Thành viên: Đinh Viết Thường : 13020607, Bùi Văn Hoàng.

Guildline for segmentation usage:

**file / folder:**

* dictionary: gồm dictionary đã được tách từ trainning data và được chia làm 5 từ điển tương ứng với những word gồm 1-2-3-4-5 tiếng, từ điển địa danh của việt nam, từ điển những ký hiệu có thể có trong dữ liệu.
* testingDataSet: gồm file chứa dữ liệu test dưới định dạng:

Sentence,segmented

Sentence2,segmented2

…

* trainedModel: lưu trữ bigram word count, unigram word count, alpha beta đã train được cho linear interpolation technique
* trainingDataSet: chứa tập dữ liệu train đã segmented và đã tách thành từng câu
* preprocessingSentence.py: xử lý tách câu thành những phrase nhỏ, phát hiện date, number, sign,url, tên riêng,...
* maximumMatching.py: tách câu bằng maximum matching
* Train.py: đếm bigram,unigram dựa trên tập train, train alpha1, alpha2 cho thuật toán linear interpolation
* conditionalProbabilityEvaluate.py: tính xác suất cho câu dựa vào linear interpolation
* dictionaryRead.py: đọc từ điển, và một số input khác.
* Segmentation.py: chứa hàm main segmentation.

Phương pháp sử dụng:

Sử dụng kết hợp longest matching và bigram

**Quá trình segment gồm những bước sau:**

*Bước 1:* tiền xử lý câu

Sử dụng regular expression để phát hiện những đối tượng Ngay-thang, url, email, số, và ký hiệu lạ, những phần đó sẽ là spliter cho câu đầu vào, tức là câu đầu vào sẽ được tách làm N-phrase chỉ gồm những từ bình thường, sau đó áp dụng bước 2 cho từng phrase đó. Tách những từ viết hoa gần nhau và coi như đó là một từ mới ví dụ như tên người Nguyễn Văn A.

*Bước 2:* Sử dụng longest matching để tìm ra cách tách đơn giản nhất cho từng phrase:

* Ta xây dựng một đồ thị có hướng có trọng số với mỗi khoảng trống các từ trong câu là một note, khoẳng cách giữa 2 note được đánh giá phụ thuộc vào số lượng từ ở giữa 2 Note

1word: 7, 2word: 4, 3word :2.5, 4word: 1, 5word: 1

* Sau đó áp dụng djikstra algorithm để tìm đường ngắn nhất đó chính là cách tách của longest matching.

Bước 3: sử dụng Ngram model để khử nhập nhằng dạng: (AB)C với A(BC):

* Tìm những phần trong câu đã tách được bước trước có dạng (AB)C với A(BC)

Ta tính xác suất của câu trong hai trường hợp và chọn trường hợp có xác suất lớn hơn. Xác xuất được xác định bằng mô hình linear interpolation giữa bigram và unigram. Estimate P(wi/wi-1)= alpha1\*P(wi/wi-1)+ alpha2\*P(wi)

Chạy thuật toán và định dạng input:

File test input định dạng :

Sentence 1

Sentenve 2

Sentence 3

…

Chay python segmentation.py và điền tên file

Sau khi chạy sẽ thu đựoc file output là gồm câu đã tách