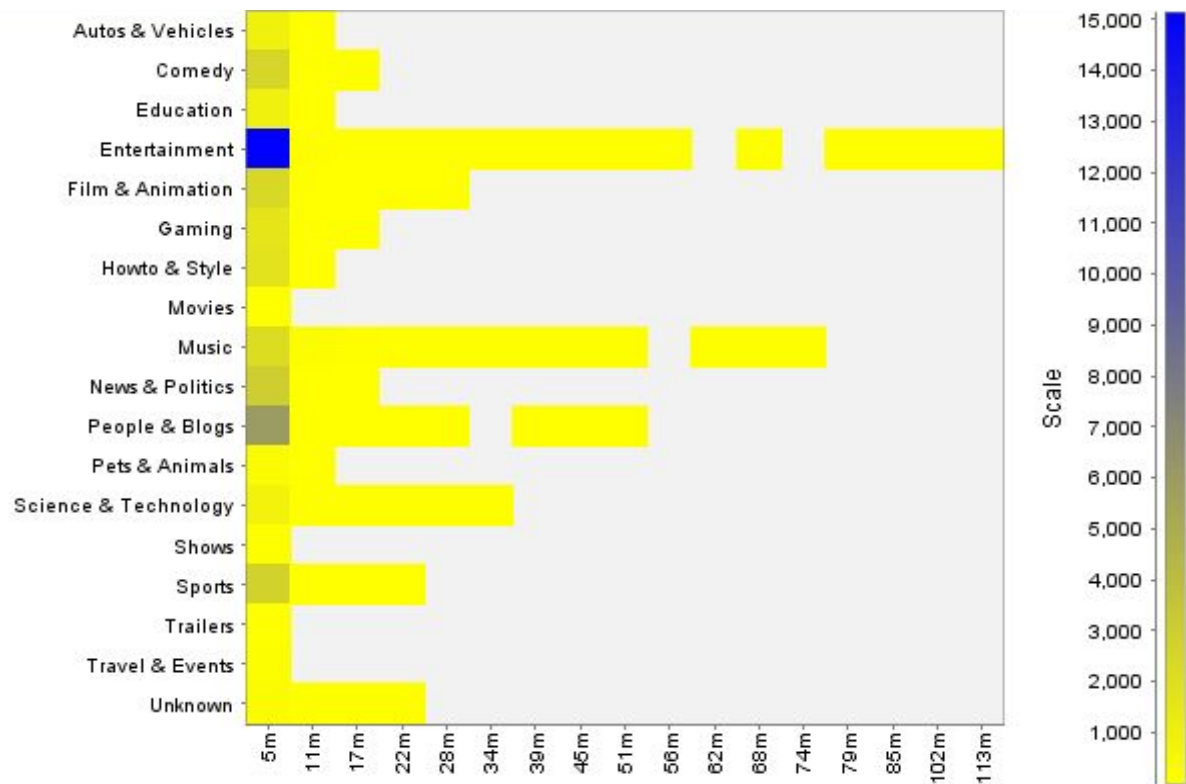
Primećuje se ponovo da su grafici sličnog oblika između zemalja, i da se razlikuju po broju pregleda - zbog prosečne dužine popularnosti klipa.



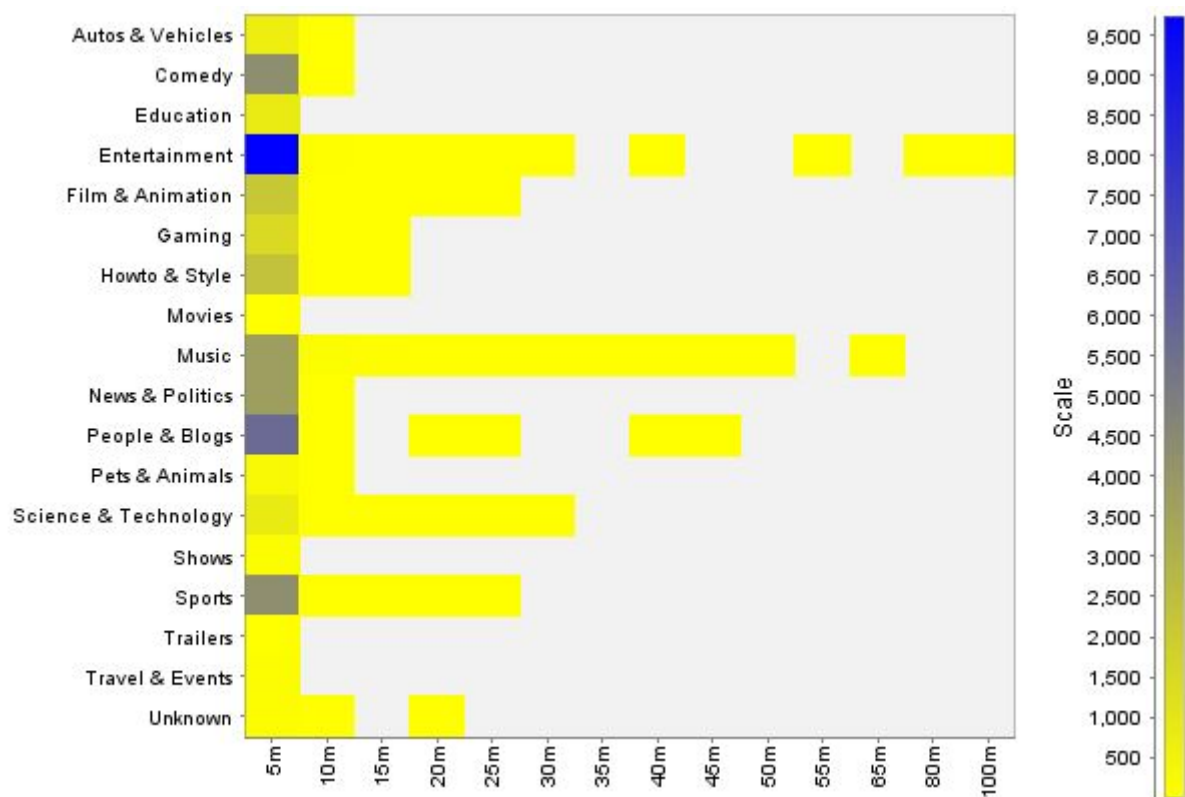
Broj pregleda po kategoriji u Velikoj Britaniji



