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:**

Артамонов С.Ю.

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

*(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:

В качестве основного набора данных было решено использовать готовый набор текстовых отзывов о 24 категориях 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. Вместе с этим, алгоритм автоматически подбирает “c” – коэффициент регуляризации, тем самым подбирая лучшую модель для каждого конкретного случая. 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. Вместе с этим получили готовый шаблон в Tableau Public, который может быть обновлён в полуавтоматическом режиме и показывает пользователю основные характеристики модели: как по подбору коэффициента регуляризации, так и по её качеству.

3.3 Описание использованных в проекте способов и технологий

В работе были использованы методы машинного обучения, а конкретно:

1. Для задачи определения тональности:
   * Метод Логистической регрессии
2. Для задачи многоклассовой классификации:
   * Метод опорных векторов
   * Случайный лес
   * Градиентный бустинг над решающими деревьями
3. Для задачи визуализации:
   * Tableau Public

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

* Пехтерев Денис – разработка модели по определению тональности
* Чертенков Влад – разработка модели по многоклассовой классификации
* Горчавкина Анастасия – разработка и изучение инструмента визуализации Tableau Public

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

Какие бы трудности нас не настигали, нам всегда с ними помогал справиться наш сплочённый командный дух.

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

* Пехтерев Денис – разобрался и научился строить модели по определению тональности, повысил навыки написания кода на Python, повысил навыки взаимодействия с Github, повысил навыки работы с внешними источниками и библиотеками в Python.
* Чертенков Влад – разобрался и научился строить модели по многоклассовой классификации, повысил навыки написания кода на Python, повысил навыки взаимодействия с Github, повысил навыки работы с внешними источниками и библиотеками в Python.
* Горчавкина Анастасия – разобралась и научилась визуализировать результаты моделей машинного обучения в Tableau Public, повысила навыки написания кода на Python, повысила навыки взаимодействия с Github, повысила навыки работы с внешними источниками и библиотеками в Python.

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



Рисунок Облако слов для категории “Books”



Рисунок Облако слов для категории “Movies and TV”

