Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.383

Чёрный
Родион Павлович

Система сбора, анализа и классификации инорфмации о культурных мероприятиях

АВТОРЕФЕРАТ
на соискание академической степени
магистра информатики и вычислительной техники

по специальности 1-40 81 04 – Обработка больших объемов информации

Научный руководитель
Сиротко С.И.
к.т.н., доцент

Минск 2020

КРАТКОЕ ВВЕДЕНИЕ

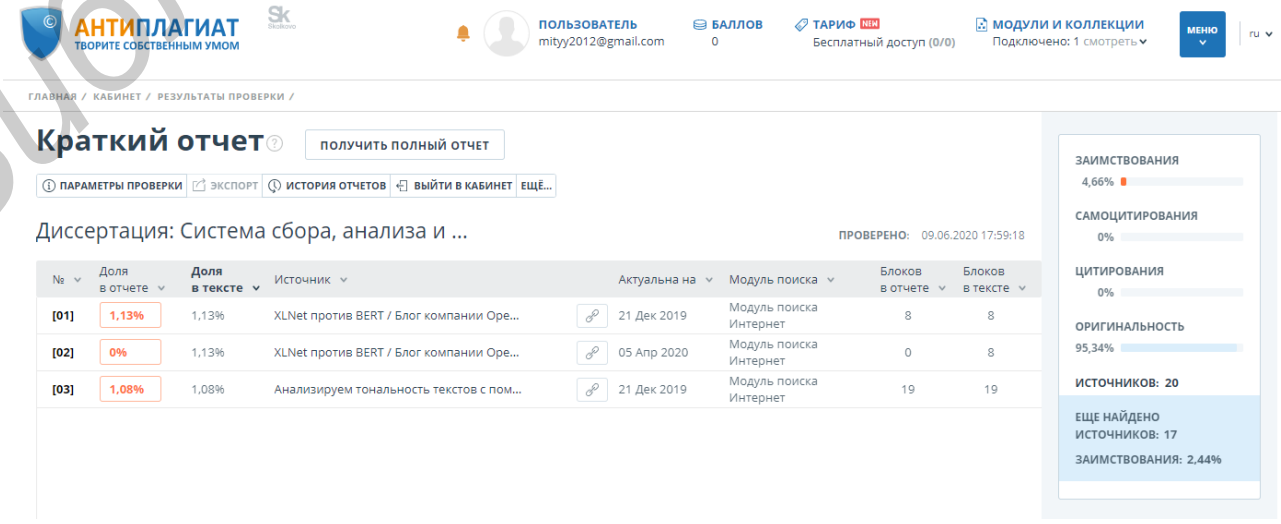
Чем крупнее город и чем более насыщенная в нем жизнь, тем больше в этом городе происходит различных событий и мероприятий. Люди собираются по интересам у устраивают различные мастер классы, лекции, обсуждения. Бары и небольшие концертные площадки дают свои помещения для выступления местным музыкантам. Авангардные артисты устраивают премьеру своей новой постановки, а какой-нибудь фонд или банк спонсируют выставку современного искусства. Таких событий: крупных и мелких, каждый день может быть десятки, а даже и сотни.

Организаторы таких мероприятий заинтересованы в том, чтобы о нем узнали: организаторам выставки нужны посетители, менеджерам артистов необходим аншлаг, чтобы получить максимум выручки. В то же время, сами зрители заинтересованы в том, чтобы узнать о выступлении их любого артиста, или открытии выставки современного искусства.

Для удовлетворения интересов обеих групп существуют так называемые “интернет афиши” – веб-сайты, на которых можно разместить объявление о предстоящем мероприятии (дату, описание, стоимость, тип и т.д.), а посетители сайта смогут ее увидеть и поделиться.

Помимо этого, объявления могут публиковаться ещё и на страницах в социальных сетях, мессенджерах, чтобы охватить еще большую аудиторию. При этом самому человеку, при большом количестве источников, становится трудно следить за всеми мероприятиями, которые происходят.

Диссертационная работа выполнена самостоятельно, проверена в системе «Антиплагиат». Процент оригинальности – 95.34%, что соответствует норме, установленной кафедрой Информатики.



ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цели и задачи исследования

Целью диссертационной работы является разработка системы, которая бы собирала информацию о планируемых в городе Минске мероприятиях, классифицировала ее, извлекала из информации необходимые метаданные и предоставляла пользователю в удобном виде и давая пользователю возможность вносить корректировки в работу системы.

