Санкт–Петербургский государственный университет

Сошникова Мария Дмитриевна

Выпускная квалификационная работа

Методы определения геопозиций пользователей бизнес-аккаунта в социальной сети Твиттер

Уровень образования: бакалавриат
Направление 01.03.02 «Прикладная математика и информатика»
Основная образовательная программа СВ.5005.2017 «Прикладная математика, фундаментальная информатика и программирование»
Профиль «Математическое и программное обеспечение вычислительных машин»

Научный руководитель:
доцент, кафедра технологии программирования, канд. техн. наук
Блеканов Иван Станиславович

Рецензент:
профессор, кафедра управления медико-биологическими системами, д-р. физ.-мат. наук
Утешев Алексей Юрьевич

Санкт-Петербург
2021 г.

Содержание

Введение

После нескольких десятилетий онлайн-покупок, процент потребителей, предпочитающих делать покупки в обычных розничных магазинах, составляет 85,7% [1]. Многие люди предпочитают покупать одежду и обувь в офлайн магазинах, где предоставляется возможность примерить различные модели и приобрести товар без ожидания его доставки. Именно поэтому небольшие бренды, изначально реализующие товары посредством онлайн-продаж, открывают офлайн магазины. При открытии офлайн магазина обычно проводятся маркетинговые исследования и анализ рынка. В данном случае может быть полезным изучить аудиторию аккаунтов бренда в социальных сетях.

Социальные сети предлагают много преимуществ для владельцев бизнеса. Располагая более чем 300 млн активных пользователей в месяц, [2] Твиттер является эффективным инструментом для поиска потенциальных клиентов и взаимодействия с ними. В последнем исследовании "Hootsuite Social Trends 2021 Survey", проводимом в конце 2020 года, 33% опрошенных выделили Твиттер как наиболее эффективную платформу для достижения бизнес-целей [3].

Информация о месте проживания заинтересованных в продукте или услуге людей помогает грамотно скорректировать бизнес-стратегию и получить больше прибыли. Основываясь на геопозиции, определенной по личным страницам пользователей, можно исследовать местоположение группы клиентов. Однако, из-за того, что при сздании аккаунта указание геопозиции не является обязательным полем для заполнения, поле с указанной пользователем геопозицией чаще всего остается незаполненным или имеет недостоверную информацию, что было подтверждено Hecht и др. в [4]. Поэтому, внимание многих исследователей направлено на поиск решений для определения местоположений пользователей на основе их аккаунтов в социальной сети. Также стоит отметить, что количество подписчиков аккаунта может быть очень большим, что делает невозможным получение информации на основе всех подписанных пользователей без использования программных реализаций алгоритмов анализа данных.

Глава 1. Постановка задачи

Целью данной работы является исследование подходов для определения геопозиций пользователей в социальной сети Твиттер, их применение для определения местоположений подписчиков бизнес-аккаунта и отображение полученных данных в удобном для восприятия формате.

В рамках данной работы были выделены следующие этапы:

- Обзор существующих решений задачи определения геопозиций пользователей

- Получение списка подписчиков бизнес-аккаунта и их постов в необходимом для нейросетевой модели формате

- Сравнение результатов работы различных моделей для определения местоположения пользователей по тексту их постов и с учетом их взаимодействия в социальной сети Твиттер

- Построение нейросетевой модели для определения геопозиций подписчиков бизнес-аккаунта

- Визуализация полученных результатов в виде тепловой карты геопозиций пользователей

Глава 2. Анализ социальной сети Твиттер

Будучи одной из самых популярных социальных сетей в интернет-пространстве, Твиттер постоянно накапливает большое количество разнородных данных с высокой скоростью. К ним относятся: 1) короткие и малоинформативные текстовые сообщения(твиты), размещенные пользователями, 2) социальный граф, который образуют собой пользователи, 3) многообразие контекстной информации как на основе пользователей, так и на основе твитов.

2.1 Контент в Твиттере

Твит – это небольшой текст, сгенерированный пользователем, длиной до 140 символов. Он может описывать любую информацию, которую пользователь захотел опубликовать, например, его настроение или информацию о мероприятиях и событиях, которые происходят вокруг него. Помимо обычных постов, пользователь может также делать ретвит, т.е. делиться постами других пользователей. Твиты и ретвиты пользователя отображаются у его подписчиков, которые могут их прочитать и отреагировать на них. При написании поста для публикации, пользователи могут включать в текст хэштеги, которые представляют собой слова или фразы не разделенные пробелами, начинающиеся с символа “#”. Наконец, в тексте твита можно также упомянуть имя другого пользователя с помощью символа “@” и его никнейма. Упомянутый пользователь будет уведомлен и может начать диалог посредством упоминания автора поста.

2.1.1 Социальный граф пользователей

Помимо публикации постов, пользователь может получать новые твиты других пользователей, подписавшись на них. Если пользователь u_i подписан на пользователя u_j, мы называем u_i подписчиком u_j. Важно заметить, что отношения имеют однонаправленный характер, т.е. из того что u_i подписан на u_j не следует то что u_j подписан на u_i. Когда взаимностью отношений можно пренебречь, мы называем u_i и u_j друзьями. Если ока-

зывается, что u_i и u_j подписаны друг на друга, мы говорим, что u_i и u_j взаимные друзья. В дальнейшем все отношения пользователей будем называть дружбой в Твиттере или просто дружбой, когда контекст понятен. Необходимо отметить, что дружба в Твиттере не подразумевает дружбы в реальной жизни. Распространенный факт, что знаменитости не знакомы с большинством своих подписчиков. Более того, даже два незнакомца, находящихся на большом расстоянии друг от друга, могут случайно стать взаимными друзьями. Однако замечено, что друзья в реальной жизни, как правило, часто общаются друг с другом в интернете [5], [6], [7], [8]. При использовании информации о дружбе пользователей в Твиттере, мы рассматриваем подписки и упоминания пользователей как равнозначные действия и строим основанный на них граф – социальный граф пользователей.

2.1.2 Контекст твитов

Твит это больше, чем короткий текст. Когда твит публикуется, он получает отметку о времени публикации. Кроме того, с преобладанием мобильных устройств с поддержкой GPS, таких как смартфоны и планшеты, пользователи могут по желанию публиковать свои текущие местоположения в виде геотегов в твитах. Наконец, пользователи могут указывать в основной информации своей страницы такую информацию, как родной город, часовой пояс и личный веб-сайт. Можно отметить, что вся вышеприведенная информация предоставляет собой контекст, который помогает лучше проанализировать твиты. Обыденные твиты пользователя могут быть интерпретированы более точно, если вся эта информация доступна. Поскольку метки времени, геотеги и основная информация со страниц пользователей служат контекстной информацией для твитов, их называют контекстом твита.

Глава 3. Обзор существующих решений

3.1 Методы, основанные на содержании твитов

На местоположение могут указывать определенные слова в содержании твитов. Например, петербуржцы более вероятно будут обсуждать матчи ФК Зенит, чем москвичи. Жители Новосибирска чаще употребляют в своей речи такие диалекты как “мультифора” и “карачинская”, а люди из краснодарского края называют баклажаны словом “синенькие” [9].

