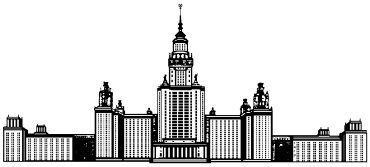
Филиал Московского государственного университета имени М. В. Ломоносова в городе Севастополе



Факультет компьютерной математики
Кафедра программирования

ДИПЛОМНАЯ РАБОТА СТУДЕНТА ГРУППЫ ПМ-501

Извлечение описаний событий предопределённых типов из потока сообщений пользователей микроблогов

Выполнил:
студент 5 курса группы ПМ-501
Е. Ю. Пышнограев

Научный руководитель:
к.ф.-м.н.
Д. Ю. Турдаков

Севастополь, 2014

Аннотация

В работе рассматривается задача извлечения описаний событий из потока сообщений пользователей сети Твиттер. Приведена постановка задачи. Рассмотрены особенности событий в сети Твиттер. Проанализированы существующие подходы к решению задачи извлечения событий. Предложен алгоритм на основе тематической модели HDP с модификацией частичного обучения. Описаны составляющие части алгоритма. Алгоритм реализован, приведен код программы, диаграмма классов и описание компонентов. Рассмотренна работа предложенного алгоритма на тестовых данных, показаны результаты. Найдена точность, полнота и F-мера разработанного алгоритма, показано что на тестовых данных предложенный алгоритм работает лучше, чем модификация существующего алгоритма New Event Detection.

Содержание

Введение . . . 3
1 Постановка задачи . . . 4
2 Обзор существующих решений рассматриваемой задачи . . . 5
2.1 New Event Detection . . . 5
2.2 Выявление событий с помощью LDA . . . 6
2.3 Information nuggets . . . 7
2.4 Тематическая модель HDP . . . 8
2.4.1 Иерархический процесс Дирихле . . . 9
2.4.2 Chinese Restaurant Franchise . . . 9
2.5 Нахождение экстремумов . . . 10
3 Исследование и построение решения задачи . . . 12
3.1 Нахождение экстремумов . . . 12
3.2 Извлечение ключевых слов . . . 12
3.3 Модификация частичной обучаемости . . . 13
3.4 Анализ сообщений темы . . . 14
4 Реализация и тестирование решения . . . 16
4.1 Описание данных . . . 16
4.2 Извлечение событий . . . 16
4.3 Оценка качества разработанного алгоритма . . . 20
5 Описание практической части . . . 21
Заключение . . . 22
Список литературы . . . 23
Приложение . . . 24

Введение

В современном мире информация быстро устаревает, поэтому способы вовремя находить нужные данные — постоянный объект для исследований. Одним из направлений в этой области является извлечение данных из платформ микроблогов.

Платформы микроблогов стали очень популярным способом размещения данных в Сети. В них можно найти сообщения пользователей практически на любую тему, начиная стихийными бедствиями и заканчивая рейтингами музыкальных исполнителей. Правильная обработка доступной информации — нетривиальная задача, которая имеет множество областей применения. Последние несколько лет эта тема активно исследуется во многих университетах мира.

Отслеживание сообщений о стихийных бедствиях в реальном времени поможет вовремя организовать спасательные операции и сохранить жизни людей [1]. Руководствуясь сообщениями пользователей микроблогов можно судить о популярности товаров и вовремя принимать экономически целесообразные решения. Можно делать предположения о рейтингах политических деятелей и эффективности рекламы на основании информации в микроблогах. Помимо перечисленных способов применения доступной информации в микроблоггинговых платформах можно привести множество других.

В данной работе в дальнейшем будет рассматриваться сервис микроблогов Твиттер (http://www.twitter.com). В нем помимо текстовой информации можно публиковать фото, видео и геотеги, что так же может быть использовано при анализе. В доступном наборе данных можно проводить анализ разных сущностей, эта работа посвящена выявлению событий среди потоков информации и извлечению их описания. Поскольку трактовка событий в сообщениях микроблогов может быть субъективной, выделим несколько свойств, которыми характеризуется событие.

Событие в первую очередь является чем-то аномальным на фоне остальных данных. Оно определяется резким изменением частотных характеристик некоторых слов в сообщениях. События в микроблогах носят “взрывной” характер, в течении нескольких часов частоты релевантных слов возрастают в десятки раз и так же быстро опускаются до нормального уровня. Примером события могут быть: стихийное бедствие, выход законопроекта на резонансную тему, получение фильмом награды на кинофестивале.

Существуют разные подходы для решения описанной задачи. В следующих двух разделах будут даны формальные определения и рассмотрены некоторые подходы для решения задачи извлечения событий.

1 Постановка задачи

Цель дипломной работы состоит из нескольких частей:

– исследовать существующие подходы по извлечению описаний событий из сообщений пользователей, выделить возникающие проблемы и рассмотреть возможные методы их решения,
– исследовать возможность применения тематических моделей для решения задачи выявления событий,
– разработать метод для извлечения описаний событий из сообщений пользователей сети Твиттер на основе иерархического процесса Дирихле,
– протестировать работу алгоритма на реальных данных.

Объект изучения этой работы — алгоритм, который по входным данным строит множество событий. В качестве данных для задач подобного рода служит корпус документов (сообщений):

$$\Omega = \left\{ D_i \mid i \in \overline{1, n} \right\}. \tag{1}$$

Будем считать, что каждый документ D_i имеет временную метку t_i. В свою очередь документ D_i определяется как упорядоченный набор слов:

$$D_i = \left\{ w_j \mid j \in \overline{1, l_i} \right\}, \tag{2}$$

при этом слова в документах принадлежат словарю V.

Событие — некоторая сущность, которая характеризуется временем возникновения и ключевыми словами. Оно вызывает резкий подъем частотных характеристик некоторых слов. Событием может быть футбольный матч и музыкальный концерт. В социальной сети Твиттер есть популярные темы, которые всегда генерируют много сообщений. Например это сообщения с ключевыми словами iphone и ipad. Но такие сообщения нельзя считать событиями. Также событиями нельзя считать еженедельные пятничные сообщения о конце рабочей недели [2].

Особенности социальной сети Твиттер состоят в следующем:

– короткие сообщения (до 140 символов),
– наличие шума и ошибок,
– большая плотность сообщений,
– “взрывной” характер событий.

2 Обзор существующих решений рассматриваемой задачи

Прежде чем составлять решение задачи были изучены существующие подходы. Все они комбинируют техники машинного обучения, обработки текстов, вероятностных графических моделей и других разделов науки. Разные методы лучше работают для одних данных и хуже для других, на их качество также влияет природа данных. Выделим несколько подходов, для того чтобы изучить общую схему подобных алгоритмов и чтобы обозначить идеи, которые могут быть использованы при составлении метода извлечения событий из сообщений сети Твиттер. После этого рассмотрим способы решения нескольких подзадач, которые будут входить в разработанный алгоритм.