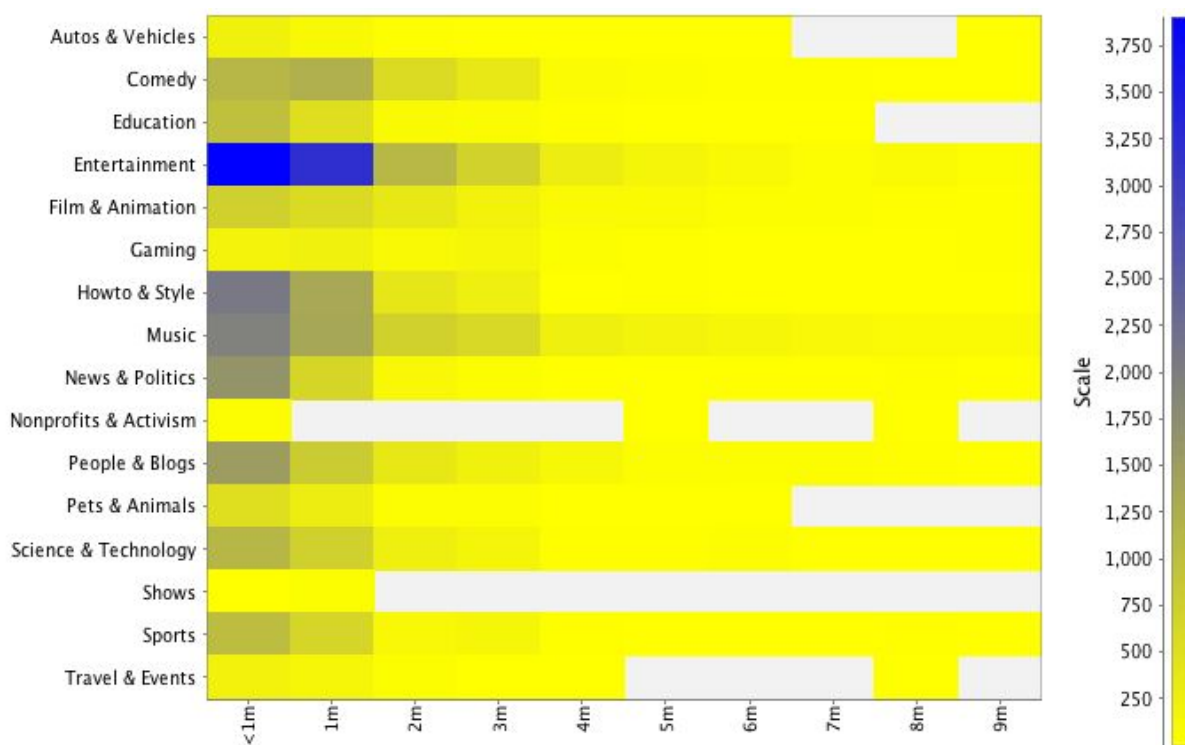
Broj pregleda po kategoriji u Kanadi



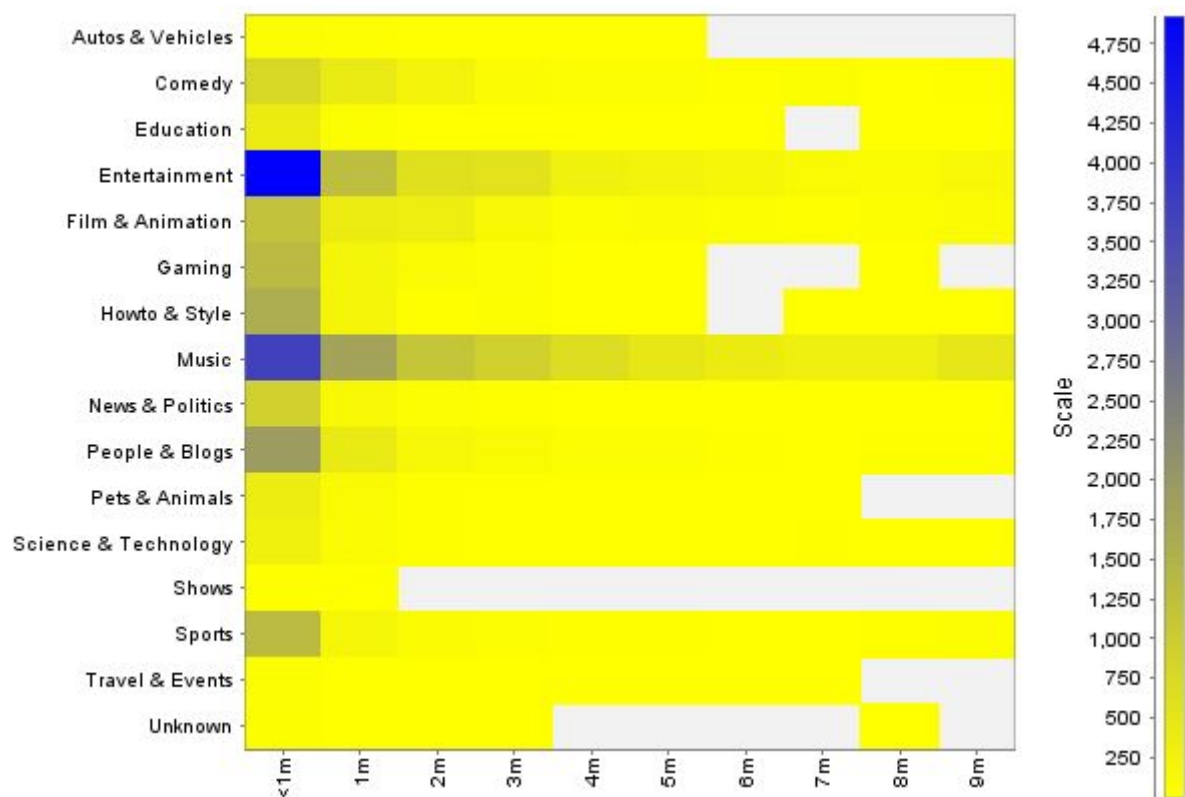
Broj pregleda po kategoriji u Nemačkoj



