Санкт-Петербургский государственный политехнический
университет
Институт компьютерных наук и технологий
Кафедра компьютерных систем и программных технологий

ДИССЕРТАЦИЯ
на соискание ученой степени
МАГИСТРА

Тема: **Полуавтоматическое извлечение часто задаваемых вопросов на основе анализа обращений в службу поддержки**

Студент гр. 63501/3 П.П. Жук

Санкт-Петербург – 2017

Санкт-Петербургский государственный политехнический
университет
Институт компьютерных наук и технологий
Кафедра компьютерных систем и программных технологий

Диссертация допущена к защите
зав. кафедрой

_____ В.М. Ицыксон

«____» _____ 2017 г.

ДИССЕРТАЦИЯ
на соискание ученой степени
МАГИСТРА

Тема: **Полуавтоматическое извлечение часто задаваемых вопросов на основе анализа обращений в службу поддержки**

09.04.01 – Информатика и вычислительная техника
09.04.01.15 – Технологии проектирования системного и прикладного программного обеспечения

Выполнил студент гр. 63501/3 _____ П.П. Жук

Научный руководитель,
м. т. т. _____ М.Х. Ахин

Рецензент,
науч. степ, науч. звание _____ Х.Х. YYYYY

Консультант по нормоконтролю,
ст. преп. _____ С.А. Нестеров

Консультант по ,
м.ф.-м.н. _____ М.С. Давыдова

Санкт-Петербург – 2017

Эта страница специально оставлена пустой.

РЕФЕРАТ

Отчет, 38 стр., 1 рис., 4 табл., 10 ист., 1 прил.

АНАЛИЗ ТЕКСТА, ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА, ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ, ЧАСТО ЗАДАВАЕМЫЕ ВОПРОСЫ

Часто задаваемые вопросы (ЧЗВ) содержат актуальную информацию о программном продукте и позволяют снизить нагрузку на отдел технической поддержки. Формирование ЧЗВ и поддержка их в актуальном состоянии требует существенных затрат от разработчика.

Описываемый в данной работе способ позволяет в автоматическом режиме выбрать наиболее релевантные для добавления в ЧЗВ вопросно-ответные пары, которые затем передаются эксперту для редактирования перед публикацией. Для этого применяются методы интеллектуального анализа текста и тематического моделирования.

Данный подход может быть применен и для других источников ИТ-дискуссий, таких как: форумы, вопросно-ответные системы. Практические результаты показывают, что используемый подход позволяет упростить формирование актуальных ЧЗВ.

ABSTRACT

Report, 38 pages, 1 figures, 4 tables, 10 references, 1 appendicies

TEXT MINING, NATURAL LANGUAGE PROCESSING, TOPIC MODELING, FREQUENTLY ASKED QUESTIONS

Frequently asked questions (FAQ) contains answers for typical user problems of the software product and helps to decrease amount of calls to user support department. Creating the FAQ and filling out it with the actual information is pretty time- and resource-consuming from the developer.

This work proposes the technique for automatic extraction of the most relevant for the adding in the FAQ question-answer pairs. Then extracted question-answer pairs should be validated or edited by the expert before publication. The technique is based on the text mining and topic modeling approaches.

It is also could be applied for the other IT-discussions sources such as forums, question-answers systems an so on. Practical results shows this technique can be used to facilitate the creation of the FAQs.

СОДЕРЖАНИЕ

СПИСОК ОБОЗНАЧЕНИЙ И СОКРАЩЕНИЙ

FAQ	Frequently Asked Questions
JSON	JavaScript Object Notation
LDA	Latent Dirichlet Allocation, скрытое размещение Дирихле
XML	eXtensible Markup Language
YAML	Yet Another Markup Language
ВОП	Вопросно-Ответная Пара
ЧЗВ	Часто Задаваемые Вопросы

ВВЕДЕНИЕ

Часто задаваемые вопросы (ЧЗВ) — список вопросов, которые часто возникают по какой-либо теме, и ответы на них, данные экспертами в соответствующей области. Программное обеспечение может сопровождаться ЧЗВ для помощи пользователям в решении распространенных проблем, например: Linux[1], Apache Lucene[2], Eclipse SWT[3].

