Čo robí filmy obľúbenými a neobľúbenými?

Otázky, ktorým sme sa venovali

Naším cieľom bolo predikovať úspešnosť filmov podľa rôznych atribútov, porovnať kvalitu týchto modelov a na základe toho zhodnotiť, ktoré z pozorovaných vlastností najviac vplývajú na to, či je film úspešný alebo nie. Úspešnosť filmu sme si pre účely tejto práce definovali ako priemerné hodnotenie, ktoré film dostal od divákov na Rotten Tomatoes.

Dátové zdroje

V projekte sme používali dáta zo stránok IMDb a Rotten Tomatoes, ktoré sme získali ako datasety zo stránky Kaggle. IMDb nám poskytol charakteristiku filmu, a Rotten Tomatoes sme potrebovali kvôli hodnoteniu divákmi. Skúmali sme filmy, ktoré boli prítomné v obidvoch datasetoch. Takých filmov bolo 8854.

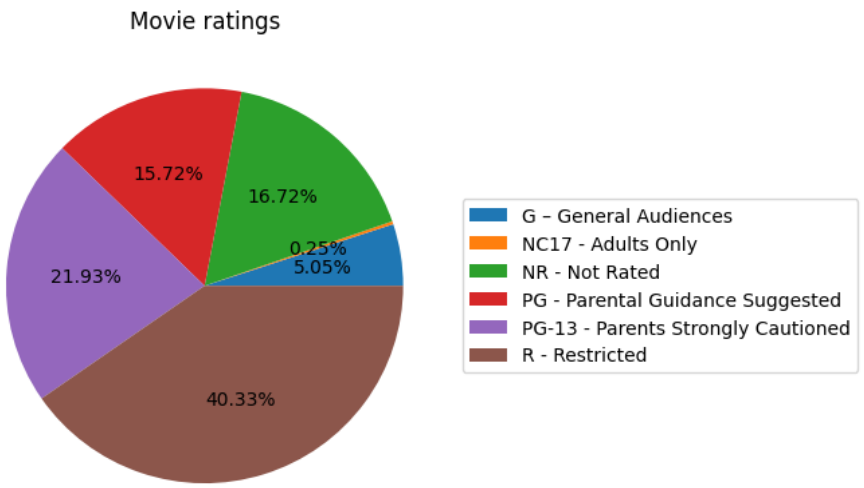
UPOZORNENIE: Keďže dáta sú sťahované s pomocou Kaggle API pri spustení je potrebné zadať prihlasovacie meno a kľúč. Tieto informácie sú uvedené v časti “Downloading data”.

Atribúty, s ktorými sme pracovali:

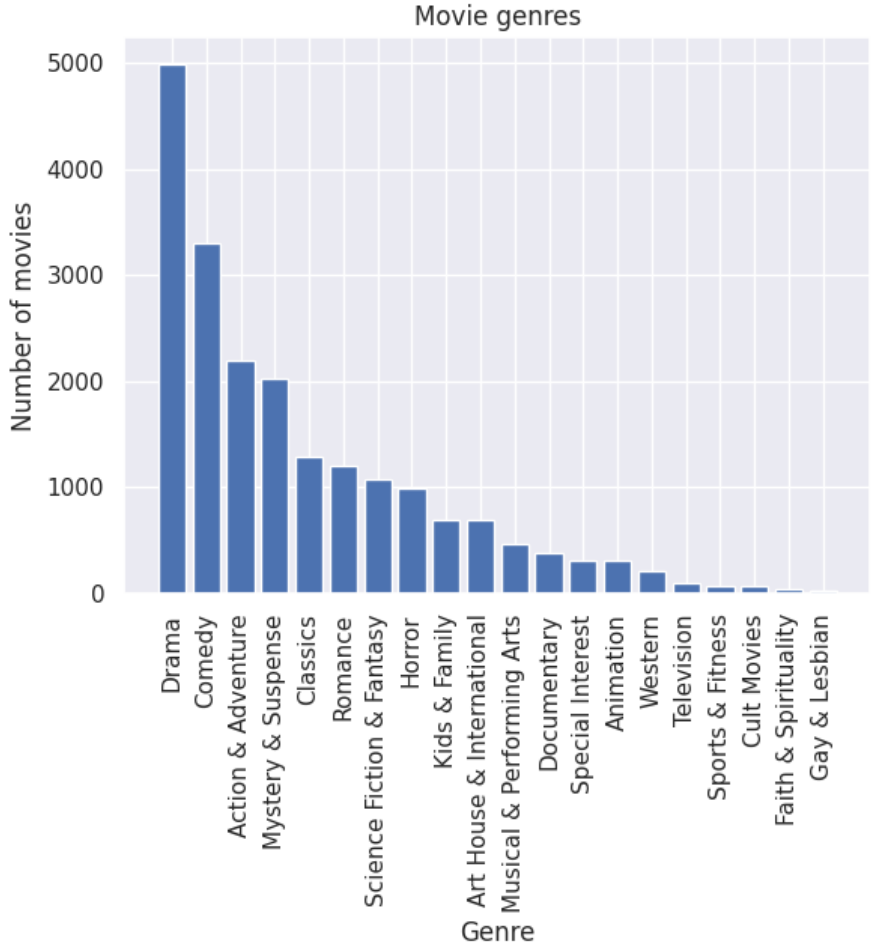
- názov filmu
- filmový rating (pre akú vekovú kategóriu je určený)
- žánre filmu
- rok vydania
- režisér
- autori
- herci
- produkčná spoločnosť
- hodnotenie divákov
- dĺžka filmu
- rozpočet
- krátky obsah filmu
- výnos
- kľúčové slová

