VietnameseTextPro

A Vietnamese Text Processing Toolkit

Version 1.0

Xuan-Hieu Phan (pxhieu@gmail.com)

University of Engineering and Technology (UET),
Vietnam National University, Hanoi (VNU)

Mục lục

1. Giới thiệu

VietnameseTextPro (Vietnamese Text Processing Toolkit) là bộ công cụ xử lý văn bản tiếng Việt được phát triển bằng Java với mục đích:

- Thực hiện các bài toán xử lý tiếng Việt nền tảng (như tách câu, tách từ, gắn nhãn từ loại, xác định cụm từ, …) với độ chính xác rất cao. Độ chính xác tách câu là 99.85% và độ chính xác tách từ là 98.86% (F1-score).
- Tốc độ thực hiện nhanh và ổn định.
- Có thể đảm bảo là nền tảng cho các ứng dụng xử lý ngôn ngữ tự nhiên, khai phá dữ liệu, thông minh doanh nghiệp (business intelligence) ở mức cao hơn.

Các mô hình tách câu, tách từ, .v.v. ở trong bộ công cụ này được huấn luyện trên các tập dữ liệu lớn sử dụng các mô hình học máy thống kê mạnh như conditional random fields (CRFs – xem FlexCRFs-0.4) và maximum entropy (maxent – MaxEnt-0.2).

1.1. Giấy phép sử dụng

1.2. Tải về

Bộ công cụ này chỉ phục vụ mục đích xây dựng các hệ thống ứng dụng thương mại.

2. Biên dịch

2.1. Cấu trúc mã nguồn

Cấu trúc của VietnameseTextPro-1.0 và mã nguồn được tổ chức như sau:

```
build                       (thư mục chứa kết quả đầu ra của biên dịch)
dist                        (thư mục chứa kết quả đầu ra của biên dịch)
docs                        (tài liệu liên quan – bao gồm cả tài liệu này)
lib                         (chứa các thư viện liên quan)
models                      (các mô hình được huấn luyện)
      dm                    (các mô hình khai phá dữ liệu – data mining)
      ml                    (các mô hình học máy – machine learning)
      nlp                   (các mô hình xử lý ngôn ngữ tự nhiên - nlp)
            en              (các mô hình xử lý ngôn ngữ tự nhiên – tiếng Anh)
            vn              (các mô hình xử lý ngôn ngữ tự nhiên – tiếng Việt)
                  vnchunker         (mô hình xác định cụm từ)
                  vnner             (mô hình nhận dạng thực thể tên)
                  vnpostagger       (mô hình xác định từ loại)
                  vnsentsegmenter   (mô hình tách câu)
                  vntokenizer       (mô hình rời rạc hóa tiếng Việt)
                  vnwordsegmenter   (mô hình tách từ tiếng Việt)
nbproject                   (các files cấu hình của dự án - NetBeans)
resources                   (các files dữ liệu, tài nguyên)
      dicts                 (các files từ điển)
            abbreviations   (các từ điển các từ viết tắt)
            endicts         (các từ điển tiếng Anh)
            lexicons        (các từ điển từ vựng)
            names           (các từ điển tên riêng)
            specialchars    (các từ điển liên quan đến các ký tự đặc biệt)
            vndicts         (các từ điển tiếng Việt)
      regexes               (các files biểu thức chính quy)
src                         (mã nguồn)
      lib                   (các gói thư viện)
            collection      (thư viện về các cấu trúc chứa – array, vector, …)
            filesystem      (thư viện về hệ thống file, thư mục)
            math            (thư viện toán học)
            optimization    (thư viện tối ưu toán học - LBFGS)
            pairs           (các cấu trúc cặp đôi)
            properties      (thư viện về các tham số, đường dẫn, …)
            statistics      (thư viện về thống kê)
            string          (thư viện xử lý xâu ký tự)
      mlearning             (các mô hình học máy)
            data            (định dạng dữ liệu cho các phương pháp học máy)
            flexcrfs        (phần suy diễn của conditional random fields)
            maxent          (phương pháp phân lớp maximum entropy)
      nlp                   (mã nguồn xử lý ngôn ngữ tự nhiên)
            vn              (xử lý ngôn ngữ tự nhiên tiếng Việt)
```

```
                vnchunker          (xác định cụm từ)
                vnner              (nhận dạng thực thể tên)
                vnpostagger        (xác định từ loại)
                vnsentsegmenter    (tách câu)
                vntextpro          (gộp các tính năng xử lý tiếng Việt)
                vntokenizer        (rời rạc hóa tiếng Việt)
                vnwordsegmenter    (tách từ tiếng Việt)
      resources                (xử lý dữ liệu và các tài nguyên)
            dicts              (xử lý các từ điển)
                abbreviations  (xử lý các từ điển viết tắt)
                specialchars   (xử lý các ký tự đặc biệt)
                vndicts        (xử lý các từ điển tiếng Việt)
            regexes            (xử lý các biểu thức chính quy)
      tests                    (các chương trình test, chạy thử)
build.xml                      (file cấu hình cho biên dịch)
vietnamesetextpro.properties   (thông tin cấu hình, đường dẫn, ...)
```

2.1. Biên dịch

Yêu cầu nền tảng phần mềm:

- Java 1.4 trở về sau
- Công cụ ant (make tool cho Java)

Biên dịch trong NetBeans:

Mở dự án VietnameseTextPro-1.0 với NetBeans 1.7 và tiến hành biên dịch trong môi trường này.

