АНТИПЛАГИАТ
ТВОРИТЕ СОБСТВЕННЫМ УМОМ

Отчет о проверке на заимствования №1

Автор: МФТИ admin@phystech.edu / ID: 211
Проверяющий: (admin@phystech.edu / ID: 211)
Организация: Московский физико-технический институт

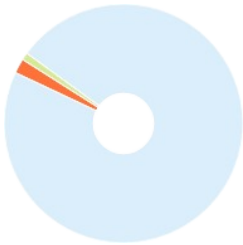
Отчет предоставлен сервисом «Антиплагиат» - http://mipt.antiplagiat.ru

ИНФОРМАЦИЯ О ДОКУМЕНТЕ

№ документа: 7603
Начало загрузки: 24.06.2019 10:14:43
Длительность загрузки: 00:01:02
Имя исходного файла: main.pdf
Размер текста: 5163 кБ
Символов в тексте: 38431
Слов в тексте: 4692
Число предложений: 246
Method of text extraction: OCR

ИНФОРМАЦИЯ ОБ ОТЧЕТЕ

Последний готовый отчет (ред.)
Начало проверки: 24.06.2019 10:15:46
Длительность проверки: 00:00:07
Комментарии: не указано
Модули поиска: Сводная коллекция ЭБС, Коллекция РГБ, Цитирование, Коллекция eLIBRARY.RU, Модуль поиска Интернет, Модуль поиска "МФТИ", Модуль поиска перефразирований eLIBRARY.RU, Модуль поиска перефразирований Интернет, Модуль поиска общеупотребительных выражений, Кольцо вузов



ЗАИМСТВОВАНИЯ	ЦИТИРОВАНИЯ	ОРИГИНАЛЬНОСТЬ
1,62%	0,88%	97,5%

Заимствования — доля всех найденных текстовых пересечений, за исключением тех, которые система отнесла к цитированиям, по отношению к общему объему документа.
Цитирования — доля текстовых пересечений, которые не являются авторскими, но система посчитала их использование корректным, по отношению к общему объему документа. Сюда относятся оформленные по ГОСТу цитаты; общеупотребительные выражения; фрагменты текста, найденные в источниках из коллекций нормативно-правовой документации.
Текстовое пересечение — фрагмент текста проверяемого документа, совпадающий или почти совпадающий с фрагментом текста источника.
Источник — документ, проиндексированный в системе и содержащийся в модуле поиска, по которому проводится проверка.
Оригинальность — доля фрагментов текста проверяемого документа, не обнаруженных ни в одном источнике, по которым шла проверка, по отношению к общему объему документа.
Заимствования, цитирования и оригинальность являются отдельными показателями и в сумме дают 100%, что соответствует всему тексту проверяемого документа.
Обращаем Ваше внимание, что система находит текстовые пересечения проверяемого документа с проиндексированными в системе текстовыми источниками. При этом система является вспомогательным инструментом, определение корректности и правомерности заимствований или цитирований, а также авторства текстовых фрагментов проверяемого документа остается в компетенции проверяющего.

№	Доля в отчете	Доля в тексте	Источник	Ссылка	Актуален на	Модуль поиска	Блоков в отчете	Блоков в тексте
[01]	0,62%	0,62%	Impact of URI Canonicalization on Meme..	http://arxiv.org	22 Мар 2018	Модуль поиска Интернет	2	2
[02]	0,23%	0,4%	https://esu.citis.ru/dissertation/MWTACQ.	https://esu.citis.ru	10 Мая 2018	Модуль поиска Интернет	2	2
[03]	0,18%	0,39%	Буслаев Павел Ильич Thesis_Buslaev.pdf	не указано	07 Июн 2018	Модуль поиска "МФТИ"	1	2
[04]	0%	0,36%	Емельянов, Алексей Владимирович Тр...	http://dlib.rsl.ru	14 Июн 2019	Коллекция РГБ	0	3
[05]	0,23%	0,32%	Downloadable Full Text	http://dspace.mit.edu	13 Окт 2018	Модуль поиска Интернет	1	2
[06]	0,11%	0,31%	moshkov_n_e_programma-klassifikacii-t...	не указано	05 Июн 2017	Кольцо вузов	1	2
[07]	0%	0,3%	69955	http://e.lanbook.com	09 Мар 2016	Сводная коллекция ЭБС	0	2
[08]	0,05%	0,27%	Ковалев, Роман Александрович Метод...	http://dlib.rsl.ru	30 Мая 2019	Коллекция РГБ	1	2
[09]	0%	0,21%	111947	http://biblioclub.ru	14 Апр 2016	Сводная коллекция ЭБС	0	1
[10]	0%	0,21%	Факультет мировой экономики и мир...	не указано	21 Мая 2012	Кольцо вузов	0	1
[11]	0,21%	0,21%	vrabie_i_v_razvitie-modeley-opisaniya-ne.	не указано	22 Мая 2019	Кольцо вузов	1	1
[12]	0%	0,17%	Multimedia on Mobile Devices 2009.	http://elibrary.ru	28 Авг 2014	Коллекция eLIBRARY.RU	0	1
[13]	0%	0,17%	shcherbakov_i_a_sravnitelnyy-analiz-algo.	не указано	05 Июн 2017	Кольцо вузов	0	1
[14]	0%	0,17%	yamaeva_s_f_bayesovskie-metody-v-ney..	не указано	31 Мая 2019	Кольцо вузов	0	1
[15]	0%	0,16%	Ефременко, Дмитрий Сергеевич Техно..	http://dlib.rsl.ru	19 Фев 2018	Коллекция РГБ	0	1
[16]	0%	0,15%	209693	http://biblioclub.ru	18 Апр 2016	Сводная коллекция ЭБС	0	1
[17]	0,46%	0%	не указано	не указано	раньше 2011	Цитирование	1	1
[18]	0,42%	0%	не указано	не указано	раньше 2011	Модуль поиска общеупотребительных выражений	6	6

