Санкт-Петербургский государственный университет

Работа допущена к защите
зав. кафедрой

_____ Блеканов Иван Станиславович

«____» _____ 2019 г.

ВЫПУСКНАЯ РАБОТА БАКАЛАВРА

Тема: Методы тематического моделирования ad-hoc дискуссий в социальных сетях

Направление: – Прикладная математика и информатика

Выполнил студент гр. 15.Б06-пу _____ Тарасов Никита Андреевич

Научный руководитель,
Кандидат технических наук, доцент _____ Блеканов Иван Станиславович

Санкт-Петербург – 2019

Оглавление

Введение

1. Актуальность

В современном мире все больше людей начинают использовать социальные сети как основной источник новостей: будь то политических, экономических или социальных. Примером может послужить недавнее исследование [1], в ходе которого было выяснено, что в США на 2018 год использование социальных медиа впервые опередило газеты, как один из основных источников новостей, и по своим значениям приближается к лидерам на данный момент - новостным сайтам и телевидению. Непрерывно растущее число пользователей, меньшая заинтересованность пользователей соцсетей в искажении фактов, скорость получения информация по крупнейшим общественным событиям (будь то записи с места происшествий или сообщения о крупных экономических событиях “из первых рук”), все это способствует дальнейшему увеличению популярности социальных сетей как источника новостей. В связи с этим как никогда актуальна задача анализа больших корпусов текстов, полученных из социальных сетей. В частности особый интерес представляет анализ так называемых ad-hoc дискуссий - дискуссий привлекающих большое число участников и касающихся определенной проблемы. Подобные дискуссии имеют свои особенности: такие как принадлежность к конкретной ситуации и случайность их формирования. Такие дискуссии носят взрывной характер, могут иметь различные закономерности распространения, вовлекать конкретных пользователей на основе аффекта - то есть формируются эмоциями а не рациональной аргументацией. [2] Однако задачи обработки и анализа подобных данных часто оказываются осложнены не только объемом данных (крупные социальные события влекут за собой дискуссии насчитывающие сотни тысяч участников и миллионы сообщений), но и специфичностью языка, используемого в социальных сетях (большое количество сокращений, возможностей интерпретации пользовательских сообщений), а также одной из основных проблем - проблемы коротких текстов. Большинство пользователей выражают свое мнение короткими сообщениями и максимальная длина пользовательских сообщений во многих социальных медиа (таких как twitter) ограничена, что ведет к относительно низкой средней длине текстов. Данные проблемы ограничивают использование традиционных мето-

дов статистического анализа ввиду большей разреженности данных. В данной работе рассматривается один из методов анализа текстовых данных - тематическое моделирование. Данный метод позволяет выделять из больших текстовых корпусов темы - вероятностные распределения слов, характеризующих ту или иную идею, мнение, аргумент. Данные распределения в свою очередь позволяют говорить о принадлежности отдельных текстов корпуса к той или иной тематике, а в определенных моделях и о распределениях тем по текстам.

2. Цель работы

Целью данной работы стала разработка методов тематического моделирования крупных пользовательских ad-hoc дискуссий в социальных сетях, позволяющие с учетом специфики анализа коротких текстов автоматически выявлять темы, которые затрагивают пользователи в своих сообщениях. Применение подобных моделей позволяет как уменьшать размерность данных , так и выделять основные идеи, аргументы и мнения. Особый интерес представляет совместное применение тематического моделирования с другими алгоритмами анализа данных как способ предварительной обработки данных (путем уменьшения размерности данных, отсеивания, так называемого, шума - слов и сообщений не оказывающих на дискуссию значительного влияния), так и совместно с другими алгоритмами для получения принципиально новых результатов. Примером подобного совместного применения может послужить использование тематического моделирования и метрик центральности пользователей - участников дискуссии. В данном случае использование мер центральности позволяет выделять наиболее влиятельных участников дискуссии, а тематическое моделирование позволяет рассматривать их основные идеи и аргументы, выявлять как данные идеи влияют на сообщения других пользователей.

3. Задачи

Основными задачами данной работы являются:

- Сбор данных и их предварительная обработка
- Рассмотрение методов тематического моделирования

- Проведение сравнительного анализа методов и метрик оценивания качества алгоритмов
- Разработка архитектуры программы
- Реализация методов и их интеграция в программный комплекс по обработке больших корпусов текстов, полученных из социальных сетей
- Визуализация результатов работы алгоритмов

Глава 1

Обзор методов

Одной из первых тематических моделей стала модель LSI (Latent semantic indexing)[3], захватывающая скрытые связи слов средствами линейной алгебры, в частности сингулярного разложения, для уменьшения размерности исходных данных и, как следствие, - выделение основных слов, образующих документы. Авторы данного метода не ставили задачи разработки универсального метода обработки текстовых данных, основным вкладом данной работы стала теоретическая основа, актуальная для дальнейших моделей анализа данных. Модель PLSA (Probabilistic latent semantic analysis), разработанная на основе идей и теорем модели LSI, однако в отличии от нее основана на смешанном разложении, в свою очередь берущим своё начало из модели скрытых классов [4]. Данный подход более принципиален, поскольку имеет прочную основу в области статистики. Наиболее распространенной тематической моделью на данный момент является модель LDA, используемая для решения широкого спектра задач: как анализа текстовых данных, так и в качестве компонента анализа графических данных.

1.1. Latent Dirichlet Allocation

В LDA полагается наличие k латентных тем, согласно которым генерируются документы, а каждая тема представляется как мультиномиальное распределение по V словам в словаре. Документ генерируется сэмплированием по этому набору тем, а затем сэмплированием по словам из этого набора. Таким образом, документ из N слов $w = (w_1, ..w_N)$ генерируется с помощью следующего процесса. Сначала θ сэмплируется из распределения дирихле $(\alpha_1, ..., \alpha_k)$, а это означает, что θ лежит в $k-1$ - размерном симплексе: $\theta_i \geq 0, \sum_i \theta_i = 1$ Затем для каждого из N слов, из мультиномиального распределения $p(z_n = i|\theta) = 0$ сэмплируется тема $z_n \in (1, ..., k)$. И наконец каждое слово сэмплируется условно по теме z_n из мультиномиального распределения $p(w|z_n)$. Из данного распределения понятно, что θ_i можно считать степенью присутствия темы в документе. Полная формула вероятности документа, получаемая из приведенных выше

распределений, определяется следующим образом:

$$p(w) = \int_{\theta} (\prod_{n=1}^{N} \prod_{z_n=1}^{k} p(w_n|z_n;\beta)p(z_n|\theta)p(\theta;\alpha)d\theta,$$

где $p(\theta;\alpha)$ - это распределение дирихле, $p(z_n|\theta)$ - мультиномиальное распределение, заданное параметром θ, а $p(w_n|z_n;\beta)$ - мультиномиальное распределение по словам. Данная модель параметризуется с помощью k-размерного вектора параметров дирихле $(\alpha_1, ..., \alpha_k)$ и матрицы $\beta = k \times |V|$ - параметров, контролирующих мультиномиальное распределение по словам [5]. Рис. 1.1. иллюстрирует

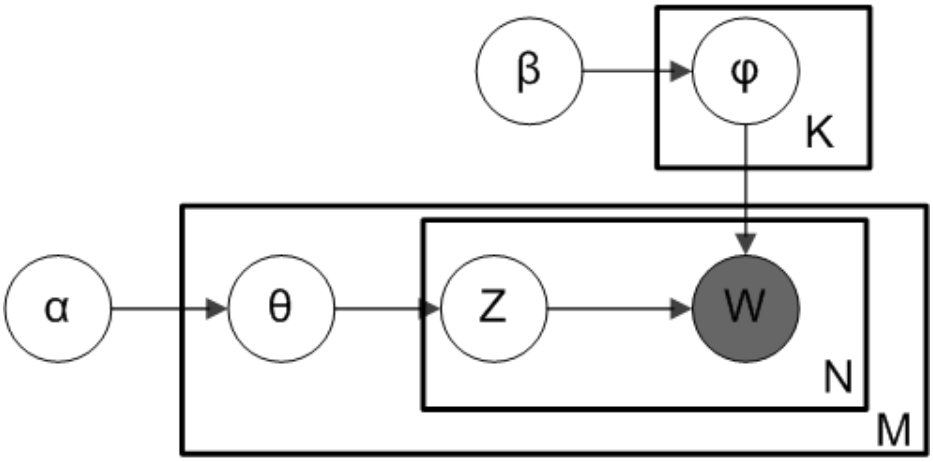

Рис. 1.1. Графическое представление модели LDA

модель LDA, при этом областями, обозначенными прямоугольниками, являются повторяющиеся элементы (тексты и темы в этих текстах). Из формулы вероятности для документов и графического представления следует еще один способ представления данной модели. Получив θ, слова получаются из следующего мультиномиального распределения $p(w|\theta) = \sum_{z=1}^{k} p(w|z)p(z|\theta)$. Данное распределение соответствует модели с использованием униграмм, в которой каждый документ генерируется одной темой, выбираемой из следующего распределения $p(w) = \sum_{z=1}^{k}(\prod_{n=1}^{N} p(w_n|z))p(z)$.

