

Style Change Detection

Зуева Надежда

January 2018

1 Аннотация

В данной статье рассматриваются методы обнаружения плагиата с помощью нейронных сетей. Традиционная проблема обнаружения плагиата [1] формулируется следующим образом. Если встречен подозрительный на плагиат документ, необходимо определить, написан ли он одним автором или содержит нелегитимные заимствования.

Для задачи поиска заимствований существует два основных подхода: обнаружение "внутренних" и "внешних" заимствований. Для поиска "внешних" заимствований мы можем опираться на внешнюю коллекцию документов, при поиске "внутренних" можем только анализировать стиль текста. Мы будем использовать генеративно-сопоставительные сети для обнаружения плагиата: ожидается, что этот метод будет более подходящим. Затем, стоит сказать о методе, который будет применен в этой статье и обозначить, почему выбрали именно его. Стоит придерживаться одинаковой терминологии, это значит, что термин "контрафакция" из аннотации нужно будет убрать.

2 Введение

Традиционным [3] алгоритмом решения этой задачи является метод классификации со спецификацией подхода в зависимости от того, известно ли число авторов или нет [3].

Мы будем использовать генеративно-сопоставительные нейронные сети [4] для выявления контрафакций, где генеративная модель порождает тексты в одном авторском стиле, а дискриминативная модель является бинарным классификатором. Обучение будет произведено на основе данных *PAN2018*, *PAN2017* и *PAN2016* [2]. Основной задачей данной статьи является имплементация, описание и тестирование алгоритма, который даст прирост в ROC-AUC-показателях для проверки антиплагиата. В соответствии с [3], мы можем выделить несколько задач, которые нам предстоит решить: когда нам дана внешняя коллекция документов и мы можем найти внешний плагиат или же, если у нас нет внешней коллекции, мы можем анализировать

только стиль данного текста (внутренний плагиат). Как мы видим, прослеживается четкая связь между описанными выше проблемами и поиском плагиата в статье, поэтому наш алгоритм, основанный на принципах GAN, может стать универсальным инструментом. До сегодняшнего дня предпринимались попытки решить эту задачу при помощи стандартных методов машинного обучения [6]. В статье [7] описываются первые попытки применения нейросетей к поиску плагиата.

3 Постановка задачи

Пусть нам дана коллекция текстовых документов D , здесь d — каждый отдельный текстовый документ. Будем решать задачу, используя generative adversarial networks[3] — здесь генеративная модель [3] порождает тексты в одном авторском стиле, дискриминативная модель [3] является бинарным классификатором, т.е. ответ — бинарен: заимствование либо есть, либо его нет.

Описание выборки

Для эксперимента используется выборка документов PAN-2018 [8], содержащая коллекцию документов с небольшим числом (до 20) авторов-кандидатов, которые используют заимствования в тексте без указания авторов другого (неизвестного) набора документов. Известные документы принадлежат нескольким темам, хотя и не обязательно одинаковы для всех авторов-кандидатов.

Предоставляется одинаковое количество документов для каждого автора-кандидата. Известные документы неравномерно распределены по авторам. Длина текста варьируется от 500 до 1000 слов. Документы представлены на следующих языках: английский, французский, итальянский, польский, испанский.

Критерий качества

Будем использовать те же критерии качества, что и те, которые применяются в соревновании RAN-2018 [8] и [9]. Пусть D — коллекция документов, а каждый ее элемент — d , а s — совокупность некоторых сегментов текста. Рассмотрим пары (s, d) — они будут представлять последовательность символов, которая помечена человеком как заимствование в документе d . $S = S = s_i$ — совокупность всех заимствованных частей текста. За пару (r, d) обозначим последовательность, помеченную алгоритмом как плагиат. $R = r_i$ — совокупность всех сегментов, которые нейросеть пометила как заимствованные. Обозначим S_R множество заимствованных частей текста, которые были обнаружены алгоритмом. R_s — части текста, отмеченные сетью, которые находят данную часть заимствований s . Рассмотрим меры качества из PAN-2018 [8]:

1. $Prec(S, R) = \frac{1}{|R|} \sum \frac{|(s_i \cup r_j)|}{|r_j|}$
2. $Rec(S, R) = \frac{1}{|S|} \sum \frac{|(s_i \cup r_j)|}{|s_j|}$

Precision характеризует долю верного распознавания плагиата ко всем выделенным частям документа, *Recall* характеризует долю правильного распознавания плагиата по отношению ко всем выделенным частям в тексте. Также нам потребуется $F1(S, R)$ мера, которая является отношением произведения *Precision* и *Recall* и их суммы: $F1(S, R) = \frac{Prec(S, R) \cdot Rec(S, R)}{Prec(S, R) + Rec(S, R)}$. Итоговая мера качества P определяется по формуле:

$$P(S, R) = \frac{F1(S, R)}{\log_2(1 + \frac{1}{|S|} \sum |R_{si}|)}$$

Формальная постановка задачи

Пусть у нас есть коллекция документов D и некоторые размеченные документы. Тогда мы можем организовать две выборки — *train* и *test*. *Train* — обучающая выборка с размеченными данными, на которой мы будем обучать нашу GAN[4].

Пусть сеть G это генеративная модель — сеть, генерирующая образцы, а D это дискриминативная модель, сеть, которая будет стараться отличить подлинники от плагиата. Таким образом, алгоритм должен выдать в качестве результата вектор, где будут построчно записаны все заимствования.

4 Базовый эксперимент

В качестве базового алгоритма приведем уже существующий, описанный в статье [8]. Целью этого базового эксперимента ставилась проверка того, что заимствованные сегменты текста имеют отличные от среднего вектора значения признаков.

В качестве признака выберем частоты встречаемости слов, т.е. каждому слову ставится в отношение некоторое число, характеризующее частоту встречаемости. Чем больше это число, тем чаще встречается это слово. Обозначим слово за w , тогда в соответствие ему ставится $freq_w = \ln \frac{n_w}{n_{max}}$, здесь n_w — число слов w , встреченных в сегменте текста, а n_{max} — число вхождений в тот же сегмент текста самого встречаемого слова. Для проверки данной гипотезы используем критерий — . Пусть m_j = среднее значение j -го признака рассматриваемого документа, а r_j это среднеквадратичное отклонение. Пусть $t_{ij} = \frac{x_j - m_j}{r_j}$, здесь t_{ij} — нормализованный признак j это i -го сегмента текста. За сегменты t_i возьмем предложения из документа. Для каждого предложения t_i строился вектор признаков t_i при помощи технологии *word_tov_{ec}* и затем подсчитаем отклонение от усредненного по всему тексту вектора t в $L1$ - метрике: $\sigma(t_i) = ||t_i - t||$, эксперимент проводится на данных PAN-2016[6].

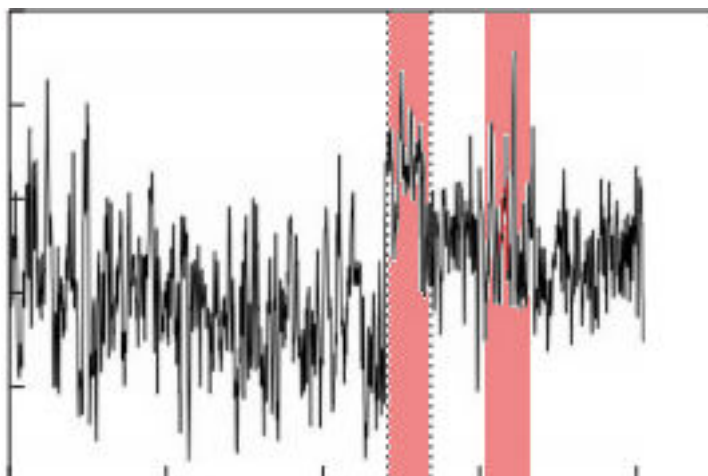


рис.1

На рисунке по оси Ox отложены сегменты текста, а по Oy – значения статистики. Красным отмечены участки, которые были помечены экспертом как заимствованные.

