**Сташа Грајић ИН54/2021**

**Теодора Јовић ИН58/2021**

Предвиђање жанра филма коришћењем неуронских мрежа на основу различитих типова податка  
(Предикција жанра)

Извештај за практично истраживање

# Увод

## Предмет истраживања

Предмет овог истраживања је ефикасност предикције жанра филма на основу нумеричких података (трајање филма, буџета, број гласова, оцене гледалаца) и наслова филма.

У филмским индустријама поставља се питање који жанр филма је уствари најисплатљивији за продукцију. На основу нумеричких података може се проценити који жанр филма покренути у продукцију у односу на буџет којим нека филмска продукција располаже. На основу броја гласова може закључити колико је тај жанр уствари популаран међу гледаоцима, док на основу оцена гледаоца може закључити колика је успешност тог жанра. На крају уколико постоји недоумица колико дуго би филм одређеног жанра требао да траје,такву предикцију можемо одредити на основу трајања филмова.

Прављењем сајта или апликације за гледање филмова, потребно је расподелити на најефикаснији начин филмове из базе у жанрове којима они припадају.

Поставља се питање како на основу само наслова филма ово може да се постигне.

Ми смо се бавили баш овим проблемима у потрази за најтачнијом предикцијом.

## Циљеви истраживања

Циљ истраживања је одговор на постављена питања и омогућавање генералној филмској индустрији да прескочи пар корака у одабиру жанрова користећи модел који смо развили и истренирали.

Генерални циљ искључиво нашег модела је да постигне високу тачност у предвиђању жанра филма на тест скупу података. Након добијања ових резултата можемо спровести анализу које оцене, трајање филма и наслови су највише допринели предвиђању одређеног жанра.

## Задаци истраживања

* 1. Препроцесирање података
  2. Анализа дистрибуције жанрова
  3. Стандардизација карактеристика
  4. Изградња и обука модела
  5. Предикција жанрова
  6. Анализа тачности модела
  7. Визуелизација резултата

## Очекивани резултати истраживања

Очекујемо да ћемо успешно развити модел,који на основу вишеслојних и рекурентних неуронских мрежа,предвиђа жанрове филма на основу произвољних примера.

Такође се надамо да ће модел давати што боље резултате за примере који нису идентични узорцима из скупа података. Ово је веома важно за предвиђање жанра на основу наслова, јер уколико је пример нека реч која није садржана ни у једном наслову скупа података, модел мора бити способан да предвиди жанр на основу свега онога што је успео да научи. Ако модел да резултат који није ни приближно сличан стварности,односно нашој претпоставци какав жанр би требао бити, модел није добро истрениран и постоје пропусти. Смањивањем пропуста требали бисмо да добијемо очекиване резултате.

Такође очекујемо да модел неће давати исте резултате за потпуно различите примере и да ће имати велику тачности при предикцији.

На крају очекујемо да ће модел узимати у обзир податке који су мање дистрибуирани у скуповима података и да се неће ослањати искључиво на жанрове који имају велику дистрибутивност.

# Методологија

## Коришћени подаци

Скуп података најбољих 100 филмова свих времена снимљених између 2010. и 2016. године на ИМДб.

Из овог скупа података коришћене су следеће колоне:

1. Оцене
2. Број гласова
3. Жанр 1
4. Жанр 2
5. Жанр 3
6. Буџет
7. Време трајања

Скуп података са сајта за претрагу и препоруку филмова корисницима, последњи пут ажуриран 2018. године.

Из овог скупа података коришћене су следеће колоне:

1. Наслов
2. Жанр

## Претходна истраживања других особа над коришћеним подацима

<https://data.world/saipranav/top-rated-english-movies-of-this-decade-from-imdb>

<https://grouplens.org/datasets/movielens/>

## Методе истраживања

**1.Рад са нумерички подацима**

**Учитавање и уређивање података**

Узимање података из одређених скупова података.

Ради лакшег руковања подацима прво сам променила имена колона на српски језик.

Жанрови у скупу података су представљени у три колоне. Пошто ми је потребна једна ознака за предикцију, спојила сам их у нову колону названу “жанр”, где су жанрови одвојени /. Мапирање жанрова је неопходно јер АНН модел захтева нумеричке податке за предикцију.

Оцене у скупу података су децималног типа, из тог разлога је потребно то и навести. Након тога извршила сам категоризацију оцена у категорије- "одличан", добар", "просечан", "лош" и додала у нову колону ради прегледнијег представљања података.