2.1 New Event Detection

Исторически первым подходом к извлечению событий принято считать NED (New Event Detection) [3]. NED предназначен для того, чтобы находить первый документ на тему, которая не встречалась раньше. Следующие документы на эту тему уже не будут новыми и не будут помечены алгоритмом. Для того, чтобы отвечать на вопрос, является ли документ новым, необходимо указать способ как определять степень сходства двух документов.

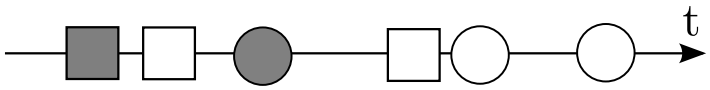


Рисунок 1. Документы, соответствующие двум разным темам. Новые документы помечены серым цветом

Для этого в алгоритме NED используется техника Incremental TF–IDF (Term Frequency – Inverse Document Frequency). TF–IDF — базовый метод выяснять насколько отдельно взятые слова характеризуют весь документ, другими словами, насколько большой вес имеет слово w в документе d. Пусть $f(d, w)$ — количество слов w в документе d. Определим значение $df_t(w)$ как количество документов, поступивших не позднее времени t, в которых встречается слово w. Используя введенные величины, можно записать значение веса определенного слова w в документе d. В момент времени t имеем:

$$\text{weight}_t(d, w) = \frac{1}{Z_t(d)} f(d, w) \cdot \log \frac{N_t}{df_t(w)}, \tag{3}$$

где N_t — общее количество документов, поступивших не позднее времени t, $Z_t(d)$ — нормализационное значение:

$$Z_t(d) = \sqrt{\sum_w \left[f(d, w) \cdot \log \frac{N_t}{df_t(w)} \right]^2}. \tag{4}$$

.

Теперь можно записать значение похожести двух документов, q и d:

$$\text{sim}_t(d, q) = \sum_w \text{weight}(w, d) \cdot \text{weight}(w, q). \tag{5}$$

Указанные формулы записаны в косинусной метрике, также могут быть использованы метрики Хеллингера, Кульбака–Лейблера и другие.

Для того, чтобы понять, является ли добавленный в момент времени t документ q новым, необходимо вычислить степень его похожести со всеми предыдущими документами. Пусть d^* — документ, максимально похожий на q:

$$d^* = \operatorname*{argmax}_d \operatorname{sim}_t(d, q). \tag{6}$$

Тогда значение

$$\operatorname{score}_t(q) = 1 - \operatorname{sim}_t(d^*, q) \tag{7}$$

может быть использовано для того, чтобы определить, является ли документ q новым. Новыми будем считать все документы q, у которых значение $\operatorname{score}_t(q)$ больше, чем пороговое значение θ_s. В обратном случае считается что существует документ d^*, достаточно похожий на q и поэтому q не является сообщением на новую тему. Для того, чтобы определить подходящее значение θ_s, можно использовать размеченный корпус и посчитать значение $\operatorname{sim}_t(d, q)$ среди документов, соответствующих одному и разным событиям.

Недостатком алгоритма NED является то, что он не учитывает отношения между словами. Например синонимичные слова в NED будут считаться совершенно разными. Для того, чтобы решить эту проблему могут быть применены тематические модели.

2.2 Выявление событий с помощью LDA

На следующем примере рассмотрим как тематические модели (topic models) могут быть использованы для задачи распознавания событий. Для этого опишем схему, по которой работает алгоритм, предложенный в статье [4].

Задача состоит в том, чтобы по данным к набору фотографий в сети Flickr распознать все события на определенную тему, проходившие в обозначенных городах. Есть несколько вариантов постановок задачи, они различаются тематикой событий и городами.

Фотографии содержат в себе описания, они будут являться документами в задаче распознования. Также фотографии могут включать геотеги. Авторы статьи предлагают разбить решение задачи на пять составных частей:

1. предобработка текстовых данных,
2. определение в каком городе сделана фотография по ее описанию,
3. распознавание темы, к которой относится фотография,
4. распознавание события, к которому относится фотография,
5. оптимизация описания события, полученного на шаге 4.

В качестве предобработки, авторы предлагают выполнить следующее: удалить стоп-слова и html теги, провести стемминг слов[1], перевести не английские слова на английский язык используя сервис Google Translate.

Так как задача упоминает отдельные города, необходимо разработать метод распознавания города по документу. Указанные в фотографии географические координаты были доступны авторам в 20% случаев, из этих координат были выявлены города используя сервис Google Tables. Используя технику TF–IDF, в описаниях фотографий с известными городами были извлечены ключевые слова. По этим ключевым

[1]стемминг (stemming) — удаление окончаний у слов для их нормализации

словам появилась возможность назначить фотографиям без геотегов “ближайший” город с точки зрения похожести описаний. Для того чтобы определять похожесть документов, был использован подход, описанный в (2.1). В случаях, когда “ближайший” город выявить не удавалось, авторами использовались следующие предположения: считалось что один и тот же фотограф не мог побывать в один день более чем в двух разных городах, и что путешествие из одного города в другой занимает как минимум два часа. Эти предположения позволили улучшить классификатор, таким образом более 97% фотографий были привязаны к городу.

Далее необходимо для каждого города кластеризовать документы по темам и рассмотреть только те из них, которые заданы в описании задачи. Для распознования тем была использована тематическая модель LDA (Latent Dirichlet Allocation) [5], для определения параметров которой применялось сэмплирование по Гиббсу [6]. LDA работает из предположения, что каждый документ D_i характеризуется случайным распределением над темами, в то время как каждая тема является мультиномиальным распределением над словами.

Оставшаяся часть — извлечение событий и их оптимизация. Для того, чтобы алгоритм выявил событие, отвечающее теме k в день d, необходимо чтобы количество документов D_i по этой теме в день d превосходило некоторое пороговое значение θ. Оптимизация событий подразумевает под собой объединение событий на одну тему в последовательные дни и разделение событий в разных городах.

Авторы статьи тестировали алгоритм на трех разных вариантах условия задачи, в таблице 1 приведены результаты по каждому из них.

Таблица 1. Точность, полнота и F-мера алгоритма при разных условиях задачи

Данные	Точность	Полнота	F-мера
1	80.98	19.25	31.10
2	91.21	77.85	84.00
3	90.76	81.91	86.11

По результатам из таблицы можно видеть что в первом случае алгоритм справляется со своей задачей существенно хуже, чем в других. Авторы объясняют это тем, что в задаче 1 необходимо было находить научные конференции, а так как они проходят на разные темы, не получалось выделить конкретный набор ключевых слов. По этой причине полнота алгоритма в случае 1 относительно низкая. В задаче 2, 3 напротив удалось обозначить необходимые ключевые слова, о чем свидетельствуют результаты.