Pre mnohé filmy však chýbala informácia o budgete a revenue, či kľúčové slová.

Pár vizualizácii o datasete:

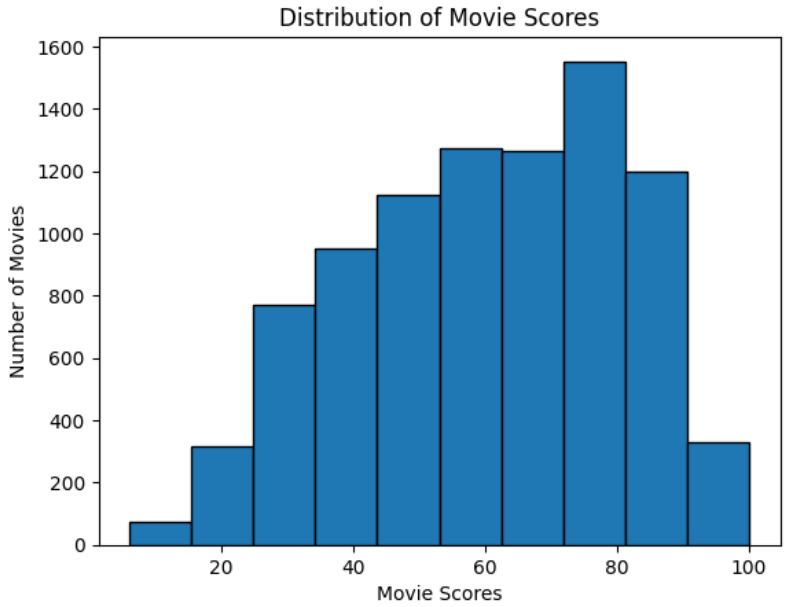


Obr. 1 : Percentuálne rozdelenie filmových ratingov v skúmanom datasete

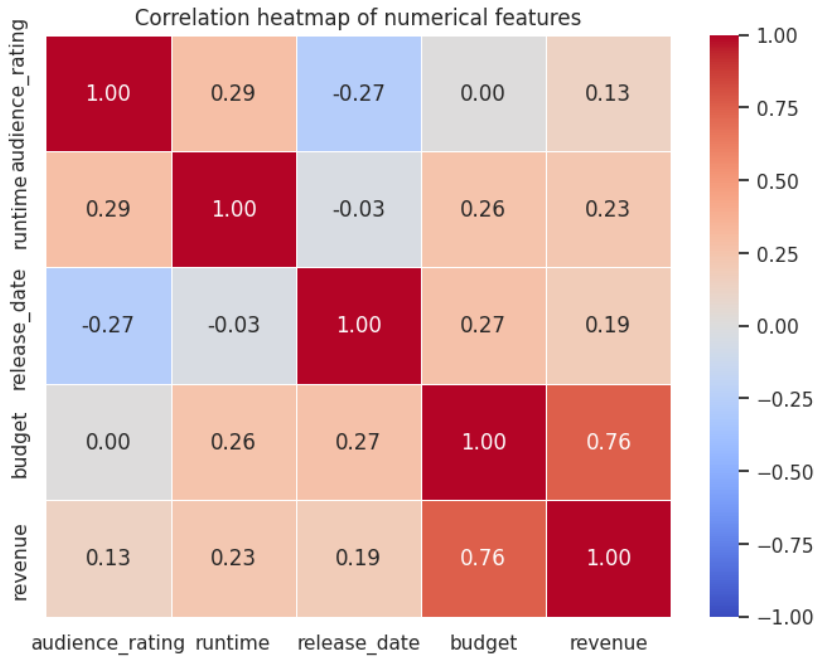


Obr. 2: Počet filmov patriacich do rôznych žánrov - jeden film väčšinou patrí do viacerých rozličných žánrov

Veľká časť filmov patrí do žánrov dráma a komédia. Naopak, každý zo žánrov Sports & Fitness, Cult Movies, Faith & Spirituality, Gay & Lesbian obsahuje menej než 100 filmov.