Број гласова у скупу података је представљен као 400 000, а ја тај начин репрезентовања података не могу да искористим у моделу и морам да избацим зарез из броја и да поставим његов тип на цео број.

Буџет у скупу података је представљен $400 000 000 и морам прво да избацим све $ и зарезе из податка и након тога да узмем све бројеве из податка који ће бити децималног типа.

Време трајања у скупу података је представљено као 134 мин, прво сам узела само бројеве из податка.Затим сам уклонила све недостајуће податке, јер је ова колона садржала неодстајуће вредности (НаН) и коначно сам овај податак поставила на тип целог броја. Након ове припреме података категоризовала сам времена трајања у категорије “кратак”, ”средњи”, ”дугачак” и убацила у нову колону намапираних података.

Након свих ових корака припреме података за тренирање,веома битан корак је ресетовање индекса новог скупа података.

**Балансирање података**

Балансирање података ми је неопходно из разлога што је скуп података веома кратак. Балансирањем се постиже да сваки жанр има приближно исти број узорака.

Креирала сам листу која садржи све жанрове из колоне ознака. Користећи бројач из библиотеке колекција пребројала сам колико пута се сваки жанр појављује.

Затим сам пронашла жанр који се највише пута појављује међу подацима. Изменила сам скуп података тако да сваки жанр има приближнији број узорака жанру са највише узорака.

**Графикони дистрибутивности**

Анализирала сам податке пре и после балансирања помоћу графикона како бих увидела промене у узорцима и колико је успешно извршено балансирање. Моја слободна процена је да блансирање није направило претерану измену у дистрибуцији, али је значајно побољшало шансе да се мање дистрибутивни жанрови нађу у предикцији.

**Припрема података**

При првом формирању модела користила сам идентични начин припреме податка као што је рађено на осмим вежбама. Међутим таква припрема података ми је давала исте резултате за различите предикције. Након тога сам у припрему убацила библиотеку за бинаризацију ознака у матрицу,што такође није давало очекиване резултате.

Кључна грешка у мојим претходним покушајима припреме података је та што нисам додала стандаризацију података. Ако подаци имају веома различите опсеге, оне са већим опсегом могу доминирати над онима са мањим опсегом у процесу учења модела,па на тај начин осигуравам да сви подаци имају једнак допринос моделу. Овакав приступ је неопходна у мом моделу који пре свега има јако мало података и где ти подаци нису равномерно дистрибуирани.

**Обучавање**

Модел сам креирала секвенцијално,односно креиран је тако да ће сваки слој бити додат један након другог.

Користила сам “Релу” функцију јер је на мом моделу давала најпрецизније предикције.

У првом моделу сам користила три слоја и прецизност је била скоро па непостојећа. Разлог из ког сам се одлучила да додам више слојева јесте тај да омогућим мом моделу да трансформише податке што више пута. Код мањег броја слојева може се десити да модел не научи неке сложеније везе између података и жанрова, већ се задржи на неким основним шаблонима за препознавање жанра. Много је значајно да се модел развија и да учи проласком кроз више слојева, као што док ми учимо не можемо одмах, након првог прегледа кода, схватити све што се дешава. У сваком проласку кроз код запажамо и схватамо нове ствари које нам раније нису биле јасне.

**Евалуација**

Да ли је могуће модел научити да савршено предвиђа? Као што људи имају своје предности и мане, тако и сваки код има своје и никад не можемо знати кад покренемо свој модел колико ће он бити тачан.

То нас доводи до самог губитка,односно мере која описује колико је добро или лоше модел предвиђао вредности током тренинга, тј. колико су предвиђања модела одступала од стварних вредности.

И до тачности,односно мере која описује колико је модел тачан у својим предвиђањима, тј. проценат тачних предвиђања које је модел направио у односу на укупан број узорака.

Губитке и тачности сам представила помоћу графикона,како бих могла анализирати ове ствари.

**Валидација(предикција)**

На самом крају је потребно проверити рад модела који сам имплеменирала.

То сам урадила на тај начин што сам осигурала да се све вероватноће мање од 0,5 које постају 0, не узимају у обзир,односно жанрови са том вероватноћом неће припадати предикцији,док све вероватноће веће од 0,5 постају 1 и оне ће припадати предикцији.

Шта значи 1 и 0? 1 представља истину, док 0 представља неистину.

Код примера које сам користила је битно да напоменем да сам жанрове декодира из њихових ознака пресликавања које су биле нумеричке, назад у оригиналне ознаке.