Основное преимущество ЧЗВ над пользовательской документацией и тематическими форумами — простота поиска необходимой информации. Однако создание качественных ЧЗВ — это нетривиальный процесс, требующий либо предугадывания потенциальных вопросов, либо ручного анализа обратной связи от пользователей. Целью данной работы является разработка метода извлечения ВОП из обращений в службу поддержки для упрощения задачи формирования ЧЗВ.

Предлагаемый способ, помимо обращений в службу поддержки, может быть использован и для других источников ИТ-дискуссий: форумов, вопросно-ответных систем. Сначала определяются часто обсуждаемые, повторяющиеся темы, для этого используется тематическое моделирование, а именно — скрытое размещение Дирихле (Latent Dirichlet allocation, LDA) [1], дополненное шагами пред- и постобработки, специфичными для ИТ-дискуссий. Далее среди обращений, относящихся к одной теме, с помощью косинусного расстояния и дополнительных фильтров проходит поиск вопросно-ответных пар (ВОП).

Извлечение ВОП — это автоматический процесс, однако перед публикацией в ЧЗВ необходимо провести дополнительный экспертный анализ, поскольку для извлеченных ВОП может потребоваться валидация, переформулирование или редактирование (например, удаление конфиденциальных данных). Таким образом, весь подход является полуавтоматическим.

Работа состоит из пяти разделов. В разделе 1 рассматриваются существующие подходы к задаче извлечения ЧЗВ, описываются различные тематические модели и способы оценки их качества. Раздел 2 посвящен постановке задачи извлечения ЧЗВ из обращений в службу поддержки. В разделе 3 описывается предлагаемый подход к решению поставленной задачи. Представлена общая схема подхода, а также по-

[1] http://tldp.org/FAQ/Linux-FAQ/index.html
[2] https://wiki.apache.org/lucene-java/LuceneFAQ
[3] http://www.eclipse.org/swt/faq.php

дробно рассмотрены каждый из этапов. В разделе 4 рассматривается разработка алгоритма. Раздел 5 посвящен оценке качества полученного решения. В этом разделе приводятся результаты тестирования алгоритма, результаты экспертной оценки, предлагаются направления для проведения дальнейших исследований.

1. АНАЛИЗ МЕТОДОВ ИЗВЛЕЧЕНИЯ ЧАСТО ЗАДАВАЕМЫХ ВОПРОСОВ

О чем глава?

1.1. Обзор существующих подходов к задаче извлечения ЧЗВ

Более менее подробный обзор статей по теме + вывод почему наш выбор именно такой

Работы [2] и [3] также используют LDA для вопросно-ответных систем. Работа [2], однако, предлагает проводить тематическое моделирование в рамках одного обращения, что может дать менее качественный результат, поскольку LDA показывает лучшие результаты на больших объемах данных. В [3] LDA используется для определения темы вновь поступивших вопросов, при этом для них не определяется ответ.

Работа [4] предлагает способ для нахождения лучшего ответа на вопрос среди уже предоставленных на примере размеченных данных со Stack Overflow. В текущей работе не используются размеченные данные, что позволяет получить более универсальное решение. В статье [5] представлен другой подход поиска ответов, связанный с использованием поисковой системы. Сначала коммментарии разделяются на 6 классов: вопрос, уточнение, ответ, отзыв на ответ, мусор. Затем используется специально настроенная поисковая система для поиска только по ответам. Основное отличие от текущей работы заключается в способе определения релевантных ответов.

1.2. Задача тематического моделировлания

Что такое ТМ и почему это основной этап

1.3. Обзор методов построения тематической модели

1.3.1. метод 1

1.3.2. метод 2

1.3.3. метод 3

Почему именно ЛДА

1.4. Вывод

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

2. ПОСТАНОВКА ЗАДАЧИ

В данной работе для анализа использовались обращения пользователей в техническую поддержку системы отслеживания ошибок YouTrack[1]. Для взаимодействия с пользователями команда YouTrack использует Zendesk[2] — систему учета и обработки пользовательских обращений.