Текст документа

Аннотация

Целью данной работы является изучение возможности определять неработающие видео по статистике просмотров. В ходе работы был предложен ряд моделей и проведены численные эксперименты. Наилучшей точности классификации удалось достигнуть с использованием классификатора из библиотеки CatBoost [13]. Наилучший классификатор показал точность в 80.3 % и полноту в 51.4 %. Точность классификации удалось улучшить на 9 % с помощью добавления в данные событий плеера и создания отдельной модели для каждого плеера. В данном случае удалось достигнуть точности классификации в 89.3 % при той же полноте. На основании данной работы рекомендуется внедрение классификатора в сервис YandeX Video, а также дальнейшее изучение свойств модели и ее последующее улучшение.

Содержание

Используемые определения

1.1 Видеоплеер

Под видеоплеером (плеером) в Данной работе подразумевается элемент web страницы, позволяющий пользователю просматривать видеозаписи (смотри рисунок 1). Как правило, один видеоплеер умеет воспроизводить ряд видеозаписей, хранящихся на сервере провайдера контента.

Помимо предоставления конечным пользователям возможности просматривать видео—
записи, видеохостинги часто предоставляют возможность другим интернет ресурсам инте—
грировать этот контент в свои страницы. Для этого, как правило, используется тэг iframe
языка разметки HTML. Данный тэг, согласно спецификации [10], позволяет web странице
открывать внутри себя вложенные страницы, например, с видеоплеером.
Кроме того ряд крупных видеохостингов предоставляют более обширное AP1 для ин—
теграции контента. В частности, некоторые плееры открывают доступ к так называемым
событиям плеера, то есть оповещениям о некоторых событиях, которые дают возможность
оценивать состояние контента. К таким событиям чаще всего относятся события старта, пау—
зы и ошибки плеера. Примером плеера, который предоставляет подобную функциональность
может служить плеер видеохостинга YouTube [17].

1.2 Неработающие видео

Под неработающим видео (рисунок 2) в данной работе поднимается любая видеозапись,
которую пользователь не может просмотреть. Видеозапись может прекратить работу по ряду
причин: отключение сервера видеохостинга, на котором данная видеозапись хранилась или
блокировка просмотра видеозаписи на основании обращения правообладателя контента.

1.3 Бинарная классификация

В данной работе рассматривается задача бинарной классификации, то есть разбиения
множества объектов на два класса. Были выбраны следующие обозначения: под классом
0 (нулевым классом, отрицательным классом) подразумевается класс работающих видео,
а под классом 1 (положительным классом) подразумевается класс неработающих видео,
детектирование которых и является целью данной работы.
При анализе предсказания модели в задаче бинарной классификации принято говорить
о четырех группах обьектов: ложно отрицательные обьекты (false negatives, F N), ложно

. . < > ЕП ВуоишЬелот
@ Tutorial: Deep Learning for Objects and Scenes - Part1
Watch later Share
CvF
› :19 107:43/13128 БЗ @ YouTube I]

Рисунок 1 — Видеоплеер YouTube.

положительные обьекты (false positives, F P), верно отрицательные обьекты (true negatives,
TN) и верно положительные обьекты (true positives, TP).
Для оценки модели в задаче бинарной классификации принято использовать метрики
точности (precision, P) и полноты (recall, R), которые вводятся следующим образом:

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}.$$

Здесь и далее в данной работе под точностью и полнотой классификации подразумеваются
I/IMCHHO ВВСIICHHBIC BbIIHC метрики.

1.4 Решающие деревья

Решающие деревья [3] — это модель машинного обучения, которая позволяет решать
задачи классификации и регрессии с помощью построения двоичного дерева, подобного
дереву поиска. В каждом внутреннем узле данного дерева находится условие, исходя из
соответствия которому для конкретного обьекта поиск спускается либо в правое, либо в
левое поддерево. В листьях данного дерева находится предсказание модели, то есть метка
класса В случае задачи классификации либо численное значение В случае задачи регрессии.
Решающие деревья являются одной из самых интерпретируемых моделей машинного
обучения, так как финальная модель может быть очевидным способом разложена В ряд ло—
гических предикатов и соответствующих им значений предсказания. Кроме того данная
модель является устойчивой к высокой размерности данных и зависимости среди признаков,
что позволяет сводить предобработку данных к минимуму при работе с данной моделью. Тем

EI]

E youtube.com

° YouTube RU

Video unavailable

This video is no longer available due to a copyright claim by National sports channel LLC.

яндекс Бразилия - боливия 3—0 обзор матча кубок америки 15.06.:

Видео

Бразилия - боливия 3-0 обзор матча кубок

америки 15.06.2019

° youtube.com 14 июня

...Боливия - Бразилия 0-3 Обзор Матча,Bra2i| VS

Bolivia 3:0, Bolivia VS Brazil 0:3,Copa America 2019.

Источник видео:

3.0 „„

O "

G? ossop MAM 031207

` Кубок Америки по футболу 2016 Аргентина -

Боливия 3-0 (14 июня 2016). Обзор матча

. hlamer.ru 28 июня 2016

Теги видео: Футбол, Евро 2016, Аргентина, Боливия.

