Analýza trendov v dĺžkach filmových (a iných) snímok

Projekt na 2-INF-150/15 Strojové učenie
zimný semester 2017/2018

Zuzana Hromcová

8. januára 2018

Abstrakt

Tento text popisuje priebeh a výsledky experimentu sledujúceho súvislosť medzi rôznymi parametrami filmu, seriálu či inej snímky a ich minutážou. V experimente sme sa snažili zistiť, aké atribúty vplývajú na výslednú dĺžku snímky a či sa tieto menia v čase. Na toto zisťovanie sme použili (generalizovanú) lineárnu regresiu, ktorú sme aplikovali na rôzne atribúty snímky, s cieľom predpovedať jej dĺžku. Úspešnosť tohto modelu sme porovnali s jednoduchou heuristikou.

1 Popis problematiky

Podľa nášho prieskumu je filmový priemysel obľúbenou aplikáciou algoritmov strojového učenia.

Hneď niekoľko riešení sme našli na predpovedanie toho, či daný film získa alebo nezíska Oscara, napríklad na základe hodnotení používateľov na IMDb a výsledkov iných veľkých ocenení (Zlaté Glóbusy, BAFTA a podobne) v [6] alebo kľúčových slov spojených s filmami v [7], a to s pomerne dobrou úspešnosťou.

Ďalším obľúbeným typom úlohy je predpovedanie obľúbenosti nie u filmových kritikov, ale u bežných divákov. Tu sa na problém možno pozerať z opačnej strany a spolu s ostatnými atribútmi filmov použiť aj výsledky ocenení, vrátane Oscarov. Takejto úlohe sa venoval napríklad [8]. Firma Netflix dokonca vyhlásila súťaž [9] o najlepší algoritmus na predpovedanie filmových ratingov podľa histórie hodnotení iných filmov používateľmi.

Nemálo riešení možno nájsť aj na problém predpovedania filmového žánru podľa rôznych jeho atribútov - hneď šiestim rôznym scenárom sa venujú analytici v štúdii [5], prípadne problému predpovedania filmového úspechu [10].

Naším pôvodným cieľom bolo taktiež predpovedať žáner filmu, avšak nie podľa hodnotení, ale podľa jeho režiséra a hercov, ktorí v ňom hrajú. To sa však, žiaľ, ukázalo ako nereálne, pretože s 45633 režisérmi a 359089 hercami, ktorí sa v našej databáze vyskytovali, by dáta mali príliš veľa atribútov.

Preto sme sa rozhodli vydať iným smerom - nie predpovedať niečo vopred neznáme, ako napríklad hodnotenie či úspech filmu, ale dĺžku filmu. Samozrejme, dĺžka filmu je známy údaj, ktorý zvyčajne nie je potrebné predpovedať, naším cieľom však bolo skúmať trendy v kinematografii. Natáčajú starší režiséri dlhšie filmy? Sú akčné filmy dlhšie ako romantické?

K tejto téme sme našli niekoľko štatistických štúdií - vývoj dĺžky filmov v čase, napríklad [12], [11] alebo [13]. My sa zameriame okrem filmov aj na iné typy snímok (seriály či televízne relácie) a aj na iné atribúty.

2 Dáta

2.1 Zdroj dát

Pri projekte sme vychádzali z dát IMDb, dostupných z [3] a [4]. Tieto dáta sú aktualizované denne a v čase realizácie tohto projektu obsahovali viac záznamy o 4.7 rôznych tituloch.

Dáta sú rozdelené v siedmich databázach, ktoré obsahujú rôzne informácie o jednotlivých tituloch, predovšetkým tieto atribúty:

- identifikátor záznamu
- typ záznamu (film, seriál, videohra)
- názov a originálny názov
- rok vydania
- krajina pôvodu
- dĺžka
- režisér, scenárista
- hlavné obsadenie
- žáner
- počet hodnotení a priemerný rating

Súčasťou databázy sú aj informácie o jednotlivých osobách - hercoch, režiséroch, scenáristoch, predovšetkým ich rok narodenia, úmrtia či najznámejšie diela.

