# **Моделирование рекламных геотематик на основе пространственных данных**

# **социальной сети Инстаграм**

# Введение

Социальная сеть *Instagram* позволяет пользователем опубликовать запись с отметкой на конкретной местности, что позволяет анализировать пространственные тематические признаки. В данной работе представлен метод определения рекламных тематик. Метод включает разделение данных, предварительную обработку текста, фильтрацию записей, тематическое моделирование и их агрегирование. Чтобы продемонстрировать данных подход, было собрано 5,46 миллиона записей в *Instagram* с геометками для Санкт-Петербурга, Россия. В результате фильтрации рекламных постов остается 2,12 млн записей, по которым было выделено 8 основных категорий рекламы. Кроме того, была предоставлена тепловая карта распределений тем и ценового сегмента предоставляемых услуг и товаров внутри города Санкт-Петербурга.

Данные, использованные в исследовании, были собраны в 2020 году и состоят из постов Instagram с геометками, опубликованных в Санкт-Петербурге, Россия в 2017-2020 годах без временной метки. Общее количество постов составляет около 5,46 миллиона, около 4,11 миллиона из которых содержат уникальные текстовые записи, около 0,73 миллиона вообще не содержат записей. Количество записей на русском языке составляет около 3,89 миллиона, на английском - около 0,24 миллиона.

# Обзор литературы

Существует множество исследований с использованием социальных сетей с геометками для различных целей: обнаружение локальных событий [1], предупреждение стихийных бедствий [2] и исследование человеческих активностей и движений [3].

Исследование социальных сетей с использованием тематического моделирования дает много преимуществ для различных областей, таких как маркетинг, образование, исследование социальных тенденций и безопасности. В работе [4] рассматривается разработка системы поддержки маркетинга продуктов социальных сетей с использованием временного фактора. Однако, данные, которые авторы рассматривали, содержали только 11945 сообщений электронной почты 147 различных пользователей без их геометок.

Использование тематического моделирования на записях с геометками социальных сети «Твиттер», для оптимизации размещения рекламных баннеров, была рассмотрена в работе [5]. Автор выявил наиболее обсуждаемые темы для каждого района Ирландии, но заметил, что «Твиттер» часто используется для выражения мыслей о повседневной жизни без указания контекста, что усложняет нахождения актуальных тем внутри социальной сети.

# Основная часть

Записи следует разделить по геометкам. Предлагаем построить сетку с заданным шагом для исследуемой локации и дать уникальный идентификатор для каждого полученного квадрата. Для каждой записи сопоставить идентификатор квадрата, которому оно принадлежит.

![Diagram

Description automatically generated]()

Рисунок 1 - Шаги метода для обнаружения категорий реклам по текстовым данным с их геометками

## Предварительная обработка данных

Методы предварительной обработки текстов в основном зависят от языка записей, однако можно выделить основные шаги предварительной обработки записей из социальных сетей (Рис. 1). Во-первых, записи нужно очистить от эмодзи, интернет-ссылок и ссылок на другие аккаунты пользователей сети.

Далее – извлечение и разделение хэштегов, характерных для *Instagram*. В данной работе проводилась двухэтапная сегментация хештегов. Первая – предварительное разделение слов с помощью регулярных выражений: добавление пробела в хэштегах между буквами и цифрами, буквами разных языков и регистров. Второй этап уже зависит от языка. В случае русского языка можно рассматривать использование пакета для Python SymSpell[[1]](#footnote-1), который разделяет слова на основе словаря и реализует разделение с использованием треугольной матрицы, кроме того, с помощью данного пакета можно исправить орфографические ошибки и опечатки (актуально для социальных сетей). Сегментация английских слов проводилась с использованием пакета Python WordSegment[[2]](#footnote-2). Так как хэштеги – это темы, заданные самим пользователем, то разделение хештегов позволяет точнее выявлять тему конкретной записи.