Футбол, Евро 2016, Аргентина, Боливия, Кубок...

Бразилия - Боливия 3:0 Обзор матча

1 5/06/2019

II youtubecom 14 июня

& Бразилия - Боливия 3-0 — Обзор Матча Кубок

Америки 15/06/2019 HD G Бразилия Боливия

Бразилия...

‘ Аргентина 3-0 Боливия - Кубок Америки

: 2016 - 3-й тур - Обзор матча

j в ok.ru 15 июня 2016

Аргентина 3-0 Боливия - Кубок Америки 2016 - 3-й

тур — Обзор матча. Аргентина 3-0 Боливия...

ситца: ЩБЕ _

зд; \\.

\= ', 4 “”/"1 3?

Сборная России вновь

Brazrl vs Bolrvra 3-0 Бразилия 3-0 Боливия обыграла Турцию в матче

Обзор матча Match review and ALL Goals...

° youtube.com 15 июня

бразилия боливия, бразилия боливия 3-0, бразилия

боливия обзор матча, неймар, коутиньо, кубок...

Бразилия - Катар 2-0 -

Аргентина 3-0 Боливия | Кубок Америки-

Рисунок 2. Пример неработающего видео на сайте www.youtube.com И на сайте YandeX Video.

› youtube.com

0 отв

Q пізіоіё

.—

—o-

Будьте в Ппюсе ° . @ Вадим Шиянов

П<

@ Отправить на устройство

Video unavailable

This video is no longer available due to a copyright

claim by National sports channel LLC.

3?

", „

_ .

‘ ,94:21 но ‹

... ›.

“ {32:05

Слуга народа - Сезон 1 - Бразилия камерун обзор

матча HD!

Li

. ‘J

ЕВ ы; 11.125

Аргентина — Никарагуа 5-1

Видеообзор матча 1/4

финала ЧМ-201 8 по

2 Я.".

: ":\ 17

05:56 HD

Чили - Аргентина О-О (neH. Аргентина Боливия 2:0

не менее Данная модель имеет И ряд существенных минусов. Так предсказание решающего дерева очень чувствительно к гиперпараметру глубины дерева. Действительно, если глубина не ограничена сверху, то модель легко переобучается (производя деление до тех пор, пока в каждом листе не останется единственный обьект обучающей выборки). Предсказания такой модели скорее всего выйдут очень шумными, а обобщающая способность модели окажется крайне низкой. В то же время слишком жесткое ограничение на глубину деления вызовет создание модели, сложности которой недостаточно для описания зависимостей в данных. В то же время, даже обладая априорным знанием оптимальной глубины дерева, нахождение оптимального дерева остается нетривиальной задачей, которой посвящен ряд статей (на— пример [8]). Все эти проблемы вместе С развитием других моделей машинного обучения привели к тому что на данный момент решающие деревья в чистом виде практически не ис— пользуются, однако они являются важным составляющим компонентом для более сложных и Эффективных моделей.

1.5 Bootstrap aggregation

Bootstrap aggregation (bagging) [2] — это метод, применяемый в моделях машинного обучения для улучшения стабильности и точности предсказания. Метод заключается в сле— дующем: пусть у нас есть выборка из п обьектов, давайте сгенерируем из нее m подвыборок размера п’ путем последовательного выбора произвольного обьекта исходной выборки. Да— вайте теперь обучим на полученных выборках модели и усредним их предсказания. Можно доказать, что в случае независимости предсказаний полученных моделей, дисперсия резуль— тирующего предсказания падает С увеличением их числа.

Изначально метод никак не связан С решающими деревьями, однако на данный момент чаще всего bagging применяется именно для моделей решающего дерева. Это связано С тем, что для таких моделей проще обеспечить независимость предсказания отдельных моделей путем дополнительной подвыборки признаков, по которым разрешено ветвиться каждому конкретному дереву в ансамбле, а также довольно мягкому ограничению глубины деревьев (предполагается, что переобученные на разных подвыборках деревья будут давать суще— ственно независимые предсказания). В таком случае также не сильно существенно строить оптимальное дерево каждый раз. Поэтому чаще всего выбор признака и значения для даль— нейшего ветвления на каждом шагу производится жадным образом, что позволяет значи— тельно ускорить время обучения модели. На этом основана модель random forest [7], ко— торая использовалась в данной работе, как базовая модель. Реализация модели была взята из библиотеки SCikit—learn [12]. С ее помощью удалось достигнуть точности предсказания неработающих видео в 72.0 % и полноты предсказания в 50.8 %.

1.6 Модель extremely randomized trees

Модель extremely randomized trees [6] — это модель, во многом похожая на модель random forest. Для ее построения также производится генерация подвыборок в соответствии с ме— тодом bagging и обучение на подвыборках моделей решающего дерева. Однако в отличии

от жадного поиска оптимального признака И значения для ветвления при обучении каждого дерева В ансамбле, как это происходит при обучении модели random forest, модель extremely randomized trees произвольно генерирует несколько предполагаемых вариантов и выбирает оптимальный среди них. Это позволяет сделать деревья В ансамбле еще более независимыми, что в ряде случаев позволяет улучшить точность предсказания. Реализация данной модели также была взята из библиотеки Soikit—learn [12]. Использование позволило увеличить точ— ность предсказания до 74.3 %, однако при этом полнота упала до 47.0 %.