2.2 Príprava dát

Ešte pred samotným trénovaním sme si museli určiť, ktoré z týchto záznamov budeme používať, nie všetky sú totiž pre náš experiment relevantné.

V prvom kroku sme všetky databázy zlúčili do jednej podľa jednoznačného identifikátora záznamov. Ako zaujímavé atribúty sme určili:

- typ snímky
- rok vydania snímky
- rok narodenia režiséra snímky
- žáner snímky
- informácia, či je snímka určená pre plnoletých divákov

V rámci predspracovania dát sme odstránili všetky záznamy, ktoré mali aspoň jednu z menovaných položiek nekompletnú. Náš výsledný dataset sa tak zredukoval zo 4.7 milióna na 360-tisíc záznamov.

Keďže typ aj žáner snímky sú kategorickými atribútmi, použili sme metódu *bag of words*. Atribút *typ* sme nahradili 10 novými atribútmi, pre každý typ zvlášť a s hodnotami 0 alebo 1 podľa príslušnosti snímky k typu. Atribút *žáner* sme podobne nahradili 28 novými atribútmi.

Po spracovaní tak dáta mali 41 atribútov - rok vzniku filmu a narodenia režiséra, 10 atribútov pre typ, 28 pre žáner a jeden atribút označujúci, či ide o film pre plnoletých.

Dáta sme zamiešali a rozdelili na trénovacie a testovacie v pomere 4:1. O normalizáciu a pridanie intercept prvku sa za nás postarali knižnice.

2.3 Štruktúra dát

Náš cieľový atribút - dĺžka snímky - sa pohybuje v pomerne širokom rozsahu od 0 do 9000 minút, v priemere 56 minút so štandardnou odchýlkou 47 minút.

Rok vydania snímky je v priemere 1984, štandardná odchýlka 26. Hodnoty sa pohybujú v rozmedzí od 1878 do 2024[1].

Rok narodenia režiséra je v priemere 1940, so štandardnou odchýlkou 28, pričom hodnoty sa pohybujú od 1830 do 2017[2].

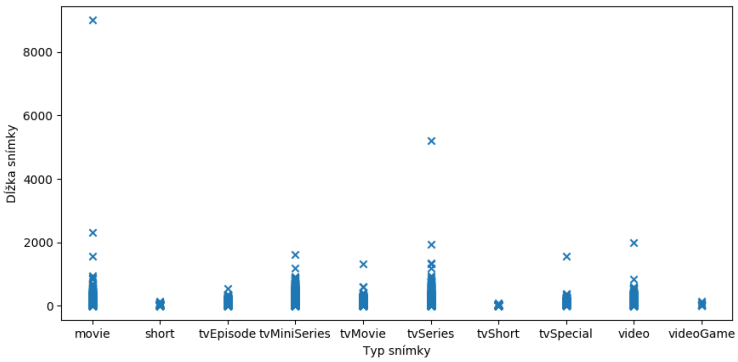
Zaujímavejším by mohol byť vek režiséra v čase vydania filmu, ten je priemerne 43, so štandardnou odchýlkou 11.

Dáta obsahujú 10 rôznych typov snímky, najpočetnejšie zastúpené sú epizódy seriálov (120-tisíc) a filmy (110-tisíc), najmenej početné videohry (8). Každá snímka má priradený jeden až tri z 28 žánrov, najširšie zastúpenie majú drámy (140-tisíc) a komédie (110-tisíc), najmenšie film-noir (800).

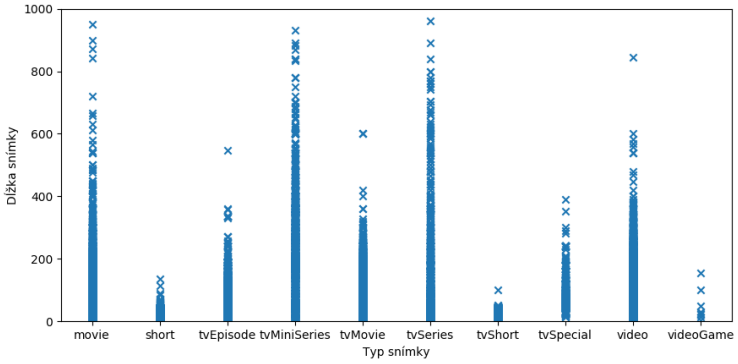
Nasledujúce grafy zobrazujú rozloženie dĺžok snímok v závislosti od jednotlivých atribútov - postupne od typu snímky (1, 2), roku jej vzniku (3, 4) a veku jej režiséra (5, 6) v čase vzniku snímky (teda rozdielu roku vzniku snímky a roku narodenia režiséra), vo väčšom aj menšom meradle.

