2 Biểu thức chính quy, chuẩn hóa text edit distance

```
User : You are like my father in some ways.
ELIZA: WHAT RESEMBLANCE DO YOU SEE
User : You are not very aggressive but I think you don't
        want me to notice that.
ELIZA: WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE
User : You don't argue with me.
ELIZA: WHY DO YOU THINK I DON'T ARGUE WITH YOU
User : You are afraid of me.
ELIZA: DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU
```

Weizenbaum (1966)

Đoạn hội thoại trên là từ **ELIZA**, một hệ thống xử lý ngôn ngữ tự nhiên đầu tiên có thể thực hiện những đoạn hội thoại với người dùng bằng cách bắt chước các câu nói trước đó của một nhà tâm lý học (Weizenbaum, 1966). ELIZA là một chương trình đơn giản sử dụng phương pháp nhanạ dạng các cụm từ như "Bạn là X" và sau đó chuyển thành câu hỏi hợp lý như "Điều gì khiến bạn nghĩ tôi là X?". Kỹ thuật đơn giản này đạt được thành công trong lĩnh vực này vì ELIZA không cần *biết* điều gì để bắt chước lại nhà tâm lý học. Trong ghi chú của Weizebaum, đây là một trong những ít chương trình hội thoại mà người nghe có thể phản ứng mà không cần biết điều gì về thế giới. Cách bắt chước của ELIZA đạt được nhiều thanh công đáng chú ý: rất nhiều người tauwơng tác với ELIZA bắt đầu tin rằng nó *thực sự* hiểu họ và vấn đề của họ, rất nhiều người tiếp tục tin vào khả năng của ELIZA mặc dù đã biết về cách vận hành của nó (Weizenbaum, 1976), và ngay cả trong thời điểm này, những chatbots như vậy cũng rất thú vị.

Tất nhiên, những chương trình hội thoại hiện đại có rất nhiều kỹ thuật; chúng có thể trả lời trả lời các câu hỏi, đặt vé máy bay hoặc tìm nhà hàng, các kỹ thuật chúng sử dụng dựa nhiều vào việc hiểu sâu sắc ý định của người dùng, chúng ta sẽ cùng tìm hiểu trong chương 25. Tuy nhiên, những phương pháp dựa vào luật như ở ELIZA hay các chabot khác đóng vai trò quan trọng trong việc xử lý ngôn ngữ tự nhiên.

Chúng ta sẽ bắt đầu với công cụ quan trọng nhất trong việc biểu diễn các luật văn bản: **biểu thức chính quy**. Các biểu thức chính quy có thể chỉ định một chuỗi chúng ta muốn tách ra từ một văn bản, từ các cụm từ "Bạn là X" như ở ELIZA, cho đến các chuỗi như $199 hay $24.99 để trích xuất giá từ một bảng trong một văn bản.

2.1. BIỂU THỨC CHÍNH QUY

Một trong những thành công lớn trong khoa học máy tính là biểu thức chính quy, một ngôn ngữ để chỉ định các chuỗi tìm kiếm.

2.1.1. CÁC BIỂU THỨC CHÍNH QUY CƠ BẢN

Dạng đơn giản nhất của biểu thức chính quy là một dãy các kí tự. Để tìm kiếm từ *woodchuck*, ta gõ `/woodchuck`. Biểu thức `/Buttercup/` khớp với tất cả các chuỗi có chứa cụm từ *Buttercup*; `grep`

RE	Example Patterns Matched
/woodchucks	interesting links to woodchucks and lemurs
/a/	Mary Ann stopped by Mona's
/!/	"You've left the burglar behind again!" said Nori

Figure 2.1: Ví dụ các biểu thức chính quy đơn giản

abc (Phuong, 2016) def

Bibliography

Le-Hong Phuong. 2016. Vietnamese named entity recognition using token regular expressions and bidirectional inference. *CoRR* abs/1610.05652. http://arxiv.org/abs/1610.05652.