Получение параметров и обучение модели. Для начала рассмотрим влияние одного документа на общее распределение. Для упрощения введем следующие обозначения : пусть $w_n^j = 1$, если w_n - j-е слово в словаре и если $z_n^i = 1$ - это i-я тема. Пусть β_{ij} обозначает $p(w^j = 1|z^i = 1)$, а $w = (w_1, ..., w_N)$, $z = (z_1, ..., z_N)$. Расписывая выражение распределения для одного документа,

получаем:

$$p(w;\alpha,\beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int_\theta (\prod_{i=1}^{k} \theta^{\alpha_i - 1})(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{|V|} (\theta_i \beta_{ij})_{n^j}^{w}) d\theta$$

Это гипергеометрическая функция, аналитическое вычисление которой невозможно. В связи с тем, что большие коллекции текстов требуют быстрого вывода и обучения, возникает необходимость использовать численные методы. В данном случае используется вариационный способ приближения функции распределения. Исходя их этого, получается следующее вариационное приближение к логарифмической функции правдоподобия:

$$p(w;\alpha,\beta) = \log \int_\theta \sum_z p(w|z;\beta) p(z|\theta) p(\theta;\alpha) \frac{q(\theta,z;\gamma,\phi)}{q(\theta,z;\gamma,\phi} d\theta \geq$$

$$\geq E_q] \log p(w|z;\beta) + \log p(z;\theta) + log(\theta;\alpha) - \log q(\theta,z;\gamma,\phi)]$$

где $q(\theta,z;\gamma,\phi) = q(\theta;\gamma)\prod_n (z_n;\phi_n)$ выбирается как полностью факторизованное вероятностное распределение, параметризуемое γ и ϕ_n таким образом, что $q(\theta;\gamma)$-распределение дирихле по γ, а $q(z_n;\phi_n)$ -мультиномиальное по ϕ_n. В этом распределении, значения в нижнем пределе вычислимы и дифференцируемы и следовательно максимизируемы по γ и ϕ для нахождения лучшего приближения $p(w;\alpha,\beta)$.

Процесс обучения состоит в выполнении следующих двух шагов до достижения желаемого результата:

$$\beta_{iw_n} \propto \exp(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j))$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}$$

,

где Ψ - это первая производная функции Γ. Имея коллекцию документов, используется EM алгоритм с вариативным шагом шагом E, максимизируя

нижнюю границу в логарифмической функции правдоподобия:

$$\log p(D) \geq \sum_{m=1}^{M} E_{q_m}[\log p(\theta, z, w)] - E_{q_m}[\log q_m(\theta, z)]$$

Шаг E пересчитывает q_m для каждого документа, используя шаги, описанные выше. Шаг M оптимизирует данное выражение с учетом параметров модели α и β. Для мультиномиальных параметров $\beta_i j$ используется следующий алгоритм обновления на шаге M:

$$\beta_{ij} \propto \sum_{m=1}^{M} \sum_{n=1}^{|w_m|} \phi_{mni} w_{mn}^{j}$$

параметры дирихле не являются независимыми друг от друга и для их оптимизации применяется метод Ньютона-Рафсона:

$$\frac{\delta l}{\delta \alpha_i} = \sum_{m=1}^{M} (\Psi(\sum_{j=1}^{k} \alpha_j) - \Psi(\alpha_i)) + (\Psi(\gamma_{mi} - \Psi(\sum_{j=1}^{k} \gamma_{mj}))$$

В вариативном EM алгоритме максимизация исходного уравнения по параметрам (α, β) чередуется с максимизацией данного уравнения по параметрам q_m до достижения сходимости.

Вычислительная сложность: Сложность по времени для LDA составляет $O(N_d K_z L_d)$, а сложность по памяти - $O(N_d K_z + N_d L_d)$. Где N_d - количество документов, K_z - количество тем, а L_d - средняя длина документа [6].

Основными преимуществами данной модели являются:

- Использование дополнительного распределения тем по документам делает данную модель относительно модели униграмм [7], в которой каждый документ задается одной темой.

- LDA является, в отличие от pLSI [4], генеративной моделью ввиду отсутствия дополнительно набора случайных переменных, задающих модель pLSI. Данный факт позволяет LDA присваивать набор тем для документа, изначально не входящего в текстовый корпус.

- Данная модель менее склонна к переобучению, ввиду отсутствия дополнительного случайного распределения $p(d, w)$

Основные недостатки модели:

- Большая чувствительность к шуму в текстовых данных, полученному в результате неточностей при предварительной обработке, ошибок в оригинальном тексте, специфики языка в социальных сетях.

- Проблемы при работе с большими корпусами коротких текстов. В данном случае на вход данного алгоритма подается разряженная матрица в которой несмотря на потенциально большое число элементов, каждый отдельный элемент имеет небольшое число параметров. А когда число тестовых документов во много раз превышает число параметров для данного алгоритма наблюдаются как ухудшение качества полученных данных, так и сложности алгоритма [8].

1.2. Biterm Topic Model

Стандартные тематические модели такие как LDA и униграммная модель получают темы, основываясь на совпадениях слов на уровне документа. Эффективность подобных методов сильно снижается в случае коротких текстов, так как вероятность совместного наличия слов в документе снижается. Данная модель решает эту проблему, моделируя процесс генерации так называемых битермов (biterms). Например, если слова «программа», «компиляция», «класс» и «приложение» часто встречаются друг с другом в одном и том же контексте, можно говорить, что они принадлежат к одной и той же тематике. Стандартные тематические модели неявно фиксируют подобные совпадения слов, моделируя генерацию слов на уровне документа. В отличии от данных подходов, BTM напрямую моделирует шаблоны словосочетания на основе битермов. Битерм представляет неупорядоченную пару слов, совместно встречающаяся в коротком контексте. В коротких текстах каждый документ рассматривается как отдельный контекстный блок и любые два разных слова в коротком текстовом документе представляют битерм. Например, в предложени «язык скомпилировал программу», есть три битерма, то есть «язык скомпилировал», «скомпилировал программу», «язык программу». Битермы, извлеченные из всех документов коллекции составляют входные данные модели BTM [9].
Основная идея BTM - получение тем коротких текстов на основе битерм во всем корпусе для решение проблемы разряженности каждого отдельного документа. Корпус документов таким образом генерируется из набора битерм, где каждый берется независимо из одной из тем. Предположим, что α и β приорные параметры распределения дирихле. Генеративный процесс в BTM можно описать следующим образом:

1. Выбрать распределение тем $\theta_i \sim Dir(\alpha)$
2. Для каждой темы z:
    a. Выбрать распределение слов для темы$\phi_z \sim Multinomial(\beta)$
3. Для каждого битерма b_i:
    a. Выбрать тему $z \sim Multinomial(\theta)$
    б. Выбрать два слова $w_i, w_j \sim Multinomial(\phi_z)$