2.3 Information nuggets

Рассмотрим подход к извлечению событий, описанный в [1]. Авторы ставят перед собой задачу составить алгоритм описания подсобытий в социальной сети Твиттер. Были использованы данные, полученные во время торнадо Joplin в 2011 году. Данные представляют из себя сообщения пользователей, содержащие хэштег #joplin, собранные 22 мая 2011 года на протяжении нескольких часов, пока плотность сообщений не стала относительно низкой. Авторы ставили цель извлекать из потока сообщений так называемые золотые самородки информации — короткие и информативные сообщения, описывающие происходящие события. По этой причине подход назван information nuggets.

Авторы видели основную проблему в том, что даже при наличии большого количества сообщений об одном событии, ими трудно пользоваться, потому что они имеют разную природу. Например это может быть сообщение очевидца о происходящем стихийном бедствии или сообщение о перечислении правительством средств на восстановление разрушенных построек. Статья предлагает делить сообщения на пять разных категорий по степени информативности:

- Персональное: информация в сообщении может быть полезна только автору и его кругу общения. Она не является интересной для людей, которые не знают автора сообщения непосредственно.
- Информативное (напрямую): сообщение может быть полезно людям вне круга общения автора, и эта информация написана прямым участником или очевидцем событий.
- Информативное (косвенно): сообщение может быть полезно людям вне круга общения автора, при этом автор пишет о том, что он слышал по телевидению, радио или любому другому источнику информации.
- Информативное (напрямую или косвенно): сообщение может быть полезно людям вне круга общения автора, но невозможно ответить на вопрос как автор связан с происходящими событиями.
- Другое: сообщение или не на английском языке, или не может быть классифицировано.

В дальнейшем рассматриваются только сообщения с информативным типом, так как они наиболее вероятно содержат полезные сведения. Затем сообщения разбиваются на подтипы по направленности информации. Всего авторами использовалось 32 разных подтипа, среди которых:

- Предостережения и советы: документ содержит предупреждение о возможном опасном происшествии.
- Жертвы и разрушения: в тексте сообщается о потерях, вызванных стихийным бедствием.
- Сбор средств: сообщения описывают пожертвования денег и предметов пострадавшим от чрезвычайного происшествия.

Для того, чтобы классифицировать сообщения по типам и подтипам, в алгоритме используется наивный байесовский классификатор, который предварительно тренируется на размеченных данных. В классификаторе используется большое количество свойств, бинарных, скалярных и текстовых. В таблице 8 приведены результаты работы алгоритма на некоторых подтипах. Для тестирования использовались размеченные вручную сообщения.

Таблица 2. Результаты работы алгоритма information nuggets

Подтип	Точность	Полнота	F-мера
Предостережения и советы	0.618	0.598	0.605
Жертвы и разрушения	0.578	0.645	0.610
Сбор средств	0.546	0.632	0.585

2.4 Тематическая модель HDP

Как было показано в разделе 2.2, для того чтобы кластеризовать сообщения по темам, можно использовать тематические модели. Было показано, что примером та-

кой модели может быть модель LDA. Но в LDA необходимо указывать число тем как параметр, поэтому рассмотрим модель HDP (Hierarchical Dirichlet process), которая автоматически находит нужное число тем. В следующем разделе при описании решения будет указано, какие модификации были внесены в модель.

2.4.1 Иерархический процесс Дирихле

Пусть $(\Theta, \mathcal{B})$ — измеримое пространство, с вероятностной мерой G_0. Пусть α_0 — положительное действительное число. Определим процесс Дирихле $\mathrm{DP}(\alpha_0, G_0)$ как случайное распределение вероятностной меры G над $(\Theta, \mathcal{B})$, такое, что для любого конечного измеримого разбиения Θ $(A_1, A_2, \ldots, A_r)$, случайный вектор $(G(A_1), G(A_2), \ldots, G(A_r))$ будет распределен согласно конечномерному распределению Дирихле с параметрами $(\alpha_0 G_0(A_1), \alpha_0, G_0(A_2), \ldots, \alpha_0 G_0(A_r))$:

$$(G(A_1), G(A_2), \ldots, G(A_r)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \alpha_0, G_0(A_2), \ldots, \alpha_0 G_0(A_r)). \tag{8}$$

Иерархический процесс Дирихле это распределение над множеством вероятностных мер над $(\Theta, \mathcal{B})$. Процесс определяет множество вероятностных мер G_j, по одному на каждую группу, и глобальную меру G_0. Глобальная мера G_0 распределена согласно процессу Дирихле с параметрами γ и базовым распределением H:

$$G_0 \mid \gamma, H \sim \mathrm{DP}(\gamma, H). \tag{9}$$

Групповые меры G_j условно независимы при заданном G_0 и имеют следующее распределение:

$$G_j \mid \alpha_0, G_0 \sim \mathrm{DP}(\alpha_0, G_0). \tag{10}$$

Гиперпараметрами HDP являются базовое распределение H и концентрационные параметры γ и α_0. Иерархический процесс Дирихле может быть использован как априорное распределение над параметрами групповых данных. Для каждого j пусть $\theta_{j1}, \theta_{j2}, \ldots$ — независимые распределенные по закону G_j случайные величины. Каждая величина θ_{ji} будет соответствовать наблюдаемому значению x_{ji}. Правдоподобие может быть записано в следующем виде:

$$\begin{aligned} &\theta_{ji} \mid G_j \sim G_j, \\ &x_{ji} \mid \theta_{ji} \sim F(\theta_{ji}). \end{aligned} \tag{11}$$

Процесс, описанные выше называется смешанной моделью для иерархического процесса Дирихле (hierarchical Dirichlet process mixture model) [8].

2.4.2 Chinese Restaurant Franchise

HDP — непараметрическая тематическая модель, процесс генерации данных по ней может быть описан в терминах процесса "chinese restaurant franchise"(CRF). В этом процессе существует сеть ресторана, в каждом из которых используется одинаковое меню. В ресторане находится неограниченное число столов, за каждым столом неограниченное число мест; каждый посетитель, заходя в ресторан, садится либо за уже занятый стол, либо за новый стол. На каждом непустом столе в ресторане заказано одно блюдо из меню, которое едят все посетители, сидящие за этим столом. Блюдо на стол заказывает первый человек, севший за этот стол.

В CRF группам соответствуют рестораны, а параметры θ_{ji} считаются посетителями. Переменные $\phi_1, \ldots, \phi_K$ составляют единое меню для сети ресторанов, они

распределены согласно базовому распределению H. Пусть ψ_{jt} — переменная, связанная со столом t в ресторане j и обозначающее блюдо, которое заказано за этим столом.

Каждому посетителю θ_{ji} соответствует стол ψ_{jt}, в то время как каждому столу ψ_{jt} соответствует одно блюдо ϕ_k. Для того, чтобы указать это соответствие, используются индексы:

- t_{ji} — индекс, связывающий посетителя θ_{ji} со столом ψ_{jt}, за которым сидит этот посетитель.
- k_{jt} — индекс, связывающий стол ψ_{jt} с блюдом ϕ_k, которое заказано для этого стола.

