Федеральное государственное автономное образовательное учреждение

высшего образования

«Национальный исследовательский университет «Высшая школа экономики»

Московский институт электроники и математики им. А.Н. Тихонова

Магистерская программа  
«Суперкомпьютерное моделирование в науке и инженерии»

О Т Ч Е Т

по проектной работе**«Анализ отзывов о лекарственных препаратах в социальных медиа»**

Выполнили студенты гр. МСКМ-181

Пехтерев Д.О.  
Чертенков В.И.  
Горчавкина А.А.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Руководитель проекта:**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*(оценка) (подпись)*

*\_\_\_\_\_\_\_\_\_\_\_\_\_*

*(дата)*

**Москва 2019**

1. Содержание

2. Общее описание проекта:

|  |  |
| --- | --- |
| Руководитель проекта | Артамонов Сергей Юрьевич, доцент департамента прикладной математики МИЭМ |
| Тип проекта | прикладной |
| Место работы проекта | Московский институт электроники и математики им. А.Н. Тихонова |

3. Содержательная часть:

3.1 Описание хода выполнения проектного задания (я не понимаю в каком виде это должно быть)

В качестве основного набора данных было решено использовать готовый набор текстовых отзывов на разные категории товаров на сайте Amazon в период с мая 1996 по июль 2014 года. Набор данных был собран в файлы по категориям профессором Стэнфордского университета Джулианом Макаули, что сыграло большую роль при оценке точности алгоритмов классификации.

На начальном этапе решалась подзадача отделения категории «Медицина и здоровье» из набора данных, содержащего отзывы по нескольким категориям. Для ее выполнения использовались три метода многоклассовой классификации: СВМ, градиентный бустинг, случайный лес. Каждый из этих методов имеет свои преимущества и недостатки, установленные нами в ходе работы, поэтому пользователь может самостоятельно выбрать метод в соответствии со своими требованиями. В основном, речь идет о выборе между скоростью и точностью.

Следующая подзадача состояла в определении тональности отзывов в рамках одной категории с помощью метода логистической регрессии. Она включает в себя несколько этапов: привидение пятибалльной системы оценивания к бинарному виду, получение равного количества положительных и отрицательных отзывов, определение коэффициента значимости слова. В заключение также была исследована точность алгоритма.

Конечные результаты визуализировались с помощью редактора инфографики Tableau Public. Для этой цели был создан личный кабинет, который обеспечивает доступ к графикам из любого устройства. Мы научились строить разные зависимости, используя разные формы визуализации. В том числе, построили облака слов для всех категорий алгоритма логистической регрессии.

3.2 Описание результатов проекта (продукта);

В результате работы мы получили готовый программный код, способный выделить отзывы, относящиеся к выбранной категории данных, и определить их тональность для конкретного средства.

3.3 Описание использованных в проекте способов и технологий

3.4 Описание своей роли в проектной команде (обязательно для группового проекта)\*

3.5 Описание отклонений и трудностей, возникших в ходе выполнения проекта\*

4. Заключение (оценка индивидуальных результатов выполнения проекта, сформированных/развитых компетенций)

5. Результат проекта (текст, фотографии, ссылки и другие подтверждающие получение результата материалы) 2

6. Приложения (при необходимости: презентация для защиты проекта, графики, схемы, таблицы, алгоритмы, иллюстрации, отзывы и т.п.).

**1. Постановка задачи**

При решении задачи анализа большого количества данных целесообразно применять удобные средства визуализации. В рамках нашего проекта было решено использовать платформу Tableau Public (public.tableau.com) – современное и удобное средство построения разного рода зависимостей с возможностью последующей публикации на веб-ресурсах. Уместно не только представить конечный результат анализа в удобной для пользователя форме, но и отслеживать правильность выполнения анализа. Основными задачами проекта являются разделение на классы и определение тональности отзывов о лекарственных средствах. Оценка алгоритма в данном случае выполняется методом построения ROC-кривой (Receiver Operator Characteristic). Площадь под кривой служит показателем точности классификатора. Чем ближе она к единице, тем лучше работает алгоритм.

