PHÁT HIỆN SAO CHÉP SỬ DỤNG STOPWORD N-GRAM

*D*x là tập những tài liệu nghi ngờ cần kiểm tra và *D*s là tập các tài liệu nguồn.

Tất cả những chỗ sao chép phải được phát hiện bao gồm 1 tập con của Ds (tài liệu chứa nội dung bị sao chép) và đường biên của đoạn văn bị sao chép (phát hiện đoạn văn sao chép trong cả tài liệu sao chép và tài liệu nguồn). Đồng thời cần có 1 số điểm gán cho đoạn văn phát hiện để đánh giá mức độ sao chép.

1. **Biểu diễn lại đoạn văn thành các stopword:**

TABLE 2. The list of 50 most frequent words of BNC corpus used in this study.

1. the	11. with	21. are	31. or	41. her
2. of	12. he	22. not	32. an	42. n't
3. and	13. be	23. his	33. were	43. there
4. a	14. on	24. this	34. we	44. can
5. in	15. i	25. from	35. their	45. all
6. to	16. that	26. but	36. been	46. as
7. is	17. by	27. had	37. has	47. if
8. was	18. at	28. which	38. have	48. who
9. it	19. you	29. she	39. will	49. what
10. for	20. 's	30. they	40. would	50. said

"These savage birds are very common in Maine, where they make great havoc among the flocks of wild-ducks and Canada grouse, and will even, when driven by hunger, venture an attack on the fowls of the farm-yard."

Đoạn text sau khi loại bỏ tất cả những từ không là stopword:

are in they the of and and will by an on the of the

Chuyển thành Stopword N-grams (SWNGs), cụ thể là 8-gram:

Nếu ta có 1 tài liệu d thì ta gọi P(n, d) là tập hồ sơ dữ liệu các SWNGs của d được sắp xếp theo thứ tự lần đầu tiên xuất hiện của SWNGs trong tài liệu d ấy.

P(8, d)={[are, in, they, the, of, and, and, will],
[in, they, the, of, and, and, will, by],
[they, the, of, and, and, will, by, an],
[the, of, and, and, will, by, an, on],
[of, and, and, will, by, an, on, the],
[and, and, will, by, an, on, the, of],
[and, will, by, an, on, the, of, the]}.

This came into existence likely **from the** deviance **in the** time-period **of the** particular billet**. As the** premier **is to be** nominated **for not** more than **a** period **of** four years, **it can** infrequently happen **that an** ample wage, fixed **at the** embarkation **of that** period, **will not** endure **to be** such **to** its end.

This probably arose **from the** difference **in the** duration **of the** respective offices. **As the** President **is to be** elected **for** no more than four years, **it can** rarely happen **that an** adequate salary, fixed **at the** commencement **of that** period, **will not** continue **to be** such **to** its end.

Ví dụ về 1 đoạn text bị đạo có sự xuất hiện các stopword tương tự như trong văn bản gốc.

Khi cố gắng sao chép 1 đoạn văn và để che giấu việc này, người đạo văn thường thay thế các cụm từ bằng từ đồng nghĩa. Rất khó thay đổi cú pháp của câu hoặc viết lại phần lớn văn bản. Stopword là các từ chức năng, chúng độc lập với nội dung đoạn văn và chúng hầu như không truyền tải thông tin ngữ nghĩa nào. Một thành phần ngôn ngữ không ổn định khi chúng có thể được thay thế bằng các từ hoặc cụm từ đồng nghĩa. Tính ổn định của từ có thể coi là sự có sẵn các từ đồng nghĩa. Như vậy, stopword là những từ có tính ổn định rất cao, do đó, chúng thường được giữ lại khi ai đó muốn đạo văn. Hơn nữa, các stopword là những từ ngắn nên thường rất khó bị viết sai chính tả, nên việc phát hiện các stopword rất dễ.