Для описания процесса, необходимо ввести следующие обозначения:

- n_{jtk} — количество посетителей в ресторане j за столом t, едящих блюдо k;
- $n_{jt\cdot}$ — количество посетителей в ресторане j за столом t;
- $n_{j\cdot k}$ — количество посетителей в ресторане j, едящих блюдо k;
- m_{jk} — количество столов в ресторане j, на которых заказано блюдо k;
- $m_{j\cdot}$ — количество столов в ресторане j;
- $m_{\cdot k}$ — количество столов, на которых заказано блюдо k;
- $m_{\cdot\cdot}$ — общее количество столов.

Запишем условное распределение θ_{ji}, где G_j исключены интегрированием:

$$\theta_{ji} \mid \theta_{j1}, \theta_{j2}, \ldots, \theta_{j,i-1}, \alpha_0, G_0 \sim \sum_{i=1}^{m_j} \frac{n_{jt\cdot}}{i-1+\alpha_0}\delta_{\psi_{jt}} + \frac{\alpha_0}{i-1+\alpha_0}G_0. \tag{12}$$

Аналогично, интегрируя по G_0, получим:

$$\psi_{jt} \mid \psi_{11}, \psi_{12}, \ldots, \psi_{21}, \ldots, \psi_{j,t-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot}+\gamma}\delta_{\phi_k} + \frac{\gamma}{m_{\cdot\cdot}+\gamma}H. \tag{13}$$

Уравнение (12) описывает выбор стола посетителем, в то время как уравнение (13) определяет выбор блюда.

Если перевести используемые термины в область тематических моделей, посетители будут соответствовать словам, рестораны документам, столы внутренним темам документов, блюда внешним темам всего корпуса [8].

2.5 Нахождение экстремумов

Рассмотрим методы, которые могут быть использованы для того, чтобы найти экстремум функции. Как было замечено выше, события в сети Твиттер носят взрывной характер и соответствуют разкому повышению частоты сообщений. В связи с этим, нахождение экстремумов может быть использовано для поиска точек во времени, в которых могло произойти событие.

Понятие экстремума в этом подразделе отличается от привычного определения в математическом анализе. Назовем экстремумом те точки, в которых плотность сообщений позволяет утверждать что в этот момент произошло событие. В первую очередь, понятие экстремума для задачи выявления событий должно быть локальным. С другой стороны не все локальные максимумы можно корректно обозначить как события. Задача в общем виде получила развитие в контексте обработки сигналов [7]. Существует множество методов, каждый из которых будет больше подходить

к определенному типу данных. Алгоритм нахождения экстремумов может меняться вне зависимости от других составных частей исходного алгоритма. Несколько возможных подходов нахождения экстремумов для функции $f(x)$ с сеточной областью определения X перечислены ниже:

- Алгоритм помечает как экстремумы все точки, в которых функция достигает максимума в некотором окне, и при этом ее значение больше некоторого заранее определенного θ. Опционально θ может зависеть от среднего значения функции и других статистических характеристик.
- Для любых двух значений аргумента x и y, где $x < y$ определим функции $T(x, y)$ (travel) и $R(x, y)$ (rise):

$$\begin{aligned} T(x,y) &= \sum_{x \le k < y} |f(k+1) - f(k)|, \\ R(x,y) &= f(y) - f(x) + \epsilon, \end{aligned} \tag{14}$$

  где ϵ — некоторое небольшое значение. Тогда значения $T(x, y)/R(x, y) > 1$ соответствуют пику функции между точками x и y. Алгоритм может находить все экстремумы, где значение T/R больше некоторого порога.
- Подход заключается в том, чтобы выделить в функции стандартную пиковую подфункцию, например функцию Гаусса. Для этого применяются согласованные фильтры. Пусть $g(x)$ — функция Гаусса с некоторыми μ и σ:

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{15}$$

  Тогда можно определить степень похожести исходной функции на функцию $g(x)$:

$$\mathrm{sim}(f, g) = \frac{(f, g)}{\|f\| \cdot \|g\|}. \tag{16}$$

  В косинусной метрике выражение примет следующий вид:

$$\mathrm{sim}(f, g) = \frac{\sum\limits_{x \in X} f(x) \cdot g(x)}{\sqrt{\sum\limits_{x \in X} f^2(x)} \cdot \sqrt{\sum\limits_{x \in X} g^2(x)}}. \tag{17}$$

  Аналогично предыдущим случаям, будем считать что функции “похожи” если значение $\mathrm{sim}(f, g)$ не меньше некоторого θ.
- Один из способов найти экстремумы — применить смешанную модель (mixture model) к входной функции. Компонентами модели могут быть нормальные распределения. Затем нужно найти параметры модели используя EM (expectation maximization) алгоритм или сэмплирование по Гиббсу. Этот метод будет корректно обрабатывать случай, когда экстремумы накладываются друг на друга или расположены очень близко.

3 Исследование и построение решения задачи

Эта работа ставит перед собой задачу разработать алгоритм нахождения событий в потоке сообщений пользователей сети Твиттер. В качестве данных были использованы сообщения с 4 июня 2013 года по 1 июля 2013 года, содержащие в себе хэштег #texas. Всего корпус включает порядка 240 тысяч сообщений и 1.5 миллиона слов. Приведем псевдокод алгоритма, а затем рассмотрим каждый шаг более подробно. Здесь и далее под частотной характеристикой сообщений подразумевается функция количества сообщений в час.

Исходные параметры: множество сообщений Ω, параметры алгоритма M, L, l_D, J, R, θ_{sim}, $\hat{\sigma}_0$, θ_{spam} параметры для HDP модели

Результат: для каждого события набор ключевых слов и частота сообщений по теме события

1 в частотной характеристике сообщений из Ω найти K экстремумов, пусть они соответствуют временам $t_1, t_2, \ldots, t_K$;
2 для каждого $t \in \{t_1, t_2, \ldots, t_K\}$ выполнять
3 извлечь M ключевых слов из сообщений в Ω в некотором радиусе R момента времени t;
4 сформировать фиктивные документы $D_0 = \left\{D_0^j\right\}_{j=1}^{J}$, каждый из которых имеет длину l_D и состоит из полученных на шаге 3 ключевых слов пропорционально их весу;
5 добавить D_0 к множеству Ω;
6 каждому сообщению из Ω назначить тему, с помощью модифицированной модели HDP (Hierarchical Dirichlet Process);
7 пусть документам D_0 была назначена тема k_0, удалить D_0 из Ω;
8 найти частотную характеристику сообщений в Ω с темой k_0;
9 если частотная характеристика определяет новое событие тогда
10 найти L ключевых слов темы k_0 из HDP;
11 добавить ключевые слова, частотную характеристику в результат;
12 конец условия
13 конец цикла

Алгоритм 1. Извлечение событий из сообщений сети Твиттер

3.1 Нахождение экстремумов

На шаге 1 основного алгоритма из частотной функции извлекаются экстремумы. Некоторые методы нахождения экстремумов были описаны в разделе 2.5. Из применение можно комбинировать со сглаживающими фильтрами если шум мешает выделить события. Для исходной задачи был выбран первый метод из списка, а именно нахождение максимума функции в некотором окне, при условии что значение функции выше пороговой величины. Данные показали что экстремумы сильно выделяются на фоне средней плотности сообщений и расположены на значительном расстоянии. Таким образом выбранный метод экспериментально обоснован для имеющихся данных.