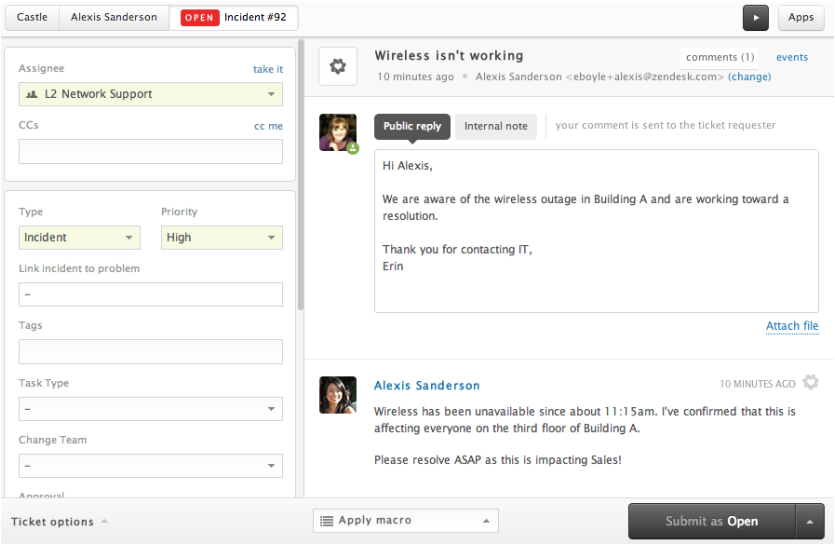

Рисунок 2.1. Пример обращения в системе Zendesk

В Zendesk обращения могут поступать из различных каналов: электронная почта, социальные сети, форма для прямой отправки обращений и так далее. Обращения состоят из комментариев и, в общем случае, представляют собой диалог между клиентом и сотрудником технической поддержки. На рисунке 2.1 приведен пример обращения в системе Zendesk. Поскольку Zendesk агрегирует все поступающие обращения, то мы не можем делать предположений о их разбиении по темам. То есть заранее неизвестно, какие из обращений относят-

[1] http://jetbrains.ru/products/youtrack/
[2] https://www.zendesk.com

ся, например, к проблемам администрирования YouTrack, а какие — связаны с пользовательским интерфейсом.

Было проанализировано 6500 обращений за период с декабря 2015 года по сентябрь 2016 года. Стоит отметить, что на момент написания статьи работа еще не была завершена, однако промежуточные результаты показывают, что подход может быть применен для генерации ЧЗВ. Из предложенных эксперту ВОП 50% было признано корректными по сравнению с 37%, полученными в работе [6].

2.1. Формулирвоание требований

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

2.2. Решаемые задачи

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

2.3. Вывод

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

3. РАЗРАБОТКА

3.1. Обзор этапов подхода

Этап 1 — подготовка обращений к отображению в ЧЗВ. Сырые данные содержат много шума: HTML разметка, заголовки электронной почты («01 июля 2001 г., 10:10 пользователь ... написал:»), приветствия, благодарности и так далее. Этот шум заметно влияет на алгоритм. Для повышения качества результатов применяется ряд эвристик предобработки, которые значительно доработаны в сравнение со статьей [6]. При этом также фильтруются обращения, которые заведомо не могут содержать вопроса или ответа.

Этап 2 — определение кластеров связанных обращений (в дальнейшем — тем) используется скрытое размещение Дирихле. LDA описывает каждую тему с помощью мешка слов (наиболее характерных терминов) и для каждого обращения определяет вероятностное распределение по темам. Затем каждому обращению сопоставляется тема с наибольшей вероятностью. После чего темы проходят через фильтр, с целью удаления незначимых с точки зрения ЧЗВ тем.

Этап 3 — формирование ВОП. Для каждого комментария в рамках обращения считается метрика близости между текстом комментария и соответствующей темой. На основе этой метрики определяются хорошо сформулированные вопросы и релевантные ответы на них.

Стоит отметить, что хотя тематическое моделирование и позволяет определить схожие по терминологии вопросы, которые фактически являются часто задаваемыми, алгоритм не ограничивается только этим и позволяет находить редкие ВОП, если они хорошо сформулированы и имеют корректный ответ. При этом большее внимание в работе уделялось поиску качественных ВОП (точность), чем поиску всех возможных ВОП (полнота). Основная мотивация такого решения заключается в желании сократить до минимума ручную часть алгоритма — валидацию и редактирование ВОП.