Предыдущие исследования по предсказанию местоположения пользователя на основе контента можно разделить на две группы: основанные на словах и основанные на местоположениях. Метод, основанный на словах, заключается в оценке вероятности принадлежности к местоположению l заданных в тексте слов w, или $p(l|w)$; в то время как метод, основанный на местоположении, заключается в поиске вероятности появления твита d в заданной локации $p(d|l)$.

3.1.1 Методы, основанные на диалектах

В начале раздела 3.1 приведены примеры слов, указывающие на местоположение в твитах пользователей. Методы, основанные на диалектах, направлены на поиск и использование таких слов для определения местоположения пользователей. Не все слова указывают на местоположение. Например, такие слова как “время” и “большой” используются очень часто. Поэтому следует использовать только местные слова, т.е. слова, которые используются пользователями, проживающими на одной территории и указывающие на их местоположение. Кроме того, информация о местоположении, подразумеваемая местными словами или их пространственным словоупотреблением, должна быть извлечена из данных, прежде чем делать прогнозы. Далее будет рассмотрено как обе эти задачи решаются в имеющихся исследованиях.

3.1.1.1 Выявление локальных слов

В литературе по информационному поиску общепринятой практикой является исключение стоп-слов из документов перед их индексированием для поиска. В случае с твитами, местные слова, по которым можно определить местоположение гораздо реже встречаются в тексте, чем обычные часто употребляемые слова. Это может привести к тому, что результаты определения местоположения станут случайными. В данном случае вместо удаления предопределенного списка стоп-слов производится устранение слов, по которым нельзя определить местоположение. Поскольку местные слова, в отличие от стоп-слов, не поддаются перечислению, большое количество усилий тратится на поиск и определение местных слов.

Методы идентификации локальных слов без учителя направлены на статистические показатели, которые непосредственно вычисляются на основе данных и указывают на локальность слова. Вдохновленные Inverse Document Frequency (IDF), Ren и др. в [10] и Han и др. в [11] предлагают Inverse Location Frequency (ILF) и Inverse City Frequency (ICF), соответственно, для измерения локальности слов. Их предположение состоит в том, что местные слова должны быть распространены в меньшем количестве мест и иметь большие значения ILF и ICF. Mahmud и др. [12] применяют ряд эвристических правил для выбора местных слов. Han и др. [13] предлагают сравнение статистических, основанных на теории информации и эвристических методов выбора локальных слов.

Методы идентификации слов с учителем также рассматриваются в ряде исследований. В [14] Cheng и др. рассмотрели проблему поиска местных слов как задачу классификации. Сначала авторы строят географическое распределение каждого слова с помощью пространственной вариационной модели представленной Backstorm и др. в [15]. Пространственная вариационная модель предполагает, что каждое слово имеет географический центр, центральную частоту C и коэффициент дисперсии α. Вероятность использования слова в месте с расстоянием d до центра пропорциональна $Cd^{-\alpha}$. После обучения модели, параметры используются как характеристики слов. Затем авторы вручную разметили 19178 слов из словаря как

локальные или не локальные. Наконец, была обучена классификационная модель и применена к остальным словам в наборе данных твита. Ryoo и Moon [16] применили вышеуказанный метод к набору данных твитов на корейском языке и достигли удовлетворительных результатов.

3.1.1.2 Сопоставление локальных слов местоположениям

После идентификации местных слов следующая проблема заключается в том, как использовать их для прогнозирования местоположения пользователей. Большинство исследователей предлагают вероятностные модели для характеристики условного распределения местоположений пользователей по содержанию их твитов, а затем конкретизируют модель для прогнозирования.

Репрезентативная вероятностная модель представлена Cheng и др. в [14]. Распределение местоположения l пользователя u с учетом содержания его твита $S(u)$, принимает вид $P(l|u) \propto \sum_{w \in S(u)} P(l|w)P(w)$. Здесь рассматривается только локальное слово w, а $P(w)$ обозначает вероятность w по всему тексту. После разложения большие усилия тратятся на оценку распределения $P(l|w)$ местоположения слова w или территориального использования слова. Сообщается, что оценка $P(l|w)$ непосредственно из всего текста является худшей. Причина в том, что некоторые слова w могут не использоваться в менее населенных местах, что не означает, что местоположение не имеет отношения к слову w. Чтобы решить эту проблему с разреженностью, необходимо задействовать методы сглаживания. Особый тип локальных слов – это названия мест в твитах. Li и др. в [17] заметили, что вероятность упоминания названий мест в твитах как может зависить от местоположения, так и может быть случайной. Таким образом, авторы делают двухуровневую оценку. Распределение Бернулли применяется для оценки того, было ли упомянуто название места случайным образом или на основе местоположения, после чего полиномиальное распределение используется для оценки вероятности публикации твита с названием места из каждого местоположения.

3.1.2 Методы, основанные на геопозициях

В нескольких исследованиях используются основанные на классификации подходы к прогнозированию местоположения пользователей. Исследователи рассматривают статистику пользователей о локальных словах как признаки, а всевозможные местоположения как метки классификации. Hecht и др. в [4] выбирают 10000 слов с наивысшими баллами `CALGARI` как местные слова. Затем пользователи представляются в виде 10000-мерных векторов частоты терминов и подаются на вход в полиномиальный наивный байесовский классификатор для обучения и прогнозирования местоположений пользователей. Похожим образом Rahimi и др. [18] применяют логистическую регрессию к векторам TF-IDF пользователей. Вместо того, чтобы выбирать локальные слова в качестве признаков, они добавляют ограничения, используя L1-регуляризацию. В [12] Mahmud и др. применяют иерархический ансамбль алгоритмов для обучения ансамблей двухуровневых классификаторов для определения местоположения с различной степенью детализации, такой как город, штат или часовой пояс. В своей расширенной работе [19] авторы также предлагают выявлять и удалять путешествующих людей из обучающей выборки для улучшения производительности классификаторов местоположения пользователей. Человек считается путешествующим, если любые два его твита были отправлены из мест с расстоянием больше 100 миль между ними.

Существуют также исследования, в которых для определения местоположений пользователей используются подходы, основанные на информационном поиске. В данном случае местоположения рассматриваются как псевдодокументы, состоящие из твитов всех пользователей, проживающих в данной локации. Рассматривая псевдодокумент пользователя, местоположение которого необходимо предсказать, наиболее схожие псевдодокументы выдаются в качестве результатов прогнозирования. В частности, Wing и др. в [20] представляют местоположение в виде сетки. Авторы оценивают языковую модель [21] для каждой сетки с ее псевдодокументом. Сглаживание Гуда-Тьюринга [22] применяется для сглаживания вероятностей невидимых слов. Расхождение Кульбака-Лейблера используется в качестве

меры подобия между псевдодокументами местоположений и пользовательскими псевдодокументами. В своей последующей работе [23] они прибегают к адаптивным сеткам, как в [24].

Помимо традиционных методов, в некоторых работах также исследуются модели глубокого обучения для прогнозирования местоположения пользователей. Продолжая свою предыдущую работу [25], Miura и др. в [26] предлагают более сложную модель. Авторы упорядочивают твиты пользователя в хронологическом порядке и применяют последовательную модель RNN для кодирования. Благодаря механизму внимания, можно получить общее представление твита, в котором содержится важная информация. Аналогичный процесс также применяется и к контексту, т.е. к описанию местоположения твита и часовому поясу. Затем комбинация трех представлений подается на слой softmax для определения местоположения пользователя. Rahimi и др. в [27] применяют многослойный перцептрон (MLP) с одним скрытым слоем для классификации местоположений пользователей. Авторы используют в качестве входных данных представление твита пользователя в виде L2-нормализованного мешка слов. Выходные данные представляют собой предопределенную дискретизированную область, сгенерированную либо k-d деревом, либо алгоритмом k-средних.