3.2 Извлечение ключевых слов

Для того, чтобы использовать на шаге 6 модель HDP с частичным обучением, необходимо найти ключевые слова события, которое мы хотим проследить с помо-

щью тематической модели. Их извлечение происходит на шаге 3. В реализации алгоритма учитывая особенности сети Твиттер был выбран простой способ извлечь M ключевых слов из набора сообщений, а именно отсортировать слова по невозрастанию частоты в некотором окне и взять первых M слов.

Этот способ показал хорошие результаты потому что события в сети Твиттер носят "взрывной" характер и частота релевантных слов во время события в десятки и сотни раз превосходит нормальную частоту этих слов.

3.3 Модификация частичной обучаемости

HDP является моделью без учителя, она не требует размеченных данных для того чтобы назначить документам соответствующие темы. Для того, чтобы контролировать процесс определения тем, необходимо определенным образом изменить модель. В алгоритме 1 требуется добиться от HDP того, чтобы она выделила ключевые слова из некоторого множества $\mathcal{K}$ в отдельную тему k_0. Есть несколько способов достичь требуемого результата, один из них описан ниже.

Составим из ключевых слов в $\mathcal{K}$ документ длины M, в котором количество вхождений каждого ключевого слова пропорционально его весу. Перед тем как находить параметры HDP для корпуса документов, добавим J сгенерированных документов в корпус [10].

Для того, чтобы извлечь параметры из модели HDP в алгоритме применяется сэмплирование по Гиббсу. В этом методе изначальные документы каким-то образом распределяются по темам, а затем итеративно каждое слово удаляется из состояния и назначается на новое место, согласно уравнениям (12) и (13). Краткий псевдокод сэмплирования по Гиббсу для модели HDP приведен ниже:

Исходные параметры: множество документов Ω, параметры для HDP модели α_0, γ, I

```
назначить всем словам внутреннюю и внешнюю тему 0;
повторять
    для каждого документа d ∈ Ω выполнять
        для каждого слова w ∈ d выполнять
            удалить w из состояния модели;
            выбрать внутреннюю тему t согласно уравнению (12);
            если выбрана новая внутренняя тема t тогда
                выбрать внешнюю тему k согласно (13);
                назначить внутренней теме t внешнюю тему k;
            иначе
                извлечь назначенную ранее внешнюю тему k для внутренней
                темы t;
            конец условия
            назначить слову w внутреннюю тему t и внешнюю тему k;
        конец цикла
    конец цикла
до тех пор, пока не прошло I итераций;
```

Алгоритм 2. Сэмплирование по Гиббсу

Для того, чтобы внести в модель частичную обучаемость, необходимо по-разному обрабатывать оригинальные документы и дополнительные. Для оригинальных документов должен сохраниться первоначальный алгоритм, а словам в дополнительном

документе всегда будем назначать зафиксированную внешнюю тему k_0. Приведем псевдокод для модифицированной HDP:

Исходные параметры: множество документов Ω, параметры для HDP модели α_0, γ, k_0, I

```
назначить всем словам внутреннюю и внешнюю тему 0;
повторять
    для каждого документа d ∈ Ω выполнять
        для каждого слова w ∈ d выполнять
            если документ d дополнительный тогда
                назначить слову w внутреннюю тему 0 и внешнюю тему k_0;
            иначе
                удалить w из состояния модели;
                выбрать внутреннюю тему t согласно уравнению (12);
                если выбрана новая внутренняя тема t тогда
                    выбрать внешнюю тему k согласно (13);
                    назначить внутренней теме t внешнюю тему k;
                иначе
                    извлечь назначенную ранее внешнюю тему k для внутренней
                    темы t;
                конец условия
                назначить слову w внутреннюю тему t и внешнюю тему k;
            конец условия
        конец цикла
    конец цикла
до тех пор, пока не прошло I итераций;
```

Алгоритм 3. Сэмплирование по Гиббсу с частичным обучением

Изменяя параметры M и J можно определять насколько сильной будет привязка к конкретным ключевым словам для искомой темы. Экспериментально установлено что в качестве параметра k_0 можно взять значение 1.

3.4 Анализ сообщений темы

Возможны варианты, когда сообщения, найденные в экстремуме не являются новым событием. Это возможно в следующих случаях:

1. уже было найдено такое же событие,
2. график функции частоты не обладает свойствами, характерными событию.

Первый случай можно выявить, найдя нормализованное скалярное произведение двух функций частоты в косинусной метрике. Если значение будет близко к единице (больше параметра θ_{sim}), события можно считать тождественными. Чтобы различать второй случай, предлагается посчитать императивное стандартное отклонение, нормированное на максимальное значение частоты. Значение стандартного отклонения лучше всего считать на сглаженных данных. Если оно будет достаточно велико ($\hat{\sigma}^2 > \hat{\sigma}_0^2$), событие нельзя считать подходящим.

Полученный метод не будет работать в случае когда событие состоит из двух достаточно удаленных пиков. В этом случае можно использовать смешанную модель, как это было описанно выше в методах нахождения экстремума. При тестировании метода на реальных данных подобных событий не было найдено [11].

Дополнительно к частотным критериям необходимо ввести проверку, является ли событие спамом или информацией, интересной только узкому кругу пользователей. С большой долей точности можно судить о принадлежности события к спаму, если проверить отношение числа сообщений к числу авторов, которые отправляли эти сообщения [12]. Пусть событие содержит n сообщений от k авторов. Тогда если $k/n < \theta_{spam}$, можно утверждать, что событие является спамом. Параметр θ_{spam} можно выяснить, посчитав отношение k/n на тестовых данных. В реализации алгоритма выбрано значение $\theta_{spam} = 0.05$.