1.7 Gradient boosting

Gradient boosting (boosting) [5] — это метод машинного обучения, позволяющий созда— вать ансамбль из простых моделей, каждая из которых обладает достаточно низкой пред— сказательной способностью, обладающий высокой точностью предсказания. Этого удается достичь путем последовательного обучения моделей: первая модель обучается на исходных данных, каждая последующая же модель учится предсказывать уже не искомое значение, а ошибку получившегося до нее ансамбля (В случае задачи регрессии, В случае задачи класси— фикации она учится предсказывать класс обьекта, но больший вес отдается тем объектам, на которых предыдущий ансамбль ошибается). В соответствии с точностью ее предсказания, получившейся модели присваивается вес и она пополняет ансамбль.

Несмотря на то, что данный 18 метод также не имеет непосредственного отношения к ре— шающим деревьям, чаще всего В качестве базовых моделей В ансамбле выбираются именно решающие деревья небольшой глубины. Реализация данной модели была взята из библиотеки CatBoost [13]. Модель показала наилучшую точность из полученных В 80.3 %. Полнота полу— чившейся модели оказалась также выше чем у моделей random forest и extremely randomized trees и составила 51.4 %. Также для данной модели был произведен эксперимент с пред— сказанием неработающих видеозаписей для одного плеера. В данном эксперименте удалось достигнуть точности В 89.3 % при той же полноте.

1.8 YandeX Video

YandeX Video [16] — это сервис, позволяющий пользователям смотреть видеозаписи, со— бранные с множества различных видеохостингов. Из—за большого объема контента и невоз— можности моментально реагировать на его изменения, хранить видео на серверах компании YandeX Не представляется возможным. Поэтому видеозаписи показываются пользователям В виде интегрированных с помощью тэга iframe видеоплееров.

1.9 YandeX Toloka

YandeX Toloka [15] — это сервис, позволяющий публиковать некоторые несложные зада— ния, которые другие пользователи могут выполнять за материальное вознаграждение. Данный сервис использовался в работе для получения экспертной разметки данных: пользователям, ВЗЯВШИМСЯ 38. ВЫПОЛНСНИС ЗЗДЗНИЯ, ПОКЗЗЫВЗЛСЯ 1118,1116 С ВІ/ІІІСОЗЗНІ/ІСВІО И требовалось ОТВС-

тить, проигрывается зта видеозапись или нет. Для увеличения точности разметки, одна и та же видеозапись размечалась несколько раз разными пользователями. На разметку отправля— лись видеозаписи, для которых за предыдущий День набралось хотя бы 100 зафиксированных просмотров. Таким образоМ за два раза была размечена выборка размероМ в 101034 видео— записи. Неработающие видеозаписи составили 1.36 % выборки.

Введение

Выбор темы обусловлен спецификой работы сервиса YandeX Video. В силу того, что контент данного сервиса выдается пользователю В виде плееров, интегрированных В сайт с помощью тзга iframe, у данного сервиса возникает необходимость вовремя детектировать неработающие видеозаписи и убирать их из выдачи. На данный момент эта задача решает— ся с помощью периодического обхода видеозаписей с помощью специального механизма, умеющего эмулировать работу клиентского web браузера, а также умеющего нажимать на кнопку начала проигрывания видеозаписи и оценивать, началось ли воспроизведение. Дан— ный механизм позволяет получать достаточно достоверный сигнал о недоступности видео, однако обладает рядом существенных минусов. К ним относится, например, невозможность проверить доступность видео В странах кроме России (так как все запросы территориально отправляются из России), то есть учесть специфику локальных блокировок сайтов и контен—

та. Также существенным недостатком такого метода является сложность масштабирования. При увеличении числа видеозаписей требуется увеличивать скорость работы обхода, а это невозможно без установки дополнительного оборудования. Для устранения выщеперечислен— ных недостатков было предложено использовать статистику просмотров для детектирования неработающих документов.

С этой целью по сессиям пользователей собирается анонимная статистика: когда пользо— ватель начинает просмотр видеозаписи, включается таймер, когда же пользователь переклю— чает видео или закрывает страницу сервиса, время просмотра записывается для дальнейшего анализа. Вместе со временем просмотра сохраняются и события плеера, если они доступны. Основной проблемой таких данных является высокая степень загрязненности. Если во время просмотра или переключения видео у пользователя пропадет интернет соединение, запись о просмотре может не прийти или содержать неверные данные. Также возможен случай, когда пользователь включает проигрывание видеозаписи, а затем отходит от компьютера. В это время таймер считает время просмотра, несмотря на то, что видео могло оказаться неработающим и так и не загрузиться.

Задачей данной работы ставится изучение возможности создания модели, которая смогла бы предсказывать неработающие видеозаписи по статистике просмотров с высокой точно— стью и приемлемой полнотой. В качестве метода исследования был выбран эксперименталь— ный подход: сбор и разметка выборки, оценка точности и полноты классификации с помощью метода перекрестной проверки и анализ предсказаний полученной модели.

Анализ предметной области

Наиболее близкой задачей к поставленной является задача Детектирования " soft 404"стра— ниц. В соответствии с протоколом HTTP [4], если при обращении к серверу клиент запра— шивает Документ, который не Доступен по той или оной причине, сервер Должен возвращать ошибку. Как правило это ошибка 403 (Forbidden Error) или 404 (Not Found). Тем не ме— нее многие интернет ресурсы стремятся предоставить пользователю более Дружественный интерфейс и вместо возврата ошибки имитируют нормальную работу, возвращая код воз— врата 200 (OK), а вместо запрашиваемого Документа отображают страницу, оповещающую пользователя о причинах недоступности контента.