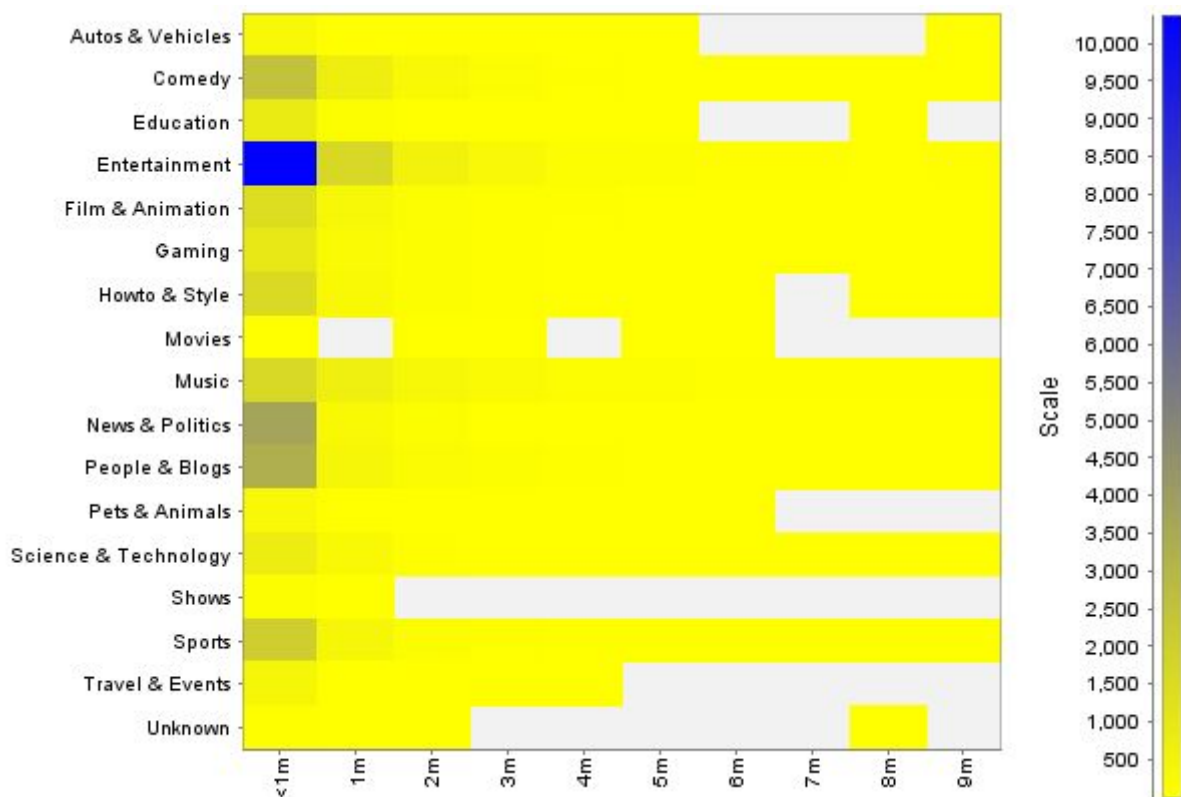
Broj pregleda po kategoriji u Francuskoj



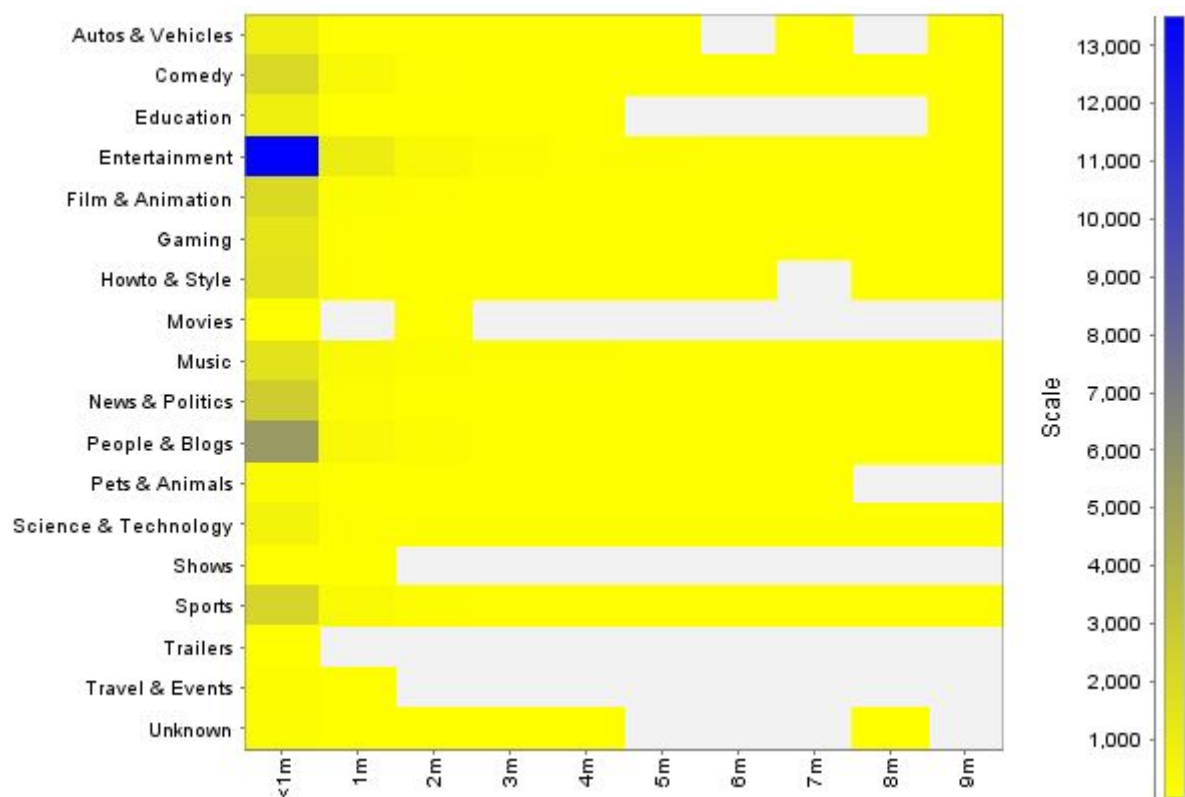
Broj pregleda po kategoriji u SAD do 10 mil.