Несмотря на то, что данные содержали записи на иностранном языке, было принято решение не рассматривать их в рамках данного проекта, так как рекламные записи в основном выкладываются на локальном языке. Несмотря на это оставалась работа перевода иностранных хештегов.

Последние два шага — это лемматизация и фильтрация стоп-слов. Методы лемматизация полностью зависят от языка. Для русского языка использовался пакет Python Pymorphy2[[3]](#footnote-3). Также для каждого языка собирается список стоп-слов, которые в последствии отбрасываются с записей.

## Фильтрация рекламных записей

Фильтрацию можно провести с помощью моделей классификации текстов, обученных на размеченных данных, или с помощью ключевых рекламных слов, отбрасывая записи, не содержащие их.

## Тематическое моделирование

Тематическое моделирование — это метод обнаружения тем и разметки текстов в заданном корпусе. Тематическое моделирование основано на матричном разложении: матрица векторов текстов раскладывается на матрицу тем (каждая тема кодируется вектором взвешенных слов) и матрицу документов (каждый документ представлен вектором взвешенных тем). В этой работе была использована техника аддитивной регуляризации тематических моделей, реализованную в пакете Python BigARTM [6]. Аддитивная регуляризация тематических моделей добавляет регуляризаторы к матричному разложению, которые помогают выделить второстепенные темы, сократить тематическую матрицу для более четкого разделения между темами или автоматически определить количество тем. Кроме того, BigARTM показывает лучшие результаты моделирования на текстах разной длины, что характерно для публикаций некоторых социальных сетей.

Для каждого квадрата был применен алгоритм тематического моделирования и выбрано оптимальное количество тем с помощью оценки когерентности разделяемых тем.

## Разделение записей на категории

После тематического моделирования на прошлом этапе получаются слова, наилучшим образом характеризующие тему конкретного квадрата. На их основе можно выделить основные категории рекламных записей. Для этого можно рассматривать полученные слова как отдельные документы и попробовать применить другое тематическое моделирование и выбрать наилучшее количество тем. В данном случае количество выделенных тем будет количеством категорий рекламных записей.

## Распределение категорий рекламы

## После выделения тем для каждого квадрата спроецируем их на географическую карту и построим тепловую карту распределения рекламных записей внутри рассматриваемого района.

## Анализ ценового сегмента

Рекламные записи обычно содержат цены на конкретные товары и услуги. С помощью регулярных выражений и последующей обработки можно извлечь почти все цены, предоставленные в записи.

Для русского языка предлагается следующий вариант извлечения цен (Рис. 2): для каждой записи применяется регулярное выражение r'(\d+[ \.\']?\d\*) ?(руб|₽|рублей)\.?[\W]'. Тогда извлекутся разделенные через точку или пробел два числа или просто одно число, после которого идет метка локальной валюты.

Далее, рассматриваем два случая: если извлеченная строка состоит из одного числа, то мы рассматриваем его как цену в записи, если из двух, то если длина первого числа меньше трех и не начинается с символа ‘0’, то мы объединяем два числа.

![Diagram

Description automatically generated]()

Рисунок 2 - Шаги метода извлечения цен с рекламных записей

# Результаты

Используя метод, описанный выше, пытаемся найти распределение рекламы внутри города Санкт-Петербург, которые можно потом использовать для других исследований или маркетинговых мероприятий.

Непустые записи были разделены на квадраты 100 на 100 метров, однако квадраты, содержащие менее 100 записей, не рассматривались. В результате было получено 6000 квадратов (Рис. 3)

Chart, histogram

Description automatically generated

Рисунок 3 - Распределение количества слов в постах

Все этапы, описанные в разделе предварительной обработки данных, были реализованы, а также помимо этого были удалены все слова короче трех символов и была добавлена проверка символа на кириллицу.