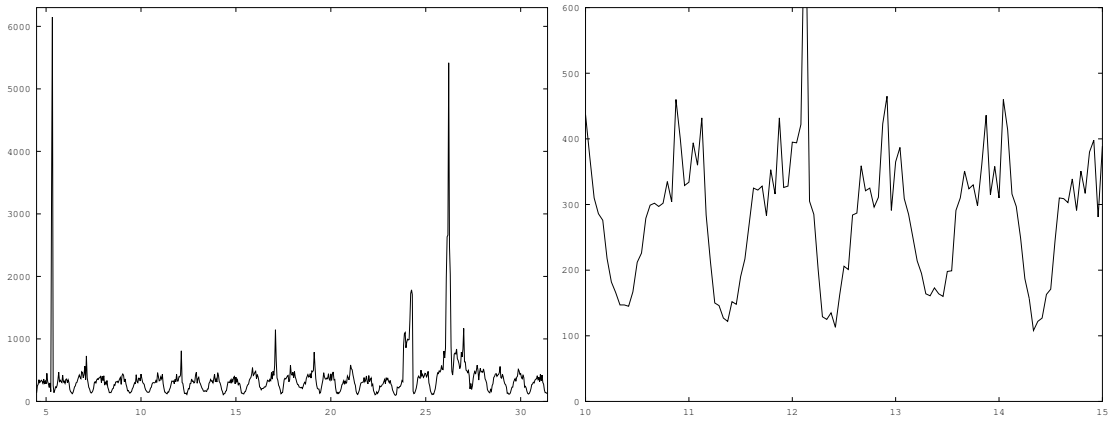
4 Реализация и тестирование решения

4.1 Описание данных

В качестве данных для задачи были использованы сообщения пользователей сети Твиттер с 4 июня по 1 июля 2013 года, содержащие хэштег #texas. Всего в корпусе находится около 240 тысяч сообщений. Цель алгоритма — найти события, показать их на графике частоты сообщений и указать ключевые слова к событиям.

Изначально данные были предобработаны: удалены знаки препинания, слова приведены к одному регистру, применен стемминг.

На рисунке 2 изображена плотность сообщений во входных данных. На графике есть несколько пиков, которые обозначают события.



а) б)

Рисунок 2. Зависимость количества сообщений в час от времени. а) Полный временной отрезок. б) Зависимость частоты от времени суток

4.2 Извлечение событий

На рисунке 3 можно видеть экстремумы, отмеченные алгоритмом 1 на шаге 1. Путем выбора метода и параметров, можно помечать больше или меньше максимумов.

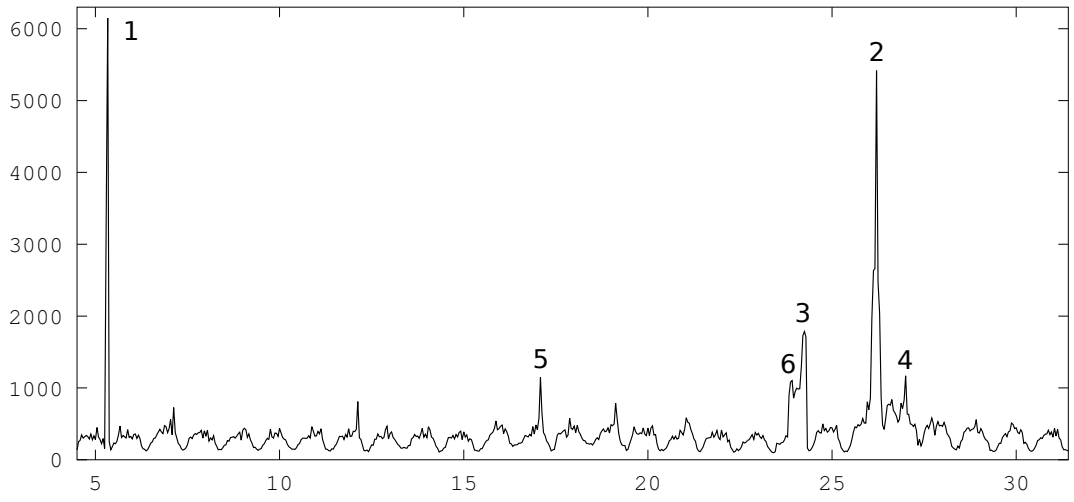


Рисунок 3. График частоты сообщений с отмеченными экстремумами

Результатом работы тематической модели на шаге 6 являются размеченные сообщения. На графиках показаны частоты сообщений, соответствующие рассматриваемым событиям.

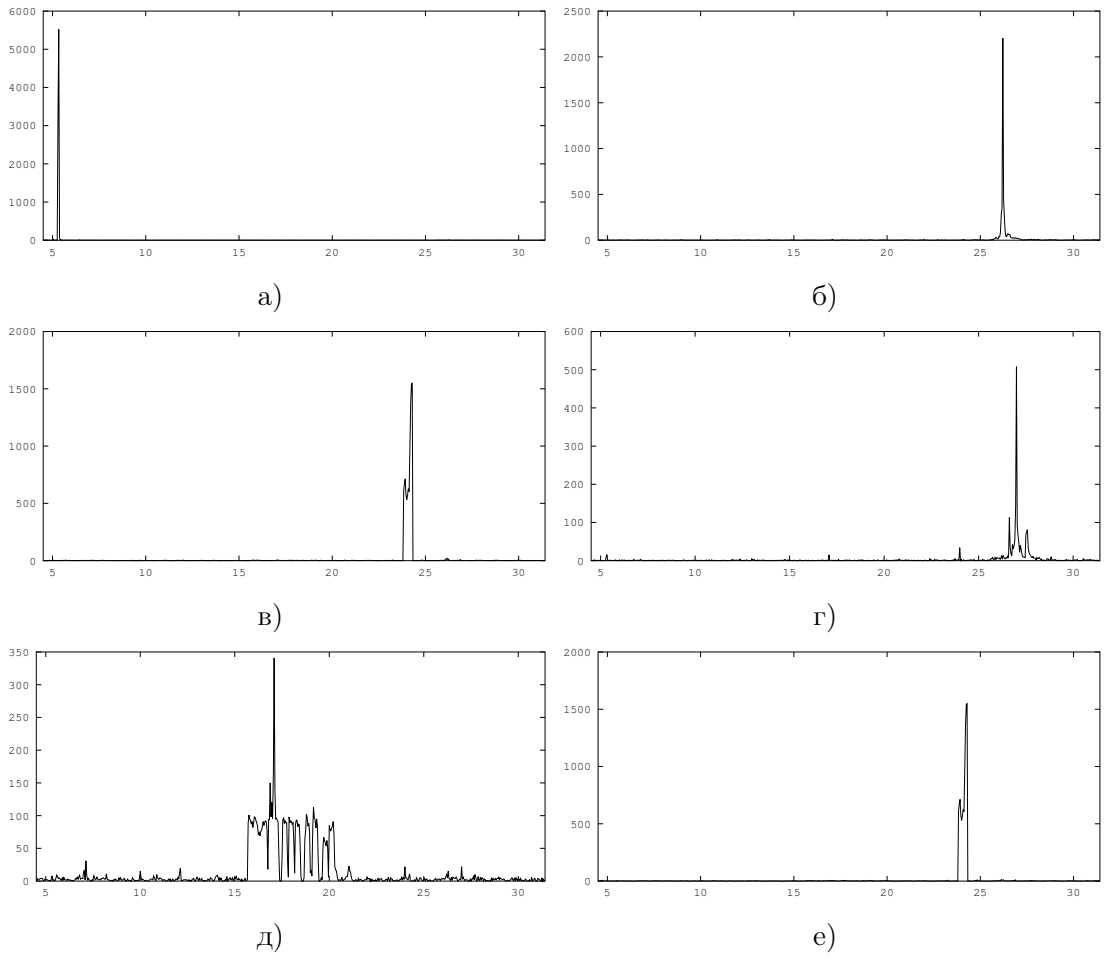


Рисунок 4. Частоты сообщений для “событий” 1 — 6. а) Событие 1. б) Событие 2. в) Событие 3. г) Событие 4. д) Событие 5. е) Событие 6.