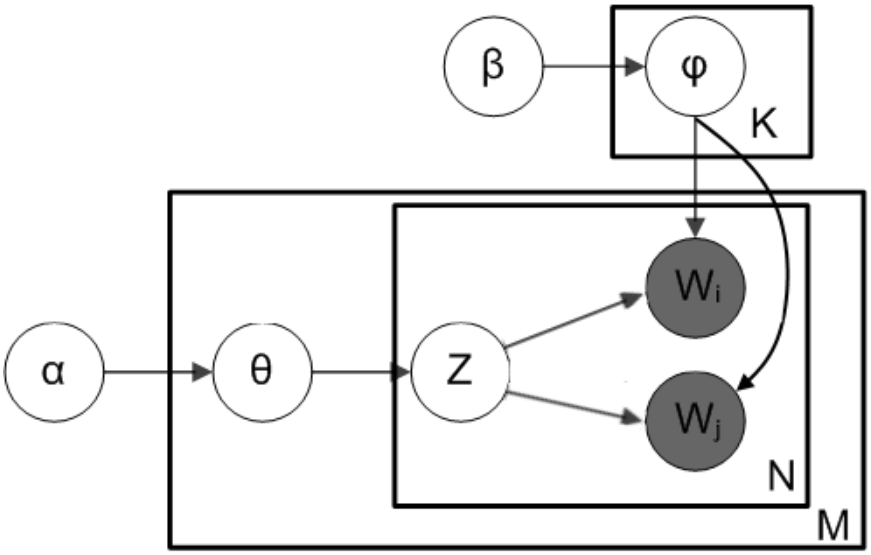

Рис. 1.2. Графическое представление модели BTM

Рис 1.2. иллюстрирует данный генеративный процесс. Исходя из него совместная вероятность отдельного битерма $b = (w_i, w_j)$ может быть выражена следующим образом:

$$P(b) = \sum_z P(z)P(w_i|z)P(w_j|z) = \sum_z \theta_z \phi_{i|z} \phi_{j|z}$$

Тогда вероятность для всего корпуса:

$$P(B) = \prod_{i,j} \sum_z \theta_z \phi_{i|z} \phi_{j|z}$$

Из данного распределения понятно, что данный алгоритм моделирует непосредственно совместно встречающиеся слова, не используя отдельных слов для определения семантики тем. Кроме того, в данном алгоритме все части корпуса агрегируются для изучения темы, что позволяет использовать более широкий набор совместно встречающихся слов для улучшения раскрытия скрытых тем.
Определение тем в документе
Одно из важнейших отличий BTM стандартных алгоритмов тематического моделирования состоит в том, что BTM не моделирует процесс генерации документов. Поэтому данная модель не позволяет напрямую получить пропорции тем для документов. Для вывода тем отдельного документа, предполагается, что пропорции документа равны математическому ожиданию пропорций би-

термов, сгенерированных из данного документа:

$$P(z|d) = \sum_b P(z|b)P(b|d).$$

В данном уравнении $P(z|b)$ можно рассчитать по формуле Байеса по параметрам, оцениваемым в BTM:

$$P(z|b) = \frac{P(z)P(w_i|z)P(w_j|z)}{\sum_z P(z)P(w_i|z)P(w_j|z)},$$

, где $P(z) = \theta_z$ и $P(w_i|z) = \phi_{i|z}$. Тогда оставшаяся проблема состоит в нахождении $P(b|d)$. В данном случае берется эмпирическое распределение битермов в документе

$$P(b|d) = \frac{n_d(b)}{\sum_b n_d(b)},$$

, где $n_d(b)$ - частота битерма b в документе d. В коротких текстах $P(b|d)$ является почти равномерным распределением по всем битермам в документе d.

Далее описывается алгоритм вывода параметров ϕ, θ [10].Подобно LDA, вывод параметров не может быть произведен аналитически. В данном случае используется семплирование по Гиббсу для вычисления приближенных значений. Семплирование по Гиббсу - это простой и широко применяемый алгоритм Монте-Карло с цепью Маркова. По сравнению с другими методами выявления скрытых переменных, такими как вариационный метод и максимальная апостериорная оценка, семплирование по Гиббсу имеет два преимущества: большая точность приближения, поскольку он асимптотически приближается к правильному распределению, большая эффективность по памяти, так как требует держать в памяти только счетчики и переменные состояния, делая данный метод более предпочтительным для крупномасштабных наборов данных. Основная идея семплирования по Гиббсу заключается в альтернативной оценке параметров путем замены значения одной переменной значением, полученным из распределения этой переменной, обусловленным значениями остальных переменных. В BTM, необходимо вычисление следующих скрытых переменных: z, ϕ и θ. Однако с техникой свертывания семплирования по Гиббсу ϕ и θ могут быть найдены с использованием сопряженных априоров α и β. Следовательно, требу-

ется только моделировать процесс присвоения тем для каждого битерма из его условного распределения с учетом оставшихся переменных. Процесс выполнения семплирования по Гиббсу может быть описан следующим образом: сначала выбираются начальные состояния для цепи Маркова (случайным образом), затем вычисляется условное распределение $P(z|z_b, B, \alpha, \beta)$ для каждого битерма $b = (w_i, w_j)$, где z_b обозначает назначения тем для всех битермов, кроме b, B - общий набор битерм. Применяя правило умножения вероятностей, получается следующая условная вероятность:

$$P(z|z_b, B, \alpha, \beta) \propto (n_z + \alpha \frac{(n_{w_i|z} + \beta)(n_{w_j|z} + \beta)}{(\sum_w n_{w|z} + M\beta)^2}$$

, где n_z - число присвоений темы z битерму b, а $n_{w|z}$ - число присвоений слова w теме z. Применяя данное выражение, производится оценка распределения слов по темам ϕ и общего распределения тем θ:

$$\phi_{w|z} = \frac{n_{w|z} + \beta}{\sum_w n_{w|z} + M\beta}$$

$$\theta_z = \frac{n_z + \alpha}{|B| + K\alpha}$$

Вычислительная сложность: сложность по времени $O(N_d L_d K_z t(L_d - 1)/2)$, сложность по памяти $O(N_d L_d (L_d - 1)(K_z + L_d)/2)$, где N_d - число документов, K_z - число тем, L_d - средняя длина документа.

Основными преимуществами данной модели являются:

- Использование преобразования исходных данных к менее разряженных позволяет решить проблему коротких текстов. BTM в связи с этим демонстрирует улучшение качества тем относительно LDA, рассматривающего темы на уровне документов как модель униграмм.

- В отличие от модели униграмм, BTM способен распознавать наличие нескольких тем в одном документе, что позволяет распознавать менее явные темы.

Основные недостатки модели:

- Большая вычислительная сложность относительно LDA. В качестве примера авторами модели BTM приводится набор данных Tweets2011, в кото-

ром средняя длина сообщений $l = 5.21$. В данном случае время выполнения одной итерации примерно в 3.5 раза больше времени данной итерации в LDA.

- Невозможность прямого выявления тем в документах, необходимость использовать дополнительный алгоритм преобразования итоговых распределений, что ведет к введению дополнительных условностей и в некоторых случаях ухудшает качество итоговой модели.

1.3. Word Network Topic Model

WNTM - еще одна модель, специализированная для работы с короткими текстами [11]. Основная идея WNTM основана на следующих наблюдениях. Когда тексты короткие, пространство “слово-документ” крайне рязрежено, а пространство “слово-слово” по-прежнему содержит большое число ненулевых элементов. Интуитивно понятно, что число слов, связанных с редкими темами, часто превышает количество документов, связанных с этими темами. Поскольку распределение по темам для каждого документа невозможно узнать точно в коротких или несбалансированных текстах, мы изучаем вместо этого распределение по темам для каждого слова. Следовательно, WNTM использует стандартный Gibbs Sampling для LDA для обнаружения скрытой группы слов (т. е. темы) и изучает распределением по темам для слов, а не тем для документов. Изучение тем слова, а не тем документа делают WNTM менее чувствительным к длине документа. Кроме того, сеть слов может быть построена с любым типом текстов, что делает модель WNTM простой и универсальной в реальных приложениях, лишая ее недостатков иных моделей, таких как mixture of unigrams и BTM.