Подобное поведение Делают задачу Детектирования недоступных Документов нетриви— альной. Первая попытка борьбы с Данной проблемой была предпринята в статье от 2004 года [1]. Авторы Данной статьи предлагают спровоцировать сервер вернуть заведомо недоступный Документ путем прибавления к имеющемуся названию искомого Документа произвольного окончания. После этого предлагается сравнивать поведение сервера в случае обращения к исходному Документу и к заведомо недоступному. Второе упоминание о проблеме soft 404 Датируется 2012 годом [9]. В отличии от первой статьи, Данная работа целиком посвящена проблеме Детектирования soft 404 страниц. Авторы этой статьи предлагают использовать классификатор, который использует в качестве Данных лексические сигнатуры, содержащи— еся В заголовке или в тексте страницы.

К сожалению ни один из предложенных способов Детектирования soft 404 Документов не может быть применен к задаче Детектирования неработающих видео, так как они оба предполагают периодический обход видеозаписей, а Данный подход обладает рядом недо— статков, о чем было сказано во введении. Кроме того важно понимать, что в отличии от Других 18 интернет ресурсов, видеохостинги при обращении к несуществующему или уже уДа— ленному Документу чаще всего предпочитают показывать страницу, полностью идентичную странице, которая была бы отображена в случае, если видео было Доступно Для просмотра. Сообщение о недоступности видеозаписи как правило располагается в самом видеоплеере. Таким образом подход, основанный на Детектировании переадресаций на заранее заготов— ленные Документы, содержащие оповещение о недоступности видео не применим так как в Данном случае 18 переадресаций не происходит. Анализ лексических сигнатур, содержащихся на странице и ее заголовке также неприменим в силу того, что и тот и Другой текст не зависит от Доступности видеозаписи, а само сообщение о возникшей проблеме находится в плеере. Таким образом Для Детектирования таких сообщений требуется технически сложный меха—

низм, который умеет эмулировать Доступ к странице из пользовательского web браузера И попытку воспроизведения видеозаписи, то есть данный подход не позволяет даже упростить существующее на данный момент решение.

12

Подготовка Данных

4.1 Исходные данные

Роль исходных данных играют записи о просмотрах видеозаписей, содержание и способ сбора которых был коротко обсужден в введении. В результате обработки сигнала, который приходит К нам из пользовательских сессий, формируется таблица, каждая строка которой содержит некоторый идентификатор плеера, идентификатор видеозаписи, а также время про— смотра и события плеера, если они доступны. Также из разметки сайта провайдера контента берется продолжительность видеозаписи. Пример таких данных представлен в таблице 1

Просмотр, с Старт Ошибка

... True

180 True

__-_ 5 True

__—

Таблица 1 — Пример исходных данных.

Более подробная информация о структуре данных и их содержании, а также о способе их сбора не публикуется, так как является интеллектуальной собственностью компании YandeX И Н6 ПОДЛСЖИТ раЗГЛЗШСНИЮ.

4.2 Выбор признаков

На данном этапе у нас есть таблица, в которой содержится множество строк, которые относятся К одной и той же видеозаписи (а в данной работе рассматривались только видео— записи, для которых набралось хотя бы 100 просмотров за день) и требуется ответить на вопрос, проигрывалась данная видеозапись или нет. Основная информация (которая меня— ется от просмотра К просмотру) находится в колонках со временем просмотра и событиями плеера. При этом время просмотра является универсальным, однако, достаточно шумным признаком по своей природе. Пользователь мог открыть видеозапись, а затем передумать ее смотреть или увидеть другую строку в выдаче и быстро переключить видео, таким образом мы получаем неоправданно короткий просмотр, даже если видеозапись работала корректно. В то же время пользователь мог открыть неработающую видеозапись и, например, отойти от компьютера. Тогда мы увидим неоправданно длинный просмотр, несмотря на то, что все 18 это время пользователю отображалось сообщение об ошибке. При этом события плеера, напро— ТИВ, ДОСТаТОЧНО НЗДСЖНЫЙ ИСТОЧНИК ИНфОРМдЦИИ, ТСМ Н6 менее ОНИ имеют РЯД СУЩССТВСННЫХ

13

недостатков: во—первых, события плеера доступны не для всех плееров, а во—вторых, каж— дому отдельному плееру характерен некоторый профиль событий для различных состояний видеозаписи, и эти профили могут существенно отличаться для двух разных плееров.

В силу высокого уровня зашумленности исходных данных было принято решение их образом усреднять. В случае с данным событий видеоплеера усреднение очевидно — если событие произошло, то данной записи ставится в соответствие единица, в противном случае ноль и в качестве признака используется среднее значение, то есть доля просмотров, В которых данное событие наблюдалось.

Из времени просмотра было сгенерировано большее количество признаков исходя из некоторых предположений о природе данных. Первое предположение заключается В том, что, если видеозапись не работает, большинство пользователей заметят это и переключать видео за довольно короткий промежуток времени, близкий к 30—40 секундам. Поэтому В качестве первой группы признаков были выбраны доля пользователей смотревших видео не менее 15, 30 и 45 секунд. Следующая идея говорит о том, что информация о том, как хорошо пользователи смотрят данное видео хранится В распределении доли просмотра контента. Для простоты и однообразности расчета распределение было введено В выборку В виде долей пользователей, досмотревших видео до 5, 10, 20, 30, 40, 50, 60, 70, 80 и 90 % длительности видеозаписи. Последнее наблюдение говорит о том, что если видеозапись попала на выдачу,