3.2. Предобработка данных

Исходные данные представляют собой 6500 обращений, собранных из различных каналов поступления обращений с помощью систе-

мы автоматизации запросов клиентов Zendesk[1]. Каждое обращение содержат ряд метаданных. В то время как использование метаданных ограничивает область применения алгоритма, это позволяет повысить его качество. В данной работе использовалась метаинформация, широко распространненная для данных такого рода: статус обращения и авторство комментария.

Мы предполагаем, что ответ на вопрос содержится в одном комментарии и не пытаемся объединить несколько комментариев для создания ответа. Для анализа использовались только обращения на английском языке.

Эвристики предобработки делятся на 2 категории: эвристики отображения и эвристики тематического моделирования. Первые предназначены для приведения обращений к виду, максимально близкому к виду ЧЗВ, вторые — применяются поверх первых и создают отдельное представление, используемое в LDA.

Из входных данных были отфильтрованы обращения только со статусом "закрыто"и "выполнено". Данные статусы говорят о том, что обращения имеют окончательный набор комментариев, в то время как другие обращения еще могут находиться в активном обсуждении.

3.2.1. Эвристики отображения

Эвристика 1 (специфичные регулярные выражения): данная эвристика направлена на удаление фрагментов, зависящих от предметной области или используемого программного обеспечения. Например: информация, добавляемая системой управления обращениями; шаблоны оформления обращений через веб-форму, содержащие дополнительные поля (имя, e-mail, компания); и так далее.

Эвристика 2 (удаление цитат электронной почты): 33% обращений созданы через электронную почту. Комментарии в таких обращениях часто цитируют предыдущее сообщение. Для удаления цитат использовалась самостоятельно разработанная библиотека email-parser[2].

Эвристика 3 (удаление общих суффиксов): большинство пользователей, как правило, имеют подпись, которая добавляется в конец каждого отправленного ими сообщения. Для удаления таких подписей предлагается следующее:

[1] www.zendesk.com

[2] https://github.com/JetBrains/email-parser

- Для всех комментариев в исходных данных попарно посчитать общий суффикс;

- У каждого комментария удалить суффикс максимальной длины;

Суффикс определяется построчно, что позволяет избежать частичного удаления абзацев с полезной информацией.

Эвристика 4 (короткие абзацы): многие сообщения начинаются со слов приветсвия и заканчиваются словами благодарности. Как правило, эти фрагменты выделены в отдельные абзацы (отделены символом новой строки) и значительно короче основной части сообщения (20-25 символов против 300-500). Данная эвристика удаляет (при наличии) один короткий начальный абзац и все короткие конечные абзацы. Абзац является коротким, если он состоит из 3 или меньше слов. Это число было определено эмпирически. Дополнительно этот шаг позволяет удалить фрагменты подписи, оставшиеся после эвристики 3.

Эвристика 5 (частые предложения): данная эвристика была взята из статьи [6] и говорит о том, что предложения, встречающиеся на всем наборе обращений более 15 раз, не содержат информации, специфичной для конкретного вопроса. Стоит отметить, что учитывание предложений, состоящих из одного слова, или игнорирование регистра текста приводит к частичному удалению предложений и, как следствие, ухудшению внешнего вида ВОП.

Как результат применения описанных выше эвристик текст комментариев часто может начинаться с нижнего регистра (ввиду удаления приветствий) и содержать лишние пустые строки, что снижает читаемость. Данные недостатки следует исправить, так как именно в таком виде ВОП будут показываться экспертам.

3.2.2. Эвристики тематического моделирования

Эвристики из данной группы применяются с целью повышения качества LDA. Поскольку при этом теряется часть информации, необходимой для отображения ЧЗВ, две версии каждого комментария должно быть сохранено, как показано в таблице 3.1.

Эвристика 6 (регулярные выражения): пользовательские данные ухудшают качество LDA. Например, тема, включающая в себя имя

некоторого пользователя, будет содержать обращения, в которых часто встречается это имя, несмотря на то, что сами обращения могут относиться к разным подсистемам. На этом этапе предлагается удалять: унифицированные идентификаторы ресурса (URI) и пути, адреса электронной почты и названия сайтов (www.mysite.com).