Obr. 3: Distribúcia hodnotení filmov v datasete



Obr. 4: Korelačná matica numerických atribútov

Pozorujeme pozitívnu koreláciu medzi hodnotením a dĺžkou filmu, či výnosom. Negatívna korelácia medzi rokom vydania a hodnotením naznačuje, že staršie filmy sú medzi divákmi obľúbenejšie. Korelácia medzi budgetom a hodnotením je nulová, čo však môže byť spôsobené tým, že tento atribút je v použitom datasete pre veľa filmov nula.

Použité nástroje, metódy a technické výzvy

Keďže medzi atribútmi a výsledným hodnotením môžu existovať komplikované vzťahy, ktoré je ťažké odpozorovať a logicky zdôvodniť, použili sme na vytváranie predikcií neurónovú sieť. S využitím nástrojov z knižnice TensorFlow sme vyskúšali dva modely:

- jeden bez skrytých vrstiev, so sigmoidou ako aktivačnou funkciou na výstupnej vrstve,
- druhý s jednou skrytou vrstvou pozostávajúcou z dvoch neurónov, s ReLU na skrytej vrstve a sigmoidou na výstupnej.

Testovali sme aj modely s väčším počtom skrytých neurónov, tieto najjednoduchšie však dávali v tomto prípade najlepšie výsledky.

Ako chybovú funkciu sme použili mean squared error (MSE). Táto funkcia viac penalizuje veľké chyby, zatiaľ čo malé nepresnosti nerobia taký problém. Pri konečnom porovnaní úspešnosti modelov sme sa pozerali na odmocninu z tejto hodnoty, teda na root mean squared error (RMSE). Toto číslo je v pôvodnej škále, je preto jednoduchšie zhodnotiť presnosť predikcií.

Pri každom trénovaní modelu sme sa snažili vyladiť rýchlosť učenia. Vyskúšali sme oba modely natrénovať s rôznymi hodnotami learning rate a vybrali sme model, ktorý dosiahol najlepšiu presnosť.

V oboch modeloch bola použitá sigmoida na výstupnej vrstve. Táto funkcia vracia hodnoty medzi 0 a 1, preto sme hodnotenie, ktoré je z rozsahu 0 až 100, preškálovali vydelením 100 a tieto hodnoty sme použili ako očakávaný výstup pri trénovaní modelu. Predikcie, ktoré z takto natrénovaného modelu dostaneme, potom jednoducho prenásobíme 100, aby sme ich vrátili do pôvodného rozsahu.

Mnohé z použitých atribútov nie sú číselné, napríklad máme pre každý film zoznam hercov, ktorí v ňom hrajú. Aby sme takéto atribúty mohli použiť ako vstupnú maticu pre neurónovú sieť, použili sme MultiLabelBinarizer a OneHotEncoder z knižnice scikit-learn. V prípade hercov takto vznikla matica, kde riadky reprezentujú filmy, stĺpce hercov, 1 znamená, že vo filme hrá daný herec, inak je tam 0. Keďže máme veľké množstvo rôznych hercov, matica spôsobovala problémy s pamäťou. Tie sme vyriešili tak, že sme takéto objekty reprezentovali ako riedke matice, pretože typicky obsahovali veľa núl.