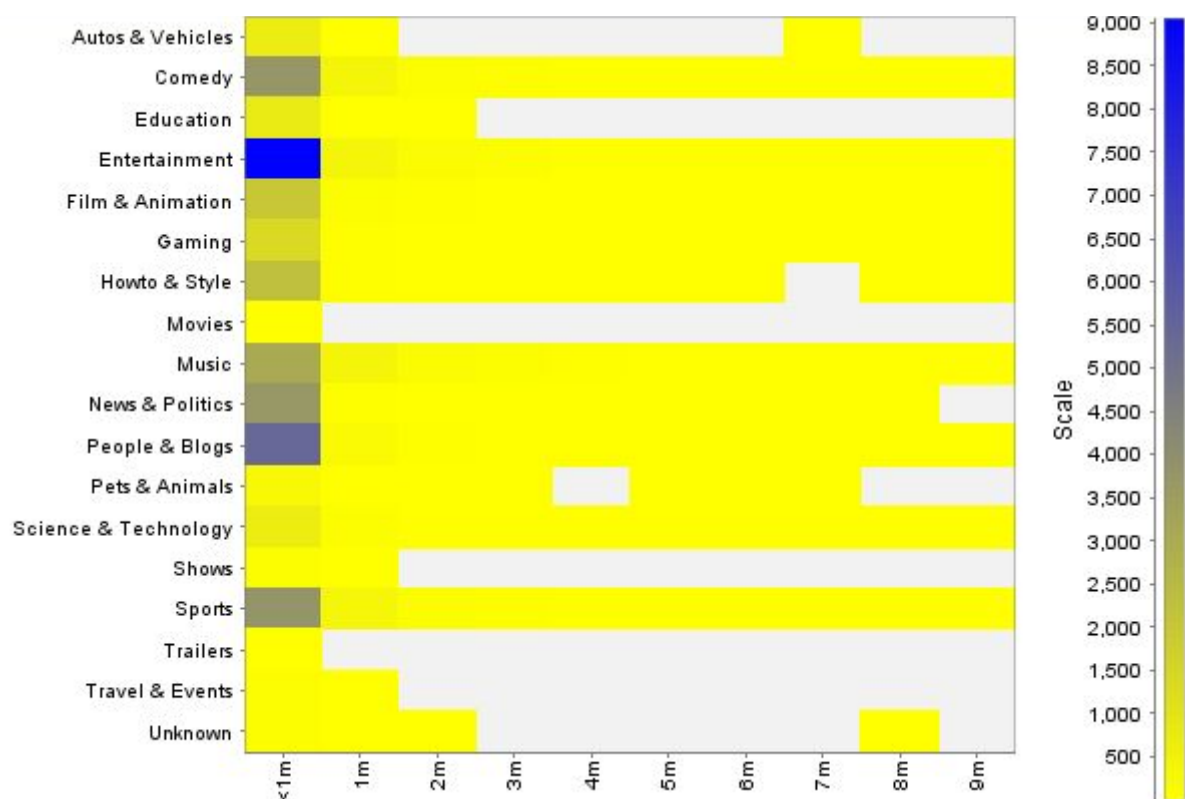
Broj pregleda po kategoriji u Velikoj Britaniji do 10mil.



Broj pregleda po kategoriji u Kanadi do 10mil.



Broj pregleda po kategoriji u Nemačkoj do 10mil.