3.2 Методы, основанные на социальном графе пользователей

Помимо публикации твитов, другие важные действия, которые пользователи совершают в Твиттере это подписка на аккаунты других пользователей и взаимодействие с друзьями. Как и содержание твитов, социальные отношения пользователей в Твиттере также могут указывать на их геопозицию. В разделе 3.2.1 будут рассмотрены методы, основанные на дружбе, использующие предположение, что друзья имеют меньшие расстояния между их геопозициями.

Более того, в исследованиях также утверждается, что социальная близость пользователей, которая основана на дружбе, взаимодействиях и других неявных признаках, более надежна для оценки местоположений,

чем дружба в отдельности. Подробнее эти исследования будут рассмотрены в разделе 3.2.2.

3.2.1 Методы, основанные на дружбе

В социологии существует понятие “гомофилия”, обозначающее тенденцию индивидов с большей вероятностью вступать в контакт с похожими на них людьми. Рассматривая данное понятие в контексте нашей задачи, местоположение пользователя наиболее вероятно совпадает с местоположением большинства его друзей. В своей модели [10] Ren и др. предположили, что чем большее количество друзей пользователя живет в том или ином месте, тем выше вероятность того, что пользователь находится в том же месте. Davis и др. в [28] используют аналогичный подход, за исключением того, что рассматривают только взаимную дружбу.

Все вышеупомянутые методы неявно предполагают, что дружба, наблюдаемая в социальной сети, подразумевает дружбу в реальной жизни и следовательно небольшое расстояние между местоположениями пользователей-друзей. Однако, это может быть далеко не так. В [29] Kong и др. обнаружили, что пара друзей живет в пределах 10 км с вероятностью 83%, если их общие друзья составляют более половины от всего списка их друзей. Вероятность снижается до 2,4%, если количество общих друзей составляет 10%. Это означает, что крепкая дружба, основанная на общем количестве друзей пользователей, может лучше указывать на дружбу в реальной жизни и, следовательно, на небольшое расстояние между геопозициями пользователей. Rout и др. в [30] также связывают вероятность того, что пользователь живет в конкретном городе с распределением непрямых дружеских отношений между пользователем и его друзьями в данной локации.

3.2.2 Методы, основанные на социальной близости

В социальной сети Твиттер упоминание – еще один вид взаимодействия пользователей. Когда пользователи упоминают друг друга или ведут беседы друг с другом, считается, что эти пользователи имеют близкие взаимоотношения или схожие интересы. Информация о таком виде дружбы

очень полезна для определения геопозиций пользователей. McGee и др. в [5] провели анализ 104214 пользователей, проживающих в США. Авторы выяснили, что помимо взаимной дружбы, основанной на подписках, упоминания и беседы пользователей также указывают на небольшое расстояние между их геопозициями. В следующей своей работе [6] McGee и др. подтвердили свои наблюдения, изучив более большой набор данных. Также были отмечены и другие наблюдения: 1) если аккаунт подписчика пользователя закрытый, т.е. другим пользователям требуется разрешение для подписки на этот аккаунт, то эти два пользователя располагаются недалеко друг от друга; и 2) аккаунты местных новостных источников располагаются в одной локации со своими подписчиками. Считая географическую близость напрямую связанной с социальной близостью, McGee и др. обучили дерево принятия решений для сопоставления социальной близости между различными пользователями десяти квантилям. Также как и McGee в [5][6], Compton и др. в [7] также используют информацию об упоминании пользователей. Авторы строят граф упоминаний пользователей и определяют неизвестные местоположения так, что пользователи, упоминающие друг друга находятся близко друг к другу. Jurgens [8] также рассматривает взаимные упоминания пользователей вместо дружбы. Rahimi и др. в [18] утверждают, что взаимные упоминания слишком редко встречаются, чтобы быть полезными. Авторы рассматривают упоминание одним пользователем другого как ненаправленное ребро в графе.

Помимо того, что упоминания и беседы пользователей могут быть показателем социальной близости, в некоторых исследованиях также показано, что данные взаимодействия могут иметь негативное влияние при определении местоположения пользователя. Например, пользователь из Екатеринбурга может быть подписан на знаменитость из Москвы и известного исполнителя из Санкт-Петербурга. Установление данного типа отношений не является следствием социальной близости между пользователем и знаменитостью, а определяется социальным влиянием знаменитостей. Данное интуитивное предположение, рассмотренное на этом примере, подтверждено несколькими исследованиями. Анализируя большой набор данных, Kwak и др. в [31] выяснили, что пользователи, с менее чем 2000 взаимными

друзьями (что является фактом того, что они маловероятно имеют большое влияние) наиболее вероятно будут географически близки к большинству из них. В работе [6] McGee и др. также описывается, что друг пользователя u, у которого много друзей и подписчиков, обычно расположен географически дальше от u. В [32] Li и др. описывают модель влияния пользователя, доказывающую вышеупомянутые предположения. Авторы представляют влияние пользователя как двумерное распределение Гаусса с центром в его местоположении, причем дисперсия распределения интерпретируется как объем его влияния. Вероятность того, что пользователь u_i подпишется на u_j, измеряется плотностью вероятности распределения влияния u_j в местоположении u_i. В расширении [17] своей более ранней работы [32] Li и др. осуществляют определение нескольких местоположений для пользователя. Главная идея данного подхода состоит в том, что у многих людей могут быть родные города, а также города, в которых люди учатся или работают, не совпадающие с их родным городом. Следовательно, пользователи могут быть подписаны не только на друзей, живущих поблизости и знаменитостей, живущих далеко, но и на коллег и одногруппников, проживающих в городе с местом работы или учебы соответственно.

3.3 Методы, основанные на контексте твитов

Mahmud и др. в [12], [33] учитывают время публикации твита, значения которого в рассматриваемом датасете представлены в формате GMT. После разделения дня на временные промежутки равной длины, пользователи рассматриваются как распределения времени публикации их твитов. Затем при обучении классификатора также учитываются сдвиги распределений, вызванные разницей в часовых поясах. Efstathiades и др. в [34] используют вероятностную модель, основанную на временном распределении географических меток, связанных с твитами, для оценки местоположения дома пользователя и рабочего места. Метод основан на наблюдении авторов, что публикация твитов пользователем в нерабочее время (например, поздним вечером) наиболее вероятно происходит из “домашнего” местоположения, в то время как публикация постов во время рабочего времени

наиболее вероятно происходит из рабочей локации. Poulston и др. в [35] также используют в своей работе геотеги, но также отмечают, что обычно пользователи активны сразу в нескольких локациях. Авторы сначала кластеризуют геотеги, а затем группу с наибольшим количеством постов определяют как “домашнюю” геопозицию. Геометрическую медиану всех точек в “домашнем кластере” принимают за координаты “домашней” геопозиции.