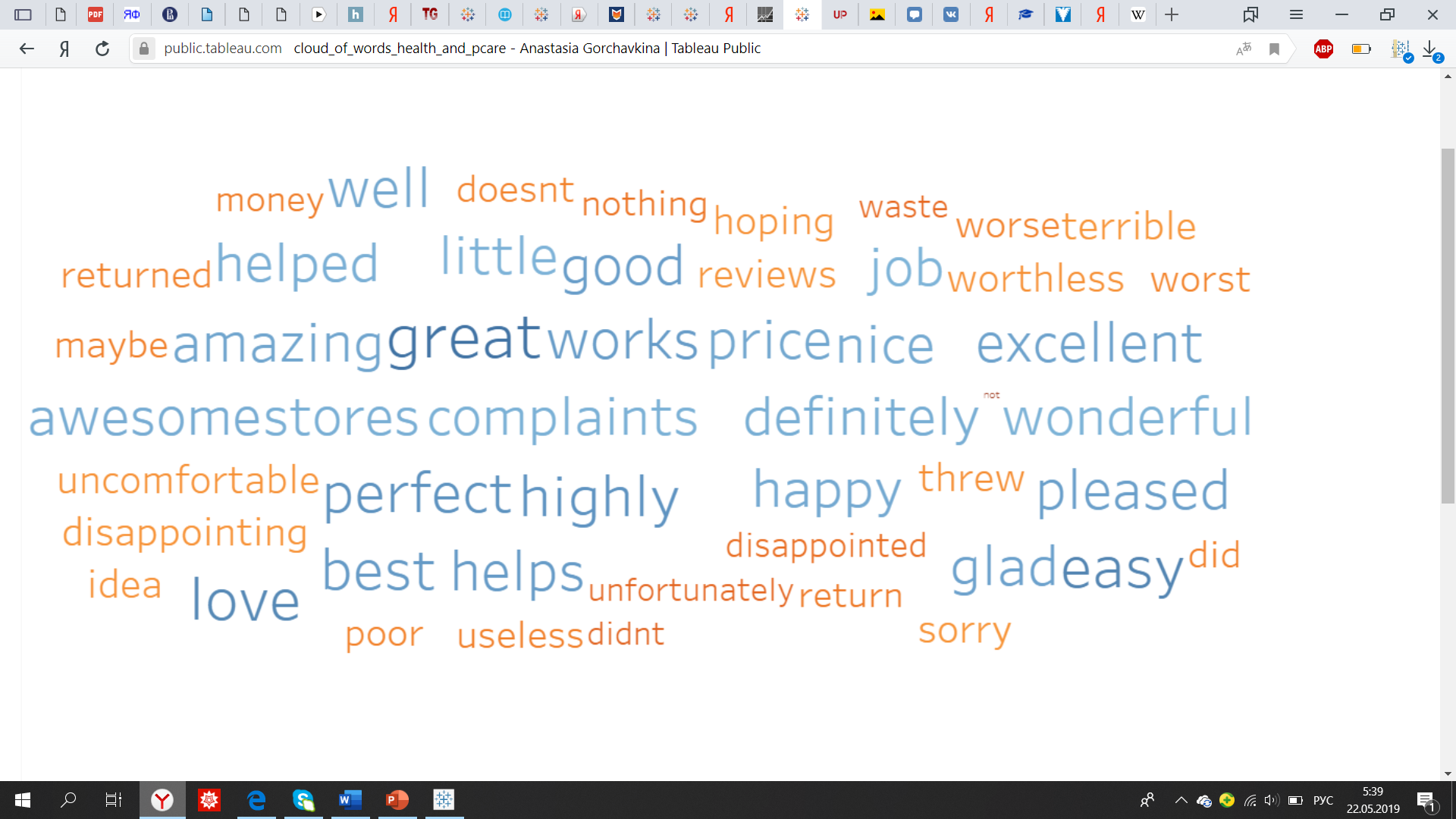


Рисунок Облако слов для категории “Health and care”

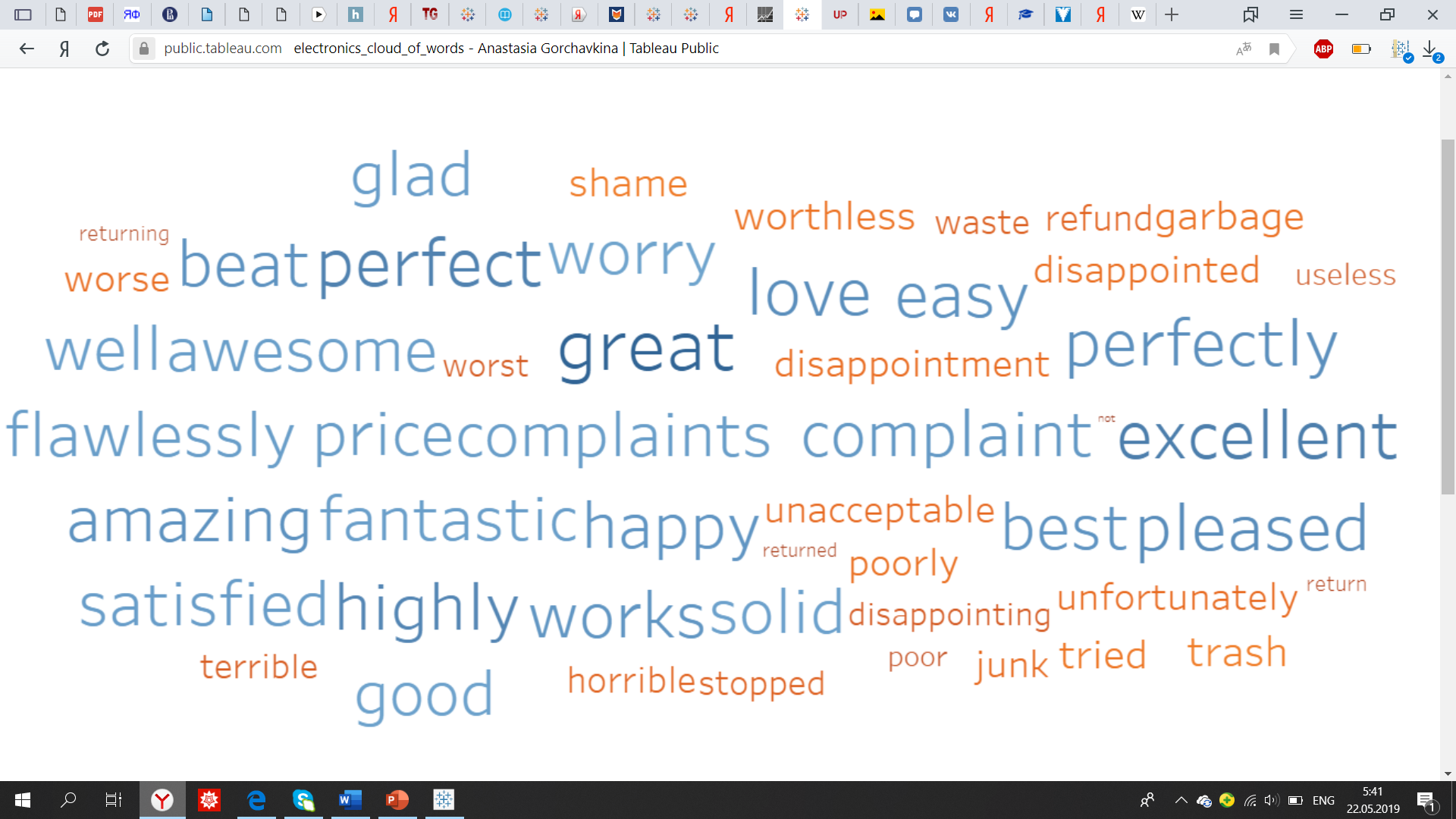


Рисунок Облако слов для категории “ Electronics”