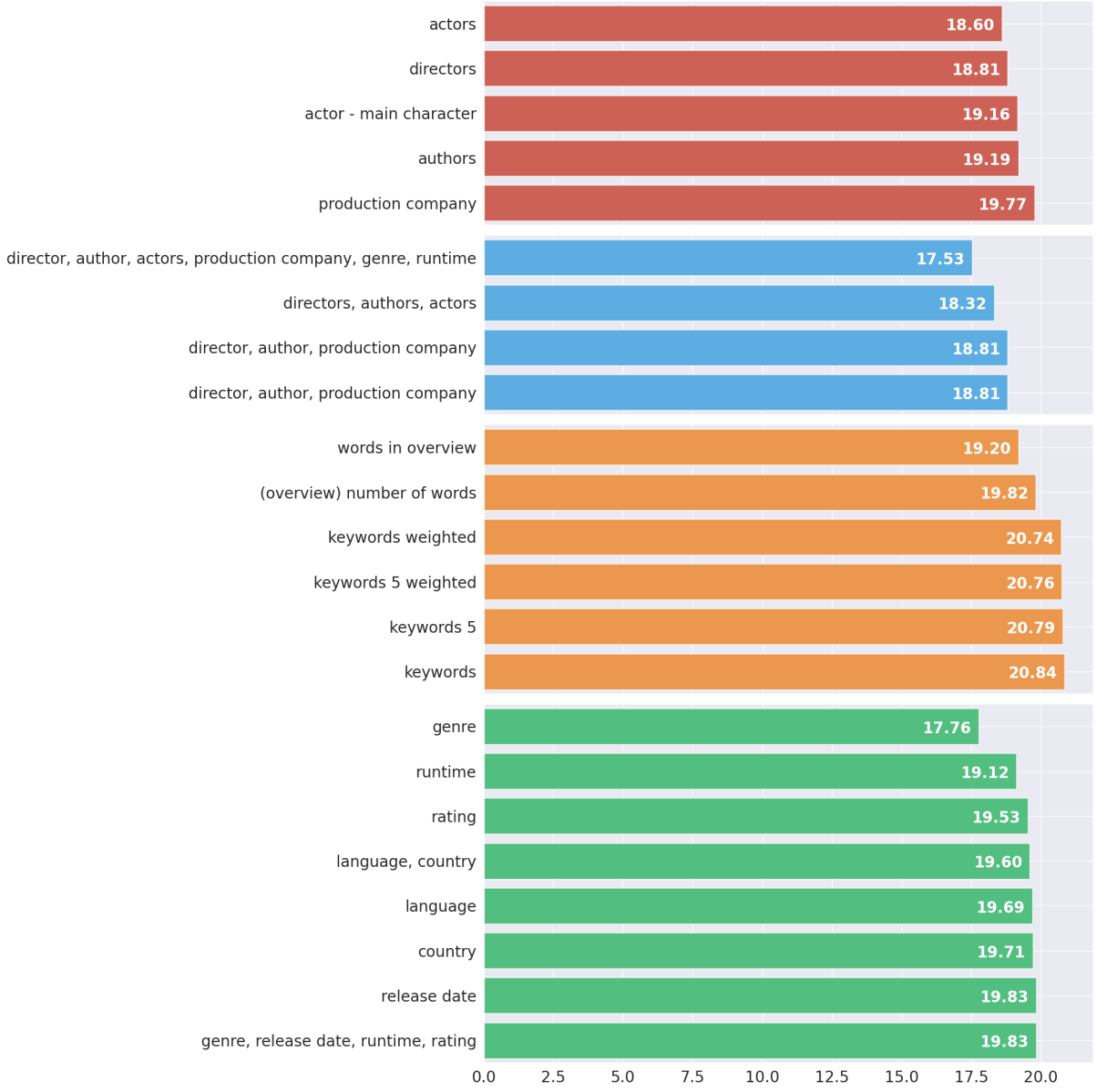
Hlavná výzva bola nájdenie správneho modelu. Skúsili sme použiť aj iné modely napr. lineárna regresia. Zistili sme, že tá nefunguje dobre na našom datasete, pravdepodobne kvôli nelineárnym vzťahom medzi atribútmi.

Výsledky našej analýzy

Pre získanie výsledkov sme porovnali RMSE z modelov s rozličnými kombináciami atribútov, pričom atribúty boli rozdelené na kategórie:

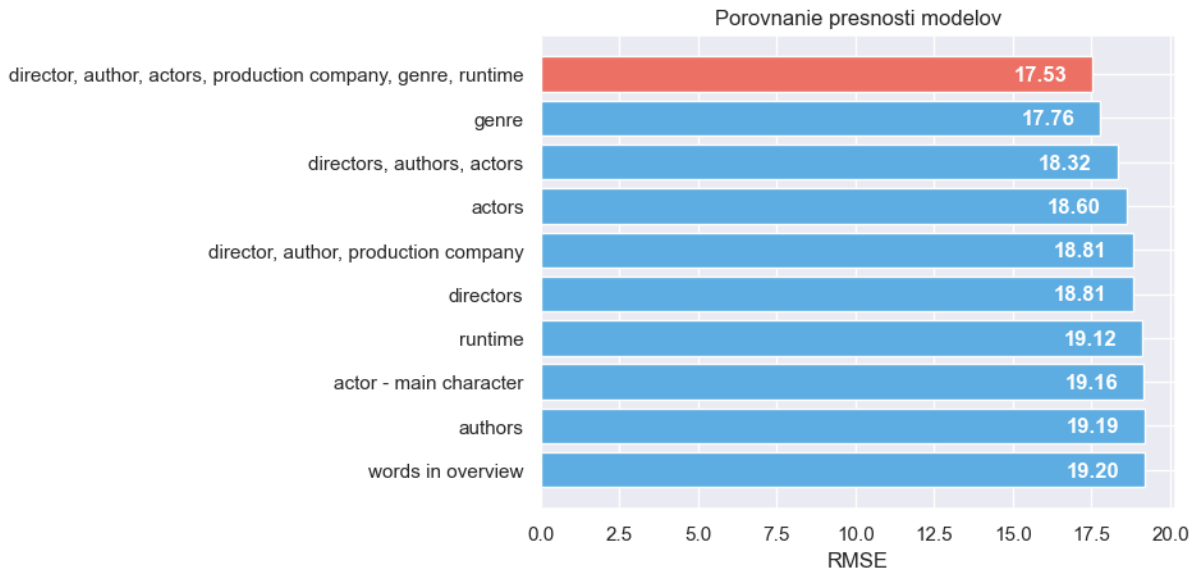
- herci
- režisér, autor, produkčná spoločnosť
- slová v obsahu, kľúčové slová zápletky
- žáner, rok vydania, trvanie, rating (pre aké vekové skupiny sú vhodné), jazyk a krajina

Náš výsledný graf vyzerá takto:



Obr. 5: Porovnanie úspešností modelov s rôznymi atribútmi

Zistili sme, že najlepšie kombinácie sú nasledovné:



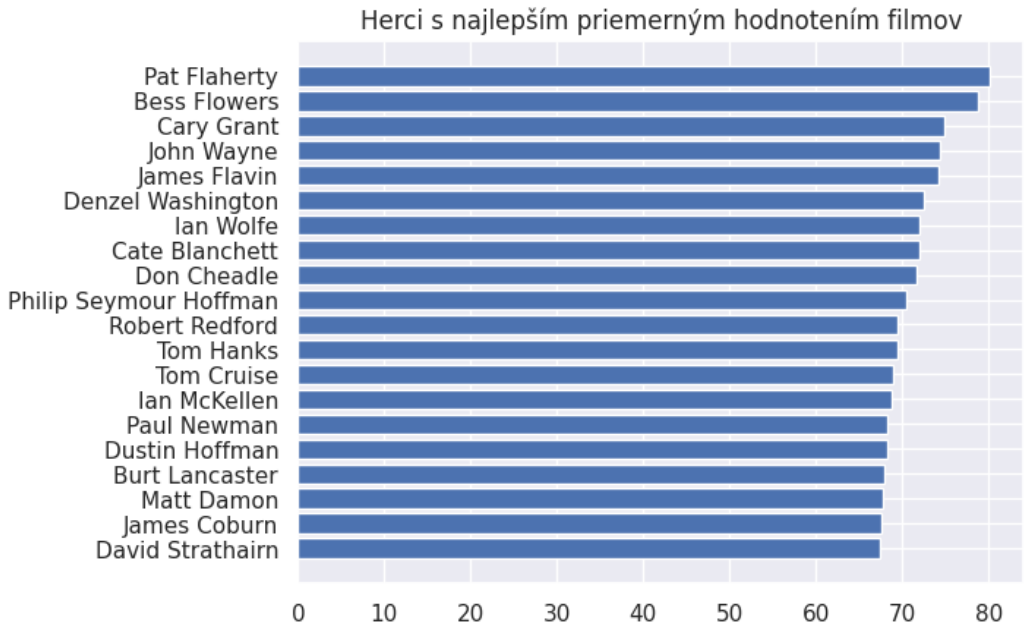
Obr. 6: Porovnanie presnosti najlepších modelov

Teda najpresnejšie predikcie vznikli pomocou kombinácie atribútov režisér, autor, obsadenie, produkčná spoločnosť, žáner a trvanie. Tento model používal na predikcie najviac informácii. Na druhom mieste skončil model s žánrami. Tretie najlepšie predikcie mal model s atribútmi režisér, autori, herci.

