Кроссязычный поиск дубликатов с использованием тематического моделирования

Е. В. Тищенко[1]***, К. В. Воронцов***[2]***, П. С. Потапова***[3]

В данной статье рассматривается задача кроссязычного поиска текстового плагиата. Целью работы является получение модели, выделяющей информацию о распределении слов в тексте независимо от их языковой принадлежности, при этом ограниченной по размеру и времени обучения для ее практического использования.

Ключевые слова: *машинное обучение, тематическое моделирование, мультиязыковой поиск дубликатов*

1 Введение

Задача поиска и распознавания плагиата является особенно важной из-за невозможности проверить все документы общей тематики на плагиат в силу размеров глобальной сети. Задача становится особенно трудной, так как возникает множество документов, являющихся переводом исходных работ. Согласно исследованиям Donald L. McCabe, [1] 36% студентов американских университетов перефразировали или копировали информацию из всемирной паутины без ссылки на источник.

Методы векторизации документов для поиска плагиата [2,3] преимущественно ограничиваются одним языком. В таком случае возникает проблема создания единообразной системы получения векторных представлений мультиязыковой коллекции документов. Для возможности применения модели за пределами научных экспериментов ставятся технические ограничения ресурсами сервера, а именно на размер модели, временную сложность обучения.

Объектом исследования являются мультиязыковые тематические модели, алгоритмы поиска документов в текстовой коллекции по словам и документам, способы векторного представления слов и документов запроса и коллекции, применяемые для поиска дубликатов независимо от языковой принадлежности.

Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова или термины образуют каждую тему. Вероятностная тематическая модель описывает каждую тему дискретным распределением вероятностей слов, а каждый документ — дискретным распределением вероятностей тем. Тематическая модель преобразует любой текст в вектор вероятностей тем.

Для решения задачи планируется использовать мультимодальную тематическую модель. Такая тематическая модель описывает документы, содержащие метаданные наряду с основным текстом. Метаданные улучшают точность определения тематики документа. В качестве модальностей используются 100 языков, а также научные рубрики. Использование языков в качестве модальностей позволяет получить векторное представление текста, независимое от оригинального языка текста, что позволяет решать проблему поиска плагиата без ограничения на язык статьи. Во время предобработки текста используется BPE токенизация — итеративная замена наиболее встречаемой пары символов на символ, который не встречается в слове. Это существенно уменьшает объем изначального словаря для практического применения модели.

Целью эксперимента является построение модели, исследование влияния регуляризации и предобработки текстовых данных на качество поиска, подбор разнообразных функций для сравнения тематических расстояних векторов а также поиск эвристик для улучшения точности предсказаний модели. В качестве обучающих данных используются статьи

с сайта Wikipedia, а также выборка научных статей из научной электронной библиотеки eLIBRARY.ru.

2 Постановка задачи

Пусть D – некоторая коллекция документов, T – множество тем документов. Кандидатом на дубликат для документа $d \in D$ обозначим такой элемент коллекции $f(d)$, что

$$f(d) = \arg\min_{d' \in D \setminus \{d\}} distance(m(d), m(d'))$$

В качестве функции *distance* может быть использована произвольная метрика векторного пространства. Функция $m(d)$– модель, осуществляющая преобразование документа в векторное пространство. Качество модели поиска дубликатов измеряется на тестовой выборке при помощи двух метрик сопоставления документов:

1. Средняя частота, с которой документ-запрос попадает в топ 10%. Для документа в тестовой выборке уже известно, какие документы являются переводом исходного. Рассматривается доля тех документов, для которых среди 10% отранжированных кандидатов на дубликат встречаются их переводы.

2. Средний процент документов в топ 10% документов-переводов, которые имеют такую же рубрику, что у документа-запроса. Аналогично предыдущей метрике, однако рассматриваются не все документы в коллекции, а лишь определенной рубрики.

Требуется решить задачу поиска дубликатов, при этом качество должно превосходить 0.9 по первой метрке, 0.3 по второй. Также необходимо ограничить размер модели до 100 Гб, а обучение модели должно производиться не более чем за 24 часа.

В качестве преобразования текста в векторном представлении используется тематическое моделирование. Тематическая модель [4] по коллекции документов строит вероятностное распределение $p(w|t)$ термов w – слов, словосочетаний и терминов в темах $t \in T$. Согласно гипотезе об условной независимости, а также формуле полной вероятности, распределение термов w в документе d является вероятностной смесью распределений термов в темах $\varphi_{wt} = p(w|t)$ с весами $\theta_{td} = p(t|d)$

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

Данное выражение можно перемисать в матричном виде. Матрица частот термов в документах $F \approx \Phi\Theta$. Так как число тем, как правило, намного больше, чем число документов, то требуется найти такое разложение, ранг которого не превосходит $|T|$.

3 Вычислительный эксперимент

В рамках эксперимента ставится задача поиска оптимальных параметров тематической модели, а также метода токенизации текста, при которых будут достигнуты требования. После выбора оптимальных параметров будет проведено исследование, как именно влияет отдельный параметр на качество модели.

3.1 Данные

3.2 Базовые модели

В качестве базовой модели используется тематическая модель, испльзующая в качестве модальностей 100 языков. Обучение происходит на основе данных Википедии. Используются тексты, посвященные 300 различным темам. При предобработке текста используется

метод BPE токенизации, причем изначальный объем словаря в 120 тысяч токенов ужат до 2 тысяч. Обучение модели занимает порядка 6 часов, а размер модели составляет 4.6 Гб. Тесты качества модели описаны ниже.

Средняя частота УДК	Средний процент УДК	Средняя частота ГРНТИ	Средний процент ГРНТИ
0.4332	Coming soon	Coming soon	Coming soon

3.3 Модель

3.4 Результаты

3.5 Абляционные эксперименты

Литература

[1] *Donald L. McCabe* Cheating among college and university students: A North American perspective// International Journal for Educational Integrity, 2005

[2] *Zdenek Ceska, Michal Toman, and Karel Jezek* Multilingual Plagiarism Detection// Artificial Intelligence: Methodology, Systems and Applications, 2008

[3] *Duygu Ataman, Jose G. C. de Souza, Marco Turchi, Matteo Negri* Cross-lingual Semantic Similarity Measurement Using Quality Estimation Features and Compositional Bilingual Word Embeddings//

[4] *Воронцов К. В.* Обзор вероятностных тематических моделей.//

Поступила в редакцию