Следующие таблицы показывают параметры частотных характеристик сообщений, полученных на шаге 8 основного алгоритма. На основании этих данных те или иные события войдут в финальный результат или будут отброшены.

Таблица 3. Жирным шрифтом помечены события с $\hat{\sigma}^2 > \hat{\sigma}_0^2$

	$\hat{\sigma}^2$
1	0.00004
2	0.00188
3	0.00831
4	0.34469
5	**3.81528**
6	0.00831

Таблица 4. Значения нормализованного скалярного произведения функций частот в косинусной метрике. Жирным шрифтом помечены значения, позволяющие считать события тождественными

1	1					
2	0.004	1				
3	0.001	0.009	1			
4	0.049	0.052	0.018	1		
5	0.011	0.022	0.027	0.050	1	
6	0.001	0.007	**1.0**	0.019	0.026	1

Таблица 5. Отношение количества уникальных авторов к общему числу соощений

	k/n
1	0.992
2	0.967
3	0.983
4	0.855
5	0.178
6	0.983

Из таблиц 3 и 4 можно видеть, что алгоритм выдаст в конечном счете только события 1, 2, 3, 4. Событие 5 будет отброшено как имеющее большое нормализованное среднеквадратичное отклонение. Событие 6 не войдет в финальный результат потому, что значение похожести функций частоты для событий 3 и 6 больше допустимой нормы:

$$\mathrm{sim}^{\cos}(f_3, f_6) = \frac{(f_3, f_6)}{\|f_3\| \cdot \|f_6\|} > \theta_{\mathrm{sim}}. \tag{18}$$

Приведем ниже таблицу 6 с описаниями событий 1, 2, 3, 4:

Таблица 6. Таблица с описаниями событий

№	Ключевые слова	Вызвавшее событие
1	year, read, prison, dragon, amazon, written, bestseller	Написанная в техасской тюрьме книга The Sword and the Dragon является бестселлером на Amazon уже более трех лет.
2	sb5, standwithwendi, abort, senat	Обсуждение сенатом закона по запрету абортов, который носит название Senate Bill 5. Wendi Davis — политик, которая боролась с принятием этого закона.
3	houston, california, england, artist, rt, watch, defjam	Хип хоп исполнитель из Хьюстон Devin the Dude объявил, что его восьмой студийный альбом будет называться One for the Road и будет выпущен в продажу в сентябре 2013 года. Альбом будет выпущен лейблом Def Jam Recordings.
4	execut, 500th, mccarthi, kimberli, 1976	В штате Техас приведен в исполнение 500-й по счету смертный приговор. Смертная казнь имеет место с 1976 года.

Ниже указаны значения параметров алгоритма 1, при которых были получены результаты этого раздела:

Таблица 7. Использованные значения параметров

M	20
L	20
l_D	50
J	10^4
R	30 минут
I	300
α_0	1.0
γ	2.0
k_0	1
θ_{sim}	0.95
$\hat{\sigma}_0^2$	3
θ_{spam}	0.05

Помимо алгоритма 1 на данных был запущен оптимизированный алгоритм New event detection, описанный в разделе 2.1. Оптимизация состоит в том, что при поиске наиболее близкого документа для заданного документа d, проверка проходила не по всем предыдущим документам, а только по 20000 последних. В результате было выяснено что NED помимо правильно помеченных сообщений также отмечает множество документов, которые не являются событием.

Это происходит потому, что сеть Твиттер содержит большое количество шума, который нередко состоит из слов, нигде больше не встречающихся. Примером

таких сообщений может быть “But #Texas ... #Why???Lol #wheredeydodatat Oh. #Dallas#Shaq #SodaShaq ??”. NED будет помечать эти сообщения новыми событиями, хотя они таковыми не являются. Алгоритм 1 лишен этого недостатка потому что на первом шаге рассматривает максимумы частотной характеристики, значение частоты сообщений в которых больше нормального уровня шума. Затем в случае если этот максимум составлен шумом или спамом, это будет выявлено на шаге 8.

4.3 Оценка качества разработанного алгоритма

Для того, чтобы иметь возможность сравнивать алгоритм с аналогами, необходимо предложить критерий качества полученного результата. В данной работе предлагается использовать точность, полноту и F-меру алгоритма как метод оценки качества алгоритмов.

Пусть алгоритм на выходе выдал информацию об n найденных событиях, из которых k событий определены верно. Тогда точностью алгоритма будет называться величина $p = k/n \in [0,\, 1]$.

Для корректного сравнения алгоритмов помимо точности, необходимо использовать полноту. Пусть в корпусе Ω содержится информация об m событиях, из них l событий были успешно отмечены алгоритмом. Тогда полнотой алгоритма называется отношение $r = l/m \in [0,\, 1]$. F-мерой алгоритма называется среднее гармоническое точности и полноты (считается, что $p \cdot q > 0$):

$$F = \left(\frac{1}{p} + \frac{1}{r}\right)^{-1} = \left(\frac{n}{k} + \frac{m}{l}\right)^{-1}. \tag{19}$$

Очевидно, что чем выше значения p, q и F показывает алгоритм на определенных данных, тем лучше он работает на этих данных.

Заметим, что для определения точности необходимо вручную проверить n событий, выданных алгоритмом, и определить насколько они достоверные, а для определения полноты просмотреть все $|\Omega|$ сообщений и выделить из них события. В связи с этим, для того, чтобы упростить определение полноты алгоритма будем считать, что событие может содержаться в корпусе только в той точке, частота сообщений в которой является максимумом в некотором окне и величина частоты больше значения θ_{freq}. Это предположение означает что нас интересуют только те события, которые активно обсуждались пользователями сети Твиттер. При тестировании решения θ_{freq} было взято средним значением частоты сообщений.

Ниже указана таблица со значениями точности, полноты и F-меры для разработанного алгоритма и для NED. По таблице можно видеть что алгоритм 1 работает на данных сети Твиттер существенно лучше.

Таблица 8. Результаты предложенного алгоритма и NED

Алгоритм	Точность	Полнота	F-мера
Алгоритм 1	0.8	1.0	0.44
NED	0.01	0.8	0.01

5 Описание практической части

Для реализации программы, описанной алгоритмом 1, был выбран язык Java. Выбор мотивирован тем, что в качестве свободной реализации для модели HDP использовалась реализация на языке Java, написанная Bleier et al [9]. Java относится к объектно-ориентированным языкам высокого уровня, что оказалось очень удобным для реализации алгоритма.