Построение сети слов. В сети совместных слов вершинами являются слова, встречающиеся в корпусе, а ребро между двумя словами указывает на наличие этих слов в одном контексте не реже одного раза. В качестве контекста рассматривается документ или скользящее окно с фиксированным размером. Что-

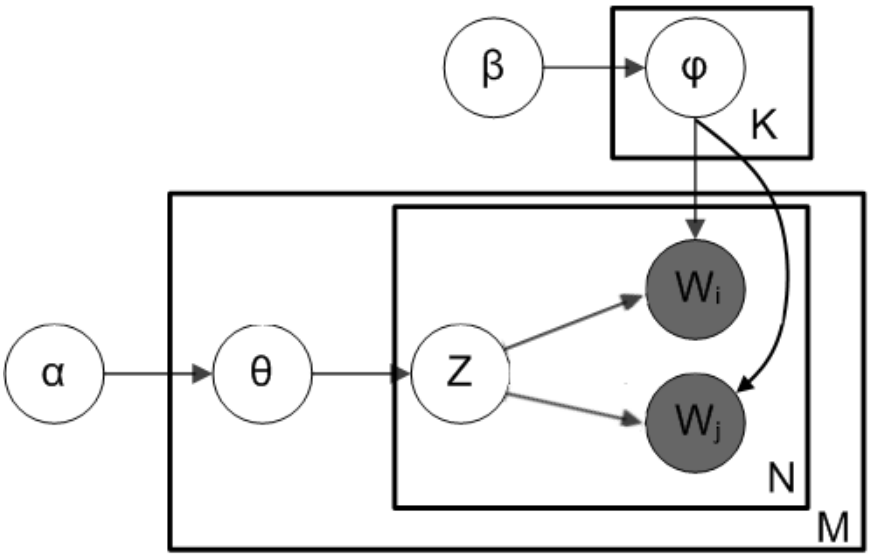

Рис. 1.3. Процесс преобразования исходных данных в сеть слов

бы ограничить размер сети слов и получить только локальный контекст для

каждого слова, используется скользящее окно фиксированного размера. Размер окна выбирается опытным и для коротких текстов составляет как правило 10 слов. Вершинам присваивается значение суммы всех смежных ей ребер. Чтобы преобразовать коллекцию документов в сеть слов, сначала фильтрация стоп слов и низкочастотных слов, а каждый документ сканируется с использованием скользящего окна. В процессе сканирования слов через документы, любые два различных слова, появившихся в одном и том же окне, будут рассматриваться совместно. Число таких пересечений аккумулируется и определяет значение грани между ними. Также стоит отметить тот факт, что в тематических моделях тему можно рассматривать как набор слов, часто встречающихся совместно, что во многом схоже с группами, образующими отдельные элементы сети слов. Таким образом, группы слов в данной модели можно принимать как темы модели LDA. В то же время, получение тем из сети слов имеет теоретическую гарантию согласованности тем. Более того, редкие темы могут образовывать компактные скрытые группы слов в сети слов, а поэтому тематическая модель, основанная на сети слов, может эффективно находить группы слов, соответствующих редким темам.

Стандартное семплирование по Гиббсу из модели LDA может быть использоваться для обнаружения групп слов в больших сетях. В это же время как для использования данного метода сначала необходимо представить сеть слов как набор псевдодокументов. Как показано на рис. 1.3, каждое слово в сети может трактоваться как псевдодокумент, состоящий из списка смежных слов. Хотя WNTM использует ту же самую выборку Гиббса как и LDA, параметры, лежащие в основе генеративного процесса для них различны: LDA обучается создавать коллекцию документов, используя темы и слова по этим темам, в то время как WNTM обучается генерировать список слов, смежный с каждым словом в сети, используя слова, принадлежащие скрытым группам слов. Таким образом WNTM изучает статистические отношения между словами, скрытые группы слов и смежные списки слов, предполагая, что каждый смежный список слов генерируется семантически определенной вероятностной моделью. Сначала предполагается, что в сети слов имеется фиксированный набор групп скрытых слов, и каждая скрытая группа слов z связана с распределением по словарю Φ_z , которое берется из априора Дирихле $Dir(\beta)$. Генеративный процесс всего псевдодокумента может быть интерпретирован следующим

образом:

1. Для каждой латентной группы слов z выбрать $\phi_z \sim Dir(\beta)$
2. Выбрать $\theta_i \sim Dir(\alpha)$, распределение латентной группы слов для смежного списка слов L_i для слова w_i
3. Для каждого слова w_i:
    а. Выбрать латентную группу слов $z_j \sim \theta_i$
    б. Выбрать смежное слово $w_j \sim \phi_{z_j}$

В WNTM распределения θ - это вероятность появления скрытых групп слов в смежном списке каждого слова, а распределения ϕ обозначают вероятность принадлежности слов к определенной скрытой группе. Так как смежный список слов для каждого отдельного слова состоит из контекстной информации, полученной из всего корпуса, WNTM моделирует распределение по скрытым группам слов для каждого слова, а не распределение по темам для каждого документа.

Как и модель, основанная на битермах [9], WNTM не моделирует процесс создания документа и не позволяет напрямую получить распределения тем по документам из семплирования по Гиббсу. Поскольку WNTM моделирует процесс генерации смежного словарного списка для каждого слова, что позволяет выявить глобальную контекстную информацию слова, то в качестве пропорций тем смежного списка слов θ_i, для слова w_i можно рассматривать как пропорции тем для w_i. Темы для каждого документа таким образом могут быть получены, используя пропорции тем для всех слов. В частности, предполагается, что математическое ожидание распределения пропорций слов, сгенерированных документом равно пропорциям тем документа:

$$P(z|d) = \sum_{w_i} P(z|w_i)P(w_i|d),$$

, где $P(z|w_i)$ равно $\theta_{i,z}$, полученному в процессе обучения модели, а $P(w_i|d)$ можно, как и в случае BTM, взять равным эмпирическому распределению слов по документам:

$$P(w_i|d) = \frac{n_d(w_i)}{len(d)},$$

, где $n_d(w_i)$ - частота слова w_i в документе d, а $len(d)$ - это длина данного документа.

Вычислительная сложность. Так как сложность LDA по времени $O(N_d K_z L_d)$, где N_d - это число документов, K_z- число тем, а L_d- средняя длина документа, сложность WNTM аналогично равна $O(N_p K_g L_p)$, где N_p - число псевдодокументов (число слов в словаре корпуса), K_g - число тем, L_p - средняя длина псевдодокумента. Тогда в случае равного числа тем и использования скользящего окна размера c, сложность WNTM по времени в $O(c^2)$ раз больше сложности LDA. Сложность по памяти так же аналогична модели LDA в случае рассмотрения псевдокументов и превышает ее в $O(c^2)$ раз.

Основными преимуществами данной модели являются:

- Использование всех документов корпуса для построения псевдодокументов позволяет рассматривать пространство "слово-слово а не стандартное для классических моделей "слово-документ решая тем самым проблему коротких текстов и разреженности данных.

- WNTM проста в реализации, так как процесс обучения данной модели во многом совпадает с процессом обучения модели LDA.

Основные недостатки модели:

- Большая чувствительность к длине текстов корпуса. Если размер скользящего окна c меньше средней длины документов, алгоритм демонстрирует сильное увеличение сложности как по памяти, так и по времени.

- Склонность алгоритма к переобучению, ввиду составления относительно небольшого числа псевдодокументов, содержащих большое число, не всегда связанных, слов.

- Аналогично модели BTM, данный алгоритм не позволяет напрямую выявлять темы в документах.

1.4. Методы оценивания качества тематических моделей

При отображении тем для пользователей каждая тема t обычно представляется в виде списка $M = 5, .., 20$ самых вероятных слов этой темы, в порядке

убывания их тематических «свернутых» вероятностей. Экспертная оценка результатов исследования предполагает, что темы низкого качества могут быть обнаружены с использованием метрик, основанных на совпадении слов в моделируемых документах. Обозначая $D(v)$ как частоту документа со словом v (то есть числа документов с хотя бы одним словом типа v), а $D(v, v')$ как совместную частоту слов v и v' (т.е. количество документов, содержащих одно или более слов v и хотя бы одно слово v'), связность темы определяется как

$$C(t, V(t)) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^t, v_l^t) + 1}{D(v_l^t)}$$

где $V(t) = (v_1^t, ..., v_M^t)$ - список из M наиболее вероятных слов для темы t [24].

- UMass coherence [25] определяет значение связности, основываясь на совпадениях слов в документах

  $$score(v_i, v_j, e) = \log \frac{D(v_i, v_j) + e}{D(v_j)}$$

  где $D(x, y)$ обозначает число документов содержащих слова x и y, а $D(x)$ обозначает число документов, содержащих слово x.