Глава 4. Используемые данные

4.1 Данные, используемые для обучения моделей

Для решения поставленной задачи были выбраны нейросетевые модели, использующие обучение с учителем. В данном случае входные данные представляют собой полностью размеченный датасет, т.е. каждому элементу в представленном наборе данных соответствует значение, которое алгоритм должен получить после выполнения. В контексте определения геопозиций пользователей, входными данными является файл, содержащий в себе имена пользователей, координаты их местоположений и коллекции их твитов. Ввиду сложности сбора геопозиций пользователей, в качестве набора для обучения был выбран датасет `Twitter-US`, представленный Roller и др. в [24] и находящийся в свободном доступе.

Данный датасет представляет собой архив твитов, собранных в течение трех месяцев с помощью `Twitter Spritzer`. При сборе учитывались только публикации, имеющие геометки, и авторы которых написали не более чем 1000 твитов за данный период. Каждый пользователь представлен совокупностью своих твитов и помечен широтой и долготой первого собранного твита с геометкой. Фрагмент с информацией об одном из пользователей из данного набора данных представлен на Рис. 1. Полученный датасет состоит из 38 млн сообщений от 449694 пользователей из Северной Америки.

```
Chamerlik       42.06486        -87.9382        Kurt Vonnegut – Writing 101 | Luke James | Blog Post | Red Room: http://t.co/
ZYVw6aiZ ||| Sitting in the most crowded Starbucks in the world, I think. #starbucks ||| @chicagotribune Lol. Why even run
this poll? ||| iTunes is offering free download off Miles Davis Quintet Live in Europe Bootleg. Get it here! http://t.co/
ytlcTZQz #jazz #milesdavis ||| @BallyFitness @AMWFitness Great, but they're hard on your teeth! :( ||| @Jazzaholic1 Thanks.
Yes. Nice town. Small feel, as if you are getting away from city but still a suburb. Event at the Ski jump here today. |||
Woman with her kid at Starbucks. Kid running around in his pajamas. #Starbucks ||| @ArtistSusan Must have something to do
with me. :) ||| RT @Samuel_Clemons: ALERT: The Deleted Star Wars Clips: you know. the part where da #ferret saves da Universe
and stuff and steals the ... ||| Seth's makes a good point: "First, make rice." http://t.co/yJrhQYhN ||| @LaurieHawley
@ThomasMarzano @HugoR But from a Buddhist perspective, what you think you "are" is your own "myth" anyway. :)
riflekind       36.4364754      -77.09788799    im fairly certain that they're not speaking english ||| I'm at Chowan
University – Marks Hall (One University Place, Murfreesboro) http://t.co/kgeyE6cT ||| I'm at Chowan University – Marks Hall
(One University Place, Murfreesboro) http://t.co/6ekSfqHZ ||| @thollux omg :( the only good part is getting to wear fancy
dresses :( ||| chicken boy was written by tegomass themselves ke ke keke ||| every time i see this i think it's nakamaru
```

Рис. 1: Фрагмент с информацией об одном пользователе из обучающего набора данных

4.2 Сбор данных о подписчиках бизнес-аккаунта

В данной работе в качестве примера бренда, перешедшего из онлайна в офлайн, рассматривается компания “Allbirds”, занимающаяся дизайном и продажей обуви и одежды [36].

Процесс сбора данных был разделен на два этапа: сначала были получены никнеймы подписчиков аккаунта, а затем осуществлялся сбор твитов каждого пользователя из полученного списка.

Для получения датасета, содержащего информацию о подписчиках аккаунта, использовался Selenium – инструмент для автоматизации действий браузера [37]. На первом этапе программа переходит на страницу со списком подписчиков аккаунта и получает html код страницы с помощью метода `page_source`. Затем в полученном html коде страницы производится поиск тегов, содержащих никнеймы пользователей. После извлечения всех никнеймов из тегов на данной странице, selenium листает страницу вниз для отображения следующей части списка подписчиков. Описанные выше действия повторяются до тех пор, пока не будут получены все подписчики данного аккаунта. В результате был получен массив, содержащий 28760 никнеймов подписчиков аккаунта “Allbirds”.

После формирования массива никнеймов, осуществляется второй этап сбора данных – получение текста твитов каждого подписчика аккаунта. Для этого Selenium переходит на каждую из личных страниц пользователей из массива никнеймов и получает html код страницы. Далее осуществляется поиск html тегов, содержащих текст постов пользователя. Полученные посты записываются в массив. Затем происходит загрузка следующей части страницы и описанные выше действия повторяются. Полученные твиты записываются в одну строку с разделителем “|||”. Далее формируется объект, состоящий из массива никнеймов и массива, содержащего строки с твитами всех пользователей. Если при переходе на страницу оказывается, что аккаунт является закрытым, т. е. доступ к постам пользователя ограничен или твиты на странице отсутствуют, то осуществляется переход на страницу следующего пользователя из массива никнеймов. После получения информации обо всех подписчиках, из данных объекта пользователи -

посты формируется датафрейм и записывается в csv файл. В полученном файле в первом столбце находятся никнеймы подписчиков, а во втором столбце строки с их твитами.

Результатом проделанных действий является файл, содержащий никнеймы подписчиков и их твиты, объединенные в строки. Для аккаунта бренда “Allbirds” были собраны посты 20082 подписчиков. Полученные данные имеют формат, схожий с датасетом для обучения модели.

Глава 5. Выбор и обучение моделей

Тексты твитов и информация о взаимодействии в социальной сети одинаково важные параметры при определении геопозиций пользователей. Некоторые пользователи публикуют много контента, но имеют не информативный или не отражающий их местоположение социальный граф. В таком случае тексты постов являются основным признаком для определения местоположения. Другие пользователи используют социальные сети в основном чтобы читать комментарии других людей и общаться с друзьями. Модели, использующие при обучении только один из приведенных признаков, не смогли бы достаточно точно определить местоположение этих пользователей. Поэтому было принято решение рассмотреть модели, использующие и информацию о социальных связях пользователей, и текст опубликованных ими постов. Далее будут представлены принципы работы двух подобных алгоритмов: GCN, в основе которого лежат графовые сверточные сети (Graph Convolutional Networks), описанные Kipf и др. в [38] и DCCA, основанном на глубоком каноническом корреляционном анализе (Deep Canonical Correlation Analysis), представленном Andrew и др. в [39].

5.1 Модель GCN

GCN представляет собой нейросетевую модель $f(X, A)$. Пусть $X \in \mathbb{R}^{|U|\times|V|}$ обозначает текстовую часть информации, состоящую из мешка слов для каждого пользователя из U, использующего слова из списка слов V, и $A \in \mathbb{1}^{|U|\times|U|}$ обозначает графовую часть, описывающую взаимодействия между пользователями. Каждый из слоев данной нейронной сети можно описать следующим образом:

$$\hat{A} = \tilde{D}^{-1/2}(A + I)\tilde{D}^{-1/2} \qquad H^{l+1} = \sigma(\hat{A}H^{(l)}W^{(l)} + b) \tag{1}$$

где $\tilde{D}$ – матрица степеней матрицы $A + I$, которая содержит в себе информацию о степени каждой вершины, т. е. количество ребер, входящих в каждую вершину, I обозначает единичную матрицу. $H^0 = X, d_{in} \times d_{out}$ матрица $W^{(l)}$ и $d_{out} \times 1$ матрица b – параметры обучаемого слоя, а σ –

нелинейная функция активации. На первый слой подается среднее значение каждого элемента и его ближайших соседей с использованием весов из матрицы $\hat{A}$ и выполняется линейное преобразование с использованием W и b, за которым следует нелинейная функция активации σ. Другими словами, для пользователя u_i, выходное значение слоя l вычисляется по формуле:

$$\overrightarrow{h_i^{l+1}} = \sigma(\sum_{j \in nhood(i)} \widehat{A_{ij}} \overrightarrow{h_j^l} W^l + b^l)$$

где W^l и b^l параметры обучаемого слоя, и $nhood(i)$ обозначает соседей пользователя u_i.