то когда—то она была обнаружена обходом, а значит В какой—то момент времени она работала.
Таким образом чаще всего мы занимаемся не поиском неработающего контента, а пытаемся поймать момент, в который видеозапись перестала работать. В таких случаях принято гово— рить о данных в форме временного ряда, то есть последовательных значениях, зависящих от времени. Однако для удобства работы с данными и расширения круга потенциальных алгоритмов классификации хотелось бы иметь данные в виде вектора фиксированной раз— мерности. Также важно помнить о шуме в данных, а значит и необходимости усреднения данных. Для соблюдения всех ограничений было принято решение воспользоваться методом скользящего среднего. В таком случае мы получаем, например, такие признаки как среднее время просмотра за 5 последних просмотров заканчивая последним записанным просмот— ром и такое же среднее за 5 просмотров заканчивая предпоследним просмотром и так далее. Хотелось бы заметить, что подход со скользящим средним применим не только к времени просмотра. Скользящее среднее также считалось для событий плеера и долей смотревших видео хотя бы 15, 30, 45 секунд и так далее.
В итоге для каждого видео мы получили вектор признаков, размерность которого состави— ла 154 признака. Более подробная информация о признаках не публикуется потому что также является интеллектуальной собственностью компании YandeX и не подлежит разглашению.
После того как были выбраны признаки и собрана выборка, была произведена разметка видеозаписей с помощью сервиса YandeX Toloka. Таким образом был собран набор из 101034 размеченных объектов.
14

Обучение классификатора

5.1 Выбор модели

Как было сказано ранее, пространство признаков имеет очень высокую размерность (154). К тому же многие из признаков сильно взаимосвязаны С Другими (действительно, Доля до— смотревших до 90 % видео не может быть выше доли досмотревших до 50 %). Несмотря на то, что с этими недостатками можно было попытаться справиться С помощью методов понижения размерности, таких как метод главных компонент [11], в данной работе было принято решение использовать модели машинного обучения, основанные на решающих де— ревьях. Это позволяет исключить из предобработки уменьшение размерности признаков, а также перестать беспокоиться о зависимостях в признаках.
В качестве базовой модели была выбрана модель random forest. Данный выбор обусловлен тем что модель зарекомендовала себя, как модель, которая показывает довольно высокие ре— зультаты практически во всех задачах. Также было замечено, что выборка в 101034 обьекта не очень велика. Также важно понимать, что среди этих объектов только 1.36 % видео— записей были размечены как неработающие. Таким образом выборка содержит довольно мало сигнала, который мы хотим детектировать. Поэтому в качестве дополнительной моде— ли была выбрана модель extremely randomized trees, которая позволяет в некоторых случаях уменьшить разброс предсказаний за счет небольшого увеличения смещения предсказания. В качестве последней модели был выбран метод gradient boosting На решающих деревьях, как потенциально наиболее точная модель из всех представленных.
Численные эксперименты производились с помощью языка прогромирования Python [14]. Реализации первых двух моделей были взяты из библиотеки SCikit—learn [12], реализация последней модели была взята из библиотеки CatBoost [13].

5 .2 Обучение модели

Для подбора гиперпараметров 1/1 получения оценки точности 1/1 полноты использовался сеточный поиск с оценкой точности 1/1 полноты с использованием 5—кратной перекрестной проверки. Для моделей random forest и extremely randomized trees подбиралось количество деревьев в ансамбле. Для классификатора CatBoost кроме того подбиралась максимальная глубина дерева.
Результаты сеточного поиска представлены в таблице 2. Графики зависимости точности 1/1 полноты от значений гиперпараметров представлены на рисунках 3, 4 и 5 . При этом в таблице
15

2 модели extremely randomized trees И классификатору из библиотеки CatBoost в таблице соответствуют по Две строки каждому. Это сделано для того, чтобы показать метрики и для

модели с наибольшей точностью, и для модели с наивысшей полнотой. Для модели random forest разница между точностью для этих двух классификаторов несущественна, поэтому представлены ТОЛЬКО значения МСТРИК ДЛЯ модели С НЗИВЫСШСЙ ПОЛНОТОЙ.

Модель Точность, % Полнота, %

RandomForestClassifier, 500 деревьев 50.8

ExtraTreesClassifier, 1100 деревьев 47.2

CatBoostClassifier, 650 деревьев глубины 3 51.4

CatBoostClassifier, 500 деревьев глубины 4 77.5 53.2

ExtraTreesClassifier, 1400 деревьев

Таблица 2 — Результаты сеточного поиска

Как можно видеть, модель extremely randomized trees действительно смогла немного уве— личить точность классификации по сравнению с результатами базовой модели, однако при этом показала меньшую полноту. Лучше всех себя показал классификатор из библиотеки CatBoost. I/IMeHHo С использованием данного алгоритма была достигнута наилучшая в дан— ном эксперименте точность классификации в 80.3 %. При этом полнота классификации у данной модели оказалась выше, чем у моделей RandomForestClassifier и ExtraTreesClassifier. Наибольшей полноты классификации в данном эксперименте (51.4 %) также удалось достичь именно классификатору из библиотеки CatBoost, правда, при других значениях гиперпара— метров.