- PMI (pointwise mutual information) [26] - использует взаимную информацию контекста документов, а не только взаимные совпадения слов. Вычисляется следующим образом

  $$PMI(w_i, w_j) = \log_2 \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

- NPMI (Normalised PMI) [27] - дополнение к модели PMI, нормализующее значения для данной метрики

  $$NPMI(w_i, w_j) = \frac{PMI(w_i, w_j)}{-\log(p(w_i, w_j))}$$

1.5. Выводы

Рассмотренные модели обладают своими достоинствами и недостатками. LDA, являясь наиболее простой моделью, не требующей преобразования исходных данных, требует наименьших временных затрат и затрат по памяти. При этом данная модель разрабатывалась для анализа полноразмерных текстов (статей, публикаций и т.д.) и не является оптимальной для обработки коротких текстов, полученных из социальных сетей. В связи с этим в дальнейшем LDA будет рассматриваться как базовая модель, эффективность которой для стандартных текстов подтверждена как теоретически, так и на многих наборах данных. Модели BTM и WNTM, предназначенные для обработки коротких текстов обладают, в теории, большей точностью, однако требуют дополнительных затрат для преобразования исходного корпуса сообщений к соответствующему виду. Дальнейший раздел посвящен анализу эффективности данных алгоритмов на реальных данных, способам оценивания качества алгоритмов тематического моделирования и визуализации полученных распределений.

Глава 2

Программный комплекс для тематического моделирования ad-hoc дискуссий

2.1. Архитектура программного комплекса

В рамках данной работы была разработана архитектура программного комплекса, включающего как реализацию методов тематического моделирования средствами Python, так и обработку начальных данных, представление результатов работы алгоритмов.

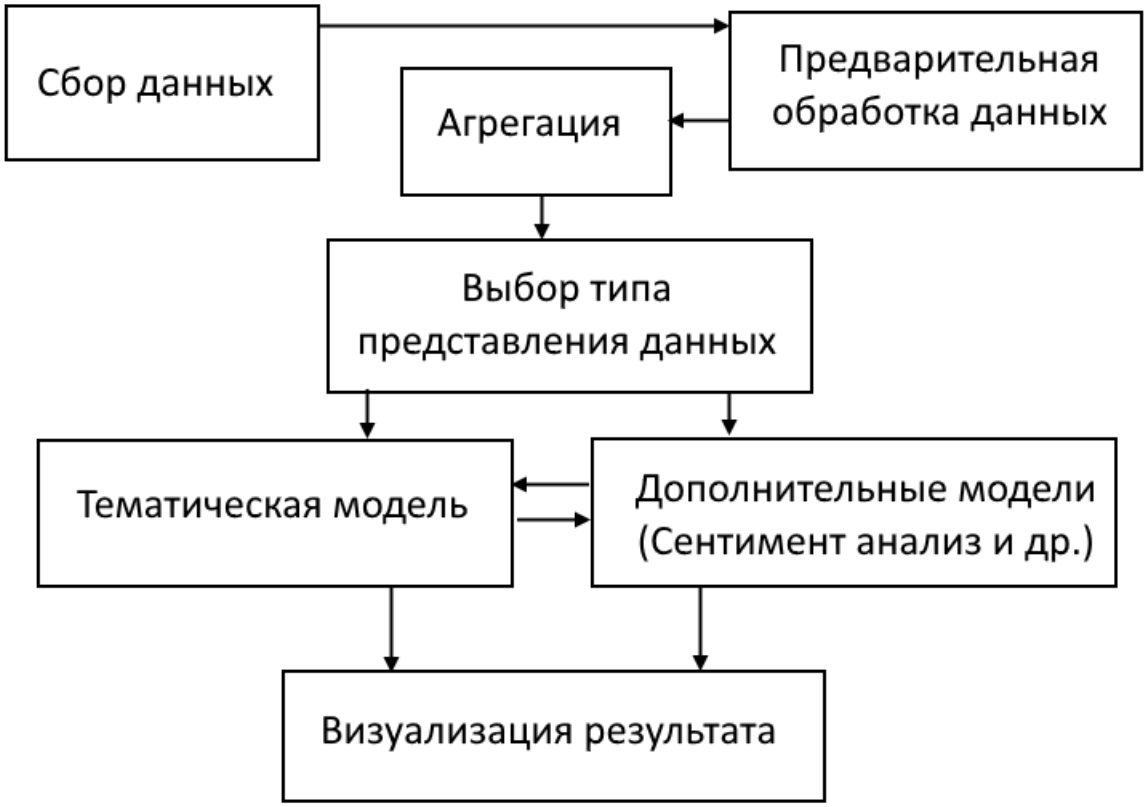

Рис. 2.1. Архитектура программного комплекса

2.1.1. Предварительная обработка данных

Обработка данных включает в себя очистку исходного файла от специальных символов, разбиение на отдельные сообщения и далее отдельные слова.

Слова каждого сообщения обрабатываются с использованием лемматизации, стемминга, стоп слова и слова встречающиеся слишком редко и слишком часто отбрасываются [12]. Лемматизация — это приведение каждого слова в документе к его нормальной форме. В русском языке нормальными формами считаются: для существительных — именительный падеж, единственное число; для прилагательных — именительный падеж, единственное число, мужской род; для глаголов, причастий, деепричастий — глагол в инфинитиве. Разработка хорошего лемматизатора (lemmatizer) требует составления грамматического словаря со всеми формами слов, либо аккуратной формализации правил языка со всеми исключениями, что является трудоемким проектом. Известные лемматизаторы совершенствуются постепенно. Их недостатком является неполнота словарей, особенно по части специальной терминологии и неологизмов, которые во многих приложениях как раз и представляют наибольший интерес. Стемминг — это более простая технология, которая состоит в отбрасывании изменяемых частей слов, главным образом, окончаний. Она не требует хранения словаря всех слов и основана на правилах морфологии языка [13]. Недостатком стемминга 12 является большее число ошибок. Стемминг хорошо подходит для английского языка, но хуже подходит для русского. Отбрасывание стоп слов необходимо как для сокращения средней длины текстов, с минимальным изменением длины словаря, так и для непосредственного улучшения качества моделей, упрощения интерпретации итоговых тем, так как данные слова (предлоги, союзы, числительные, местоимения, некоторые глаголы, прилагательные и наречия) встречаются в текстах практически любой тематики и при этом в отрыве от остальных слов предложения несут минимальную смысловую нагрузку. Отбрасывание частых слов необходимо также ввиду их наличия в большинстве документов и как следствие - большинстве тем (такими словами являются например ключевые слова, заданные на этапе сбора данных). Редко встречающиеся слова отбрасываются на основании того, что они не характеризуют отдельные темы, в случае больших объемов данных часто являются словами не относящимися к рассматриваемой дискуссии или словами с невыявленными на предыдущем этапе грамматическими ошибками, добавление которых в словарь может привести к дублированию его элементов. Отбрасывание слишком редких или частых слов для отдельных документов, в случае коротких текстов, способно ухудшить качество модели, так как в случае когда средняя длина текстов небольшая каж-

дое отдельное слово больше влияет на итоговые темы, в этой связи частые и редкие слова рассматриваются не в пределах отдельных документов, а в пределах всего корпуса.

2.1.2. Способы представления данных

Для представления исходных сообщений в вектора с численными компонентами для проведения дальнейшего анализа возможно использование различных моделей представления текстовых данных. Модель мешка слов (Bag of Words) является одним из наиболее распространенных способов представления текстовых данных [14]. В подобной модели каждый параметр связан с каждым словом в словаре: значение данного параметра либо ноль, если слово не появляется в предложении, либо отлично от нуля в противном случае. Данное значение обычно вычисляется как оценка TF-IDF, то есть произведение количества вхождений слова в предложение (частота термов, TF) на значение усиливающее вклад редких слов (обратная частота документов, IDF) [15]. Также данная модель может быть расширена для учета n-грамм, то есть последовательных словосочетаний, так что порядок слов в предложениях (по крайней мере, локально) будет использован моделью.

Второй подход, рассмотренный как вариант для представления предложений, заключается в использовании деревьев синтаксического анализа (constituency parse tree), которые естественно кодируют структуру предложения, описывая грамматические связи между частями предложения через дерево. Сходство между подобными структурами можно использовать с ядрами деревьев (Tree Kernels). Ядро дерева состоит из меры сходства между двумя деревьями, которая учитывает количество общих подструктур или фрагментов. Разные определения фрагментов требуют применения разных мер для использования подобных структур [16].

Третий подход к представлению предложений основан на вложенных словах (Word Embeddings), популярный метод, разработанный на основе анализа текстовых данных средствами глубокого обучения. Нейронные сети, такие как CNN и LSTM, могут обрабатывать текстовый ввод путем преобразования его в последовательность идентификаторов (по одному для каждого отдельного слова), которые в свою очередь используются для обучения нейросети и получения векторных представлений для слов и предложений [17].