[1] Ešte nevydané snímky

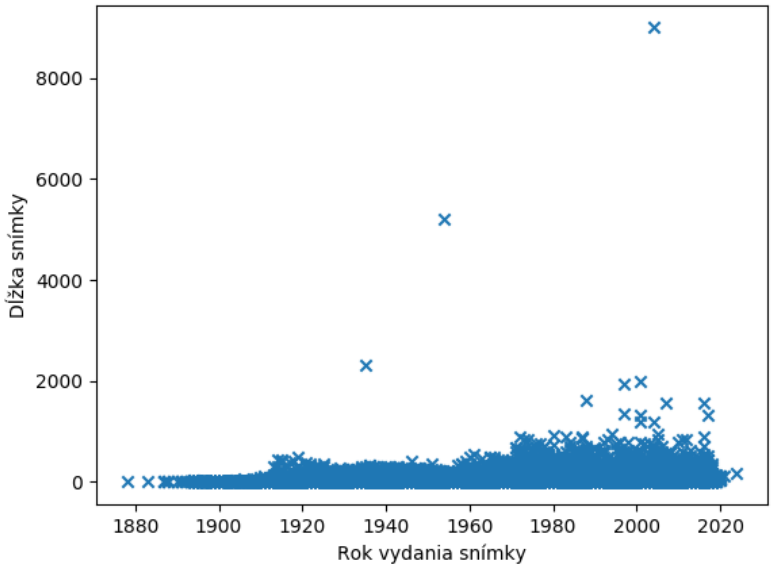
[2] Očividne ide o chybu v dátach IMDb, takýto rok narodenia má uvedený napríklad Kieran Sullivan



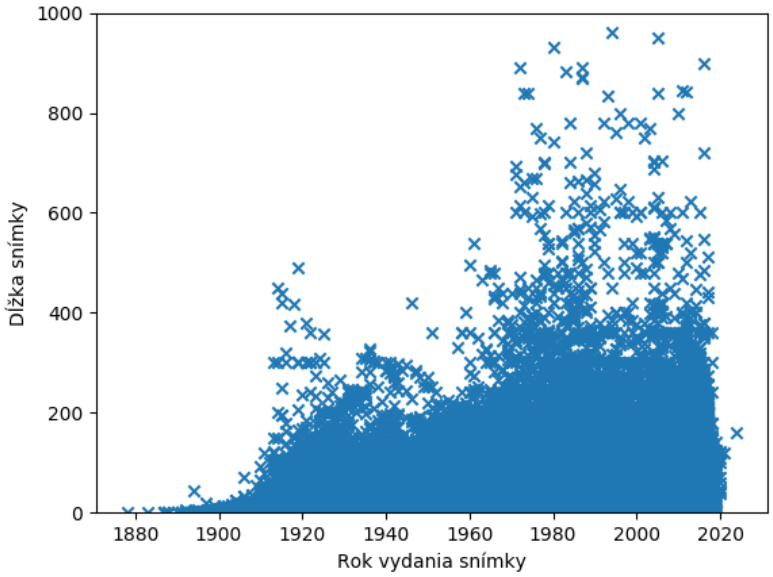
Obr. 1: Závislosť dĺžky od typu snímky (celý pohľad)



