### Word location:

Trong phương pháp này, chúng ta xem các câu liền kề nhau như là một phân đoạn văn bản. Vậy một phần của văn bản hay toàn bộ văn bản cũng có thể được xem là một phân đoạn. Trong thí nghiệm này, tất cả các công việc ta sẽ xét trên các phân đoạn văn bản.

Một văn bản được chia thành nhiều phân đoạn. Trong mỗi phân đoạn ta sẽ có các từ, và thứ tự của nó.



***Với wi, 0<= i<= n-1, là thứ tự của từ trong phân đoạn.***

Vị trí của từ trong văn bản được xác định



***Dể dàng nhận thấy rằng, 0< Wli<1.***

Thực ra , trong một phân đoạn văn bản. Chỉ có xác định một số hữu hạn **m** các từ **wk** và mỗi từ thì có **nk** hiển thị (*Xuất hiện nhiều lần tại những vị trí khác nhau trong văn bản)*.

Ta có một công thức như sau.



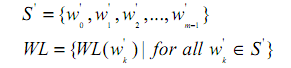
***Nghĩa là: Số từ đếm được trong văn bản, là tổng của tất cả các từ phân biệt có trong văn bản đó và số lần xuất hiện của chính nó.***

Đối với một từ **wk** thì có **nk** hiển thị. Do đó ta nói từ **wk** có một tập các giá trị định vị vị trí.



Đặt **S’** là tập hợp các từ phân biệt có trong văn bản. WL là tập hợp địa chỉ của chúng.

Ta có công thức.



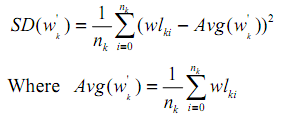
Và đây là dạng khác của biểu diển một đoạn văn bản.



### Thống kê:

Trong văn bản tự nhiên. Sự lặp lại của một từ luôn thể hiện sự phân bổ không đồng đều. Một từ có thể xuất hiện nhiều ở một đoạn nào đó, nhưng ở chổ khác thì hiếm khi xuất hiện.

Đặt **SD** ( chỉ số dàn trải : spread degree) để biểu thị điều đó.

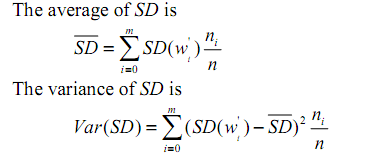


Từ công thức của một đoạn văn bản



Cho thấy, mỗi phân đoạn văn bản sẽ là tập hợp của một số từ (dĩ nhiên là mỗi từ sẽ có 1 tập hợp vị trí của nó)

Chúng ta có thể đo được sự phân bổ của từ trong một phân đoạn văn bản bằng cách kiểm tra sự phân bổ của chỉ số **SD.**



Giá trị **SD trung bình** và **VarSD** biểu thị đặc trưng phân bổ của một từ xác định.

Do đó có thể dùng nó như một luật để phân loại văn bản thường và văn bản đã có ẩn dữ liệu.

### Ví dụ cụ thể cho phương pháp xác định trên như sau:

Ta có một đoạn văn bản như sau:

**Wherever you go ,whoever you will be. I will be right here waiting for you .**

**(số từ n= 15).**

Thì tập S của ta sẽ là

**S = {‘wherever’, ‘you’, ‘go’, ‘whoever’, ‘you’, ‘will’, ‘be’, ‘i’, ‘will’, ‘be’, ‘right’, ‘here’, ‘waiting’, ‘for’, ‘you’ }**

Ta có các giá trị wl được xác định như bên dưới.

* + **wl0 = 0/15 = 0 (‘whatever’)**
  + **wl1 = 1/15 = 0.067 (‘you’)**
  + **wl2 = 2/15 = 0.133 (‘do’)**
  + **wl4 = 4/15 = 0.266 (‘you’)**
  + **wl14 = 14/15 = 0.933 (‘you’)**

Tập **S** bên trên chứa tất cả các từ xuất hiện trong văn bản.

Ta định nghĩa ra tập **S’**, là tập các từ xuất hiện trong văn bản, và không lặp lại

**S’ = {‘wherever’, ‘you’, ‘go’, ‘whoever ’, ‘will’, ‘be’, ‘i’, ‘right’, ‘here’, ‘waiting’, ‘for’ } (Số từ m=11)**

Và ta có tập hợp các giá trị định vị của từ được ví dụ như sau

* + **WL(w’0 ) = { 0 } (‘wherever’)**
  + **WL(w’1 ) = {0.067, 0.266, 0.933 } (‘you’)**

Từ những giá trị chứa trong tập hợp các giá trị **WL** trên, ta xác định được giá trị định vị trung bình cho các từ (**riêng biệt**) có trong văn bản. Và từ đó tính được chỉ số dàn trải của từ đó trong văn bản.

Ví dụ đối với từ **You**, thì có

**WL(w’1 ) = {0.067, 0.266, 0.933 }**

**Avg = (0.067+0.266+0.933 )/3 = 0,421.**

**=> SD = 0.12**

Sau khi có các giá trị **SD** thì ta sẽ tiến hành tính hai giá trị tiếp theo là Giá trị **SD trung bình** và **varSD**. Đây là hai giá trị cuối cùng, biểu diển kết quả của phương pháp.

Hai giá trị này dùng cho mô hình phát hiện.