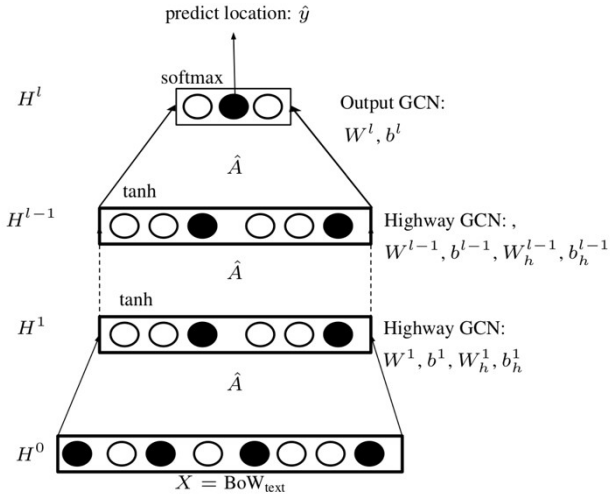

Рис. 2: Архитектура нейросетевой модели GCN с проверяющими уровнями (W_h^i, b_h^i)

Чтобы контролировать количество информации о соседях, которое передается узлу, на каждом слое используются проверяющие уровни, используемые также в магистральных сетях (Highway networks). В магистральных сетях, как описано Srivastava в [40], выходное значение слоя суммируется с его входным значением с контролирующими весами $T(\overrightarrow{h^l})$):

$$T(\overrightarrow{h^l}) = \sigma(W_t^l \overrightarrow{h^l} + b_t^l) \qquad \overrightarrow{h^{l+1}} = \overrightarrow{h^{l+1}} \circ T(\overrightarrow{h^l}) + \overrightarrow{h^l} \circ (1 - T(\overrightarrow{h^l}))$$

где $\overrightarrow{h^l}$ – входное значение для слоя $l+1, (W_t^l, b_t^l)$ – контролирующие веса и переменные смещения, $\circ$ - поэлементное умножение и σ – функция Сигмоида. Архитектура полученной нейронной сети представлена на Рис. 2.

5.2 Модель DCCA

Используя описанные выше матрицы X и $\hat{A}$ в качестве входных данных, CCA, представленная Hotelling в [41] и ее глубокая версия DCCA, описанная Andrew и др. в [39] подбирают веса моделей $f_1(X)$ и $f_2(\hat{A})$ таким образом, чтобы корреляция между выходными значениями была максимальной:

$$\rho = corr(f_1(X), f_2(\hat{A})) \tag{2}$$

Полученные представления $f_1(X)$ и $f_2(\hat{A})$ являются сжатыми представлениями двух частей информации о пользователях (текстовой и графовой) с уменьшенным некоррелированным шумом. Новые представления отлично описывают взаимодействие пользователей с точки зрения социального графа и языковые особенности пользователей с точки зрения текста, а их объединение представляет собой разновидное представление данных, которое может быть передано на вход другим моделям.

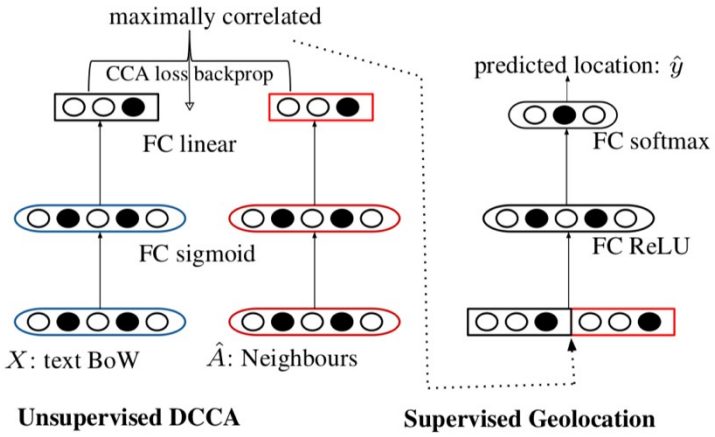

Рис. 3: Архитектура модели DCCA

В начале алгоритма DCCA для каждого из переданных на вход представлений данных (функции f_1 и f_2 из (2)) производится уменьшение размерности с помощью многослойного перцептрона. Полученные значения используются для оценки стоимости CCA:

$$\begin{aligned} &\text{максимизировать :} \quad tr(W_1^T \Sigma_{12} W_2) \\ &\text{при условии :} \qquad\quad W_1^T \Sigma_{11} W_1 = W_2^T \Sigma_{22} W_2 = I \end{aligned}$$

где Σ_{11} и Σ_{22} ковариации двух выходных значений, а Σ_{12} их кросс-ковариация. Веса W_1 и W_2 представляют собой линейные проекции выходных значений многослойных перцептронов, которые используются для оценки стоимости CCA. Задача оптимизации решается с помощью SVD с обратным распространением ошибки для подбора параметров двух многослойных перцептронов. После обучения выходные значения обеих нейронных сетей объединяются в одно представление, содержащее данные разных видов.

5.3 Обучение моделей

5.3.1 Предварительная обработка данных

Для обучения моделей используется датасет `Twitter-US`, описанный в данной работе ранее. Для формирования графовой части входных данных, строится матрица $\hat{A}$, описанная в уравнении (1), основанная на упоминаниях пользователями друг-друга в постах. Два пользователя связаны ($A_{ij} = 1$), если один упомянул другого или они вдвоем упомянули третьего пользователя. Текстовая часть входных данных представляет собой BoW модель, основанную на постах пользователей с применением TF, IDF и L2 нормализации.

5.3.2 Переход к задаче классификации

Рассматриваемые модели решают задачу классификации, однако изначальная постановка задачи, определение геопозиции пользователей, не является таковой. Поэтому перед началом работы необходимо перевести

координаты пользователей в метки классов. Для этого используется алгоритм построения k-d-дерева - специальной структуры данных, которая позволяет разбить двумерное пространство на меньшие части посредством сечения этого пространства прямыми.

Основываясь на обучающей выборке, пространство, образованное координатами пользователей, было разбито на непересекающиеся области. Размер и количество областей ограничивались посредством определения числа точек, которые могут попасть в ту или иную область. При этом признак принадлежности той или иной точки к одной из областей говорит о принадлежности данной точки к классу этой области пространства. Теперь мы можем выбрать в качестве координаты, соответствующей каждому из классов, координату медианы всех точек, относящихся к данному классу.

Таким образом, координаты пользователей могут быть переведены в метки классов посредством выбора ближайшей медианной точки. А по метке класса можно определить координаты некоторой точки пространства, которая соответствует медиане всех точек, относящихся к данному классу. В дальнейшем эти координаты будут использоваться при визуализации полученных результатов.