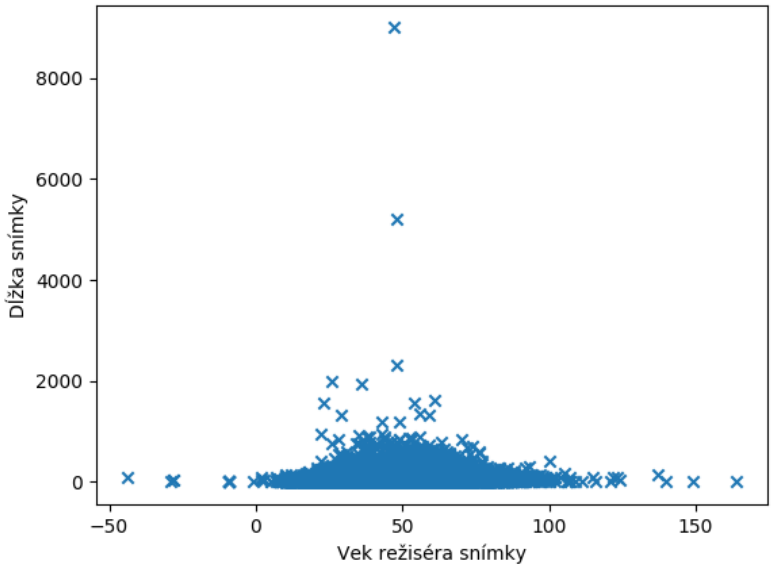
Obr. 2: Závislosť dĺžky od typu snímky (časť)



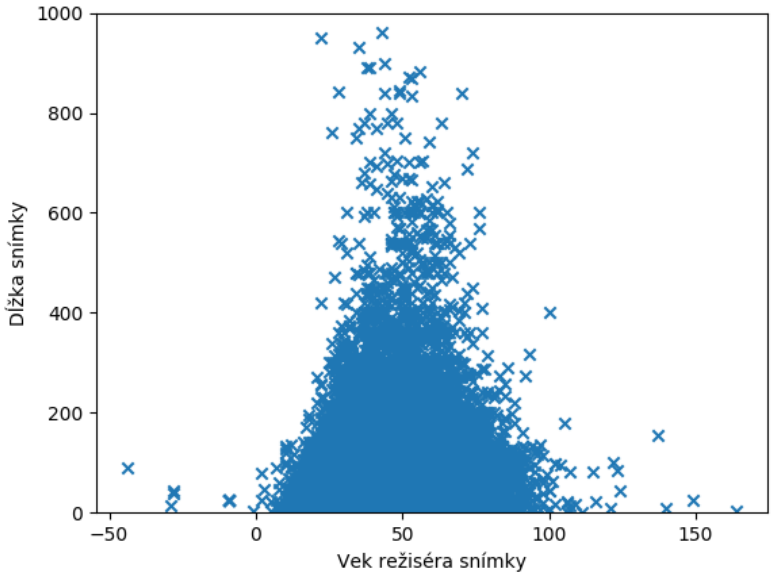
Obr. 3: Závislosť dĺžky od roku vydania snímky (celý pohľad)



Obr. 4: Závislosť dĺžky od roku vydania snímky (časť)



Obr. 5: Závislosť dĺžky snímky od veku režiséra (celý pohľad)



Obr. 6: Závislosť dĺžky snímky od veku režiséra (časť)

3 Návrh experimentu

Na riešenie tohto problému sme použili lineárnu regresiu, ktorej vstupom boli popísané atribúty a výstupnou hodnotou dĺžka filmu, ako aj generalizovanú lineárnu regresiu, s atribútmi rozvedenými do polynómov vyšších stupňov.

Na porovnanie sme použili jednoduchú heuristiku, ktorá každej snímke priraďuje priemernú dĺžku snímky príslušného typu. Keďže každá snímka má priradený práve jeden typ, takéto priradenie je jednoznačné.

Vďaka tomu bolo možné porovnať, či ostatné atribúty, ako napríklad rok narodenia režiséra, majú na výslednú dĺžku snímky výrazný vplyv, alebo vieme rovnako dobré výsledky dosiahnuť naším heuristickým prístupom.

4 Implementácia

Na implementáciu sme používali jazyk Python, predovšetkým jeho knižnice numpy na pracovanie s dátami a sklearn na samotné trénovanie modelu.