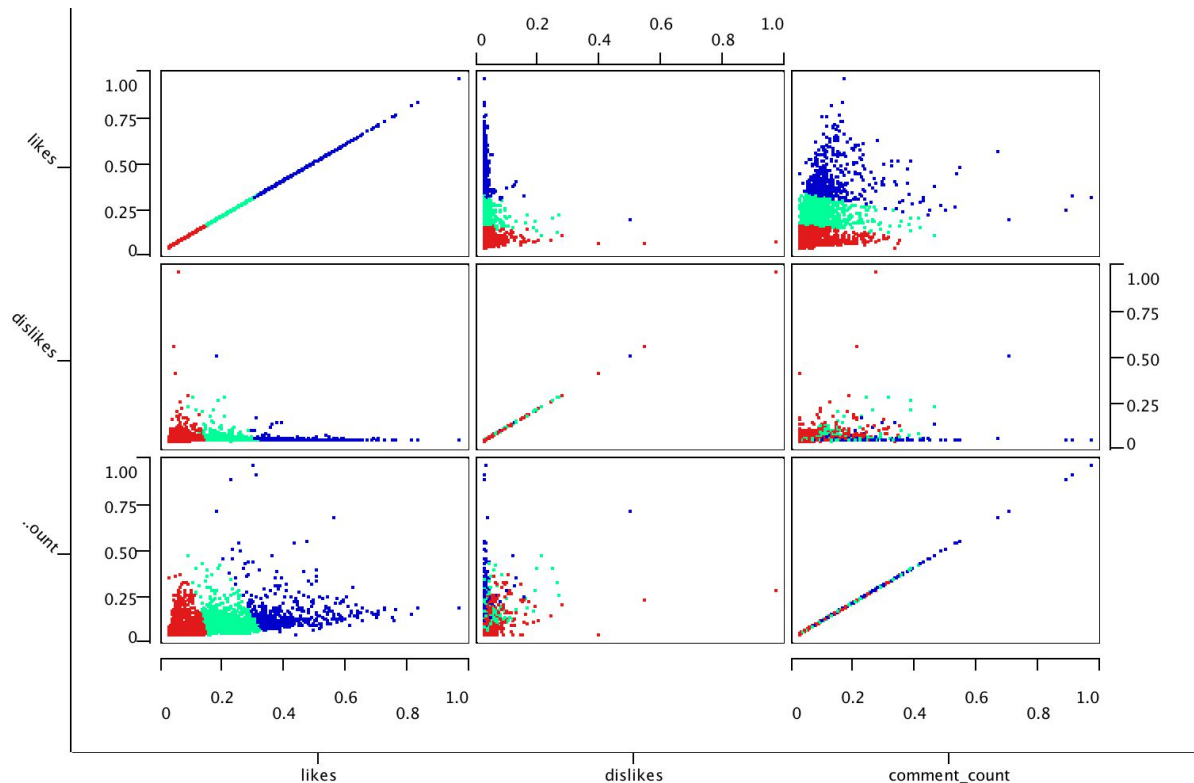


Broj pregleda po kategoriji u Francuskoj do 10mil.

Klasterovanje

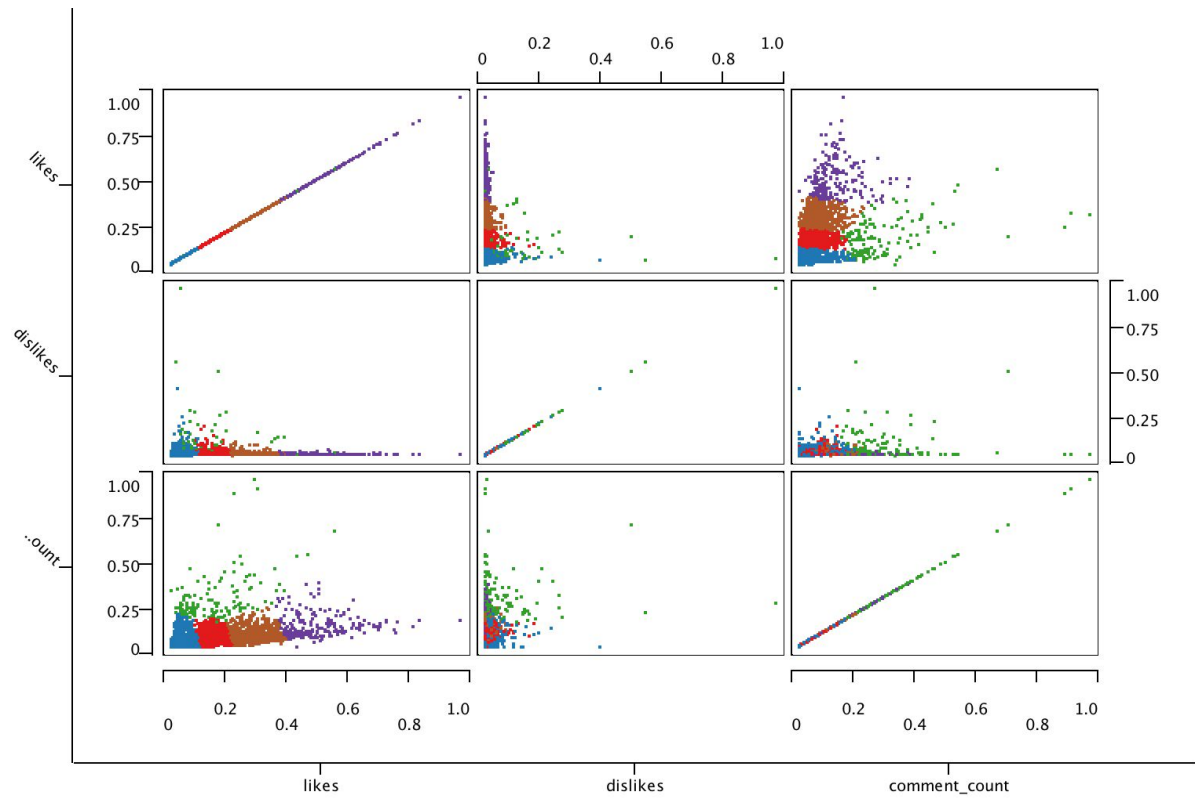
Rađeno je klasterovanje klipova K-Means algoritmom uzimajući u obzir broj komentara, pozitivnih i negativnih ocena klipa. Brojevi se uzimaju relativno na broj pregleda tako da ocenjujemo prosečnu reakciju gledalaca klipa.

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 klasteruje 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.

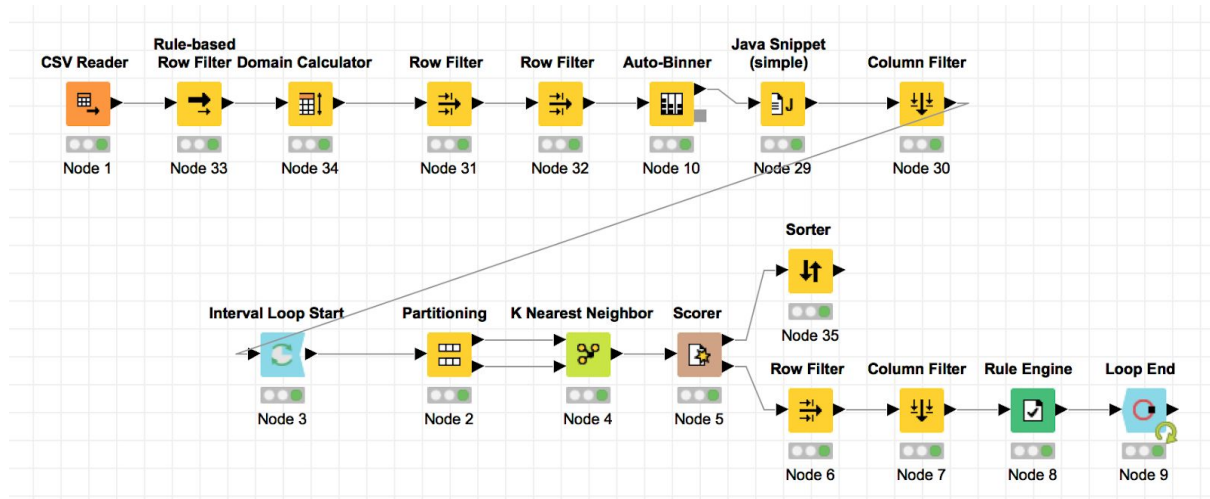


K-Means klasterovanje u 5 klastera

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.

