**CHÚ Ý:**

Tao đã up project lên git, mọi người clone về rồi chạy thử.

Tao đã tách được các n-gram.

Output hiện tại là string chứa các n-gram của file, các n-gram cách nhau bởi kí tự xuống dòng;

Làm tương tự ta sẽ được 2 chuỗi chứa n-grams chủa 2 file nguồn và file cần kiểm tra.

**Nhiệm vụ của Trí:**

Chỉ ra được thứ tự xuất hiện của các n-grams trong file nghi ngờ xuất hiện ở vị trí nào trong file nguồn. Cụ thể, nếu n-gram thứ nhất trong file ngờ giống với n-gram thứ 9 trong file nguồn thì mọi người phải lưu được cặp giá trị (1,9);

Gợi ý, có thể dùng mảng 2 chiều để lưu, cụ thể:

0 i

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

J

Arr1[i] chứa các n-grams của file nghi ngờ, mỗi phần tử arr[k](0<=k<=i) là 1 string chứa 1 n-gram;

Trương tự, arr2[j] là của file nguồn;

Output arr[i, j];

0->i chỉ thứ tự các n-grams trong file nghi ngờ;

0->j chỉ thứ tự các n-grams trong file nguồn;

For(int a=0, a<=I, a++)

For(int b=0, b<=j, b++){

If(arr1[a]==arr2[b]) arr[a,b]=1;

Else arr[a,b]=0;

}

Trên đây chỉ là gợi ý, mọi người nên tìm hiểu thêm để có được ý tưởng tốt nhất;

**Nhiệm vụ của Mẫn:**

Giả sử có output của Trí, khoanh vùng các đoạn giống nhau, đọc bài báo để hiểu cách khoanh vùng đoạn. Sau đó tính độ tương tự của các đoạn được khoanh vùng ở trên. Sau đó phải xuất được có bao nhiêu đoạn giống nhau, chỉ số tương tự của các đoạn đó.

---HẾT---