2. **Trích xuất tập tài liệu nguồn:**

Bước đầu tiên là trích xuất được tập tài liệu nguồn. Công việc này sẽ là lấy tài liệu nghi ngờ đem so sánh với bất kì tài liệu nguồn nào để phát hiện bất kì điểm chung nào đáng ngờ. Chúng ta sẽ tách tài liệu ngờ và kho tài liệu nguồn thành các SWNGs, sau đó đem so sánh để tìm ra bất kì điểm chung nào giữa 2 tài liệu. nếu tài liệu nguồn nào có bất kì điểm chung nào với tài liệu ngờ thì sẽ được đưa vào tập tài liệu kiểm tra gọi là Drx.

Trong việc trích xuất n-gram từ tài liệu nguồn để đi so sánh, việc chọn n rất quan trọng. Vì chọn n nhỏ thì kết quả sẽ kém chính xác vì sẽ rơi vào trường hợp trùng lặp ngẫu nhiên như ví dụ sau:

> ***The*** *minutes* ***of the*** *committee record* ***the*** *motion* ***of*** *appreciation* ***to the*** *owners. Mr. Robert Bell* ***of the*** *old printing firm* ***of that*** *name made…*
>
> *…****the*** *Fathers* ***of the*** *Church;* ***the*** *aesthetic mysticism* ***of*** *Plotinus, reborn* ***to*** *its*

*greatest triumphs, during **the** classic period **of** German thought. Through **the** midst **of** these variously erroneous theories, **that** traverse…*

Nhưng đối với các đoạn văn bị sửa đổi cao, thì việc chọn n lớn sẽ rất khó tìm được chuỗi SWNGs giống nhau. Vì vậy, việc chọn giá trị n phù hợp rất quan trọng.

Một việc quan trọng nữa là sự trùng hợp ngẫu nhiên các SWNGs khi các SWNGs này chỉ chứa các stopword phổ biến sau: {the, of, and, a, in, to}. Ví dụ sau sẽ cho thấy rõ điều này.

***The** minutes **of the** committee record **the** motion **of** appreciation **to the** owners. Mr. Robert Bell **of the** old printing firm **of that** name made…*

*…**the** Fathers **of the** Church; **the** aesthetic mysticism **of** Plotinus, reborn **to** its greatest triumphs, during **the** classic period **of** German thought. Through **the** midst **of** these variously erroneous theories, **that** traverse…*

Sự xuất hiện của các trường hợp như trên sẽ làm kết quả của chúng ta bị sai lệch 1 cách đáng kể. Vì vậy, chúng ta cần có ràng buộc về sự xuất hiện của các stopword phổ biến này trong SWNGs.

Gọi C={the, of, and, a, in, to} là tập hợp những stopword phổ biến nhất trong tiếng anh. Gọi Dx, Ds lần lượt là tập các tài liệu nghi ngờ và tài liệu nguồn, dx∈Dx và ds∈Ds là tập các tài liệu nghi ngờ và tài liệu nguồn đem đi so sánh. P (n1, dx) và P (n1, ds) là tập hợp các SWNGs của dx và ds. Khi đó, sự trùng khớp giữa các tài liệu này được phát hiện khi đáp ứng tiêu chí sau:

$$\exists g \in P(n1,dx) \cap P(n1,ds): member(g,C) < n1-1 \wedge maxseq(g,C) < n1-2 \quad (1)$$

g ở đây là 1 SWNG.

member(g,C): số lượng stopword trong g xuất hiện trong C,

maxseq(g,C): độ dài lớn nhất chuỗi stopword liên tiếp đến khi gặp 1 stopword không xuất hiện trong C.

với điều kiện trên thì n1 càng lớn càng tốt.

Ví dụ: P(8, dx)={ [the, of, the, the, of, to, the, of],

[of, the, the, of, to, the, of, the],

[the, the, of, to, the, of, the, of],

[the, of, to, the, of, the, of, that]}.