Biên dịch với ant:

Đứng ở thư mục chủ của VietnameseTextPro-1.0 và tiến hành biên dịch bằng dòng lệnh:

```
$ ant
```

Kết quả của quá trình biên dịch:

Kết quả biên dịch là file `VietnameseTextPro-1.0.jar` trong thư mục `./dist`

3. Sử dụng VietnameseTextPro-1.0

Phần này sẽ hướng dẫn sử dụng các tính năng của VietnameseTextPro-1.0 để thực hiện các tác vụ xử lý tiếng Việt như tách câu, tách từ, .v.v.

3.1. Tách câu với VietnameseTextPro

Để thực hiện tách câu cho một đoạn văn bản, chúng ta cần đoạn mã sau:

```
1: VnSentSegmenter sentSegmenter = new VnSentSegmenter();
2: sentSegmenter.init();
3: String text = "<đoạn văn bản tiếng Việt cần tách câu (encoding UTF-8)>";
4: List<String> sents = sentSegmenter.segment(text);
```

Trong đó:

- Dòng 1: thực hiện tạo một đối tượng mô hình xử lý tách câu
- Dòng 2: khởi tạo và nạp (load) mô hình tách câu vào bộ nhớ
- Dòng 3: `text` là đoạn văn bản tiếng Việt cần tách câu, encoding phải là UTF-8
- Dòng 4: thực hiện tách câu cho đoạn văn bản `text`, trả về một danh sách các câu trong `sents`.

3.2. Tách từ với VietnameseTextPro

Để thực hiện tách từ cho một câu tiếng Việt, chúng ta cần đoạn mã sau:

```
1: VnWordSegmenter wordSegmenter = new VnWordSegmenter();
2: wordSegmenter.init();
3: String sent = "<câu văn bản tiếng Việt cần tách từ (encoding UTF-8)>";
4: String segmentedSent = wordSegmenter.segment(VnTokenizer.tokenize(sent));
```

Trong đó:

- Dòng 1: thực hiện tạo một đối tượng mô hình xử lý tách từ
- Dòng 2: khởi tạo và nạp (load) mô hình tách từ vào bộ nhớ
- Dòng 3: `sent` là câu văn bản tiếng Việt cần tách từ, encoding phải là UTF-8
- Dòng 4: thực hiện tách câu cho câu văn bản `sent`, trả về một câu trong đó các từ đã được tách `segmentedSent`. Lưu ý, trước khi tách từ thì cần thực hiện rời rạc hóa bằng phương thức tĩnh `tokenize` của lớp `VnTokenizer`.

3.3. Thực hiện tách câu, rời rạc hóa, và tách từ

Để thực hiện tách câu, rời rạc hóa, và tách từ cho một văn bản tiếng Việt, chúng ta cần thực hiện đoạn mã sau:

```
1: VnTextPro vnTextPro = new VnTextPro(true, true, true);
2: vnTextPro.init();
```

```
3: String text = "<văn bản tiếng Việt cần tách câu và tách từ (encoding UTF-8)>";
4: List<String> segmentedSents = vnTextPro.segmentText(text);
```

Trong đó:

- Dòng 1: thực hiện tạo một đối tượng mô hình xử lý tiếng Việt (bao gồm cả tách câu - `true`, rời rạc hóa - `true`, và tách từ - `true`).
- Dòng 2: khởi tạo và nạp (load) mô hình tách từ vào bộ nhớ
- Dòng 3: `text` là văn bản tiếng Việt cần tách câu, tách từ (encoding UTF-8)
- Dòng 4: thực hiện tách câu, rời rạc hóa, và tách từ cho văn bản `text`, trả về một danh sách các câu trong đó các từ đã được tách `segmentedSents`.

3.3. Chạy thử tách câu, rời rạc hóa, và tách từ với dòng lệnh

Đứng ở thư mục gốc của VietnameseTextPro-1.0, chúng ta có thể chạy thử các tính năng tách câu, tách từ bằng các câu lệnh sau:

- Tách câu:

```
$ java –classpath dist/VietnameseTextPro-1.0.jar tests.VnSentSegmenterTest
```

- Tách từ:

```
$ java –classpath dist/VietnameseTextPro-1.0.jar tests.VnWordSegmenterTest
```

- Tách câu + tách từ:

```
$ java –classpath dist/VietnameseTextPro-1.0.jar tests.VnTextProTest
```