Row ID	20m	40m	61m	81m	102m	122m	143m	163m	184m	204m	225m
20m	12061	60	8	2	0	0	0	0	0	0	0
40m	20	58	13	6	4	1	1	0	0	0	0
61m	2	7	23	5	2	0	0	0	0	0	0
81m	0	0	1	0	0	0	0	0	0	0	0
102m	0	0	0	0	1	3	0	1	0	0	0
122m	0	0	0	0	0	0	1	0	0	0	0
143m	0	0	0	0	0	0	0	1	0	0	0
163m	0	0	0	0	0	0	0	0	1	0	0
184m	0	0	0	0	0	0	0	0	0	0	0
204m	0	0	0	0	0	0	0	0	0	1	0
225m	0	0	0	0	0	0	0	0	0	0	2

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).

Row ID	1m	2m	3m	4m	5m	6m	7m	8m	9m
1m	6805	675	164	23	8	6	4	2	4
2m	375	1084	354	43	24	13	6	5	4
3m	64	259	804	111	64	33	32	15	10
4m	5	18	32	80	12	9	3	12	5
5m	4	10	16	11	57	13	10	5	6
6m	0	0	9	5	10	36	12	2	5
7m	0	0	5	3	9	8	23	19	3
8m	0	6	4	2	2	2	4	24	5
9m	0	0	2	0	3	3	1	5	18

Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifity	F-me...	Accuracy
1m	6805	448	3390	886	0.885	0.938	0.885	0.883	0.911	?
3m	804	586	9551	588	0.578	0.578	0.578	0.942	0.578	?
2m	1084	968	8653	824	0.568	0.528	0.568	0.899	0.547	?
9m	18	42	11455	14	0.562	0.3	0.562	0.996	0.391	?
7m	23	72	11387	47	0.329	0.242	0.329	0.994	0.279	?
5m	57	132	11265	75	0.432	0.302	0.432	0.988	0.355	?
4m	80	198	11155	96	0.455	0.288	0.455	0.983	0.352	?
6m	36	87	11363	43	0.456	0.293	0.456	0.992	0.356	?
8m	24	65	11415	25	0.49	0.27	0.49	0.994	0.348	?
Overall	?	?	?	?	?	?	?	?	?	0.775

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 pogada opseg pregleda sa preciznošću od 2 miliona.

Row ID	02m	04m	06m	08m	10m
02m	9105	607	100	17	8
04m	181	713	155	58	25
06m	14	53	190	44	28
08m	2	12	14	87	21
10m	2	5	9	12	67

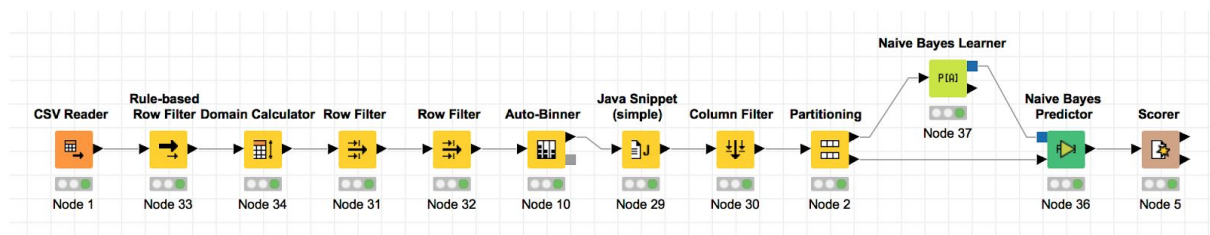
Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifity	F-me...	Accuracy
02m	9105	199	1493	732	0.926	0.979	0.926	0.882	0.951	?
04m	713	677	9720	419	0.63	0.513	0.63	0.935	0.565	?
10m	67	82	11352	28	0.705	0.45	0.705	0.993	0.549	?
08m	87	131	11262	49	0.64	0.399	0.64	0.989	0.492	?
06m	190	278	10922	139	0.578	0.406	0.578	0.975	0.477	?
Overall	?	?	?	?	?	?	?	?	?	0.881

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.

D Accuracy	k
0.773	1
0.784	2
0.777	3
0.782	4
0.78	5
0.779	6
0.778	7
0.778	8
0.774	9
0.775	10

Naivni Bajesov algoritam



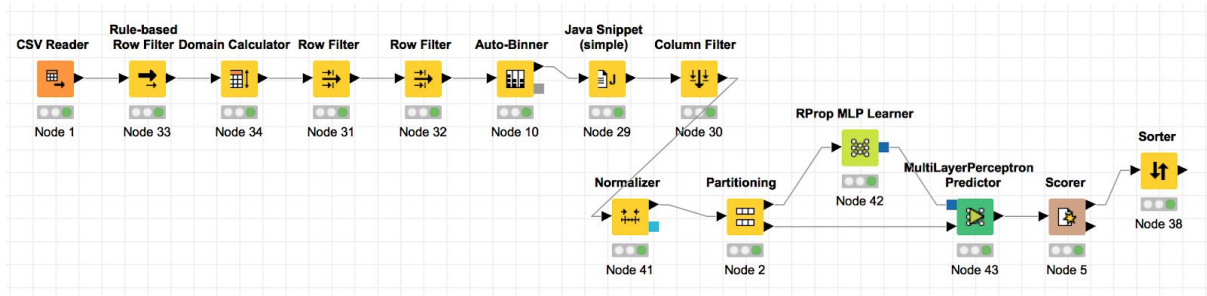
Primenjena je ista podela na 10 kategorija kao za KNN. Naivni Bajesov algoritam se ponašao značajno loše od KNN algoritma. Osim preciznosti od oko 70%, primetili smo da nijedan klip u kategoriji od 4 miliona nije tačno klasifikovan. Kada smo detaljnije analizirali statistiku klasifikatora primetili smo 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.