Celý proces sme rozdelili do niekoľkých krokov:

1. čistenie a predspracovanie dát
2. rozdelenie dát na trénovaciu a testovaciu množinu
3. trénovanie pomocou (generalizovanej) lineárnej regresie
4. aplikovanie heuristiky

Na záver sme porovnali výsledky dosiahnuté jednotlivými metódami.

4.1 Lineárna regresia

Na lineárnu regresiu, rovnako na pridanie intercept prvku a normalizáciu dát, sme použili už hotový model LinearRegression z knižnice sklearn.linear_model.

Úspešnosť sme vyhodnocovali pomocou koeficientu determinácie R2. Ten možno vypočítať ako $1 - u/v$, kde $u = \Sigma(y_{true} - y_{pred})^2$ a $v = \Sigma(y_{true} - \overline{y_{true}}))^2$. Hodnota y_{true} predstavuje hľadanú a hodnota y_{pred} predikovanú cieľovú hodnotu.

Hodnota tohto koeficientu je 0, ak model predpovedá vždy očakávanú hodnotu y a 1 pri najlepšej zhode, môže ňou však byť aj ľubovoľne malá (záporná) hodnota v závislosti od toho, aké zlé sú predpovede modelu.

Pomocou tohto koeficientu sme určovali trénovacie aj testovacie skóre. Pre ilustráciu, pre jedno z premiešaní dát sme pri trénovaní sme pri 10-násobnej cross-validácii dosiahli skóre 0.557 (0.518 bez nej), pri testovaní skóre 0.508.

Okrem lineárnych funkcií sme vyskúšali rozšíriť množinu hypotéz aj na kvadratické funkcie a polynómy vyšších stupňov. Keďže išlo o výpočtovo aj pamäťovo náročnú operáciu, pracovali sme len s 500 trénovacími a 100 testovacími príkladmi a len do štvrtého rádu. Väčšia trieda hypotéz však nepriniesla lepšiu presnosť, respektíve nepriniesla prakticky žiadnu presnosť, čo nás pri takom

malom počte trénovacích príkladov vôbec neprekvapilo. Zvýšenie tohto počtu už však bolo výpočtovo nereálne.

Nasledujúca tabuľka ilustruje naše výsledky pri použití generalizovanej lineárnej regresie - priam ukážkový prípad preučenia.

Stupeň	Trénovacie skóre	Testovacie skóre
1	0.674	$-1.26 * 10^{23}$
2	0.673	$-3.56 * 10^{26}$
3	0.788	$-6.81 * 10^{24}$
4	0.938	$-1.62 * 10^{23}$

Keď sa pozrieme na výsledky lineárnej regresie, teda optimálne koeficienty, môžeme usúdiť, že pre určenie dĺžky snímky je najdôležitejší práve jej typ. Tieto koeficienty sa pohybovali rádovo v 10^{12}. Žánre majú oveľa menšiu dôležitosť, rádovo 10^{1}, ale taktiež prinášajú nejakú informáciu. Niektoré žánre majú napríklad koeficienty záporné, teda tendenciu dĺžku snímky znižovať - napríklad muzikál, správy, talk-show či reality-show.

4.2 Heuristika

Dáta sme v trénovacej časti rozdelili na menšie časti podľa toho, akého typu boli skúmané snímky. Ako sme videli pri vizualizácii dát, práve toto by mohol byť ten rozhodujúci atribút. Pre každú takúto snímku sme vypočítali jej priemernú dĺžku, a tú sme potom v testovacej časti používali ako odhad pre všetky snímky tohto typu.

Typ snímky	Priemerná dĺžka
tvMiniSeries	166.58
movie	91.10
tvSpecial	87.93
tvMovie	82.81
video	75.65
tvSeries	51.30
tvEpisode	38.76
videoGame	32.67
tvShort	21.86
short	13.91

Skóre sme počítali rovnako ako v predošlej časti - ako koeficient determinácie. Pre rovnaké premiešanie a rozdelenie dát ako v predošlom prípade sme v trénovacej časti dosiahli skóre 0.445, v testovacej 0.437.

