Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection

Trích xuất các tài liệu nguồn:

Từ tài liệu ngờ và mỗi tài liệu trong kho dữ liệu, ta trích xuất ra các k-shingle tương ứng với các tài liệu đó. Gía trị n này thường được lấy là 3 hoặc 4. Sau đó, độ tương tự của 2 tài liệu được tính theo công thức

$$J(A,B)=|shingles\ of\ A \cap shingles\ of\ B| / |shingles\ of\ A \cup shingles\ of\ B| \quad (1)$$

J(A, B) >=0.1 thì ta đưa B và danh sách các tài liệu nguồn Dx.

Phân tích:

Ở bước này, ta phân tích chi tiết từng câu trong tài liệu ngờ dq với tài liệu nguồn dx∈Dx.

Dầu tiên, cả dq và dx sẽ được phân tách thành các câu Sq và Sx dựa vào dấu chấm cuối câu. Để đánh giá mức độ tương quan giữa 2 câu, ta định nghĩa một yếu tố tương quan giữa 2 từ như sau:

$$\mu_{q,x} = 1 - \prod w_k \in Sx(1 - F_{q,k})$$

trong đó wk là 1 từ thuộc Sx,

Fq,k chỉ mức độ tương tự của 2 từ, bằng 1 nếu 2 từ giống nhau, 0.5 nếu đồng nghĩa và 0 nếu 2 từ mang nghĩa khác nhau hoàn toàn. Tập những từ tương tự của wq được lấy từ kho dữ liệu từ vựng WordNet. Khi đó, độ tương tự của sq so với sx sẽ tính bằng công thức:

$$Sim(sq, sx) = (\mu 1,x + \mu 2,x + \ldots + \mu q,x + \ldots + \mu n,x) / n$$

Với n là số từ của câu ngờ sq.

Lưu ý, các câu đem đi tính độ tương đồng đều đã được loại bỏ các stopwords và đưa về dạng nguyên mẫu.

Ví dụ, S1=“the teacher gives each student a text that he authored”

S2=“a textbook authored by the instructor is given to his pupils”

Sau khi loại bỏ các stopword, S1=”Teacher give student text authored.”

S2=”Textbook authored instructor given pupils.”

Sim(S1, S2)=(0.5+1+0.5+0.5+1)/5=0.7

Sim(S2, S1)=(0.5+1+0.5+1+0.5)/5=0.7

Một trường hợp khác,

S1= “This car consumes a lot of oil.”

S2=“The engine of this car is of poor quality and consumes a lot of petrol.”

Sau khi loại bỏ các stopword,

S1=”car consumes oil”

S2=”engine car poor quality consumes petrol”

Sim(S1, S2)=(1+1+0.5)/3=0.83

Sim(S2, S1)=(0+1+0+0+1+0.5)/6=0.42

Chính vì có sự khác nhau giữa Sim(s1, s2) với Sim(s2, s1) nên tác giả đưa ra ngưỡng α=0.65 xem như là ngưỡng tối thiểu để đạt sự tương tự giữa 2 câu. Khi đó, 2 câu s1 và s2 được xem là tương tự nhau nếu Sim(s1, s2) và Sim(s2, s1) đều lớn hơn hoặc bằng α.

Vì sử dụng câu làm đơn vị so sánh nên sau khi có output của tất cả các câu trong cả 2 tài liệu, ta có thể gom các câu tương tự liên tiếp hoặc cách nhau không quá 100 kí tự để đánh dấu các đoạn sao chép.