Table "Spec_name" - Rows: 9										Spec - Columns: 9		Properties	Flow Variables
Row ID	1m	2m	3m	4m	5m	6m	7m	8m	9m				
1m	6512	681	162	22	15	4	0	1	3				
2m	690	984	499	63	31	8	8	18	7				
3m	37	349	601	137	74	54	25	17	11				
4m	0	0	1	0	4	4	3	0	0				
5m	0	14	51	27	29	27	18	21	9				
6m	1	7	38	14	13	6	19	11	9				
7m	2	12	15	5	10	12	9	7	5				
8m	10	4	12	7	5	1	3	5	5				
9m	1	1	11	3	8	7	10	9	11				

Table "default" - Rows: 10											Spec - Columns: 11		Properties	Flow Variables
Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensiti...	Specifi...	F-me...	Accuracy				
3m	601	789	9435	704	0.461	0.432	0.461	0.923	0.446	?				
1m	6512	741	3388	888	0.88	0.898	0.88	0.821	0.889	?				
2m	984	1068	8153	1324	0.426	0.48	0.426	0.884	0.451	?				
5m	29	160	11173	167	0.148	0.153	0.148	0.986	0.151	?				
8m	5	84	11393	47	0.096	0.056	0.096	0.993	0.071	?				
6m	6	117	11294	112	0.051	0.049	0.051	0.99	0.05	?				
9m	11	49	11419	50	0.18	0.183	0.18	0.996	0.182	?				
7m	9	86	11366	68	0.117	0.095	0.117	0.992	0.105	?				
4m	0	278	11239	12	0	0	0	0.976	NaN	?				
Overall	?	?	?	?	?	?	?	?	?	0.708				

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.

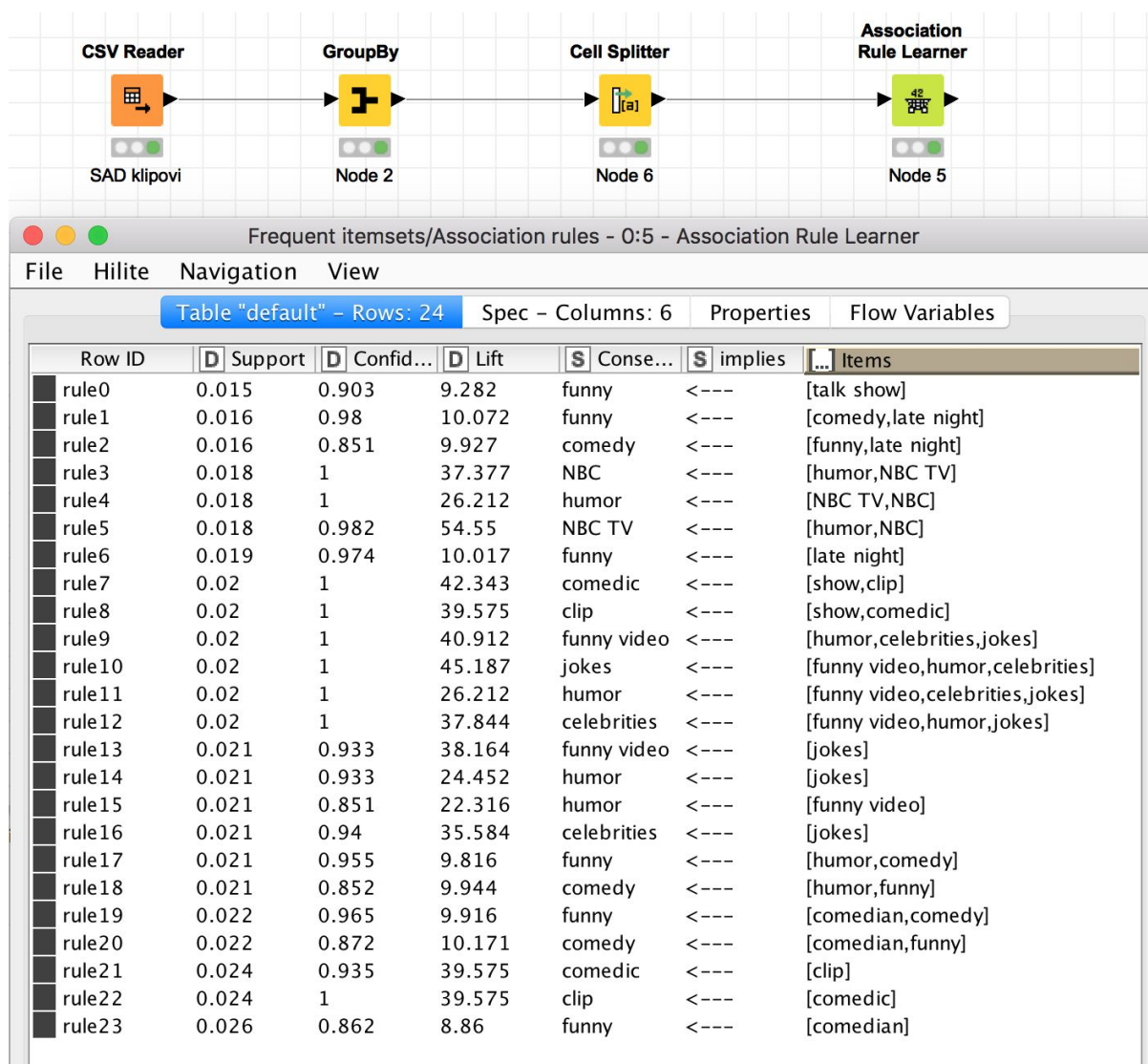
Row ID	I 1m	I 2m	I 3m	I 4m	I 5m	I 6m	I 7m	I 8m	I 9m
1m	6801	759	143	12	9	2	1	0	3
2m	432	1022	397	45	27	2	2	12	4
3m	19	271	840	205	131	86	53	50	28
4m	0	0	0	3	0	0	2	1	0
5m	0	0	5	4	7	7	11	3	2
6m	0	0	2	5	7	10	7	12	2
7m	0	0	1	1	4	2	3	1	1
8m	0	0	2	3	4	12	11	5	12
9m	1	0	0	0	0	2	5	5	8