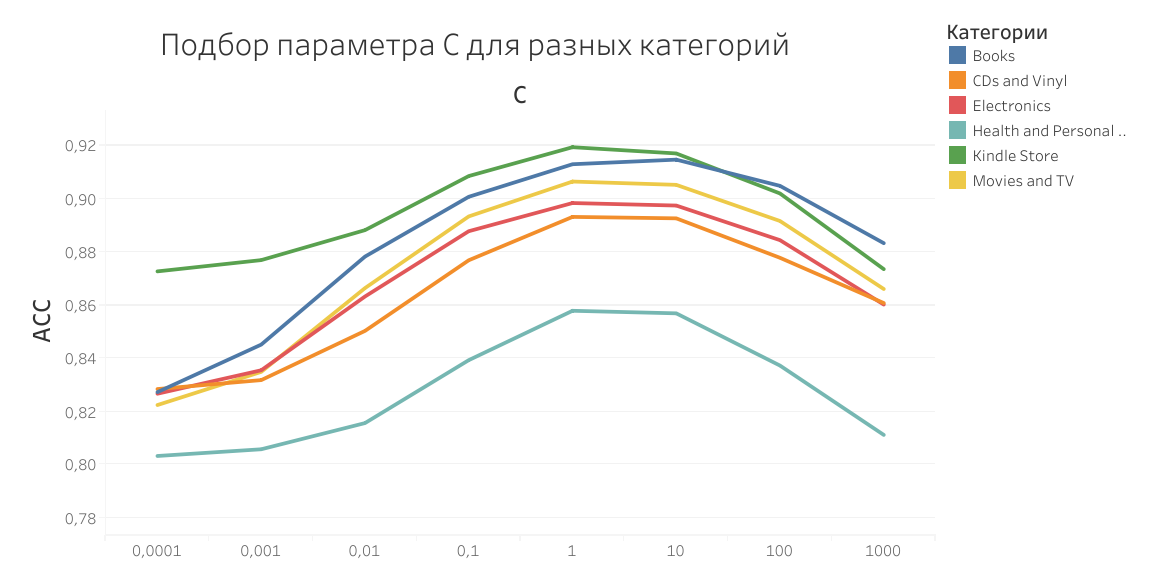


Рисунок Подбор коэффициента "c" для модели логистической регрессии

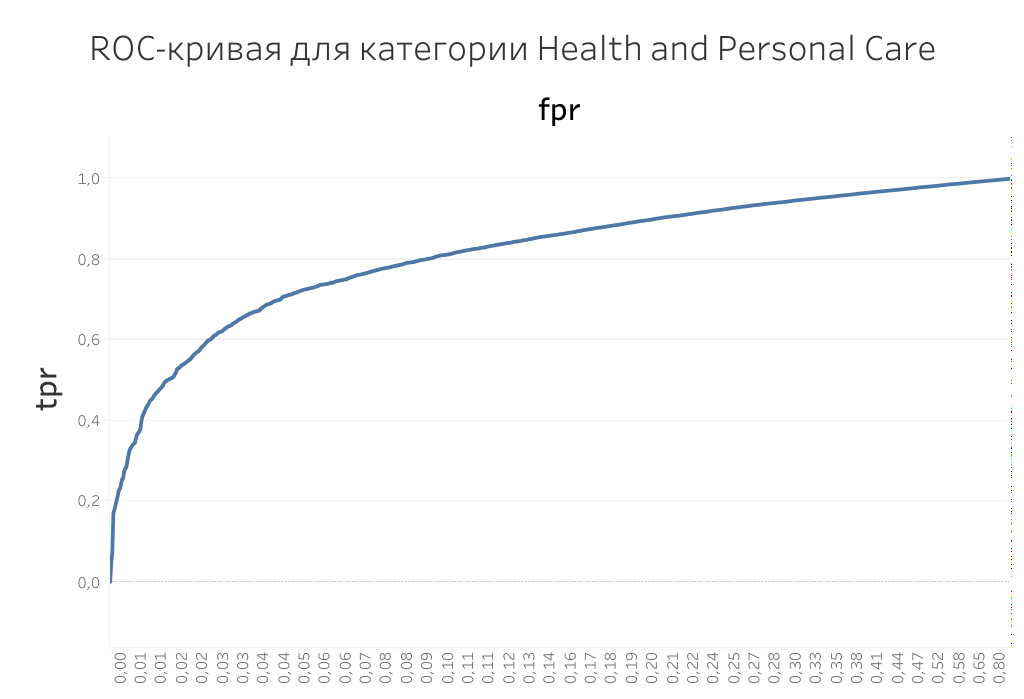


Рисунок ROC-кривая для категории товаров Healt and Personal Care

Код программы по определению тональности находится в приложении 1.

Код программы по многоклассовой классификации находится в приложении 2.

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

Приложение 1. Код по определению тональности

Приложение 2. Код по многоклассовой классификации

Презентация для защиты проекта.