Эвристика 7 (удаление длинных абзацев): абзацы естественной речи для ИТ-дискуссий редко превышают 800 символов, в то время как длина машинно сгенерированного текста (логи, трассировки, код) часто больше этого значения.

Эвристика 8 (абзацы с пунктуацией): было установлено, что абзацы длиной больше 200 символов и содержащие более 6% символов пунктуации также являются машинно сгенерированными. Ограничение на минимальную длину абзаца позволяет избежать ложных срабатываний. В качестве символов пунктуации использовались следующие символы:

’‘-=|+,;:(){}[]<>%$@&_*

Эвристика 9 (удаление стоп-слов): удаляются наиболее частые слова английского языка, которые не помогают в определении темы в виду своего общего назначения. К ним стоит добавить слова, часто используемые в анализируемой области (’java’ или ’class’ для обсуждения разработки на Java).

Таблица 3.1. Эффект применения эвристик

Версия для ЧЗВ	Версия для LDA
I want to configure youtrack over SSL but not able to find any solution or article on the subject	configure SSL solution article subject

3.2.3. Фильтрация обращений

После применения эвристик некоторые из комментариев могут оказаться пустыми, в то время как другие могут не содержать ответа от технического специалиста. Из таких обращений не удастся извлечь ВОП. Воспользуемся метаинформацией об авторстве и отфильтруем обращения, имеющие не пустой первый комментарий (в версии для

LDA) и не менее одного не пустого комментария от сотрудника технической поддержки.

Обращения, содержащие длинную нить обсуждения (более 6 комментариев), вероятно, имеют одну из следующих проблем: вопрос плохо сформулирован, вопрос слишком специфичен, ответ недостаточно полон и содержится в нескольких комментариях. Такие обращения необходимо удалить.

Спицифичные для предметной области обращения, например: обращение закрытое по причине слияния с другим обращением, и так далее — также необходимо удалить.

3.3. Тематическое моделирование

Тематическое моделирование [7] позволяет (а) сгруппировать схожие обращения по темам (например, одна тема может касаться почтовой интеграции, а другая — вопросов о продлении подписки для пользователей), (б) охарактеризовать каждую тему списком терминов — мешком слов.

3.3.1. Скрытое размещение Дирихле

В работе использовался метод скрытого размещения Дирихле и его реализация на Java [8]. LDA работает с любыми текстовыми документами, поэтому далее будет использоваться термин 'документ' для описания обращения, как совокупности его комментариев.

LDA — это вероятностная тематическая модель, не требующая размеченных данных для обучения, однако требуещая указания количества моделируемых тем. LDA описывает каждую тему t, как вероятностное распределение по всем словам из входных данных (ϕ_t). Каждый документ d описывается вероятностным распределением по темам (θ_d). Цель LDA - максимизировать функцию (1) путем оптимизации ϕ и θ:

$$P(\theta, \phi) = \prod_{t=1}^{T} P(\phi_t) \prod_{d=1}^{D} P(\theta_d) \prod_{w=1}^{W_d} P(Z_{d,w}|\theta_d) P(N_{t,w}|\phi_t) \qquad (1)$$

где T — количество тем, D — количество документов, W_d — количество различных слов в документе d, $Z_{d,w}$ — определяет принадлеж-

ность слова w к документу d и $N_{t,w}$ — принадлежность слова w к теме t.

Распределения ϕ и θ в свою очередь зависят от гиперпараметров α и β соответственно. Реализация LDA в [8] поддерживает автоматическую оптимизацию этих параметров с использованием сэмплирования по Гиббсу [9]. Программисту остается лишь указать количество тем.

Определение количества тем — нетривиальная задача. Используемая метрика - перплексия [1], показывает сходство между терминами документов и их темой (меньше - лучше), но не отражает семантическую связь, поэтому так важен этап экспертной оценки. Тем не менее, перплексия позволяет определить минимальное число тем, при котором обращения начинают разделяться на четко выраженные подтемы. Критерием является значительное замедление скорости падения перплексии с ростом числа тем. Раздел **??-??** показывает как можно избавиться от расфокусированных тем, поэтому точное определение количества тем не требуется. Основываясь на перплексии, для построения тематической модели было выбрано количество тем, равное 250.