5.3.3 Описание параметров обучаемых моделей

Для модели GCN используются проверяющие слои, контролирующие количество информации о соседях, передаваемое узлу. При построении модели используется 3 слоя размерности 300. Гиперпараметр k-d дерева, контролирующий максимальное количество пользователей в каждом кластере, устанавливается равным 50 [42].

Для модели DCCA для нейронных сетей, обучаемых без учителя, используется скрытый слой sigmoid с размером 1000 и линейный выходной слой с размером 500. Функция потерь для CCA является функцией, максимизирующей корреляции выходных данных. Многослойный перцептрон, обучаемый с учителем, имеет один скрытый слой размерности 300 [42].

5.3.4 Используемые метрики

Для оценки моделей используются метрики, основанные на расстоянии. При определении местоположения подписчиков основной задачей является получить геопозицию каждого из пользователей. Пусть s обозначает одного пользователя, а S – набор всех пользователей, для которых осуществляется определение местоположения. Для каждого значения s модель определяет геопозицию $l(s)$. Ожидается, что предсказанное значение $l(s)$ будет совпадать или находится рядом с истинной локацией $l^*(s)$. Независимо от выбора уровня, на котором производится определение геопозиции (на уровне города, региона, страны и др.) все истинные и предсказанные значения местоположений могу быть представлены в виде координат.

Error Distance (ED) определяется как расстояние Евклида между истинными и предсказанными координатами:

$$ED(s) = dist(l(s), l^*(s))$$

Поскольку оценки производятся для набора пользователей, мы можем вычислить среднее или медиану всех расстояний ошибок (ED), чтобы посмотреть на результаты на уровне всего набора. Эти результаты описываются метриками `Mean Error Distance` и `Median Error Distance`:

$$MeanED = \frac{1}{|S|} \sum_{s \in S} dist(l(s), l^*(s))$$

$$MedianED = \underset{s \in S}{median}\{dist(l(s), l^*(s))\}$$

Помимо метрик `Mean Error Distance` и `Median Error Distance`, существует еще одна широко распространенная метрика на уровне всего набора, имеющая название `Distance-based Accuracy` или `Acc@d` в краткой записи. Определяется допустимое значение ошибки расстояния d и в дальнейшем любое предсказание, ошибка расстояния которого не превосходит d, является приемлемым. Метрика `Acc@d`, вычисленная на всем наборе,

определяет долю приемлемых предсказанных значений:

$$Acc@d = \frac{|\{s \in S \colon ED(s) \leq d\}|}{|S|}$$

Обычно в качестве допустимого значения ошибки расстояния d берется 161км или 100 миль.

5.3.5 Результаты обучения моделей

При обучении моделей GCN и DCCA в качестве входных данных передаются предварительно обработанные данные, описанные в разделе 5.3.1 данной работы. Оценка качества работы моделей производится с помощью метрик, описанных в разделе 5.3.4. В ходе обучения моделей, на тестовой выборке данных были получены результаты, представленные в Таблице 1.

Таблица 1: Сравнение результатов работы моделей разного типа

Название модели	**Acc@161**	**MeanED**	**MedianED**
Модели, основанные на тексте			
Rahimi et al. [43] model	54	554	120
Wing and Baldridge [23] model	48	686	191
Модели, основанные на социальном графе			
Rahimi et al. [45] model	54	705	116
Модели, основанные на тексте и графе			
DCCA	58	516	90
GCN	62	485	71
Miura et al. [44] model	61	481	65
Rahimi et al. [43] model	61	515	77

Таким образом, видно, что модель, основанная на графовых сверточных сетях (GCN), показала более хорошие качественные результаты, чем модель, основанная на глубоком каноническом корреляционном ана-

лизе (DCCA). В сравнении с другими моделями, GCN и DCCA показали конкурентоспособные результаты на данном наборе данных. Также важно отметить, что результаты комбинированных моделей превосходят результаты моделей, основанных на одном виде данных.

На основании полученных результатов было принято решение выбрать для определения геопозиций пользователей модель GCN. Предобученная на датасете `Twitter-US` модель GCN была сохранена в файл формата pkl и впоследствии использована для получения координат местоположений пользователей бизнес-аккаунта “Allbirds”.

Для этого предварительно обработанные данные, представляющие собой csv файл с никнеймами и текстом постов пользователей были переданы в модель. С помощью настроенных в процессе обучения весов модели, для собранных подписчиков были сделаны предсказания их геопозиций. В 5.3.2 описывается каким образом решение задачи классификации транслируется в пространственные координаты. Таким образом был получен список координат в следующем формате: {"lat": lat_coordinate, "long": lon_coordinate, "count": coordinates_count}, где `coordinates_count` – количество пользователей, отнесенных к классу, соответствующему данной координате. Результаты были представлены в таком формате, так как именно этот формат используется в процессе визуализации – построении тепловой карты.

Глава 6. Визуализация полученных геопозиций

Тепловые карты используются для визуального представления числовых данных. Они применяются в основном по двум причинам, одна из которых простое восприятие информации, передаваемой с помощью цветов, а другая – наглядный способ представления данных, не являющихся непрерывными. Тепловые карты обычно используются для создания отчетов и обобщения большого количества данных [46].

Для отображения полученных координат будем использовать одну из наиболее популярных библиотек для отображения карт, `Leaflet` [47]. Она помогает легко отображать растровые карты, добавлять дополнительные слои поверх основного, разработчикам, не знакомым с ГИС.

На html странице карта располагается в блоке `<div>` с `id="map"`. Для ее построения в функцию `mapConstructor` передается массив, состоящий из объектов, содержащих значения широты и долготы пользователя, а также наибольшее значение совпадающих координат. Также задается базовый слой, поверх которого будет построена тепловая карта, и некоторые другие параметры для ее отображения, такие как вид экстремума и радиус, используемые для раскраски.

На основе файла с информацией о полученных геопозициях подписчиков аккаунта "Allbirds"в социальной сети Twitter получилась тепловая карта распределения пользователей, представленная на Рис. 4.

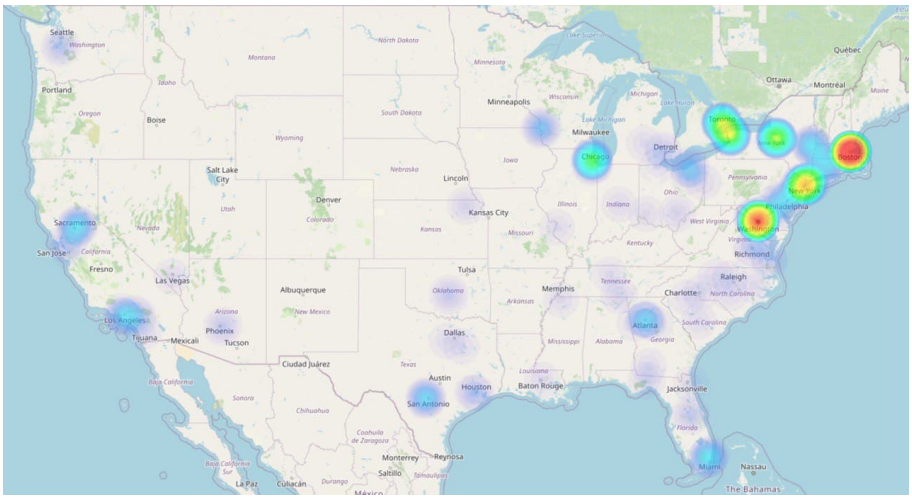

Рис. 4: Полная тепловая карта

Так как построенная карта предоставляет возможности изменения масштаба и перемещения, то можно поближе ознакомиться с полученным результатом. При описании раскраски карты было указано определение цветов в соответствии с локальными экстремумами, поэтому при приближении карты происходит перераспределение цветов. Рассмотрим данное явление на двух разных участках карты, представленных на Рис. 5 и Рис. 6.