Можно заметить, что заимствованные части документов имеют характерные выбросы из области средних значений x . Но также мы можем наблюдать выбросы там, где их быть не должно, т.е. ошибки первого рода. Выходит, что этот признак работает недостаточно хорошо и надо искать другие, более совершенные методы.

5 Алгоритм

Введем основные понятия[2], которые будем использовать в нашем алгоритме

1. Модель. Предлагаемый алгоритм работает с частотными признаками, предоставляющими описание текста. В качестве такого признака выбран признак частоты встречаемости слов. Исходный текст подвергается предобработке: удаляются служебные символы, все буквы переводятся в нижний регистр. Также из текста удаляются стоп-слова.
2. Сегментирование текста. Текст разбивается на предложения. Затем формируется разбиение текста на сегменты t_i : если длина очередного предложения меньше минимальной длины сегмента l_{segt} , к этому предложению добавляется следующее за ним — процесс повторяется, пока длина сегмента t_i не превысит заданную минимальную длину. Минимальная длина сегмента l_{segt} является настраиваемым параметром алгоритма.
3. Построение статистики и детектирование аномалий. Для каждого сегмента t_i запускается генеративно-состязательная сеть, в которой ге-

неративная модель G генерирует тексты в одном авторском стиле, а **дискриминативная модель** D старается отличить тексты без плагиата от тех, где плагиат присутствует.

На основе отработки алгоритма для каждого t_i выдается бинарный ответ, содержится ли в этом куске текста плагиат или нет. Затем, по итогам такой обработки, для вектора T , который содержит в себе все куски текста:

$$T = \langle t_1, \dots, t_n \rangle$$

Также строится еще один вектор

$$T_{answers} = \langle a(t_1), a(t_2), \dots, a(t_n) \rangle, \text{ здесь } a(x) - \text{алгоритм поиска плагиата.}$$

Рассмотрим подробнее G и D :

- (a) G — генеративная модель. Для генерации текстов будем использовать *скрытую марковскую цепь*: имея ряд признаков (часто употребляемые слова, синтаксические конструкции и тому подобное) будем стараться угадать, написан ли текст одним автором или несколькими. Для обработки этой модели используем [10], который и будет возвращать нам сгенерированный текст.
- (b) D — дискриминативная модель, которая представляет собой бинарный классификатор [11], т.е. возможны только два ответа $-0, 1$. Будем ставить 0, если в фрагменте отсутствует плагиат и 1 — если авторов больше одного. В качестве алгоритма классификации возьмем RandomForestClassifier — он плохо переобучается и устойчив к случайным выбросам и поможет в поиске стилистических разладок.

6 Вычислительный эксперимент

Для вычислительного эксперимента в качестве обучающей выборки использовались данные соревнования **PAN-2018** (корпуса 1-7), а тестирование происходило на данных корпусов 8-10.

Затем надо подобрать оптимальные параметры и запустить с ними алгоритм.

7 Список литературы

1. Stein, B., Barrón Cedeño, L.A., Eiselt, A., Potthast, M., Rosso, P.: Overview of the 3rd international competition on plagiarism detection. In: CEUR Workshop Proceedings. CEUR Workshop Proceedings (2011)
2. <http://pan.webis.de/clef18/pan18-web/author-identification.html>
3. <https://pdfs.semanticscholar.org/1011/6d82a8438c78877a8a142be47c4ee8662138.pdf>
4. <https://arxiv.org/pdf/1701.06547.pdf>

5. Zechner, M., Muhr, M., Kern, R., Granitzer, M.: External and intrinsic plagiarism detection using vector space models. Proc. SEPLN. vol. 32 (2009)
6. Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Clustering by authorship within and across documents. CEUR Workshop Proceedings (2016)
7. <https://pdfs.semanticscholar.org/c70e/7f8fbc561520accda7eea2f9bbf254edb255.pdf>
8. <http://pan.webis.de/clef18/pan18-web/author-identification.html>
9. <http://www.mathnet.ru/links/21c7959c3887dcf64bc0f1b5913c81be/ia487.pdf>