5 Výsledky a porovnanie

Pre popísaný jeden prípad zamiešania a rozdelenia dát bol teda výsledok nášho modelu lepší ako úspešnosť heuristiky, ale nie oveľa. Typ snímky je teda z hľadiska určovania jej dĺžky významným faktorom, ale ostatné atribúty vedia tento odhad vylepšiť.

Aby sme sa presvedčili, že typ snímky je naozaj významný faktor, aplikovali sme na tie isté dáta lineárnu regresiu ešte raz, ale tentokrát sme z atribútov vynechali informáciu o type snímky. Dosiahnuté trénovacie skóre bolo 0.325, testovacie 0.316, čo je značné zhoršenie oproti obom technikám, ktoré typ brali do úvahy.

Na potvrdenie týchto výsledok sme experiment zopakovali ešte niekoľkokrát.

Najprv sme zachovali rozdelenie dát medzi trénovaciu a testovaciu množinu. Testovaciu množinu sme rozdelili na 10 častí (po približne 30-tisíc príkladov) a na každú z nich sme aplikovali každý z algoritmov - lineárnu regresiu (R), lineárnu regresiu s vynechaním atribútu typ (RT) a heuristiku (H). Tabuľka zobrazuje trénovaciu a testovaciu chybu vo všetkých prípadoch.

#	Train (R)	Test (R)	Train (H)	Test (H)	Train (RT)	Test (RT)
1	0.609	$-8 * 10^{18}$	NAN	NAN	0.308	0.315
2	0.597	0.507	0.506	0.436	0.382	0.316
3	0.579	$-3 * 10^{18}$	NAN	NAN	0.364	0.316
4	0.603	0.508	0.519	0.437	0.376	0.316
5	0.247	$-2 * 10^{19}$	NAN	NAN	0.156	0.315
6	0.564	0.506	0.478	0.435	0.360	0.316
7	0.611	$-8 * 10^{18}$	NAN	NAN	0.381	0.315
8	0.598	0.508	0.515	0.438	0.371	0.316
9	0.572	0.506	0.490	0.437	0.357	0.316
10	0.600	$-3 * 10^{19}$	NAN	NAN	0.381	0.316

Približne polovica hodnôt v stĺpci Test (R) je pomerne alarmujúca, v týchto prípadoch bola natrénovaná hypotéza zrejme značne nesprávna. Domnievame sa, že to je spôsobené nerovnomerným zastúpením jednotlivých typov snímok v týchto oklieštených datasetoch - najmä tých najzriedkavejších. Keďže ide o taký dôležitý atribút, mohlo to spôsobiť zlé natrénovanie.

Tento predpoklad dokazujú aj výsledky heuristiky v tých istých prípadoch - skóre je NAN, zrejme preto, lebo niektorý z typov sa v trénovacej množine nevyskytol, takže sme v testovacej fáze nemohli predpovedať jeho dĺžku.

Pri vynechaní typu snímky (stĺpce Train RT a Test RT) už zrejme tento problém nemáme. Na druhej strane, natrénovaná hypotéza je vo všetkých prípadoch celkom slabá oproti iným prístupom.

Z týchto výsledkov je zrejmé, že náš model má vyššiu úspešnosť ako jednoduchá heuristika, ale aj to, aký dôležitý atribút je typ snímky.

Na záver si dovolíme predstaviť ešte jeden experiment, aby sme demonštrovali výsledky nášho modelu. Všetky dáta sme odznova premiešali a rozdelili na trénovaciu a testovaciu množinu (290-tisíc a 73-tisíc príkladov) a na všetkých týchto dátach vyskúšali všetky tri algoritmy, pričom túto procedúru sme zopakovali 10-krát.

