1. **Giới thiệu Text segmentation:**

* Text segmentation là quá trình phân chia đoạn thành những đơn vị có ý nghĩa như: từ, câu. Các đơn vị này sẽ là những đơn vị cơ bản sử dụng cho những bước xử lý sau này.
* Được sử dụng trong cả quá trình tinh thần của con người (nghe, nói, đọc, viết, ghi nhớ…) và ngôn ngữ máy tính
* Vấn đề khó khăn: trong khi một số ngôn ngữ có sự phân cách rõ ràng giữa các từ (ví dụ như khoảng trắng trong tiếng Anh) thì lại có một số ngôn ngữ thì không có sự phân cách rõ ràng hoặc có những ký tự không có trong các văn bản ngôn ngữ
* Gồm: tokenization và Sentence segmentation

1. **Tokenisation:**
   * Là quá trình phân đoạn văn thành từng từ bằng các word boundaries (ranh giới từ). Để phục vụ cho thuật ngữ tính toán, thì các từ được xác định được gọi là các tokens. Đối với ngôn ngữ có word boundaries rõ ràng thì tokenization còn được gọi là word segmentation.
   * Trong tiếng Anh và nhiều ngôn ngữ khác sử dụng chữ cái latin thì khoảng trắng là word boundary. Mặc dù nó cũng có giới hạn bởi vì sự đang dạng của ngôn ngữ: từ đồng nghĩa, các loại từ (động từ, danh từ, tính từ…)…
   * Tùy vào mỗi ngôn ngữ sẽ có word boundaries khác nhau, nếu như không có nó thì quá trình word segmentation sẽ gặp khó khăn. Tiếng Trung Quốc, Nhật không có word boundaries rõ ràng trong một câu. Tiếng Việt Nam có âm tiết nhưng lại không có sự phân cách rõ ràng cho một cụm từ.
2. **Sentence segmentation**:
   * Là quá trình phân chia văn bản thành những đơn vị xử lý lớn hơn từ (câu).
   * Trong tiếng Anh và một số ngôn ngữ khác sử dụng dấu chấm để kết thúc một câu. Tuy nhiên, có một vấn đề gặp phải trong tiếng Anh đó là từ viết tắt không phải là kết thúc một câu nhưng vẫn sử dụng dấu chấm. Ví dụ: “Mr. Tri is a docter.”
   * Cũng như Tokenisation thì không phải tất các ngôn ngữ viết đều chứa các ký tự kết thúc câu để xác định ranh giới của một câu.
3. **Text segmentation trong Stanford CoreNLP C#:**

* Stanford CoreNLP hỗ trợ tokenizes văn bản. Nó băt đầu như là một PTB-style tokenizer (một dạng phân đoạn từ cho tiếng Anh) nhưng nó được mở rộng xử lý cho cả các ngôn ngữ khác và các văn bản web bị nhiễu. Tokenizer lưu vị trí bắt đầu (CharacterOffsetBeginAnnotation) và vị trí kết thúc (CharacterOffsetEndAnnotation) của một token.

|  |  |  |
| --- | --- | --- |
| **PROPERTY NAME** | **ANNOTATOR CLASS NAME** | **GENERATED ANNOTATION** |
| tokenize | TokenizerAnnotator | TokensAnnotation (list of tokens), and CharacterOffsetBeginAnnotation, CharacterOffsetEndAnnotation, TextAnnotation (for each token) |

Tham khảo:

<https://s3.amazonaws.com/tm-town-nlp-resources/ch2.pdf>

<https://stanfordnlp.github.io/CoreNLP/tokenize.html#options>

<https://sergey-tihon.github.io/Stanford.NLP.NET/StanfordCoreNLP.html>

Link model: <https://nlp.stanford.edu/software/stanford-corenlp-full-2016-10-31.zip>?