1) Содержание набора данных  
Исходный набор содержит множество различных категорий. Применяя три метода классификации (метод опорных векторов, случайный лес, градиентный бустинг над решающими деревьями), можно определить относительное содержание каждой из категорий и представить его в виде круговой диаграммы.

2) Определение наиболее подходящего метода  
Для каждого из методов в пункте 1) строятся графики подбора параметра «c» (визуализация процесса обучение) и ROC-кривая. Сравнивая площади под ROC-кривыми, можно выбрать наиболее подходящий алгоритм.

3) Тональность  
Результат выбранного алгоритма используется в качестве входных данных для определения тональности отзывов методом логистической регрессии. В качестве заключительного результата работы также представляются график подбора параметра «с» и ROC-кривая.

Слайд 1.

Правильно подобранное графическое представление позволяет уловить характерные зависимости данных без детального углубления в их структуру. Для этой задачи мы используем платформу Tableau Public (public.tableau.com) – бесплатный редактор инфографики, позволяющий представить большинство типов данных в доступной форме.

* Вариативность форм представления (таблицы, линии, гистограммы, круговые диаграммы, распределение на карте и т.д.)
* Возможность работы с разными популярными форматами файлов – txt, xlsx, pdf, JSON и др.
* Обработанные зависимости сохраняются в личном кабинете пользователя, что обеспечивает к ним доступ с разных устройств. Общий объем файлов ограничивается 10 Гб.
* Результат работы можно публиковать на веб-страницах. При этом пользователь может интерактивно взаимодействовать с полотном: выделять зависимости, приближать изображение.

Слайд 2.

Перед вами типичный интерфейс Tableau. Для более детального представления изучаемых зависимостей к графику можно прикреплять дополнительные рисунки, текст, а также гиперссылки. Соотношение занимаемой площади элементов и их порядок задается пользователем.

Слайд 3.

Если алгоритм машинного обучения обрабатывает текстовые данные, полезно иметь представление о них. В частности, для алгоритма логистической регрессии мы строим облака слов – изображение некоторого множества слов в соответствии с их критерием значимости.

Для каждой категории данных получено изображение, где значимость слова определяется его близостью к центру, размером и интенсивностью оттенка. На слайдах представлено несколько из них для сравнения. На первом слайде совпадений крайне мало, хотя обе категории тесно связаны с общим понятием «досуг». На втором слайде совпадений намного больше, хотя категории и кажутся и более отдаленными. Из этого можно сделать вывод, что нельзя выделить какое-то единственное множество слов разумных объемов, используемое при описании чего-либо.

Слайд 4.

Приведем несколько примеров визуализации процессов обучения. На графике изображено изменение точности работы алгоритма в зависимости от параметра С для разных категорий слов. Видно, что максимальная точность достигается в примерно одной и той же области значений. Исключение составляет линия категории Books.

Слайд 5.

Слайд 1

Для задачи визуализации мы используем платформу Tableau Public – бесплатный редактор инфографики, позволяющий представить большинство типов данных в доступной форме. Его основные преимущества:

• Вариативность форм представления (таблицы, линии, гистограммы, круговые диаграммы, распределение на карте и т.д.)

• Возможность работы с разными форматами файлов

• Обработанные зависимости сохраняются в личном кабинете пользователя, что обеспечивает к ним доступ с разных устройств.

• Результат работы можно публиковать на веб-страницах.

Перед вами типичный пример выполненной работы в Tableau. К графику можно прикреплять дополнительные рисунки, текст, а также гиперссылки.

Слайд 2

Если алгоритм машинного обучения обрабатывает текстовые данные, полезно иметь представление о них. Для этого строят так называемые облака слов. Изображение, где значимость слова определяется его близостью к центру, размером и интенсивностью оттенка. В частности, перед вами изображено облако слов для алгоритма логистической регрессии категории «Медицина и здоровье». Также мы научились использовать Tableau Public для построения других зависимостей. Например, ROC-кривых и графиков подбора параметра С. На данный момент мы знаем какую структуру должны иметь данные, их объемные ограничения, а также опробовали способы построение наиболее популярных форм визуализации.