|  |  |
| --- | --- |
| Họ và tên: Nguyễn Hữu Kiệt  MSSV: 20183571  Mã lớp Project 1: 699554 | Giảng viên hướng dẫn: TS. Trần Vĩnh Đức |

**Báo cáo tuần 1-5**

**Tên đề tài: Phép biến đổi BURROWS WHEELER**

**và ứng dụng nén file text Tiếng Anh dung lượng lớn.**

1. Lý do chọn đề tài:

2. Nội Dung:

2.1. Phép biến đổi Burrows Wheeler (BWT) xuôi:

2.1.1. Giới thiệu:

- Phép biến đổi Burrows Wheeler (gọi tắt là BWT) là một thuật toán để tổ chức lại dữ liệu text được sử dụng trong nén dữ liệu độ mất mát ít.

- Một xâu text sau khi qua BWT sẽ trở thành một xâu có khả năng nén tốt hơn nhiều so với xâu ban đầu.

2.1.2. Chi tiết:

2.1.2.1. Thuật toán:

- Bảng sau mô tả cách thực hiện BWT xuôi:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Transformation** | | | | |
| **1. Input** | **2. All rotations** | **3. Sort into lexical order** | **4. Take the last column** | **5. Output** |
| ^BANANA| | ^BANANA|  |^BANANA  A|^BANAN  NA|^BANA  ANA|^BAN  NANA|^BA  ANANA|^B  BANANA|^ | **A**NANA|^B  **A**NA|^BAN  **A**|^BANAN  **B**ANANA|^  **N**ANA|^BA  **N**A|^BANA  **^**BANANA|  **|**^BANANA | ANANA|^**B**  ANA|^BA**N**  A|^BANA**N**  BANANA|**^**  NANA|^B**A**  NA|^BAN**A**  ^BANANA**|**  |^BANAN**A** | BNN^AA|A |

- BWT được hoàn thành bằng cách sắp xếp tất cả các xâu dịch trái (hoặc phải cho kết quả tương đương) xoay vòng theo thứ tự từ điển rồi lấy 1 kí tự cuối cùng cuối mỗi xâu.

2.1.2.2. Giải thích:

- Xâu dịch trái xoay vòng: là xâu được dịch qua bên trái một số kí tự kể từ xâu ban đầu và các kí tự bị tràn ra ngoài sẽ được điền ngược lại vào đằng sau. Ví dụ:

Ta có xâu: BANANA^

Vậy ta sẽ có xâu dịch trái xoay vòng:

1 kí tự sẽ là: ANANA^B

2 kí tự sẽ là: NANA^BA

….

Xâu cuối: ^BANANA

- Sắp xếp tất cả các xâu đó theo thứ tự từ điển rồi lấy kí tự cuối cùng mỗi xâu ta được xâu sau khi qua BWT.

- Ta nhận thấy sau khi biến đổi các kí tự giống nhau sẽ được đẩy lại gần nhau. Tại sao?

- Điều này dễ thấy bởi trong Tiếng Anh nói riêng và các ngôn ngữ khác nói riêng thì trong 1 văn bản các kí đôi sẽ được lặp đi lặp lại rất nhiều.

Đơn cử như trong Tiếng Anh, chỉ riêng đuôi “tion” trong 1 văn bản có thể xuất hiện rất nhiều lần, điều đó đồng nghĩa với việc các kí tự đôi đi kèm nhau như: “ti”, “io”, “on”… cũng xuất hiện nhiều lần.

Các kí tự đôi như vậy sau khi thực hiện BWT thì sẽ được đẩy lại gần nhau bởi khi thực hiện dịch trái xoay vòng ta đã tách các kí tự đôi ra, đẩy 1 kí tự lên đầu còn 1 kí tự xuống cuối, vì vậy mà các kí tự đôi như “ti” sẽ trở thành “i…t” và khi thực hiện sắp xếp các xâu bắt đầu bằng “i” sẽ được đẩy lại gần nhau đồng nghĩa với các kí tự “t” ở cuối cũng được đẩy lại gần nhau.

Từ đó, ta được 1 xâu mới có khả năng nén tốt hơn nhiều so với xâu cũ.

Vậy tại sao ta không lấy các kí tự đầu mỗi xâu, không phải lấy vậy sẽ được 1 xâu mới có khả năng nén tốt hơn rất rất nhiều sao?

Đơn giản vì nếu lấy kí tự đầu mỗi xâu thì không có í nghĩa gì (vì nó sẽ thành xâu kiểu “aaaaaaaaaaaaabbbbbbbbbbbcccccccccc…”, cùng với đó là ta không thể giải ngược được các xâu này để ra xâu ban đầu trong khi nếu lấy các kí tự cuối ta hoàn toàn có thể làm được (BWT ngược).

Danh mục tài liệu tham khảo:

1. <https://en.wikipedia.org/wiki/Burrows%E2%80%93Wheeler_transform>
2. …