**2.Рад са текстуалним подацима**

**Учитавање података**

Из одређеног скупа података узимамо одређене колоне за називе филмова и њихове жанрове.

Следећи кораци су:

- приказујемо првих пар редова из скупа података

- смањујемо скуп података на 1500 јер оригиналан има преко 80 000 података

- уклањамо недостајуће вредности

- делимо жанрове (који су се оригинално налазили у једној колони одвојени '|') у листу жанрова

Након свих ових корака приказује се график са жанровима.

**Балансирање података**

На почетку рачунамо колико пута се појављује сваки жанр у оригиналном скупу података и налазимо максималан број филмова у било којем жанру.

Током балансирања уколико неки жанр има мање филмова од максималног, узимамо узорке са понављањем да би смо дошли до тог броја.

На крају спајамо све узорке у један скуп података и потом опет приказујемо график за нову расподелу жанрова.

**Токенизација**

Креирамо токенизер који ће користити највише 5000 речи (разбијамо текстове у мање, токене) и конвертоваће све у мала слова и делити текстове по размаку, и користимо функцију из библиотеке за класификација текста, превођење и анализе секвенци како би све секвенце биле исте дужине.

Ово омогућава моделу да интерпртира и процесира текстуалне податке.

Претварамо текстуалне податке у нумеричке секвенце које модел може да користи за учење.

Затим следи подела података која је исто урађена као и код колегинице из истих разлога.

**Креирање и тренирање модела - РНН (рекурентна неуронска мрежа)**

На почетку, као и до сада, креирамо секвенцијални модел и додајемо слојеве:

* “Embedding” слој - омогућава претварање сваке речи у речнику у густи вектор фиксне величине (у нашем случају 64). Овај слој омогућава моделу да учи значење речи и њихове међусобне односе - боље разумевање контекста и семантике филмова
* “SpatialDropout1D” слој - помаже у спречавању прекомерног прилагођавања, тако што насумично искључује делове неуронске мреже током тренирања
* “Bidirectional LSTM” слој - овај слој помаже моделу да ухвати контекстуалне информације из целокупне секвенце наслова
* “Dense” слој - активациона функција је “танх” која враћа вредности у опсегу -1 до 1, може помоћи у случајевима када су циљне вредности нормализоване или када је потребно представити позитивне и негативне корелације. Овај слој омогућава моделу да комбинира информације из претходних слојева и генерише коначне предикције жанрова.

Затим компајлирамо модел са функцијом која мери просечну квадратну разлику између предвиђених вредности модела и стварних вредности из скупа података као губитком и адам као оптимизатором.

Користимо заустављање тренинга ако се валидациони губитак не смањи у току епоха, и онда тренирамо модел и приказујемо графиконе губитка и тачности током тренинга за тренинг и валидациони скуп.

**Предикција**

Дефинишемо неколико наслов за тестирање и сваки наслов конвертујемо у нумеричку секвенцу ( користећи речник токенизера ), затим се секвенца прилагођава да буде исте дужине као секвенце коришћене за тренирање. Модел генерише вероватноће за сваки жанр на основу улазне секвенце. Одабиру се топ 3 жанра са највећим вероватноћама. Укључују се само жанрови са вероватноћом већом од 0.2. На крају приказујемо резултате за сваки наслов.

# Резултати

## Приказ резултата

Резултате предикције приказујемо на два начина:

* 1. Излаз на конзоли

Током имплементације модела пратили смо дешавање онога што радимо путем конзоле тако што смо исписивали:

* + - Учитан скуп података
    - Коначан скуп података
    - Дистрибуцију жанрова пре балансирања
    - Дистрибуцију жанрова након балансирања
    - Предикција над примерима
  1. Графикони
     + График дистрибуције жанрова пре балансирања
     + График дистрибуције жанрова након балансирања
     + График дистрибуције жанрова у поређењу са бројем узорака
     + График губитка при тренингу и валидацији
     + График тачности при тренингу и валидацији

## Тумачење резултата

* Излаз на конзоли

Да бисмо били сигурни да су резултати тестних примера довољно прецизни и очекивани,потребно је да прво протумачимо наш скуп података.

Након што смо у потпуности разумели како наш скуп податка функционише и какве он то податке садржи,можемо бити сигурни у тачности резултата које нам је модел пружио.Што је више различитих жанрова заступљено,то можемо бити сигурнији да је само балансирање података успешно извршено и да није дошло до губљења узорака на дну дистрибуције.

