ПРИМЕНЕНИЕ МЕТОДОВ ВЕРОЯТНОСТНОГО ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ ДЛЯ АНАЛИЗА ДЕФЕКТОВ В ПРОГРАММНОМ ОБЕСПЕЧЕНИИ

Романов А. А., Иванин Н. С.
Факультет компьютерных систем и сетей, Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: antekromanov@gmail.com, nikivnik@gmail.com

Тематическое моделирование – способ статистического анализа текстов, предназначенный для выявления скрытых тем в коллекции документов. В данной работе рассматривается применение методов вероятностного тематического моделирования для обработки текстовых данных в системах отслеживания ошибок программного обеспечения. В качестве используемого алгоритма вероятностного тематического моделирования было отдано предпочтение методу латентного размещения Дирихле.

Введение

Одной из основных задач обработки естественного языка является автоматическое извлечение тем, которые содержатся в коллекциях текстовых документов. Примером таких документов являются описания дефектов в системах отслеживания ошибок программного обеспечения. Дефектами в системах данного типа называются не только ошибки и неполадки найденные в программном обеспечении, но также пожелания и вопросы пользователей. Знание тематик дефектов, характерных для конкретного программного обеспечения, важно для более глубокого понимания проблем проекта, учета пожеланий пользователей и повышения общего качества разработки и поддержки программного обеспечения. Однако проанализировать все дефекты проекта вручную трудно а иногда и не возможно. Выявление тем дефектов можно автоматизировать, используя методы вероятностного тематического моделирования. Наиболее популярным и зарекомендовавшим себя алгоритмом является метод латентного размещения Дирихле (latent Dirichlet allocation – LDA). В настоящей работе рассматривается задача построения тематической модели для коллекции дефектов с использованием алгоритма LDA.

I. Сбор и предварительная обработка данных

Для анализа был собран корпус, состоящий из более 29 тыс. описаний дефектов проекта Apache Spark. Выбор был сделан на основе того, что система отслеживания ошибок данного проекта находится в открытом доступе и предоставляет REST API для извлечения данных. Качество исходных текстов существенно влияет на результаты тематического моделирования, поэтому описания дефектов были предобработаны, а именно:

- все символы приведены к нижнему регистру;
- удалены знаки препинания и табуляции;
- удалены стоп-слова английского языка;
- удалены тэги форматирования;
- удалены адреса электронной почты и гиперссылки;
- произведена лемматизация слов.

Современные алгоритмы машинного обучения работают с документами, которые представлены в виде векторов признаков. В качестве векторной модели используется «мешок слов». Текст в данной модели рассматривается как неупорядоченное множество слов. Каждому слову сопоставляется некий вес, отражающий его значимость. Вектор же формируется при упорядочивании всех уникальных слов в пространстве. Размерность вектора определяется числом уникальных слов во всей коллекции и является постоянной для всех документов[1]. Для оценки значимости слов была использована схема $tf - idf$ [2]:

$$tf - idf(w, d, D) = tf(w, d) \times \log \frac{|D|}{|d_w|},$$

где $tf(w, d)$ – отношение числа вхождений слова w к общему числу слов документа d,

$|D|$ – число документов в коллекции D,

$|d_w|$ – число документов из коллекции D, в которых встречается слово w.

II. Построение тематической модели

Метод латентного размещения Дирихле предложен Дэвидом Блеем в 2003 году. Он основан на вероятностной модели [3]:

$$p(d, w) = \sum_{t \in T} p(d)p(w|t)p(t|d),$$

при дополнительных предположениях:

векторы документов $\theta_d = \big(p(t|d) : t \in T\big)$ порождаются одним и тем же вероятностным распределением из параметрического семейства распределений Дирихле $\mathrm{Dir}(\theta, \alpha)$, $\alpha \in \{R^{|T|}$ на нормированных $|T|$–мерных векторах;

векторы тем $\phi_t = \big(p(w|t) : w \in W\big)$ порождаются одним и тем же вероятностным распределением из параметрического семейства распре-

делений Дирихле $\mathrm{Dir}(\theta, \beta)$, $\beta \in \mathbb{R}^{|W|}$ на нормированных векторах размерности $|W|$.

Инструментом реализации LDA выбрана библиотека gensim языка Python. Для построения модели коллекция документов была разделена на обучающую и контрольную выборки в отношении 4:1. Визуализация тематических моделей осуществлялась с помощью библиотеки pyLDAvis. Данная библиотека отображает тематическую модель в виде интерактивной пузырьковой диаграммы (см. рис. 1).