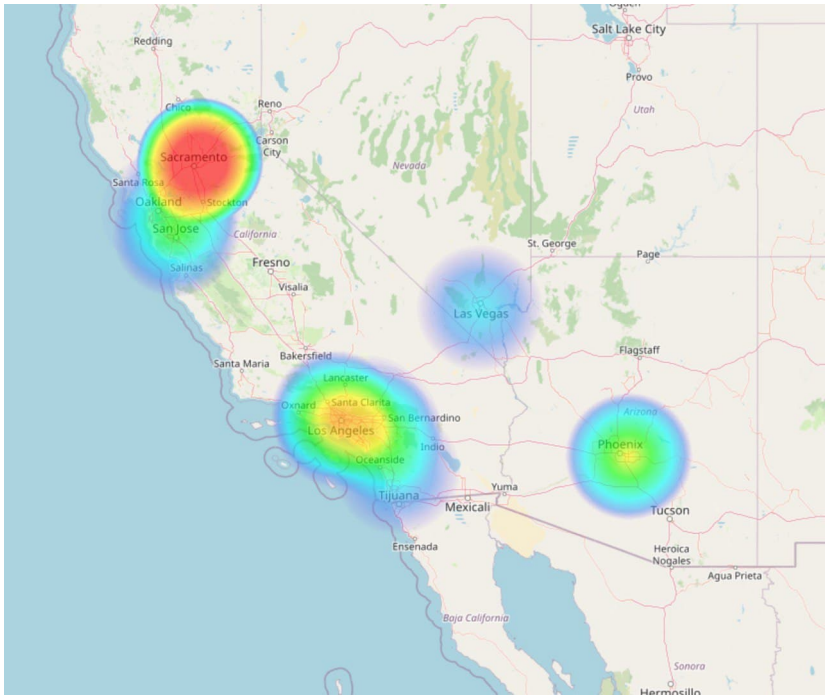

Рис. 5: Приближенный фрагмент тепловой карты

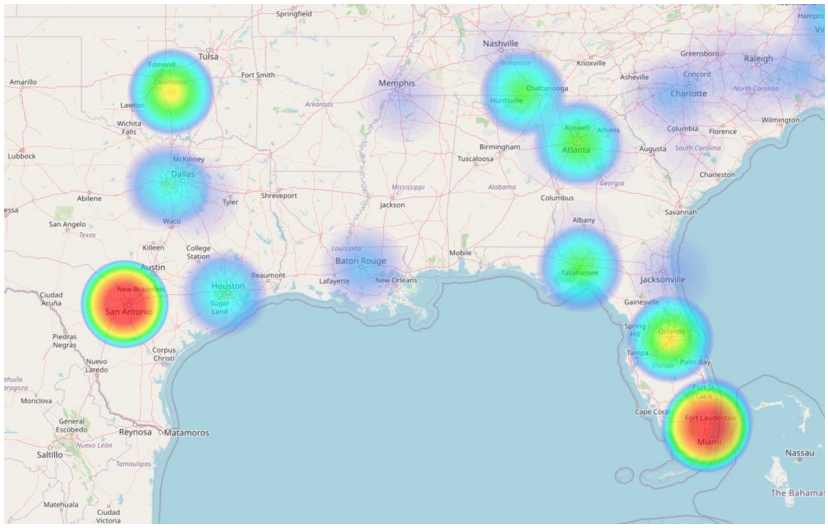

Рис. 6: Приближенный фрагмент тепловой карты

Выводы

В рамках данной работы было выявлено множество методов для определения геопозиции по аккаунту пользователя. При этом различные алгоритмы используют информацию разного типа, а их комбинации позволяют получить более точное решение поставленной задачи.

Проведенные эксперименты показали, что нейросетевые модели способны справляться с задачей определения геопозиции пользователя с достаточно высокой точностью. Поэтому данные методы можно применять на реальных неразмеченных данных для определения геопозиций подписчиков бизнес-аккаунта.

Предложенный способ визуализации позволяет представить данные в удобном для восприятия и дальнейшего анализа виде. Изучая различные участки построенной тепловой карты, можно легко сравнивать плотности распределения подписчиков в разных областях.

Заключение

В рамках данной работы был проведен обзор существующих подходов для определения геопозиций пользователей в социальной сети Твиттер. Были рассмотрены различные методы, работающие с разнотипными данными пользователей.

Далее были отобраны доступные для обучения размеченные данные и построена система по сбору информации о подписчиках бизнес-аккаунта. Основываясь на особенностях полученных данных, были выбраны алгоритмы, наиболее подходящие для решения конкретной задачи. Над ними был проведен ряд экспериментов, который позволил выявить метод, который лучше всего справляется с поставленной задачей. Применяя данный метод к реальным неразмеченным данным о пользователях, можно получать координаты их местоположения с достаточной точностью.

Полученные геопозиции подписчиков бизнес-аккаунта были представлены в виде тепловой карты, что позволило наглядно изучить их расположение в рассматриваемой области. Реализованная таким образом визуализация помогает анализировать местоположения потенциальных клиентов и определять места их максимального сосредоточения.

Таким образом, достигнутые в рамках данной работы результаты позволяют считать поставленную задачу выполненной. Рассматриваемые в работе алгоритмы могут быть успешно применены для определения местоположения пользователей социальной сети Твиттер. Исходный код программы, соддержащий в себе все этапы работы, представлен в открытом репозитории GitHub [48].

Список литературы

[1] The Unexpected Rise of the Online to Offline Movement in Retail // BEACHHEAD URL: https://medium.com/beachhead-network/the-unexpected-rise-of-the-online-to-offline-movement-in-retail-80b430680fb8 (дата обращения: 20.05.21).

[2] 10 Twitter Statistics Every Marketer Should Know in 2020 // Oberlo URL: https://www.oberlo.com/blog/twitter-statistics (дата обращения: 30.05.2020)

[3] Social Trends 2021 Survey // Hootsuite URL:https://www.hootsuite.com/pages/social-trends-2021 (дата обращения: 20.05.21).

[4] B. Hecht, L. Hong, B. Suh, and E. H. Chi, "Tweets from justin bieber's heart: The dynamics of the location field in user profiles,"

[5] J. McGee, J. A. Caverlee, and Z. Cheng, "A geographic study of tie strength in social media," in Proc. ACM Conf. Inf. Knowl. Manage., 2011, pp. 2333–2336.

[6] J. McGee, J. Caverlee, and Z. Cheng, "Location prediction in social media based on tie strength," in Proc. ACM Conf. Inf. Knowl. Manage., 2013, pp. 459–468.

[7] R. Compton, D. Jurgens, and D. Allen, "Geotagging one hundred million twitter accounts with total variation minimization," in Proc. IEEE Int. Conf. Big Data, 2014, pp. 393–401.