После проведения экспериментальных проверок различных методов представления исходных данных, для дальнейшего использования был выбран метод мешка слов (bag of words) ввиду не только наибольшей обобщенности в практическом применении, что является немаловажным фактором в случае анализа данных имеющих различную структуру, но и меньшей чувствительности к исходным данным, что немаловажно в случае анализа коротких текстов.

2.1.3. Способы агрегации

Один из способов решения проблемы коротких текстов - использование различных способов агрегации отдельных сообщений в группы по тому или иному признаку. В качестве основных методов агрегации данных из социальных сетей выделяются:

- Базовая схема (без агрегации) - По умолчанию каждый твит обрабатывается как отдельный документ, обучая выбранную модель на всех данных.

- Агрегация по автору - Объединение твитов в соответствии с автором является стандартным способом агрегирования данных в Twitter [18] и доказано [19] превосходит базовый метод. Для использования данного метода, для каждого автора строится документ, объединяющий все опубликованные твиты.

- Burst-score wise Pooling - Тенденция в Твиттере [20] (иногда упоминаемая как актуальная тема) состоит из одного или нескольких теров в период времени,в который объем сообщений, затрагивающих данные слова в данный период времени превышает некоторый ожидаемый уровень активности. С целью для выявления тенденций в постах в данных можно обнаружить необычные «всплески» частоты использования слов. Предполагается, что M - это набор всех сообщений в наборе твитов, R - набор из одного или нескольких термов (потенциальная трендовая тема), к которым необходимо присвоить значение, а $d \in D$ представляет один день в наборе из D дней. Затем мы $M(R, d)$ определяется как подмножество сообщений в M таких, что каждое сообщение содержит все слова из R, сообщения опубликованного в течение дня d. Пусть $Mean(R) = \frac{1}{|D|} \sum_{d \in D} M(R, d)$. Соответственно $SD(R)$ является стандартное отклонение $M(R, d)$ в тече-

ние дней $d \in D$. Затем значение тенденции определяется как:

$$burst - score(R, D) = \frac{|M(R, d) - Mean(R)|}{SD(R)}$$

Сообщения имеющие значение $burst - score$ выше определенного порогового значения агрегируются в одну группу.

- Temporal Pooling - Когда происходит крупное событие, большое количество пользователей выкладывают сообщения о событии в течение короткого периода времени. Для использования подобных закономерностей, агрегируются все сообщения, опубликованные в течение определенного периода времени.

- Hashtag-based Pooling - Хэштег в Twitter - это строка символов с предшествующим символом #. Во многих случаях хэштеги можно рассматривать как тематические маркеры, указывающие на контекст твита или как основную идея, выраженная в сообщении. Таким образом все сообщения содержащие одинаковые хэштеги записываются в один документ.

В данной работе рассматриваются методы без агрегации, а также агрегации основанной на временной метке сообщений. Несмотря на наилучшее улучшение алгоритмов при использовании хэштегов для агрегации данных, использование данного метода ограничено не только отсутствием хэштегов в большом числе сообщений в стандартных данных, но и методом сбора данных, в котором события и как следствие сообщения пользователей изначально рассматриваются как набор хэштегов.

2.2. Эксперимент

2.2.1. Постановка эксперимента

В данной работе модели были протестированы на данных, собранных по трем ad-hoc дискуссиям в социальной сети Twitter: Беспорядки в западном Бирюлево (Россия) - Октябрь 2013, Беспорядки в Фергюсоне (США) - Август 2014, Террористический акт в редакции Charlie Hebdo (Франция) - Январь 2015. Данные были собраны на основе хэштегов в сообщениях пользователей. Сбор

данных производится по заранее выявленным ключевым словам и тегам, характеризующим рассматриваемую дискуссию. Заранее задается также и временной промежуток в течении которого рассматриваются записи. Сообщения пользователей собираются собственной программой без использования API социальных сетей, что позволяет осуществлять процесс с большей точностью, а также универсализировать его при рассмотрении сразу нескольких социальных сетей. Итоговый документ состоит из сообщений пользователей и соответствующих им время отправки и имя пользователя.

Бирюлево [21]

- Общее число сообщений: 10215
- Общее число пользователей, принявших участие в дискуссии: 11429
- Временной период сбора данных: 1.10.2013 - 31.10.2013
- Число пользователей, опубликовавших сообщения в рассматриваемый временной промежуток: 3574

Фергюсон [22]

- Общее число сообщений: 193812
- Общее число пользователей, принявших участие в дискуссии: 169677
- Временной период сбора данных: 22.08.2014 - 31.08.2014
- Число пользователей, опубликовавших сообщения в рассматриваемый временной промежуток: 70018

Charlie Hebdo [23]

- Общее число сообщений: 505069
- Общее число пользователей, принявших участие в дискуссии: 952615
- Временной период сбора данных: 07.01.2015 - 10.01.2015
- Число пользователей, опубликовавших сообщения в рассматриваемый временной промежуток: 238491

По дискуссиям в рамках данной работы требовалось с помощью разработанного программного комплекса применить различные методы тематического моделирования и оценить их качество с помощью заданных метрик.

2.2.2. Результаты эксперимента

Существует множество способов представления полученных тематик, как статическими списками, так и интерактивно, с возможностью дополнительной обработки финальных данных (группируя данные по датам, пользователям и так далее). Для каждого набора данных выбираются несколько наиболее связных тем (тем сфокусированных на определенных словах, имеющих более высокую вероятность появления в данных темах). В результате работы программного комплекса каждым алгоритмом были получены темы представленные в табл. 2.1 – 2.3

Таблица 2.1. Таблица тем для набора данных Бирюлево

Topic 1	Topic 2	Topic 3	Topic 4
		LDA	
migrant	warehouse	broadcast	riot
Zeynalov	work	live	Manezhka
police	man	moscow	moscow
murder	boutique	photo	migrant
Sherbakov	moscow	find	block
		WNTM	
Moscow	news	Moscow	police
event	Sherbakov	OMON	authorities
Russia	migrant	Zeynalov	killer
riot	murder	arrest	russian
mayhem	killer	Sherbakov	meetings
		BTM	
citizen	OMON	Sherbakov	russian
police	warehouse	Zeynalov	government
local	police	killer	riot
riot	arrest	arrest	migrant
Moscow	killer	moscow	news

Значения оценок качества работы методов по реальным пользовательским ad-hoc дискуссиям Фергюсон, Шарли, Бирюлево представлены на рис. 2.1 – 2.3

Результаты работы модуля интерактивной визуализации в виде тепловых временных карт представлен на рис. 2.4, в виде графического представления всех тем - на рис. 2.5.

Таблица 2.2. Таблица тем для набора данных Charlie Hebdo

Topic 1	Topic 2	Topic 3	Topic 4
		LDA	
policia	die	police	islam
Paris	satire	shooting	religion
terroristas	cartoonist	attack	youngest
sospechosos	frankreich	suspects	local
ataque	attentater	update	extrimists
		WNTM	
attack	suspects	cartoonists	media
french	police	support	cartoons
today	two	editor	toxic
terror	attack	respond	image
killed	hostage	journalism	caricatures
		BTM	
police	french	victims	muslims
suspects	shooting	solidarity	islam
hostage	gunman	attack	must
killed	dead	France	say
breaking	killed	jesuischarlie	religion

Таблица 2.3. Таблица тем для набора данных Фергюсон

Topic 1	Topic 2	Topic 3	Topic 4
		LDA	
police	MikeBrown	militarization	Miami
life	black	police	overtown
surrender	brown	law	America
must	racism	reason	vote
dissa	justice	end	jail
		WNTM	
MikeBrown	CNN	must	movement
amp	cops	surrender	speak
police	Times	police	join
black	black	dissa	support
people	shooting	see	now
		BTM	
must	join	community	pd
police	movement	Miami	look
surrender	now	support	closer
dise	speak	lot	MikebBrown
click	die	overtown	msnbc