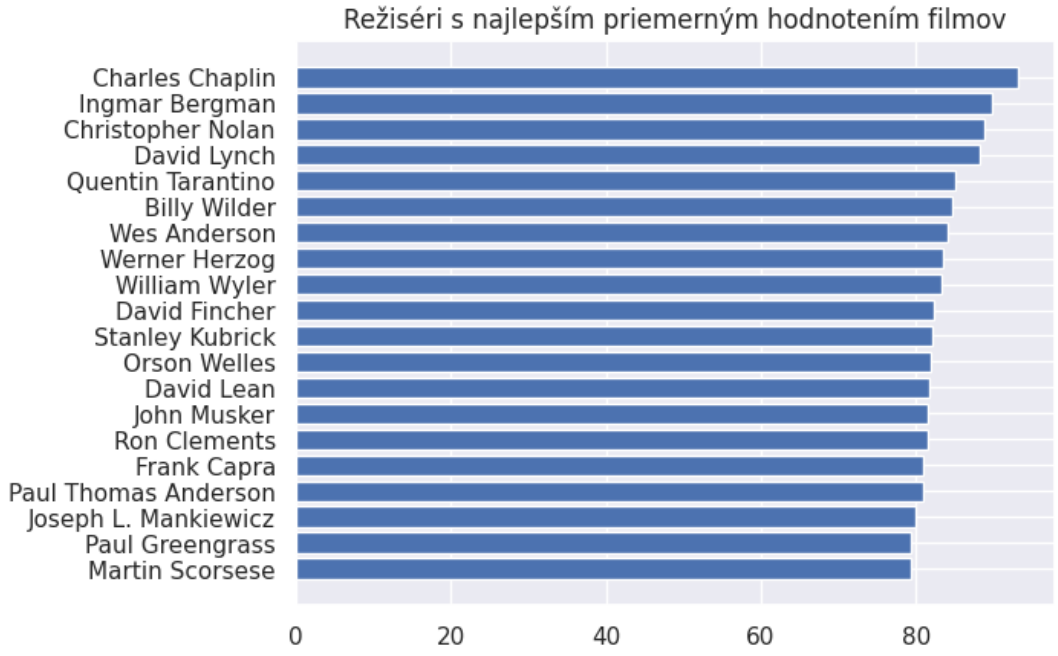
Z týchto výsledkov môžeme predpokladať, že herci majú na úspešnosť filmu väčší vplyv, než režisér, autor, či štúdio. Lepšie predikcie boli vytvorené pomocou slov v popise filmu, než podľa kľúčových slov, čo však môže byť spôsobené tým, že ich bolo pre každý film viac. Podobne to môže platiť aj pre spomínaných hercov. Mierne lepšie výsledky sme z kľúčových slov dosiahli pridaním váh - kľúčové slová na prvých miestach, teda dôležitejšie, dostali vyššiu váhu.

Je však potrebné podotknúť, že pred rozdelením na testovacie a trénovacie dáta dataset permutujeme a teda výsledky nemusia byť vždy identické.

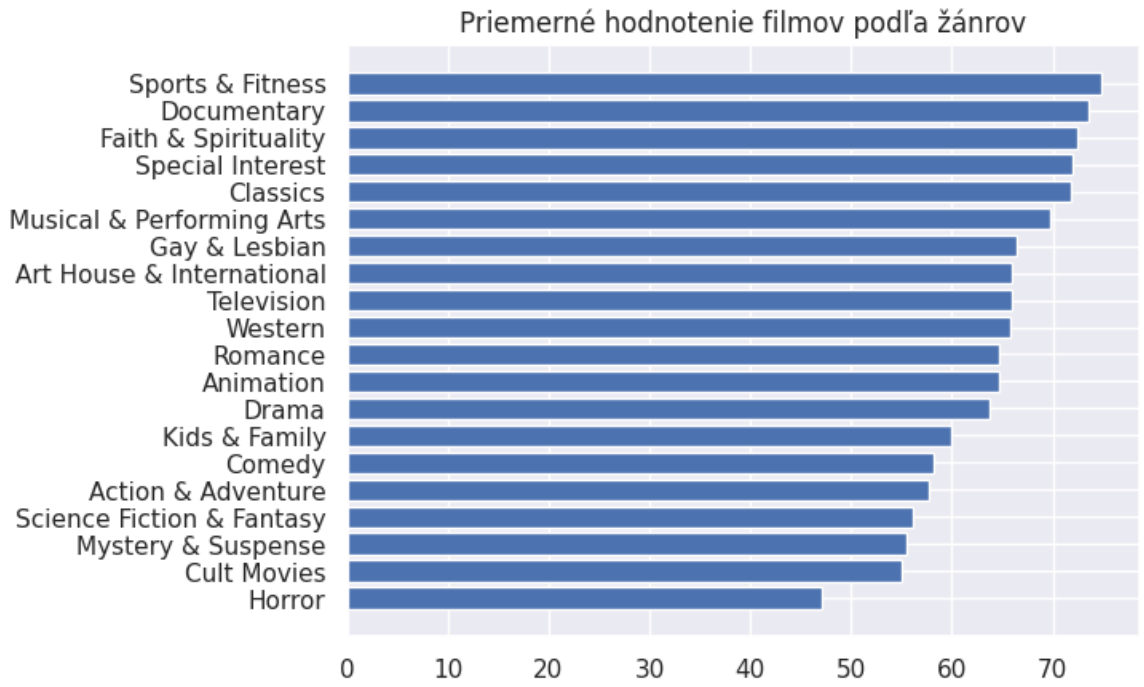
Vytvorili sme aj grafy o tom, ktorí herci, režiséri, žánre a slová majú najlepšie, prípadne najhoršie hodnotenie. Keďže pre niektorých hercov či režisérov môžeme mať napríklad len jeden film, ktorý mal práve vysoké hodnotenie, porovnávali sme len tých, s ktorými sa v našom datasete spája aspoň zvolený počet filmov.