P(8, ds)={ [the, of, the, the, of, to, the, of],

[of, the, the, of, to, the, of, the],

[the, the, of, to, the, of, the, of],

[the, of, to, the, of, the, of, that]}

Như vậy, g1=[the, of, the, the, of, to, the, of]: member(g1,C)=8>8-1 (vi phạm)

Tương tự,…

3. **Tìm ra đường biên của đoạn:**
   Sau khi có được tập tài liệu nghi ngờ và tập tài liệu nguồn, bước tiếp theo, chúng ta cần tìm được ranh giới các đoạn văn bị đạo trong cả tài liệu ngờ và tài liệu nguồn. Nghĩa là cần tìm ra các đoạn nghi ngờ.
   Trong trường hợp copy toàn văn, việc này khá dễ dàng vì cùng 1 chuỗi SWNGs thì các stopword sẽ xuất hiện cùng 1 thứ tự, tạo thành 1 đường chéo như trong biểu đồ sau.

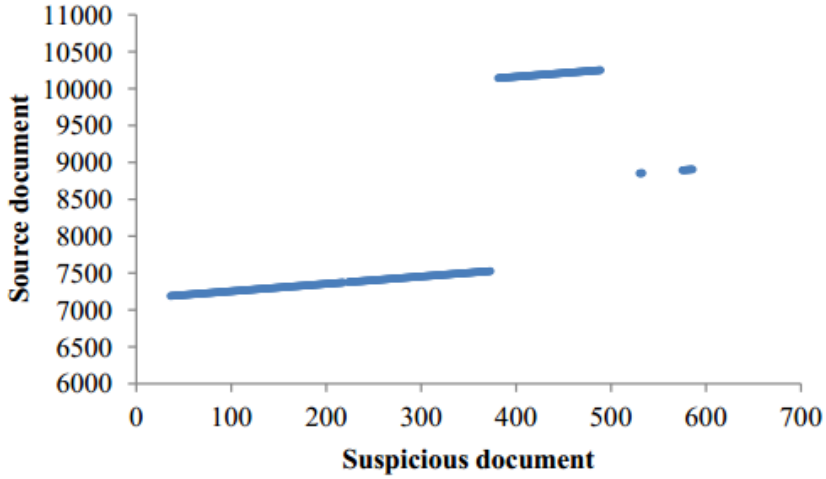

FIG 8. Scatter plot of the matched *n*-grams in verbatim plagiarism cases where the plagiarized passages are next to each other.

Nhưng trong trường hợp đoạn văn bản được sửa đổi lại thì khi ta chọn những SWNGs dài thì sẽ xuất hiện những khoảng trống giữa 2 SWNG. Nghĩa là trong 2 chuỗi SWNG ấy sẽ có những chỗ mà stopword không giống nhau. Và những khoảng trống này

xuất hiện càng nhiều nếu ta chọn giá trị n càng cao.

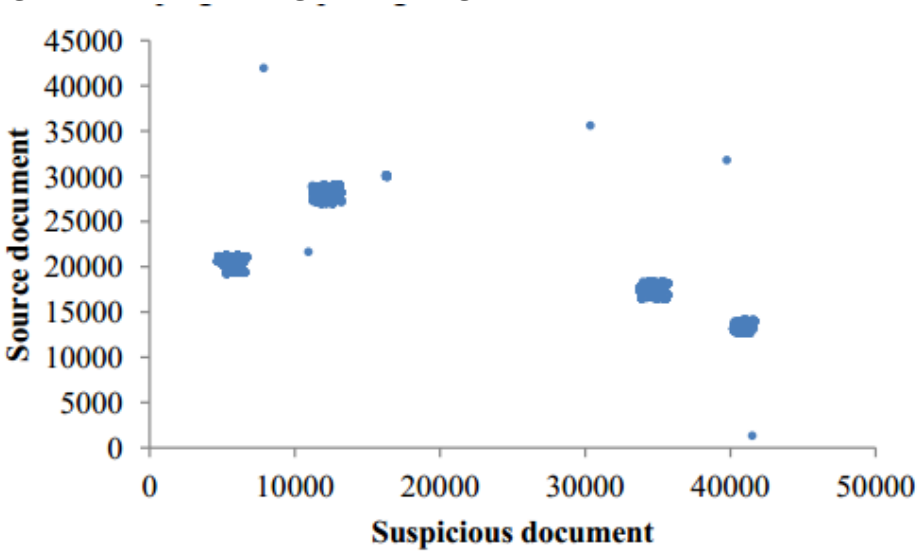

FIG 9. Scatter plot of the matched *n*-grams in cases where the plagiarized passage is significantly modified.