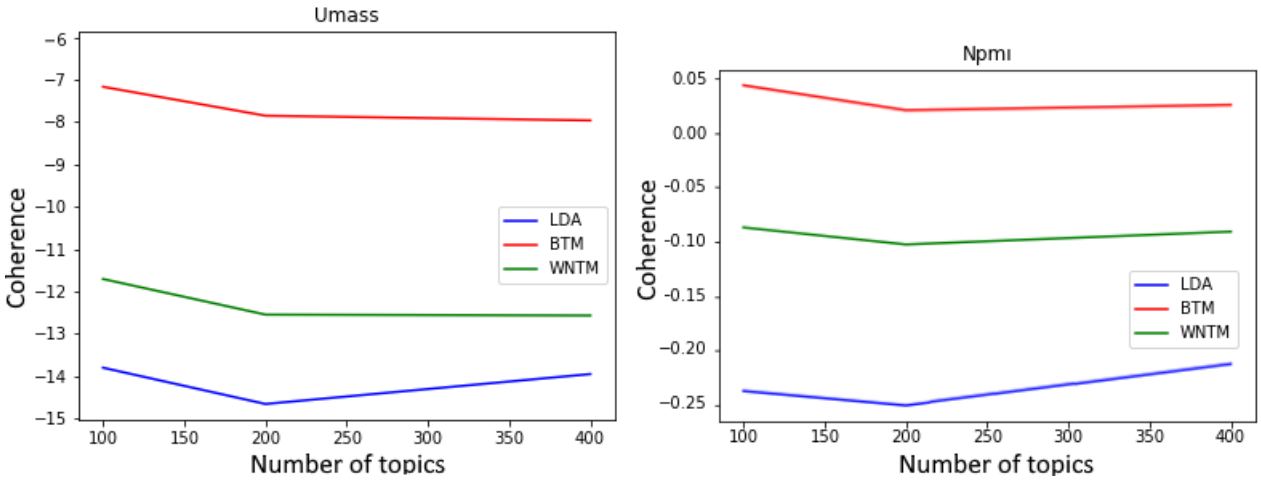

Рис. 2.2. Связность тем для набора данных Charlie

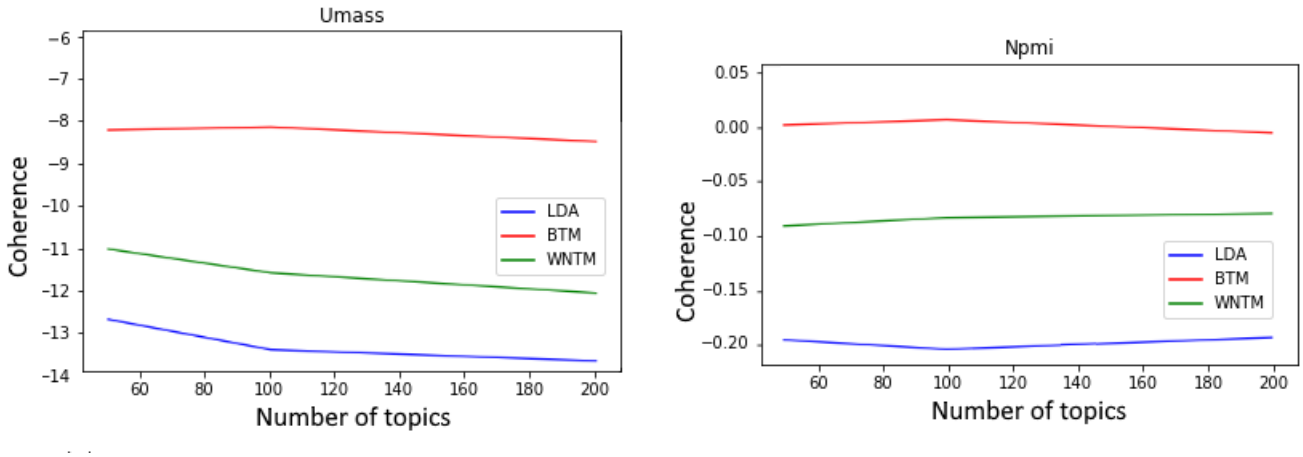

Рис. 2.3. Связность тем для набора данных Фергюсон

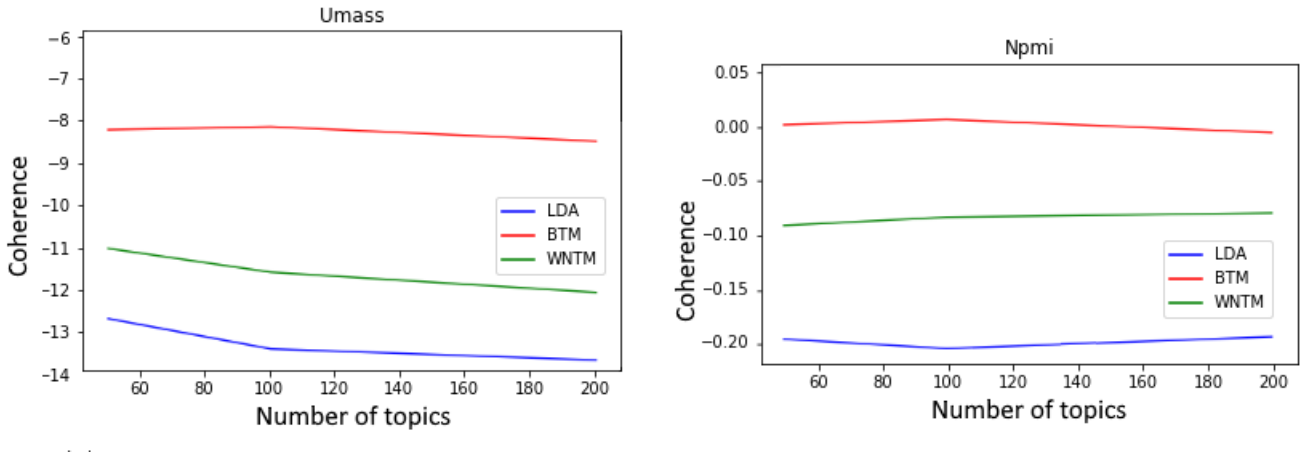

Рис. 2.4. Связность тем для набора данных Фергюсон

Рис. 2.5. Тепловая карта, показывающая наиболее актуальные темы по дням

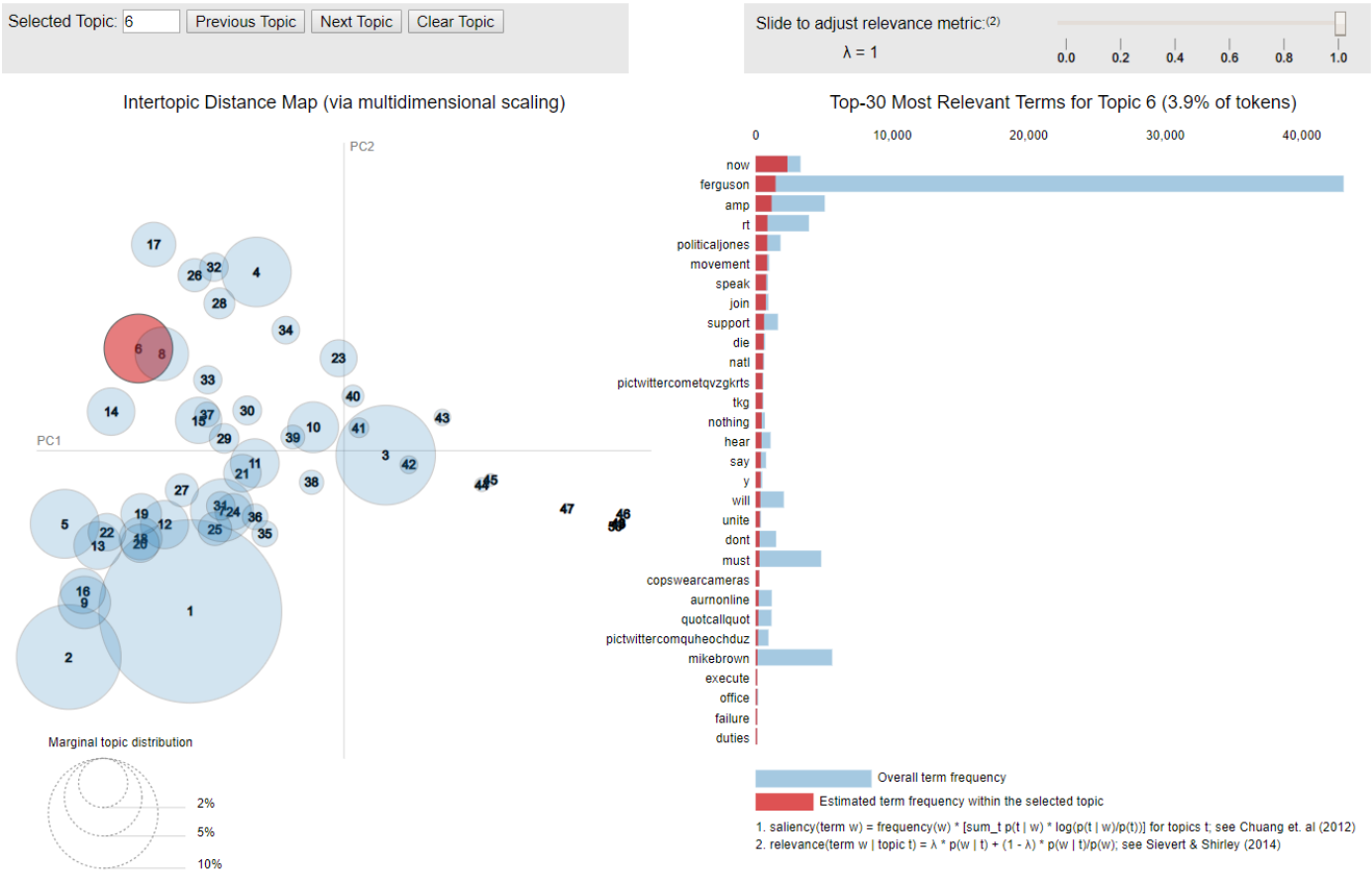

Рис. 2.6. Модуль LDAVis используемый для интерактивной визуализации [28]

2.2.3. Выводы

В результате анализа данных ad-hoc дискуссий было выяснено, что тематическая модель Biterm Topic Model оказалась наиболее эффективной и стабильной из всех рассмотренных вариантов, основываясь на всех мерах связности тем. В свою очередь базовая модель Lda не специализируемая на работе с короткими текстами без дополнительной модификации в виде предварительной обработки данных показывает худшие показатели в сравнении с моделями, созданными для работы с данными из социальных сетей в основном из-за разреженности данных. Однако темы, определенные с помощью этой модели, могут все еще будут использоваться для анализа ключевых моментов дискуссий, аргументы, а также в качестве основы для других задач анализа данных.

Заключение

2.3. Результаты работы

В рамках данной работы был проведен сравнительный анализ различных моделей тематического моделирования. Были рассмотрены как модели специализированные для анализа коротких текстов, так и модели с доказанной эффективностью для различных типов данных. В результате был разработан программный комплекс для тематического моделирования крупных пользовательских ad-hoc дискуссий в социальных сетях, основанный на использовании методов тематического моделирования LDA, BTM , WNTM, проведен сравнительный анализ данных методов на 3 реальных ad-hoc дискуссиях в социальной сети Twitter. В рамках данного комплекса были имплементированы программы для построения соответствующих моделей, вычисления связности тем и визуализации итоговых данных. Результаты исследования можно использовать как независимый инструмент анализа дискуссий так и в составе программного комплекса по анализа подобных данных, включающий реализацию методов вычисления центральности пользователей и построения тем для наиболее влиятельных пользователей, анализа их влияния на различные группы участников дискуссии, сентимент анализа, выявляющего окрас пользовательских сообщений (негативные, позитивные, нейтральные), для выявления тем в отдельных типах сообщений, что, как пример, позволяет говорить о темах, ключевых слов и аргументах, затрагиваемых в сообщениях имеющих негативный окрас.

2.4. Перспективы развития

Дальнейшая работа включает продолжение работы над системой сравнительного анализа, в частности сравнения оценки экспертов и предложенных метрик качества моделей. Первая часть данной работы, проделанная совместно с факультетом журналистики СПБГУ показала, что модель WNTM показывает наилучшие результаты, несмотря на лучшие показатели связности для модели BTM. Также планируется разработка дополнительных моделей, рассмотрение иных генеративных моделей, в частности моделей использующих генеративно-состязательные сети для обучения модели, задающей темы для сообщений.

Список литературы

1. Elisa S. Social media outpaces print newspapers in the U.S. as a news source. — 2018.
2. Blekanov I., Bodrunova S. Power Laws in Ad Hoc Conflictual Discussions on Twitter. — 2018.
3. Thomas K., Peter W., Darrell L. An Introduction to Latent Semantic Analysis. — 1998.
4. Hofmann T. Probabilistic Latent Semantic Analysis // Twenty-Second Annual International SIGIR Conference. — 1992.
5. Latent dirichlet allocation / M. David, Y. Andrew, I. Michael, L. John // Journal of Machine Learning Research. — 2003. — С. 993—1022.