В результате тематического моделирования для каждого документа определяется вероятностное распределение по темам $\theta_{d,t}$. Мы сопоставляем каждому обращению одну тему — тему с максимальной вероятностью. Однако, если для обращения d вероятность каждой темы $\theta_{d,t} < 0.25$, то такое обращение не имеет четко выраженной темы и для него не будет определяться ВОП.

3.4. Формирование пар вопрос-ответ

Процесс получения ЧЗВ из смоделированных тем состоит из трех шагов: фильтрация неинформативных тем, определение пар вопрос-ответ, удаление расфокусированных тем.

3.4.1. Дополнительная фильтрация

Данный этап не обязателен и предполагает некоторые априорные знания о данных, а также то, что алгоритм уже запускался ранее. Используемые в работе данные (обращения в техническую поддержку) могут содержать большое количество типичных, повторяющихся обращений с шаблонными ответами (просьбы о сбросе пароля; вопросы о недоступности сервиса и так далее). Такие обращения являются частыми, но не несут полезной информации для ЧЗВ.

После построения тематической модели можно найти мешки слов для таких тем. Следует оставить 10 наиболее значимых слов для каждой из них. При последующих запусках LDA удаляются темы, для которых выполняется условие: хотя бы половина из топ-10 слов темы совпадают с одной из фильтруемых тем. Темы проверяются на частичное совпадение, поскольку LDA недетерминирован и мешки слов могут незначительно отличаться от запуска к запуску.

Данный фильтр позволяет избавиться от шаблонных ВОП, однако может негативно влиять на метрики качества (см. раздел **??**), поскольку удаляемые таким образом темы четко выражены и содержат большое количество обращений.

3.4.2. Определение вопросов и ответов

Для определения вопросов и ответов для каждого комментария вычисляется метрика близости с соответствующей темой. Для этого использовалось косинусное расстояние [10]:

$$\cos(e,t) = \frac{\sum_{i=1}^{n} t_i e_i}{\sum_{i=1}^{n}(t_i)^2 \sum_{i=1}^{n}(e_i)^2} \tag{2}$$

где t - вектор, соответствущий мешку слов темы, e - вектор, соответствующий словам комментария в представлении для LDA. Чем больше значение косинуса, тем сильнее комментарий связан с темой.

Комментарий выбирается в качестве *вопроса* при выполнении трех условий: (а) это первый комментарий в обращении; (б) косинус комментария и темы ($\cos(Q,T)$) выше порога 0.15; (в) длина комментария не превышает 1000 символов (более длинный текст комментария говорит о слишком специфичном для конкретного пользователя вопросе).

Условия выбора комментария в качестве *ответа*: (а) это не первый комментарий в обращении; (б) не является комменатрием инициатора обращения; (в) косинусное расстояние с темой ($\cos(A,T)$) выше порога 0.15 и максимально среди других кандитатов на ответ.

Обращения в службу поддержки, типично представляют собой диалог с итеративным уточнением деталей, предоставлением дополнительной информации и попытками дать окончательныый ответ. В качестве вопроса выбирается только инициирующий комментарий, поскольку все последующие комментарии пользователя не будут содер-

жать полной информации о проблеме. Ответом считается комментарий сотрудника технической поддержки, наиболее совпадающий с темой по используемым терминам. Таким образом, обеспечивается сходство терминологии между вопросом и ответом для найденных ВОП.

3.4.3. Удаление расфокусированных тем

Ввиду отсутствия возможности точно определить моделируемое число тем (см. **??-??**) возможны случаи, когда реальное количество тем будет меньше или больше смоделированного. В первом случае полученные после LDA темы будут слишком общими, что приведет к большим различиям между терминологий темы и приналежащими ей обращениями и, как следствие, пониженному количеству и качеству найденных ВОП. Во втором — создадутся фантомные темы, терминология которых будет плохо совпадать с реальными данными. Удалим такие расфокусированные темы за счет введения минимальной доли ВОП (3) со значением 0,1.

$$\frac{|QAPairs|}{|tickets|} > threshold \tag{3}$$

Образованные ВОП упорядочиваются (в рамках каждой темы или глобально) с использованием гармонического среднего между $\cos(Q,T)$ и $\cos(A,T)$. Гармоническое среднее (4) для получения высокого значения требует, чтобы все составляющие были высоки, таким образом, гарантируется, что и вопрос, и ответ имеют высокое качество.