#	Train (R)	Test (R)	Train (H)	Test (H)	Train (RT)	Test (RT)
1	0.540	0.591	0.430	0.508	0.314	0.371
2	0.556	0.584	0.431	0.500	0.349	0.329
3	0.603	0.330	0.517	0.585	0.379	0.206
4	0.544	0.589	0.430	0.505	0.341	0.330
5	0.589	0.349	0.503	0.303	0.370	0.217
6	0.535	0.602	0.428	0.518	0.336	0.376
7	0.546	0.580	0.432	0.498	0.342	0.365
8	0.543	0.615	0.427	0.527	0.341	0.386
9	0.547	0.574	0.432	0.495	0.344	0.357
10	0.538	0.604	0.429	0.517	0.337	0.381

Ako ilustruje tabuľka, naše závery neboli náhodné a potvrdili sa vo viacerých prípadoch.

6 Záver

V tomto experimente sme skúmali závislosť dĺžky snímky od rôznych parametrov. Na predpovedanie dĺžky sme použili lineárnu regresiu a jednoduchú heuristiku, pričom náš model dosahoval mierne lepšie výsledky ako spomínaná heuristika.

Overili sme, že najvýznamnejším parametrom z vybraných je podľa očakávaní typ snímky - teda informácia, či ide o film, správy alebo inú snímku.

Domnievame sa, že nízke zlepšenie presnosti modelu oproti popisovanej heuristike bolo spôsobené tým, že samotné dáta neobsahovali všetky potrebné atribúty na to, aby bolo možné dostatočne spoľahlivo určiť dĺžku snímky, preto ani väčšia trieda hypotéz nepomôže. S týmito vstupnými atribútmi nevieme rozlíšiť dva filmy toho istého žánru natočené tým istým režisérom v tom istom roku, hoci jeden mohol mať napríklad väčší rozpočet a dlhšiu minutáž a druhý naopak.

Projekt teda môžeme považovať za úspešný - došlo k zlepšeniu úspešnosti oproti heuristike - avšak podľa nášho názoru bol málo ambiciózny.

Vo všeobecnosti by sme druhýkrát určite volili zaujímavejšiu a flexibilnejšiu tému, podľa možností s menším celkovým počtom atribútov.

Táto téma nepriniesla veľa variability vo výsledkoch a najmä sme nemohli naplno využiť skúsenosti získané z predmmetu. Naproti našim očakávaniam bol významný najmä jeden atribút a ostatné boli v úzadí. Výsledky sme nezlepšili ani použitím generalizovanej lineárnej regresie, takže sme sa venovali predovšetkým základnej lineárnej regresii.

A čo by sme urobili inak s aktuálnym projektom?

Pri tejto úlohe by sme určite použili knižnicu IMDbpy, o ktorej existencii sme sa dozvedeli neskôr. Tá by mohla uľahčiť prácu s prípravou dát a určite by bolo jednoduchšie experimentovať s ďalšími kombináciami, možno by sme sa dostali k ďalším zaujímavým atribútom.

Literatúra

[1] Dokumentácia knižnice scikit-learn,
http://scikit-learn.org/stable/documentation.html

[2] Dokumentácia knižnice numpy,
http://docs.scipy.org

[3] Databáza filmov IMDb,
https://datasets.imdbws.com/

[4] Popis položiek databázy filmov IMDb,
http://www.imdb.com/interfaces/

[5] Predikcia filmových žánrov podľa hodnotení používateľov,
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3573840/

[6] Predpovedanie víťazov Oscarov,
https://blog.bigml.com/2017/02/23/predicting-the-2017-oscar-winners/

[7] Predpovedanie víťazov Oscarov,
http://oscarpredictor.github.io/oscar.html

[8] Predpovedanie IMDb ratingu,
http://www.rpubs.com/caiomiyashiro/imdb_rating_regression

[9] Predpovedanie Netflix ratingu,
https://netflixprize.com/index.html

[10] Predpovedanie úspechu filmu,
http://cs229.stanford.edu/proj2013/EricsonGrodman-APredictorForMovieSuccess.pdf

[11] Štatistiky o dĺžkach filmov,
http://www.businessinsider.com/are-movies-getting-longer-2016-6

[12] Štatistiky o dĺžkach filmov,
http://www.randalolson.com/2014/01/25/movies-arent-actually-much-longer-than-they-used-to

[13] Štatistiky o dĺžkach filmov,
http://stat405.had.co.nz/project/project-01-f.pdf