Chính vì vậy, điều kiện (1) sẽ không còn phù hợp nữa. Chúng ta cần tạo ra các SWNGs ngắn hơn để có thể so sánh chi tiết hơn giữa tài liệu ngờ, nhưng chúng ta cũng cần tránh nhiễu do sự trùng hợp ngẫu nhiên các stopword phổ biến. Chính vì vậy, chúng ta vẫn phải sử dụng điều kiện (1) nhưng "lỏng" hơn, nghĩa là:

$$g \in P(n2, dx) \cap P(n2, ds) \wedge member(g, C) < n2 \qquad (2)$$

[Original]

"**Since its first discovery by non-indigenous people in the midnineteenth century, Yosemite Valley has held a special, even religious, hold on the American conscience because its beauty makes it an incomparable valley and one of the grandest of all special temples of nature.**

While Yosemite holds a special grip on the western mind, perceptions about the Valley have evolved over time due to changing politics, migration patterns and environmental concerns as man has become more attuned to his relationship and impact on nature."

Dãy các stopword của đoạn văn gốc:

its by in the has a on the its it an and of the of all of a on the the have to and as has to and on

1 [its by in the]

2 [by in the has]

3 [in the has a]

4 [the has a on]

…

26 [has to and on]

[Plagiarism]

"**Since its first found by non-indigenous people in the 18th century, Yosemite Valley has** contain an enigmatic**, even religious, hold on the American conscience since its landscape makes it an incomparable valley and one of the grandest of all special temples of nature**. Native Americans have lived the Yosemile region for as 8000 years. The first people that we have record of was a band of Native Americans that called the Valley "Ah-wah-nee" and themselves the "Ahwahnechee". **While Yosemite holds a special grip on the western mind, perceptions about the Valley have evolved over time due to changing politics, migration patterns and environmental concerns as man has become more attuned to his relationship and impact on nature**."

its by in the has an on the its it a and of the of all of have the for as the that have of was a of that the and the a on the the have to and as has to his and on

1 [its by in the]

2 [by in the has]

3 [in the has an]

4 [the has a on]

…

14 [the of all of]

…

33[a on the the]

…

41 [to his and on]

M(dx, ds){(1,1), (2,2), (7,7), (8,8), (9,9) (10,10) (11,11) (12, 12) (13, 13) (14, 14) (33, 18) (34, 19) (35, 20) … (41, 26)}

for mi ∈ M(dx,ds) :

if abs(mi − mi+1) > θg : "Boundary Detected"

else: "Boundary not detected"

Tính điểm tương tự:

$$Sim(t_x, t_s) = \frac{|P_c(n_c,t_x) \cap P_c(n_c,t_s)|}{\max(|P_c(n_c,t_x)|,|P_c(n_c,t_s)|)}$$

Khi Sim>ngưỡng θc do mình đặt ra thì kết luận đạo.

Đặt ngưỡng θc =0.5

Với ví dụ ở trên, chúng ta tìm được 2 đoạn nghi ngờ,

Sim(tx1, ts1)=10/14=0.74>0.5

Sim (tx2, ts2)=8/8=1>0.5

Một vấn đề khó khăn là các trích dẫn toàn văn rất ngắn. Lý tưởng là không nên kết luận những đoạn trích dẫn này là đạo văn. Chính vì vậy, chúng ta cần có 1 ngưỡng θL qui định độ dài của đoạn văn bị đạo. Nếu đoạn văn có sự tương đồng và độ dài > θL thì mới kết luận đạo văn.