Для последующей фильтрации было испробовано несколько подходов: обучение алгоритма классификации на основе TF-IDF Vectorizer, а также составление множества рекламных слов, по которому остаются записи, содержащие данные слова. Алгоритм классификации, обученный более чем на 2000 размеченных записей, показал высокие результаты на отложенной выборке (*86%* доля верно определенных). Однако, модель не смогла обобщить полученные признаки на остальные записи. С помощью последнего подхода выделено *2,12* млн записей рекламного характера. В список рекламных слов входят такие термины как: покупка, клиент, цена, маникюр, директ и т. д. (всего 53 слова). Помимо этого, были удалены все записи, у которых доля символов кириллицы были меньше, чем 50% всего содержимого. Это было сделано для того, чтобы убрать записи на других языках.

После фильтрации записей остался 1641 квадрат. На каждом квадрате запускается алгоритм тематического моделирования, описанный на шаге 2.3. В результате выявлено 4089 уникальных тем. Количество тем в одном квадрате варьируется от 2 до 35. Сохранены наиболее значимые слова для определения темы в каждом квадрате.

Второе тематическое моделирование для определения отдельных категорий рекламных записей на темах, выделенных после первого тематического моделирования, проводилась на трех разных алгоритмах: LDA (Latent Dirichlet allocation), LSI (Latent Semantic Indexing), NMF (Non-Negative Matrix Factorization). Далее для определения оптимального количества категорий была исследована зависимость когерентности от выбранного количества тем в документе (Рис. 4). Для оценивания когерентности использовалась поточечная взаимная информация (PMI).

Graphical user interface, application

Description automatically generated

Рисунок 4 - Зависимость когерентности от количества тем для каждой из модели

После детального сравнения была выбрана модель NMF и обучена на 8 категориях. Слова, наиболее характеризующие конкретную тему, показаны на таблице 1. Темы были интерпретированы следующим образом:

1. Салоны красоты
2. Маникюр
3. Одежда
4. Услуги красоты
5. Фитнес
6. Дизайн квартир/домов
7. Доставка товаров
8. Услуги фотостудий

Вследствие, в каждом квадрате было выделено от 1 до 4 отдельных категорий. Видно, что некоторые темы пересекаются по словам, скорее всего первая тема является рекламой самих салонов, а четвертая тема больше по конкретным услугам.

Таблица 1. Выявленные категории рекламы.

|  |  |
| --- | --- |
| № темы | Слова |
| 1 | работа, кожа, директ, хороший, свой, самый, мастер, делать, запись, процедура, уход, лицо, писать, крем, волос, средство, маска, масло, косметика, рука |
| 2 | дизайн, маникюр, ноготь, гельлак, спб, педикюр, гвоздь, покрытие, шеллак, гель, питер, красивый, лак, проспект, заказ, френч, выравнивание, аппаратный, район, наращивание |
| 3 | одежда, размер, доставка, заказ, цена, наличие, магазин, платье, спб, директ, мода, москва, купить, женский, костюм, стиль, цвет, россия, сумка, шоурум |
| 4 | красота, бровь, мастер, волос, запись, спб, процедура, окрашивание, работа, наращивание, кожа, ресница, красивый, цвет, маникюр, ноготь, лицо, питер, салон, делать |
| 5 | свой, самый, цвет, новый, тело, хороший, любить, большой, цена, петербург, красивый, купить, работать, делать, россия, скидка, еда, спб, руб, рубль |
| 6 | петербург, дом, свой, интерьер, работа, спб, новый, место, хороший, большой, дизайн, питер, санкт, город, проект, квартира, кухня, дизайнер, друг, знать |
| 7 | новый, свой, подарок, хороший, заказ, спб, фото, сделать, профиль, писать, работа, большой, любой, самый, делать, директ, доставка, друг, питер, ссылка |
| 8 | питер, спб, причёска, запись, макияж, модель, фото, стилист, студия, фотосессия, фотограф, санкпетербург, визажист, красота, директ, петербург, свадебный, детский, красивый, свадьба |

После полученных рекламных категорий была построена тепловая карта рекламных записей в г. Санкт-Петербурге с возможностью фильтрации по определенной категории (Рис. 5). Карта была построена с помощью пакета Python Ipyleaflet [[4]](#footnote-4)с радиусом каждой точки 20 пунктов. Оказалось, почти все категории рекламы сосредоточены в центральном и близких к ним районов.