Row ID	I TrueP...	I FalseP...	I TrueN...	I False...	D Recall	D Precisi...	D Sensiti...	D Specifity	D F-me...	D Accuracy
3m	840	550	9296	843	0.499	0.604	0.499	0.944	0.547	?
1m	6801	452	3347	929	0.88	0.938	0.88	0.881	0.908	?
2m	1022	1030	8556	921	0.526	0.498	0.526	0.893	0.512	?
9m	8	52	11456	13	0.381	0.133	0.381	0.995	0.198	?
6m	10	113	11371	35	0.222	0.081	0.222	0.99	0.119	?
5m	7	182	11308	32	0.179	0.037	0.179	0.984	0.061	?
8m	5	84	11396	44	0.102	0.056	0.102	0.993	0.072	?
7m	3	92	11424	10	0.231	0.032	0.231	0.992	0.056	?
4m	3	275	11248	3	0.5	0.011	0.5	0.976	0.021	?
Overall	?	?	?	?	?	?	?	?	?	0.755

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)

Row ID	D Support	D Confid...	D Lift	S Conse...	S implies	Items
rule0	0.02	0.969	10.398	funny	<---	[humor,comedy]
rule1	0.022	0.932	9.994	funny	<---	[talk show]
rule2	0.022	1	10.729	funny	<---	[comedian,late night]
rule3	0.023	0.986	30.189	clip	<---	[show,comedic]
rule4	0.023	0.986	32.075	comedic	<---	[show,clip]
rule5	0.024	1	34.689	NBC	<---	[humor,NBC TV]
rule6	0.024	0.987	23.514	humor	<---	[NBC TV,NBC]
rule7	0.024	0.987	41.072	NBC TV	<---	[humor,NBC]
rule8	0.024	1	34.689	NBC	<---	[NBC TV]
rule9	0.028	1	32.521	funny video	<---	[humor,celebrities,jokes]
rule10	0.028	1	23.832	humor	<---	[funny video,celebrities,jokes]
rule11	0.028	1	31.857	celebrities	<---	[funny video,humor,jokes]
rule12	0.028	1	32.521	jokes	<---	[funny video,humor,celebrities]
rule13	0.029	0.927	22.094	humor	<---	[funny video]
rule14	0.029	0.908	29.534	funny video	<---	[celebrities]
rule15	0.029	0.927	29.534	celebrities	<---	[funny video]
rule16	0.029	0.927	30.15	funny video	<---	[jokes]
rule17	0.029	0.927	30.15	jokes	<---	[funny video]
rule18	0.029	0.908	21.643	humor	<---	[celebrities]
rule19	0.029	0.927	29.534	celebrities	<---	[jokes]
rule20	0.029	0.908	29.534	jokes	<---	[celebrities]
rule21	0.029	0.899	10.712	comedy	<---	[comedian,funny]
rule22	0.029	0.957	10.267	funny	<---	[comedian,comedy]
rule23	0.029	0.91	10.844	comedy	<---	[funny,late night]
rule24	0.029	1	10.729	funny	<---	[comedy,late night]
rule25	0.03	0.853	10.167	comedy	<---	[comedian]
rule26	0.03	0.99	30.289	clip	<---	[comedic]
rule27	0.03	0.931	30.289	comedic	<---	[clip]
rule28	0.032	0.908	9.744	funny	<---	[comedian]
rule29	0.032	0.99	10.622	funny	<---	[late night]

Pravila pridruživanja za klipove iz GB skupa

Row ID	D Support	D Confid...	D Lift	S Conse...	S implies	Items
rule0	0.02	0.969	10.398	funny	<---	[humor,comedy]
rule1	0.022	0.932	9.994	funny	<---	[talk show]
rule2	0.022	1	10.729	funny	<---	[comedian,late night]
rule3	0.023	0.986	30.189	clip	<---	[show,comedic]
rule4	0.023	0.986	32.075	comedic	<---	[show,clip]
rule5	0.024	1	34.689	NBC	<---	[humor,NBC TV]
rule6	0.024	0.987	23.514	humor	<---	[NBC TV,NBC]
rule7	0.024	0.987	41.072	NBC TV	<---	[humor,NBC]
rule8	0.024	1	34.689	NBC	<---	[NBC TV]
rule9	0.028	1	32.521	funny video	<---	[humor,celebrities,jokes]
rule10	0.028	1	23.832	humor	<---	[funny video,celebrities,jokes]
rule11	0.028	1	31.857	celebrities	<---	[funny video,humor,jokes]
rule12	0.028	1	32.521	jokes	<---	[funny video,humor,celebrities]
rule13	0.029	0.927	22.094	humor	<---	[funny video]
rule14	0.029	0.908	29.534	funny video	<---	[celebrities]
rule15	0.029	0.927	29.534	celebrities	<---	[funny video]
rule16	0.029	0.927	30.15	funny video	<---	[jokes]
rule17	0.029	0.927	30.15	jokes	<---	[funny video]
rule18	0.029	0.908	21.643	humor	<---	[celebrities]
rule19	0.029	0.927	29.534	celebrities	<---	[jokes]
rule20	0.029	0.908	29.534	jokes	<---	[celebrities]
rule21	0.029	0.899	10.712	comedy	<---	[comedian,funny]
rule22	0.029	0.957	10.267	funny	<---	[comedian,comedy]
rule23	0.029	0.91	10.844	comedy	<---	[funny,late night]
rule24	0.029	1	10.729	funny	<---	[comedy,late night]
rule25	0.03	0.853	10.167	comedy	<---	[comedian]
rule26	0.03	0.99	30.289	clip	<---	[comedic]
rule27	0.03	0.931	30.289	comedic	<---	[clip]
rule28	0.032	0.908	9.744	funny	<---	[comedian]
rule29	0.032	0.99	10.622	funny	<---	[late night]

Pravila pridruživanja za klipove iz CA skupa