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{4}$$

После этого ВОП передаются эксперту для валидации и редактирования перед публикацией.

4. РЕАЛИЗАЦИЯ

4.1. XML

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

You can use all kinds of abbreviations that don't mean anything, but add a false sense of importance and significance to your work. Some of these abbreviations are:

- eXtensible Markup Language (XML)
- JavaScript Object Notation (JSON)
- Yet Another Markup Language (YAML)

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Таблица 4.1. Решетка замечательности аббревиатур

$$\text{XML} < \text{JSON} < \text{YAML}$$

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer

adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Листинг 4.1. Пример описания аспектов в AspectJ

```
aspect A {
  pointcut fooPC(): execution(void Test.foo());
  pointcut printPC(): call(void System.out.println(String));

  before(): cflow(fooPC()) && printPC() {
    System.out.println("Hello, world!");
  }
}
```

4.2. JSON

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent

lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

5. ОЦЕНКА КАЧЕСТВА

Для оценки качества алгоритма использовались следующие метрики: перплексия, количество ВОП, косинус между вопросом и темой, косинус между ответом и темой. Таблица 5.1 показывает влияние этапов предобработки и параметров алгоритма на значения метрик.

Видно, что применение эвристик предобработки и фильтров (за исключением фильтра тем) положительно влияет на косинус вопросов и ответов. Изменение параметров в большую сторону позволяет находить меньшее количество более качественных, с точки зрения косинусного расстояние, ВОП.

Экспертная оценка проводилась только для оптимального набора параметров. В качестве эксперта выступил разработчик YouTrack, которому было предложено 20 ВОП с наибольшим значением гармонического среднего. Эксперту необходимо было оценить, какие из предложенных ВОП подходят для публикации в ЧЗВ. Результаты представлены в таблице 5.2. Доля подходящих для публикации ВОП составила 50%, для [6] данный показатель составил 37%.

Да момент написания статьи работа еще не была завершена, планируется дополнительная оптимизация с целью повышения доли подходящих для публикации ВОП и получение дополнительных экспертных оценок.

Таблица 5.1. Влияние эвристик и параметров на медианные значения метрик

Исследуемый параметр	**Старое значение**	**Новое значение**	**Перплексия**	**Количество ВОП**	$\cos(Q,T)$	$\cos(A,T)$
Оптимальные параметры	-	-	1864	357	0.396	0.413
Эвристики отображения(??-??)	вкл	выкл	1971	361	0.371	0.395
Эвристики тематич. моделир.(??-??)	вкл	выкл	2016	384	0.369	0.391
Фильтр обращений(??-??)	вкл	выкл	1924	403	0.370	0.388
Фильтр тем(??-??)	вкл	выкл	1791	472	0.393	0.416
Порог выбора темы LDA(??-??)	0.25	0.0	1853	376	0.366	0.404
Порог выбора темы LDA(??-??)	0.25	0.4	1871	302	0.399	0.421
Минимальное значение косинуса(??-??)	0.15	0.0	1860	394	0.337	0.359
Минимальное значение косинуса(??-??)	0.15	0.3	1864	288	0.387	0.419
Минимальная доля ВОП(??-??)	0.1	0.0	1869	376	0.384	0.407
Минимальная доля ВОП(??-??)	0.1	0.2	1850	324	0.391	0.417

Таблица 5.2. Экспертная оценка

Категория ВОП	Количество	Доля, %
Общее количество	20	100
Подходит для публикации без редактирования	6	30
Подходит для публикации с редактированием вопроса или ответа	4	20
Не подходит для публикации. Некорректный вопрос или ответ	10	50

ЗАКЛЮЧЕНИЕ

На момент написания статьи работа еще на была завершена. Однако промежуточные резульаты показывают улучшение относительно оригинальной работы [6] — 50% одобренных экспертом ВОП против 37% в [6].