Для достижения поставленной цели необходимо решить следующие задачи:

1) Разработать алгоритм дедупликации объявлений о мероприятиях.
2) Обучить модель классификации мероприятий.
3) Разработать алгоритм извлечения необходимых метаданных из текста объявления мероприятия.
4) Разработать алгоритм учета правок пользователей.
5) Разработать архитектуру системы и предусмотреть ее автономность.
6) Развернуть систему в облачном сервисе.

Объектом исследования является обработка и анализ естественного языка.

Предметом исследования является объявления о проводимых в городе и за его пределами различного рода мероприятий.

Основной *гипотезой*, заложенной в основу диссертационной работы, является невозможность пользователям без дополнительных средств поддержки получать всю информацию о мероприятиях из одного источника, так как она распределена по множеству подобных веб-сайтов и других сервисов.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, четырех глав, заключения, списка использованных источников и приложения. В первой главе представлен анализ предметной области, формулирование гипотезы и предъявление списка требований, которым должна соответствовать система.

Вторая глава посвящена обзору и разработке алгоритмов, моделей, которые будут использоваться в системе. А также произведена проверка выдвинутой гипотезы.

Третья глава посвящена разработке архитектуры системы, чтобы она соответствовала всем сформулированным требованиям. Четвертая глава описывает различные аспекты непосредственной реализации системы: выбор технологий, инфраструктуры, разворачивание системы, а также демонстрация ее работы. В приложении приведен исходных код основных модулей системы.

Общий объем работы составляет 68 страниц. Из которых основного текста – 42 страниц. 20 рисунков на 8 страницах, 5 таблиц на 3 страницах, список использованных источников из 13 наименований на 2 страницах и одно приложение на 13 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

В **первой главе** была выдвинута гипотеза о том, что полный объем информации о мероприятиях распределен между различным сервисами–афишами, а также социальными сетями, что неудобно для конечного пользователя. А также был сформирован список требований, предъявленый к будущей системе.

Во **второй главе** было рассмотрено математическое обеспечение, которое будет использовано при разработке системы: алгоритмы, модели машинного обучения. Также была подтверждена гипотеза, выдвинутая в первой главе.

В **третьей главе** была описана архитектура системы, соответствующая принципам микросервисной архитектуры, а также обеспечивающая автономный характер работы системы.

В **четвертой главе** описываются технические аспекты реализации: выбор технологий, библиотек, фреймворков, а также среды для развертывания системы.

ЗАКЛЮЧЕНИЕ

В ходе работы над магистерской диссертацией была разработана система, которая собирает информацию о мероприятиях с различных источников, классифицирует ее и предоставляет пользователю в удобном виде. Формат ее работы автономный, то есть весь цикл задач она выполняет самостоятельно, в

определенные временные промежутки. Такой принцип работы позволяет сократить ручную поддержку системы до минимума, а вместо этого продолжать разрабатывать новые модули для системы и улучшать существующие.

Для достижения такой автономности были применены алгоритмы машинного обучения:

- Нейронные сети – для классификации документов;
- Алгоритмы дедупликации документов.

Также был применен разработанный алгоритм автоматического учета правок, позволяющий учитывать изменения, вносимые самим пользователями.

Данные модели несовершенны, но уже показывают удовлетворительные результаты работы. Но ввиду того, что система предоставляет возможность пользователям корректировать результаты работы алгоритмов, то система имеет большой потенциал для улучшения качества работы моделей.

Система разрабатывалась с целью сократить количество источников информации о мероприятиях, которые использует пользователей, до одного – разработанной системы. К сожалению, на данный момент система не покрывает собой все имеющиеся источники мероприятий, проводимых в Республике Беларусь. Поэтому, прежде чем выставлять систему на всеобщее обозрение, необходимо максимизировать количество собираемой системой информации.

И хотя система еще не готова к полноценному внедрению, она, тем не менее, является удовлетворительно работающим прототипом и доказательством концепции в целом, что можно использовать для привлечения, например, новых энтузиастов в команду разработки.

Помимо этого, за время работы над системой удалось углубить знания в сфере машинного обучения и, в частности, в сфере обработки естественного языка. Для решения задачи классификации воспользовались одним из наиболее современных подходов – ULMFiT. А при развертывании системы в облаке освоил инструменты и сервисы, которые предоставляет платформа Azure.