Obr. 7: Porovnanie hercov, ktorí majú najlepšie priemerné hodnotenie filmov, v ktorých hrajú



Obr. 8: Porovnanie režisérov

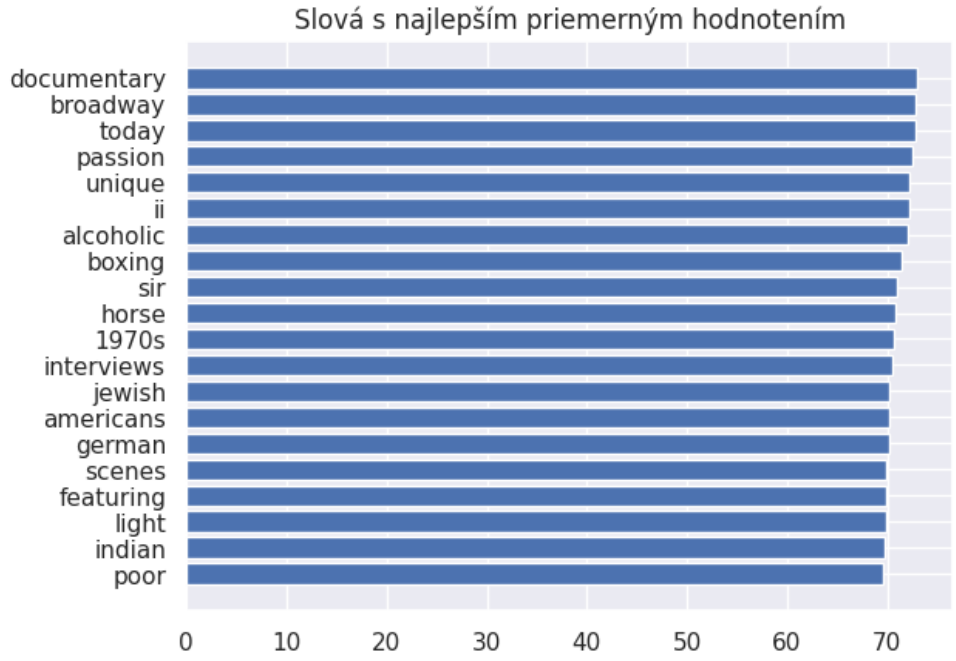


Obr. 9: Porovnanie hodnotení filmov podľa žánrov

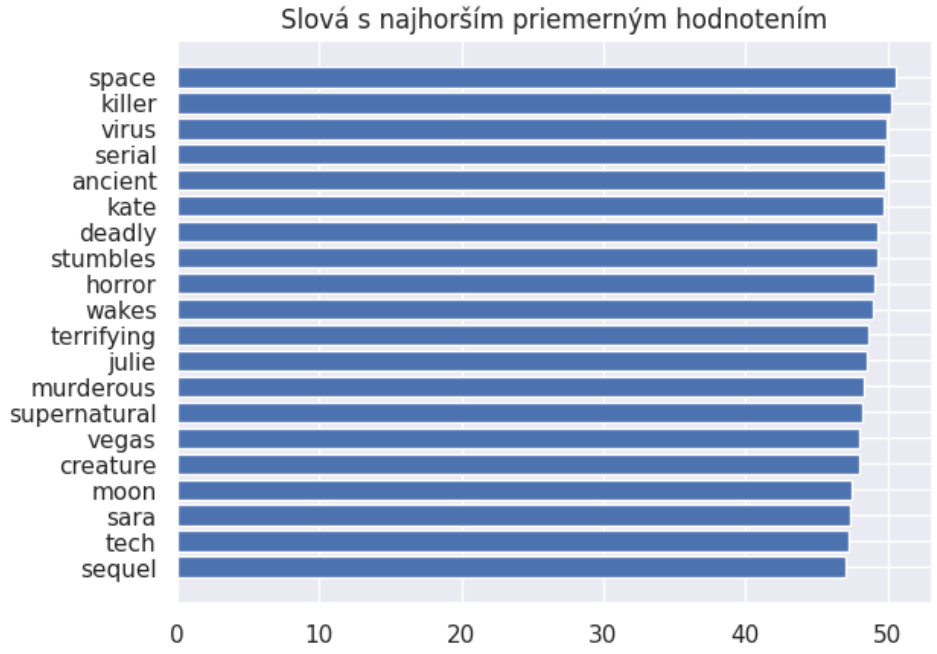
Treba podotknúť, že z niektorých žánrov máme v datasete málo filmov, preto ich priemerné hodnotenie môže mať nižšiu výpovednú hodnotu. Rozdiel v priemernom hodnotení medzi žánrami by mohol byť spôsobený aj odlišným publikom - diváci niektorých žánrov môžu byť náročnejší a dávať systematicky nižšie hodnotenia.

