

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-состязательные 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-состязательные 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. Неизвестные 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 . $S = \{s_i\}$ — 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 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>