5 .3 Анализ модели

После численного эксперимента был произведен анализ наилучшей модели. В результате этого анализа были выявлено, что события плеера, а именно усредненные события старта и ошибки, обладают наибольшим весом в финальной модели. При этом несмотря на акку— ратность и информативность такого сигнала в целом, события плеера доступны не для всех видеозаписей, а также характерный профиль событий различается от плеера к плееру. Это значит, что модель может давать сильно смещенные предсказания для некоторых плееров. В связи с этим было предложено добавить в данные идентификатор плеера, как категори— альный признак, что позволило бы модели корректировать ожидаемый профиль событий в зависимости от их источника. В данном эксперименте было принято решение ограничиться только моделью, которая показала наилучший результат, то есть классификатором из биб— лиотеки CatBoost. Результаты эксперимента представлены в таблице 3. Графики зависимости точности и полноты от значений гиперпараметров представлены на рисунке 6.

Как можно видеть, точность классификации упала на 7.7 процентных пункта до 72.6 %, однако полнота возросла на 14.9 процентных пункта и составила 68.1 %. Это говорит скорее всего о переобучении модели на новый категориальный признак, то есть модель просто запомнила, что видеозаписи некоторых плееров чаще всего не работают и прекратила ПОПЫТКИ IICCTHO ИХ КЛаССИфИЦИРОВЗТЬ.

0.7195

0.7190

0.7185

0.7180

precision

0.7175

0.7170

0.7165

600 800 1000 1200 1400 1600 1800 2000

n_estimat0rs

0.508

0.507

0.506

recall

0.505

0.504

600 800 1000 1200 1400 1600 1800 2000

n_estimat0rs

Рисунок 3 — Результаты сеточного поиска ДЛЯ модели RandomForestClassifier

17

Рисунок 4 — Результаты сеточного поиска ДЛЯ модели ExtraTreesClassifier

18

1800

1800

2000

2000

Рисунок 5 — Результаты сеточного поиска ДЛЯ модели CatBoostClassifier

19

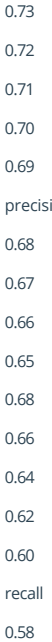


РИСУНОК 6 — РСЗУНЬТаТЬІ ВКСНСрІ/ІМСНТВ. С КЗТеГОРИЗЛЬНЫМ ПРИЗНЗКОМ.

20

Модель	Точность, %	Полнота, %
CatBoostClassifiCr, 1250 Деревьев глубины 2		
CatBoostClassifiCr, 800 Деревьев глубины 5		

Таблица 3 — Результаты эксперимента с категориальным признаком.

5.4 Одноплеерная модель

Так как Добавление идентификатора плеера в вектор признаков не позволил увлеичить точность предсказания, было принято решение попробовать обучать отдельную модель для каждого плеера. Это позволило бы модели выучивать характерный для данного плеера про— филь событий и использовать это для наиболее увеличения точности предсказания. В таком случае возникла проблема недостаточности данных в исходной выборке. Действительно, она

содержит всего лишь 101034 обьекта, которые более или менее равномерно распределены по 87 различным плеерам. Кроме того неработающих видео среди них всего лишь 1.36 %. Таким образом данных из исходной выборки недостаточно, чтобы можно было обучить по отдельной модели для каждого плеера. Поэтому было принято решение попробовать размечать выборку с помощью данных истории обхода, который принимал решение о работоспособности видео—записей. В таком случае мы получаем очень достоверный сигнал о том, какие видеозаписи не работали. Однако возникает обратная проблема: у нас нет достоверной информации о том, какие работали. Для решения данной проблемы был выбран следующий подход: известно, что среди видеозаписей, которые представлены на выдаче поиска, не работает примерно 1 %. Исходя из этого, при составлении выборки мы считаем количество достоверно неработавших видео исходя из данных обхода и размечаем эти видеозаписи. Остальные видео сортируем в убывающем порядке по среднему времени просмотра и отбираем столько, чтобы баланс классов составлял 99 к 1. Таким образом мы получаем размеченную выборку большого раз—мера, так как история обхода доступна за довольно длительный промежуток времени. В данном эксперименте также было принято решение ограничиться только классификатором из библиотеки CarBoosr. Результаты сеточного поиска представлены в таблице 4. Графики зависимости точности и полноты от значений гиперпараметров представлены на рисунке 7.

Модель	Точность, %	Полнота, %
CatBoostClassifiCr, 1550 Деревьев глубины 3		
CatBoostClassifiCr, 1850 деревьев глубины 5	54.7	

Таблица 4 — Результаты сеточного поиска для модели обученной на данных обхода.

Эксперимент показал прирост точности классификации в 9.1 процентных пункта. Сама точность составила 89.3 %. Полнота получившейся модели совпала с полнотой классифи—катора, обученного на данных размеченных С помощью сервиса YandCX Toloka. Наилучшая полнота, которая была достигнута в данном эксперименте (54.7 %) и вовсе выше, чем в случае экспертной разметки данных.

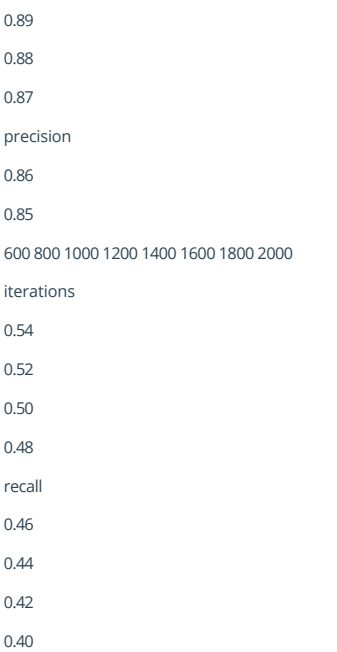


Рисунок 7 — Результаты сеточного поиска ДЛЯ модели обученной На Данных обхода

Анализ результатов

Наилучшей точности предсказания в 89.3 % удалось достигнуть с помощью одноплеерной модели. При этом такая точность уже позволяет говорить о том, что 18 предсказание модели является достаточным критерием для прекращения показа видеозаписи. При этом модель обладает достаточной полнотой в 51.4 %. Таким образом данную модель можно использо—вать в качестве дополнения для уже существующего решения в виде обхода видеозаписей. К