Дальнейшие исследования могут проводиться в следующих направлениях:

- увеличение полноты решения;
- улучшение качества детектирования машинно-сгенерированного текста;
- обработка мультиязычных данных: (а) моноязычные обращения; (б) мультиязычные обращения.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Blei David M., Ng Andrew, Jordan Michael. Latent Dirichlet allocation // Journal of Machine Learning Research. — 2003. — Vol. 3.
2. Celikyilmaz A, Hakkani-Tur D, Tur G. Lda based similarity modeling for question answering // Proceedings of the NAACL HLT 2010 Workshop on Semantic Search. — 2010.
3. Liu M, Liu Y, Yang Q. Predicting best answerers for new questions in community question answering // Proceedings of the 11th International Conference on Webage Information Management. — 2010.
4. Tian Qiongjie, Zhang Peng, Li Baoxin. Towards Predicting the Best Answers in Community-Based Question-Answering Services // Proceedings of the International AAAI Conference on Web and Social Media. — 2013.
5. Gottipati S, Lo D, Jiang J. Finding relevant answers in software forums // Proceedings of the Automated Software Engineering Conference. — 2011.
6. Henß Stefan, Monperrus Martin, Mezini Mira. Semi-automatically Extracting FAQs to Improve Accessibility of Software Development Knowledge // Proceedings of the International Conference on Software Engineering (ICSE). — 2012.
7. Коршунов Антон, Гомзин Андрей. Тематическое моделирование текстов на естественном языке // Труды Института Системного Программирования РАН. — 2012. — Т. 23.
8. McCallum Andrew Kachites. MALLET: A Machine Learning for Language Toolkit. — 2002. — http://mallet.cs.umass.edu.
9. Heinrich Gregor. Parameter estimation for text analysis // Technical report, Fraunhofer IGD. — 2005.
10. Gomaa Wael H., Fahmy Aly A. A Survey of Text Similarity Approaches // International Journal of Computer Applications. — 2013. — Vol. 68.

ПРИЛОЖЕНИЕ А
ЛИСТИНГИ

Листинг А.1. Исходный код класса Main

```
1  package executable
2
3  import executable.setup.commonOptionsMap
4  import executable.setup.outerResourcesDir
5  import executable.setup.performAction
6  import executable.setup.resourcesDir
7  import org.apache.commons.cli.Option
8  import org.apache.commons.cli.Options
9  import org.jetbrains.zkb.db.DBReader
10 import org.jetbrains.zkb.lda.lda
11 import java.io.File
12
13 fun main(args: Array<String>) =
14         performAction(
15                 defineOptions(),
16                 args,
17                 listOf(
18                         setOf("f", "dicts", "mp", "nt") to { cmd ->
19                             processData(
20                                     cmd.getOptionValue("f"),
21                                     cmd.getOptionValue("dicts"),
22                                     cmd.getOptionValue("mp"),
23                                     cmd.getOptionValue("nt").toInt(),
24                                     inResourceDir = cmd.hasOption("ird"
                                        ),
25                                     verbose = true
26                             )
27                         }
28                 )
29         )
30
31
32 private fun defineOptions(): Options {
33     val options = Options()
34     options.addOption(commonOptionsMap["f"])
35     options.addOption(commonOptionsMap["dicts"])
36     options.addOption(commonOptionsMap["ird"])
37     options.addOption(commonOptionsMap["mp"])
38     options.addOption(specificOptionsMap["nt"])
39     return options
40 }
41
42 private val specificOptionsMap = mapOf(
43         "nt" to Option("nt", "num-topics", true, "Number of topics for
               LDA"
44         )
45 )
46
47
```

```
48  private fun processData(
49          sourceFilename: String,
50          dictionariesDir: String,
51          mongoPropFilename: String,
52          numTopics: Int,
53          inResourceDir: Boolean = false,
54          verbose: Boolean = true
55  ) {
56      val defaultPath = if (inResourceDir) resourcesDir else ""
57
58      val model = lda(
59              numTopics,
60              File("$defaultPath$sourceFilename"),
61              File("$defaultPath$dictionariesDir").listFiles().toList(),
62              iterations = 2000,
63              alpha_t = 0.1
64      )
65
66      DBReader("$defaultPath$mongoPropFilename", verbose).use { reader ->
67          model.printClusters(
68                  reader.readTicketThreads(),
69                  outDir = File(
70                          "${outerResourcesDir}out${File.separator}" +
71                                  "${System.currentTimeMillis()}_${
                                        numTopics}"
72                  ),
73                  verbose = verbose
74          )
75      }
76  }
```