Далее были извлечены цены с рекламных записей описанный на шаге 2.6. После применения регулярного выражения записей с ненулевым количеством цен осталось 308240 записей, которые содержали в общем 559417 цен. В основном в записях указывают три ценовых значения. Исследуя распределение цен до 10000 рублей, можно заметить «пики» графика рядом с круглыми значениями. Далее можно построить тепловую карту минимальных, медианных и максимальных цен в квадратах (Рис. 7). Можно заметить, что в дальних от центра районов сосредоточились медианные и максимальные цены, а минимальные цены встречаются больше в центре города.

# Map Description automatically generated

Рисунок 5 – Тепловая карта категории "Фитнес" в городе Санкт-Петербург



Рисунок 6 - Описание извлеченных цен а) Распределение количества цен в одной записи б) Распределение цен до 10000 рублей

A screenshot of a map

Description automatically generated with medium confidence

Рисунок 7 - Распределение цен внутри квадратов а) Минимальные цены б) Медианные цены в) Максимальные цены

# Заключение

Это исследование реализует общую схему обнаружения географически рассредоточенных категорий реклам из данных социальных сетей. Описан подход в общем виде для модификации алгоритма и дальнейшего применения на данных других языков, районов или социальных сетей. Для использования данного метода требуются только текстовые записи и их геометки, что существенно облегчает сбор данных.

В будущем планируется улучшить классификатор рекламных записей путем увеличения размеченных данных или использованием более продвинутых технологий классификаций текстов. Не все рекламные категории были выделены, так как большинство рекламных записей в *Instagram* состоят их постов салонов красоты, что не позволяет в полной мере анализировать более мелкие категории реклам. В работе были проанализированы только числовые значения цен, далее планируется добавить анализ цен текстового вида.

# Список использованных источников

1. C. Кисилевич. Событийный анализ действий и поведения людей с использованием фотоколлекций Flickr и Panoramio с геотегами / М. Крстаич, Д. Кейм, Н. Андриенко и Г. Андриенко // Международная конференция по визуализации информации: сб. статей. - Лондон, 2010, C. 289–296.
2. Т. Сакаки (T. Sakaki). Землетрясение потрясает пользователей Twitter: обнаружение событий в реальном времени с помощью социальных датчиков / М. Оказаки (М. Okazaki) и Ю. Матсуо (Ю. Matsuo) // Международная конференция «World wide web»: сб. статей. – Нью Йорк, 2010. – С. 851-860.
3. Р. Юрдак (R. Jurdak). Исследование движений людей с помощью твиттера / К. Чжао (K. Zhao), И. Лю (J. Liu), М. Абужауд (M. AbouJaoude), М. Камерон (M. Cameron) и Д. Ньют (D. Newth) // PloS one. – 2015. – №10. – С. 1-16.
4. М. Нгуенг (M. Ngueng). Анализ социальных сетей на основе тематического моделирования / Т. Хо (T. Ho) и П. До (P. Do) // Международная конференция «Вычислительные и коммуникационные технологии»: сб. статей. – Ханой, 2013. – C. 119-122
5. M. Кул. (M. Cool). Тематическое моделирование: офлайн-реклама на основе местоположения с использованием Twitter // Дипломная работа. – Дублин, 2018. – С. 1-20
6. К. Воронцов. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. – 2014. – №3. – C. 193-202.

1. https://pypi.org/project/symspellpy/ [↑](#footnote-ref-1)
2. https://pypi.org/project/wordsegment/ [↑](#footnote-ref-2)
3. https://pypi.org/project/pymorphy2/ [↑](#footnote-ref-3)
4. https://pypi.org/project/ipyleaflet/ [↑](#footnote-ref-4)