[8] D. Jurgens, "That's what friends are for: Inferring location in online social media platforms based on social relationships," in Proc. Int. Conf. Weblogs Social Media, 2013, pp. 273–282.

[9] Бычок или хабарик? Диалекты регионов России // ДискурсURL:https://discours.io/articles/culture/bychok-ili-habarik-dialekty-regionov-rossii (дата обращения: 21.05.21).

[10] K. Ren, S. Zhang, and H. Lin, "Where are you settling down: Geo-locating twitter users based on tweets and social networks," in Proc. Asia Inf. Retrieval Symp., 2012, pp. 150–161.

[11] B. Han, P. Cook, and T. Baldwin, "Geolocation prediction in social media data by finding location indicative words," in Proc. Conf. Comput. Linguistics: Tech. Papers, 2012, pp. 1045–1062

[12] J. Mahmud, J. Nichols, and C. Drews, "Where is this tweet from? inferring home locations of twitter users," in Proc. Int. Conf. Weblogs Social Media, 2012, pp. 511–514.

[13] B. Han, P. Cook, and T. Baldwin, "Text-based twitter user geolocation prediction," J. Artif. Intell. Res., vol. 49, no. 1, pp. 451–500, 2014

[14] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: A content-based approach to geo-locating twitter users," in Proc. ACM Conf. Inf. Knowl. Manage., 2010, pp. 759–768

[15] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak, "Spatial variation in search engine queries," in Proc. Conf. World Wide Web, 2008, pp. 357–366.

[16] K. Ryoo and S. Moon, "Inferring twitter user locations with 10 km accuracy," in Proc. World Wide Web Conf. Companion Volume, 2014, pp. 643–648.

[17] R. Li, S. Wang, and K. C.-C. Chang, "Multiple location profiling for users and relationships from social network and content," Proc. VLDB Endowment, vol. 5, no. 11, pp. 1603–1614, 2012. in Proc. Conf. Human Factors Comput. Syst., 2011, pp. 237–246

[18] A. Rahimi, D. Vu, T. Cohn, and T. Baldwin, "Exploiting text and network context for geolocation of social media users," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technol., 2015, pp. 1362–1367.

[19] J. Mahmud, J. Nichols, and C. Drews, "Home location identification of twitter users," ACM Trans. Intell. Syst. Technol., vol. 5, no. 3, pp. 47:1–47:21, 2014

[20] B. P. Wing and J. Baldridge, "Simple supervised document geolocation with geodesic grids," in Proc. Annu. Meeting Assoc. Comput. Linguistics: Human Language Technol., 2011, pp. 955–964.

[21] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1998, pp. 275–281.

[22] I. J. Good, "The population frequencies of species and the estimation of population parameters," Biometrika, vol. 40, no. 3–4, pp. 237–264, 1953.

[23] B. Wing and J. Baldridge, "Hierarchical discriminative classification for text-based geolocation," in Proc. Conf. Empirical Methods Natural Language Process., 2014, pp. 336–348.

[24] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge, "Supervised text-based geolocation using language models on an adaptive grid," in Proc. Joint Conf. Empirical Methods Natural Language Process. Comput. Natural Language Learn., 2012, pp. 1500– 1510

[25] Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma, "A simple scalable neural networks based model for geolocation prediction in twitter," in Proc. Workshop Noisy User-Generated Text, 2016, pp. 235–239.

[26] Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma, "Unifying text, metadata, and user network representations with a neural network for geolocation prediction," in Proc. Annu. Meeting Assoc. Comput. Linguistics, 2017, pp. 1260–1272.

[27] A. Rahimi, T. Cohn, and T. Baldwin, "A neural model for user geolocation and lexical dialectology," in Proc. Annu. Meeting Assoc. Comput. Linguistics, Volume 2: Short Papers, 2017, pp. 209– 216.

[28] C. A. Davis Jr, G. L. Pappa, D. R. R. de Oliveira, and F. de L Arcanjo, "Inferring the location of twitter messages based on user relationships," Trans. GIS, vol. 15, no. 6, pp. 735–751, 2011

[29] L. Kong, Z. Liu, and Y. Huang, "SPOT: Locating social media users based on social network context," Proc. VLDB Endowment, vol. 7, no. 13, pp. 1681–1684, 2014.

[30] D. Rout, K. Bontcheva, D. Preotiuc-Pietro, and T. Cohn, "Where's ¸ @wally?: A classification approach to geolocating users based on their social ties," in Proc. ACM Conf. Hypertext Social Media, 2013, pp. 11–20.

[31] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in Proc. Conf. World Wide Web, 2010, pp. 591–600.

[32] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, "Towards social user profiling: Unified and discriminative influence model for inferring home locations," in Proc. ACM Conf. Knowl. Discovery Data Mining, 2012, pp. 1023–1031

[33] J. Mahmud, J. Nichols, and C. Drews, "Home location identification of twitter users," ACM Trans. Intell. Syst. Technol., vol. 5, no. 3, pp. 47:1–47:21, 2014.

[34] H. Efstathiades, D. Antoniades, G. Pallis, and M. D. Dikaiakos, "Identification of key locations based on online social network activity," in Proc. IEEE/ACM Conf. Adv. Social Netw. Anal. Mining, 2015, pp. 218–225.

[35] A. Poulston, M. Stevenson, and K. Bontcheva, "Hyperlocal home location identification of twitter profiles," in Proc. ACM Conf. Hypertext Social Media, 2017, pp. 45–54.

[36] Allbirds // URL: `https://www.allbirds.com` (дата обращения: 25.04.2021)

[37] Selenium // URL: `https://www.selenium.dev/documentation/en` (дата обращения: 21.04.2021)

[38] Kipf, T.N. and M. Welling, 2017. Semisupervised classification with graph convolutional networks. International Conference on Learning Representations (ICLR)

[39] Andrew, G., K. Bilmes and K. Livescu, 2013. Deep canonical correlation analysis. International Conference on Machine Learning, pp. 1247–1255.

[40] Rupesh Kumar Srivastava, Klaus Greff, and Jurgen ¨ Schmidhuber. 2015. Highway networks. arXiv preprint arXiv:1505.00387.

[41] Harold Hotelling. 1936. Relations between two sets of variates. Biometrika, 28(3/4):321–377

[42] Rahimi, A., T. Cohn and T. Baldwin, 2018. Semi-supervised User Geolocation via Graph Convolutional Networks. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (volume 1), Association for Computational Linguistics, pp: 2009–2019.

[43] Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2017b. A neural model for user geolocation and lexical dialectology. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), pages 207–216, Vancouver, Canada

[44] Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2017. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1260–1272, Vancouver, Canada

[45] Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2015a. Twitter user geolocation using a unified text and network prediction model. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics — 7th International Joint Conference on Natural Language Processing (ACLIJCNLP 2015), pages 630–636, Beijing, China.

[46] Akshay S., Akash M. R., Sai Ananda Krishnan G., Comparative Analysis of Heat Maps over Voronoi Diagram in Eye Gaze Data Visualization // 2017 IEEE International Conference on Intelligent Computing and Control(I2C2)

[47] Leaflet // URL: `https://leafletjs.com` (дата обращения: 01.06.2020)

[48] Репозиторий GitHub URL: `https://github.com/Marysosh/TwitterGeoHeatmap` (дата обращения: 26.05.2021)