сожалению, невысокая полнота все же не позволяет полностью заменить существующий ме—ханизм без ущерба для качества выдачи. Также данный подход обладает рядом недостатков. Во—первых, данный метод требует дополнительного механизма создания дополнительных моделей при добавлении на выдачу новых плееров. Во—вторых, Также данный подход не применим для плееров, у которых не доступны события плеера, так как исключение их из признаков ухудшает качество предсказания. Тем не менее для всех остальных плееров дан—ный метод способен улучшить существующий механизм блокирования видеозаписей. Так, например, можно блокировать показ видеозаписей, предсказанных как неработающие, а для остальных применять уже существующий подход.

С другой стороны, при классифицировании видеозаписей всех плееров удалось достиг—нуть лишь более скромной точности в 80.3 %, что не позволяет применять данный механизм из—за того, что в таком случае из выдачи пропадет достаточно большое количество работаю—щих видеозаписей.

В дальнейшем планируется внедрение одноплеерной модели в сервис YandeX Video. Ана—лиз и последовательное улучшение точности и полноты его предсказания. А также изучение других возможностей предсказывать неработающие видеозаписи по статистике просмотров и, в частности, по истории времени просмотра, например, с помощью техник, более специ—фИЧНЫХ ДЛЯ Э.НЗЛИЗЭ. временных РЯДОВ.

23

Заключение

Данная работа была посвящена изучению частного случая проблемы soft 404, а именно детектированию неработающих видеозаписей. Так как ни один из ранее предложенных спо—собов решения Данной проблемы не подходит в данном случае из—за специфики задачи, было предложено использовать статистику просмотров для детектирования неработающих видео—записей. Благодаря такому подходу удалось обучить классификатор из библиотеки CatBoost, который показал точность классификации 80.3 % и полноту 51.4 %. Также был предложен способ ограничиться одним плеером, что позволило увеличить точность классификации на 9 процентных пунктов при неизменной полноте. Таким образом был построен классификатор, обладающий точностью предсказания 89.3 % и полнотой 51.4 %.

24

[1]

[2]

[3]

[4]

[5]

[6]

[7]

[8]

[9]

[10]

[11]

[12]

СПИСОК литературы

ZiV Bar—Yossef et al. “Sic transit gloria telae”. In: Proceedings of the 13th conference on World Wide Web - WWW ’04. ACM Press, 2004. DOI: 10. 1145/988672 . 988716.

Leo Breiman. “Bagging predictors”. In: Machine learning 24.2 (1996), pp. 123—140.

Leo Breiman et a1. Classification And Regression Trees. Routledge, Oct. 2017. DOI: 10 . 1201/9781315139470.

Roy T. Fielding and Julian Reschke. Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content. 1 RFC 7231. June 2014. DOI: 10 . 17487/RFC7231. URL: https : //rfc - editor.org/rfc/rfc7231.txt.

Jerome H. Friedman. “Greedy function approximation: A gradient boosting machine 11 .” In: The Annals ofStatistics 29.5 (Oct. 2001), pp. 1189—1232. DOI: 10 . 1214/aos/1013203451.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. “Extremely randomized trees”. In: Ma-

chine Learning 63.1 (Mar. 2006), pp. 3—42. DOI: 10.1007/810994—006—6226— 1.

Tin Kam Ho. “Random decision forests”. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. IEEE Comput. Soc. Press , 1995. DOI: 10 . 1109/icdar.1995.598994.

Petr Masa and Tomas Kocka. “Finding Optimal Decision Trees”. In: Advances in Soft Computing. Springer Berlin Heidelberg, pp. 173—181. DOI: 10. 1007/3- 540- 3352 1 - 8_17.

Luis Meneses, Richard Furuta, and Frank Shipman. “Identifying “Soft 404” Error Pages: Analyzing the Lexical Signatures of Documents in Distributed Collections”. In: Theory and Practice of Digital Libraries. Springer Berlin Heidelberg, 2012, pp. 197—208. DOI: 10 . 1007/978-3-642—33290-6_22.

Scott O’Hara et a1. HTML 5.3. W3C Working Draft. https://www.w3.org/TR/2018/WD—htm153—20181018/. W3C, Oct. 2018.

Karl Pearson. “ LIII. On lines and planes of closest fit to systems of points in space ”. In: The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2.11 (Nov. 1901), pp. 559—572. DOI: 10 . 1080/14786440109462720.

F. Pedregosa et a1. “Scikit—learn: Machine Learning in Python”. In: Journal of Machine Learning Research 12 (2011), pp. 2825—2830.

25

[13]

[14]

[15]

[16]

[17]

Liudmila Prokhorenkova et a1. “CatBoost: unbiased boosting with categorical features”. In: (June 28, 2017). arXiv: http : //arxiv . org/abs/1706 . 09516v5 [cs .LG].

Pythan Pragmmming Language. 2019. URL: https : //www . python . org.

Yandex TMO/ca. 2019. URL: https : //toloka . yandex . com.

Yandex VideO. 2019. URL: http : //yandex . ru/video/.

Y0uTube Player AP! Reference f0r ifmme Embeds. May 2018. URL: https : / / developers . 300319.coш/yonгnьe/iiraше_api_reierence.

26