* Графикони

Да бисмо били сигурнији у испуњеност балансирања података,представили смо графиконе који приказују промену у дистрибутивности узорака.Самом упоредбом графикона пре и после балансирања,само једним погледом је уочљива огромна промена.

Анализирали смо тачност помоћу графикона који показују колико узорака је изгубљено,а колико узорака је тачно за предикцију.Овакава анализа нам је дала увид у то колико смо стварно добро истренирали наш модел.

# Закључак

## Анализа испуњења циљева истраживања

Модел је испунио наше циљеве истраживања и дао поприлично добру тачност над примерима предикције, при тренирању на скупу података.

## Анализа остварења очекиваних резултата истраживања

Модел успешно предвиђа жанрове на основу тест примера.

Подаци су успешно избалансирани,иако је наше очекивање било да ће много више бити избалансирани,морали смо узети у обзир да код скупа податка који је коришћен подаци су веома неравномерно дистрибуирани.Због овог нашег избора које ћемо скупова података користити,наш модел је уствари успео избалансирати податке и боље него што смо очекивали. Овакав приступ подацима је нама дао много прецизније и тачније резултате.

Модел је успео да предвиди жанрове и оних података који нису били идентични узорцима, иако се при овом предвиђању већински ослањао на више дистрибуиране узорке, ипак је узимао у обзир и оне мање дистрибуиране.

Битно је напоменути да су резултати које је давао били они које смо и ми слободном претпоставком приметили да би требали бити. Овим понашањем модел нам је показао да смо му дали довољно слојева на основу којих може да проширује своје знање и да смо успели да га истренирамо да везује сличне узорке и да прави обрасце на основу којих може да даје предикцију.

Веома је важно напоменути да претходни модели нису давали тачне предикције. Анализом понашања самог модела и преправљањем пропуста могли смо да дођемо до закључка који делови модела требају да се промене. Огромну улогу су играли које слојеве и колико слојева је коришћено при тренирању, која количина скупа података је коришћена у коначном скупу података, колико епоха је укључено у тренирање, какви примери су имплементирани и који постотак предикованих жанрова је прихваћен.

Проблем на који смо наилазили приликом прављења модела је тај да је модел давао исте резултате за различите примере, а такво понашање нисмо очекивали и такво понашање није било пожељно. Тај проблем смо решили преправљањем горе наведених пропуста.

## Могућности за примену истраживања у пракси

Сматрамо да овакав модел може да се искористи у филмским индустријама и међу свим људима који желе да се баве сајтовима за гледање филмова.

Продуценти пре самог снимања филмова могу на основу нашег модела да одреде какав жанр им је најиплатљивији.

Сваку реч која обичном човеку падне на памет се може сврстати у неки жанр филма, а наш модел управо то и омогућава.

## Идеје за побољшање и разраду истраживања

Сваки модел има мане и сваки модел се може побољшати.Конкретно наш модел би могао бити много тачнији у предикцији уколико бисмо користили неке друге скупове података.Сматрамо да први скуп података са само 120 редова није довољан за прецизну предикцији,па бисмо у будућности посветили много више времена у анализи скупова података пре одабира које ћемо корисити.Наравно пошто први пут правимо овакав модел и први пут се генерално бавимо овакавим стварима,нисмо могли знати на какве проблеме ћемо наићи.Сама тема рада са неуронским мрежама је много комплексна и за нас је тотално ново подручије и имамо још много тога да научимо да бисмо успешно правили моделе са великом тачношћу на разним скуповима података.

Нешто што бисмо испробале у будућим пројектима је рад са различитим типовима слојева,јер свака неуронска мрежа даје различите резултате на основу слојева које користимо.

Побољшање које би вероватно допринело тачнијој предикцији је додавање више скупова података,јер кад би се спојила минимално 3,4 скупа података у један,сам модел би имао много више узорака за тренирање.

На овакав модел би било занимљиво додати и обрнуто истраживање.На пример,да на основу унетог жанра он предвиђа какав наслов би могао да се искористи,или колико је тај жанр исплатљив за продукцију.

Овај модел би могао чак да сем вишеслојних и рекурентних неуронских мрежа садржи и конвулационе неуронске мреже.Помоћу конвулационих мрежа предикција жанра би се могла радити и на основу слика.Такав модел би обухватио и рад са текстом и рад са сликама и био би много сложенији и потпуњенији.

# Литература

1. http://www.acs.uns.ac.rs/sr/node/237/7216640
2. https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/
3. ...