Vidíme, že najlepšie hodnotenia majú filmy o športe a dokumentárne, naopak najhoršie hodnotenia majú horory. To sa prejavilo aj na slovách s najlepším a najhorším priemerným hodnotením.

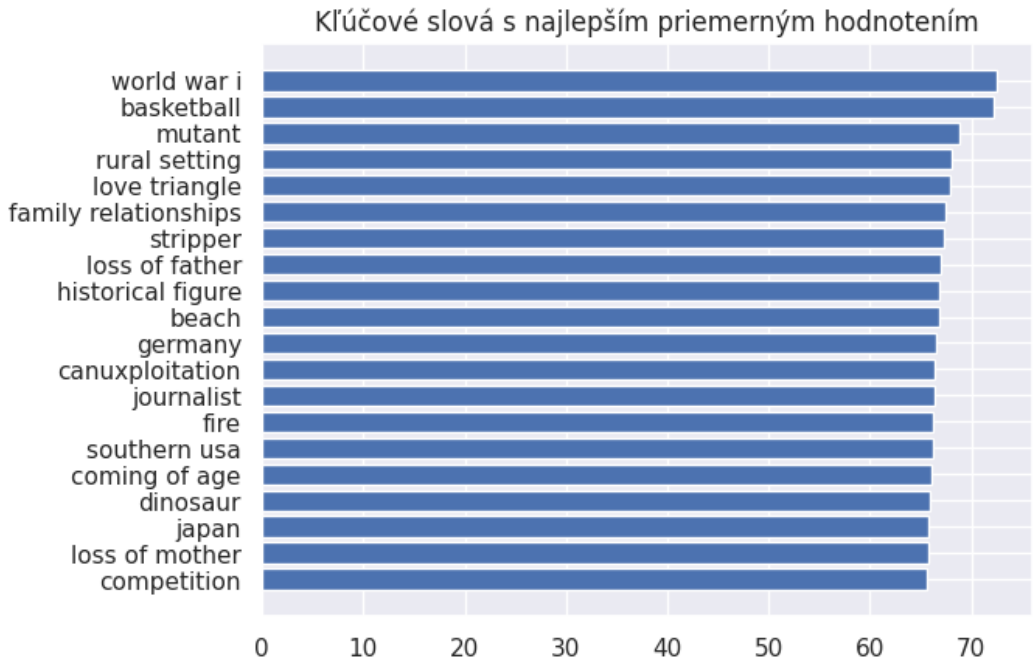
Kľúčové slová sú označenia, ktoré boli filmom priradené podľa témy - môžu to byť aj slovné spojenia. Slová sú tie, ktoré sa vyskytli v popise filmu. Aj tu sme použili podmienku na minimálny počet výskytov, aby sme nedostali len slová z filmu s najlepším hodnotením.



Obr. 10: Porovnanie slov v obsahu s najlepším hodnotením



Obr. 11: Porovnanie slov v obsahu s najhorším hodnotením



Obr. 12: Porovnanie kľúčových slov s najlepším hodnotením

Obr. 13: Porovnanie kľúčových slov s najhorším hodnotením