6. Deng C., Xiaofei H., Jiawei H. Training Linear Discriminant Analysis in Linear Time // IEEE 24th International Conference on Data Engineering. — 2008.
7. Mitchell T. Text classification from labeled and unlabeled documents using em. Machine learning / K. Nigam, A. McCallum, S. Thrun, T. Mitchell. — 2000. — URL: https://doi.org/10.1023/A:1007692713085 5404v1.
8. Blekanov I., Tarasov N., Maksimov A. Topic modeling of conflict ad hoc discussions in social networks. // ACM International Conference Proceeding Series. — 2018. — С. 122—126.
9. BTM: Topic Modeling over Short Texts / C. Xueqi, Y. Xiaohui, L. Yanyan, G. Jiafeng // IEEE Transactions on Knowledge and Data Engineering. — 2014. — С. 2928—2941.
10. On smoothing and inference for topic models / A. Asuncion, M. Welling, P. Smyth, Y. Teh // In Proceedings of the 25th Conference on UAI. — 2009.
11. Yuan Z., Jichang Z., Ke X. Word Network Topic Model: A Simple but General Solution for Short and Imbalanced Texts. — 2014.
12. Vorontsov K. Topic models rewiev //. — 2019.
13. Introduction to Information Retrieval / D. Christopher, Raghavan, Prabhakar, Schütze, Hinrich // Cambridge University Press. — 2009.

14. Leopold E., Kindermann J. Text categorization with support vector machines. How to represent texts in input space? // Machine Learning. — 2002. — С. 423—444.

15. Sebastiani F. Machine learning in automated text categorization // ACM Comput Surv. — 2002. — С. 1—47.

16. Moschitti A. Efficient convolution kernels for dependency and constituent syntactic trees // . In: Frnkranz J, Scheffer T, Spiliopoulou M (eds) ECML. — 2006.

17. Efficient estimation of word representations in vector space / T. Mikolov, K. Chen, G. Corrado, J. Dean // e. arXiv preprint arXiv:13013781. — 2013.

18. Twitterrank: finding topic-sensitive influential twitterers. / J. Weng, P. Lim, J. Jiang, Q. He // In WSDM'10. — 2010. — С. 261—270.

19. Hong L., Davison B. Empirical study of topic modeling in Twitter. // 1st ACM Workshop on Social Media Analytics. — 2010.

20. Improving LDA topic models for microblogs via tweet pooling and automatic labeling / M. Rishabh, S. Scott, B. Wray, X. Lexing // Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. — 2013. — С. 889—892.

21. Bodrunova S., Blekanov I., Litvinenko A. Please Follow Us: Media roles in Twitter discussions in the United States, Germany, France, and Russia. // Journalism Practice. — 2018.

22. Bodrunova S., Smoliarova A., Blekanov I. Mediatization of twitter? Traditional and online media in ad hoc discussions on inter-ethnic conflicts. // American Communication Journal. — 2017.

23. Content Sharing in Conflictual Ad-Hoc Twitter Discussions: National Patterns or Universal Trends? / S. Bodrunova, A. Smoliarova, I. Blekanov, A. Litvinenko // Communications in Computer and Information Science. — 2017.

24. Douven I., Meijs M. . Measuring coherence // Synthese. — 2007. — С. 405—425.

25. Optimizing semantic coherence in topic models. / D. Mimno, H. Wallach, E. Talley, M. Leenders, A. McCallum // In Proceedings of the Conference on Empirical Methods in Natural Language Processing. — 2011. — С. 262—272.

26. Automatic Evaluation of Topic Coherence / D. Newman, H. Jey, K. Grieser, T. Baldwin // HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — 2010. — С. 100—108.

27. Aletras N., Stevenson M. Evaluating topic coherence using distributional semantics / In Proceedings of the 10th International Conference on Computational Semantics (IWCS'13) Long Papers. — 2013. — С. 13—22.

28. Carson S., Kenneth E. LDAvis: A method for visualizing and interpreting topics // ACL Workshop on Interactive Language Learning, Visualization, and Interfaces. — 2014.