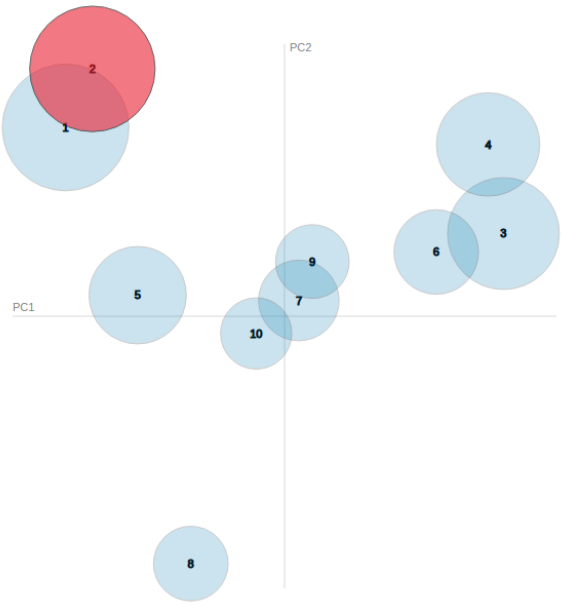

Рис. 1 – Визуализация полученной оптимальной тематической модели

Каждый пузырек на диаграмме представляет тему. Чем больше его размер, тем больше распространена данная тема. Хорошая тематическая модель должна иметь довольно большие непересекающиеся пузырьки, разбросанные по всей диаграмме. Также для каждой темы pyLDAvis показывает упорядоченный список наиболее характерных слов (см. табл. 1).

Таблица 1 – Примеры тем полученной оптимальной модели

№	Топ-5 релевантных слов темы
1	table, partition, select, hive, column
2	class, write, output, generate, implementation
3	raw, size, array, want, dataset
4	support, model, well, vector, documentation
5	spark, anonfun, run, apply, scala
6	value, key, type, function, string
7	info, executor, task, driver, run
8	file, none, set, default, build
9	code, test, follow, example, time
10	use, add, datum, create, need

III. Оценка качества и выбор оптимальной модели

Общепринятой мерой качества вероятностной тематической модели является перплексия контрольной выборки. Это мера несоответствия или «удивлённости» модели $p(w|d)$ терминам w, наблюдаемым в документах d коллекции D:

$$P(D, p) = \exp\Big(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(w|d)\Big)$$

Чем меньше эта величина, тем лучше модель p предсказывает появление терминов w в документах d коллекции D [4]. Перплексия измеряется по контрольной выборке документов, не используемых для построения модели. Это позволяет избежать занижения оценки в результате переобучения. Для поиска оптимальных параметров алгоритма LDA был использован подход, называемый grid search. Для этого алгоритм создания модели запускался с различными комбинациями значений параметров и в результате из множества построенных моделей, была выбрана модель с наименьшей перплексией. В результате подбора параметров было выявлена модель с оптимальным числом тем $T = 10$, с перплексией $P = -8.0917$.

Заключение

В рамках данной работы решена задача автоматического извлечения тем из текстовых описаний дефектов программного обеспечения. С помощью REST API осуществлен сбор дефектов из системы отслеживания ошибок, полученные данные преобразованы для построения тематической модели с использованием инструментов обработки естественного языка. Были подобраны оптимальные параметры для тематической модели, использованы средства визуализации.

Перспективой для улучшения тематической модели является использование не только описания дефектов, но и дополнительной информации, например, приложений и комментариев к дефектам. Актуальными направлениями для дальнейшего исследования могут быть определение изменения тем дефектов со временем и построение иерархического каталога тем.

Полученная тематическая модель может быть использована в качестве плагина, расширяющего функциональность системы отслеживания ошибок.

1. Manning C. D. Introduction to Information Retrieval / C. D. Manning, P. Raghavan, H. Schutze. – Cambridge: Cambridge University Press, 2008, – 496 p.
2. Lan M. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization / M. Lan [et al.]. // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2009. – Vol. 31, P. 721–735.
3. Blei D. M. Latent Dirichlet allocation / D. M. Blei, A. Y. Ng, M. I. Jordan // Journal of Machine Learning Research – 2003. – Vol. 3, P. 993–1022.
4. Вероятностное тематическое моделирование [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf. – Дата доступа: 08.10.2019.