Приведем ниже диаграмму классов для модуля, в котором находится логика алгоритма и краткое описание классов программы.

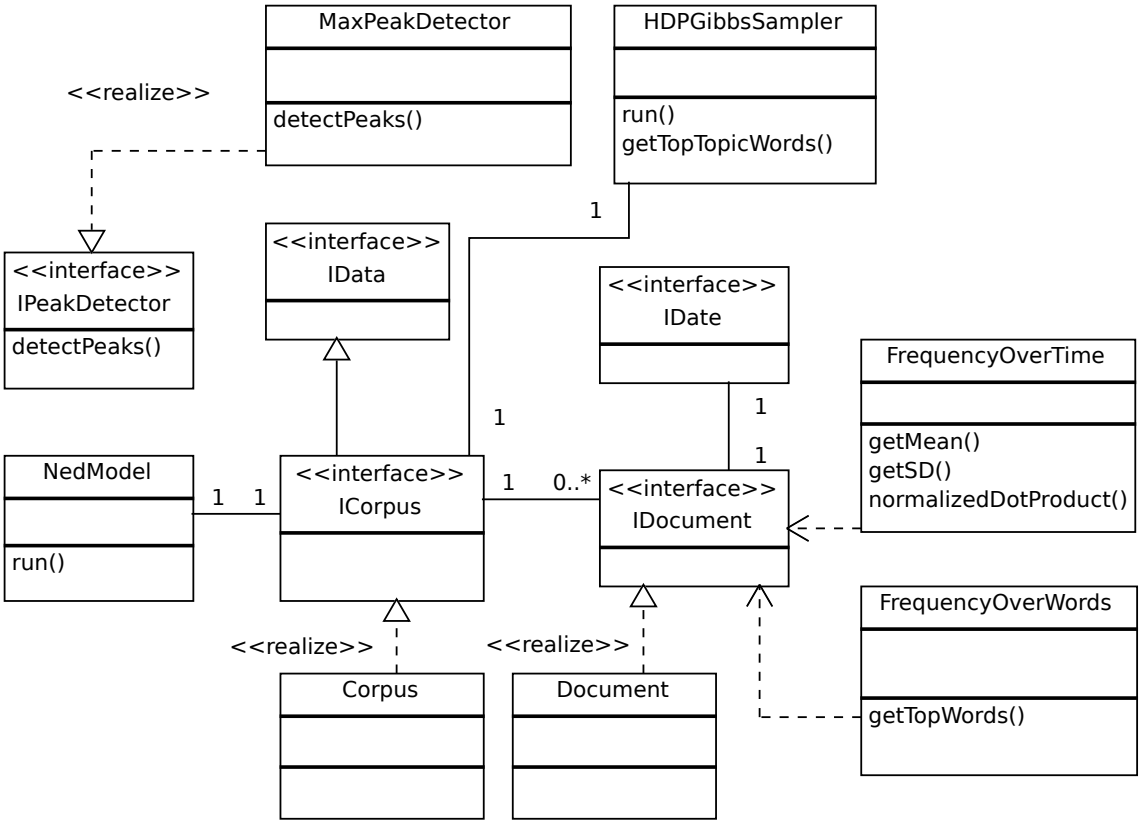


Рисунок 5. Диаграмма классов для пакета hdp

- CollectionUtils: утилитарный класс для работы с коллекциями.
- Corpus: класс, содержащий в себе документы для задачи.
- Date: класс, работающий с датой. Используется при работе с частотой сообщений.
- DateUtils: утилитарный класс для работы с датой.
- Document: документ, характеризуется словами, датой, автором и темой.
- FrequencyOverTime: класс, работающий с частотной характеристикой сообщений во времени.
- FrequencyOverWords: класс, который извлекает ключевые слова из документов.
- HDPGibbsSampler: модель HDP с модификацией частичного обучения. Размечает документы соответствующими темами.
- Main: здесь находится точка входа программы, в методе main реализуется разработанный алгоритм.

– MaxPeakDetector: класс, находящий экстремумы в частотной функции сообщений.
– MiscUtils: утилитарный класс.
– NedModel: модифицированная модель NED, используется для сравнения работы с разработанным алгоритмом.
– Pair: утилитарный класс для работы с парами объектов.
– ProbUtils: утилитарный класс для работы с вероятностями.
– WordProp: утилитарный класс, используется в деталях реализации.

Полный текст программы, реализующей разработанный алгоритм можно найти в приложении.

Заключение

В ходе выполения дипломной работы было достигнуто:

– изучены существующие методы извлечения описаний событий из коротких пользовательских сообщений,
– изучена возможность применения тематических моделей для этой задачи,
– разработан и проанализирован алгоритм извлечения описаний событий из социальной сети Твиттер на основе иерархического процесса Дирихле,
– алгоритм был реализован, параметры подобраны экспериментальным путем,
– тестирование показало, что алгоритм успешно распознает события, имеющиеся во входных данных.

Дальнейшее улучшение алгоритма можно проводить в следующих направлениях:

– исследовать работу метода на данных с большим временным отрезком, добавить возможность обработки событий с несколькими удаленными максимумами в функции частоты,
– добавить возможность извлечения подсобытий,
– использовать информацию об авторах, геотеги сообщений и другие дополнительные данные для более точного извлечения описаний событий.

Список литературы

1. Imran, Elbassuoni, Castillo, Diaz and Meier. Extracting Information Nuggets from Disaster-Related Messages in Social Media // 2013.

2. Xun Wang, Feida Zhu, Jing Jiang, Sujian Li. Real Time Event Detection in Twitter // 2011.

3. Thorsten Brants, Francine Chen, Ayman Farahat. A System for New Event Detection // 2003.

4. Konstantinos N. Vavliakis, Fani A. Tzima, and Pericles A. Mitkas. Event Detection via LDA for the MediaEval2012 SED Task // 2012.

5. David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation // 2003.

6. Tom Griffiths. Gibbs sampling in the generative model of Latent Dirichlet Allocation.

7. Girish Keshav Palshikar. Simple Algorithms for Peak Detection in Time-Series.

8. Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, David M. Blei. Hierarchical Dirichlet Processes // 2005.

9. Yee Whye Teh. Hierarchical Bayesian Nonparametric Models with Applications // 2009.

10. Ramnath Balasubramanyan, William W. Cohen, Matthew Hurst. Modeling corpora of timestamped documents using semisupervised nonparametric topic models.

11. David Blei. COS 424: Interacting with Data // 2008.

12. Sasa Petrovic, Miles Osborne, Victor Lavrenko. Streaming First Story Detection with application to Twitter // 2010.

Приложение

В приложении приведен исходный код разработанного алгоритма.