AUTHOR: Huy Pham Bui Nhat

TÀI LIỆU TỔNG HỢP KIẾN THỨC

ĐỒ ÁN GIỮA KÌ MÔN CTDL&GT NÂNG CAO (CS523.M21.KHCL)

❖ N-gram:

- Một **mô hình ngôn ngữ** (**Language model - LM**).
- Xét **chuỗi liên tiếp n từ** (thường nhỏ 5 hơn từ).
- Tính toán dựa trên tần số xuất hiện của chuỗi n từ liên tiếp (xác suất thống kê).
- Ý nghĩa: Đo mức độ giống với ngôn ngữ bản xứ của một cụm từ hoặc câu $\boldsymbol{w} = w_1 w_2 \dots w_n$ dựa trên xác suất $p(\boldsymbol{w})$ của cụm từ/câu đó.

$$\boxed{p(\boldsymbol{w}) = p(w_1)p(w_2|w_1)p(w_3|w_2 w_1) \dots p(w_n|w_1 \dots w_{n-1})}$$

- Lưu ý:
  - Chuỗi xác suất trên sẽ bất khả thi khi n lên đến hàng chục từ.
    ➔ Cách giải quyết là sử dụng **giả sử Markov** (**Markov assumption**)

$$\boxed{p(w_i|w_1 \dots w_{i-1}) = p(w_i|w_{i-n+1} \dots w_{i-1})}$$

  - Một vd cho thấy tại sao có thể sử dụng **Markov assumption**, **Markov assumption** cho phép ta có thể ước lượng xác suất trên

Corpus:

" This is the house that Jack built.

This is the malt

That lay in the house that Jack built.

This is the rat,

That ate the malt

That lay in the house that Jack built.

This is the cat,

That killed the rat,

That ate the malt

That lay in the house that Jack built. ”

VD nhỏ về việc tính chuỗi “this is the house” so với 2 câu đầu trích ra từ ngữ liệu trên với mô hình **bigram** (**n = 2**) (**mô hình Markov bậc 1** (**first-order Markov model**))

Corpus:

“ This is the house that Jack built.

This is the malt “

$$p(this\ is\ the\ house) = p(this)p(is|this)p(the|is)p(house|the)$$

$$p(this) = \frac{c(this)}{c("allword")} = \frac{2}{11}$$
$$p(is|this) = \frac{c(this\ is)}{c(this\ \dots)} = \frac{2}{2} = 1$$
$$p(the|is) = \frac{c(is\ the)}{c(is\ \dots)} = \frac{2}{2} = 1$$
$$p(house|the) = \frac{c(the\ house)}{c(the\ \dots)} = \frac{1}{2}$$
(c là viết tắt của phép đếm "$count$")

$$p(this\ is\ the\ house) = \frac{1}{11} \approx 0.091 = 9.1\%$$

Với ví dụ trên ta thấy có một vấn đề: Dù rằng với ngữ liệu nhỏ và xét 1 chuỗi chắc chắn có trong ngữ liệu nhưng lại cho ra 1 xác suất so với thực tế khá nhỏ. Vấn đề này xuất phát ở tần suất xuất hiện của $p(this)$. Tần suất này được tính so với tổng thể của ngữ liệu (theo mô hình **unigram**) và với cách tính như thế thì nó rất bất công cho tần suất xuất hiện này, vì vậy ta có 1 cách khắc phục cho trường hợp trên như sau:

$$p(this) = p(this| < s >)$$
($< s >$ là kí hiệu cho biết bắt đầu của một câu, là một token giả)

Vậy tần suất $p(this)$ được tính lại như sau:

Corpus:

“ This is the house that Jack built.

This is the malt “

$$p(this) = p(this| < s >) = \frac{c(< s > this)}{c(< s > \cdots)} = \frac{c(\_this)}{c\_ \ldots)} = \frac{2}{2} = 1$$

$$p(this\ is\ the\ house) = \frac{1}{2} = 50\%$$

Ngoài ra, ta còn thêm vào xác suất trên 1 tần suất khác với một token giả khác là </s> cho biết kết thúc một câu. Vậy xác suất của câu trên được tính toán lại như sau:

$$p(< s > this\ is\ the\ house\ </s >)$$
$$= p(this| < s >)p(is|this)p(the|is)p(house|the)p(</s > |house)$$

Corpus (để tránh trường hợp xác suất = 0, ta thêm câu "Jack in the house"):

" <s> This is the house that Jack built </s>.

<s> This is the malt </s>

<s> Jack in this house </s> "

$$p(this| < s >) = \frac{c(< s > this)}{c(< s > \cdots)} = \frac{2}{3}$$
$$p(is|this) = \frac{c(this\ is)}{c(this \ldots)} = \frac{2}{3}$$
$$p(the|is) = \frac{c(is\ the)}{c(is \ldots)} = \frac{2}{3}$$
$$p(house|the) = \frac{c(the\ house)}{c(the \ldots)} = \frac{1}{2}$$
$$p(</s > |house) = \frac{c(house\ </s >)}{c(house \ldots)} = \frac{1}{2}$$

$$p(< s > this\ is\ the\ house\ </s >) = \frac{2}{27} \approx 7.41\%$$

Với ngữ liệu ở trên, bạn đã thấy rằng với một ngữ liệu nhỏ thì xác suất có thể bằng không với phép tích chuỗi như trên. Trường hợp hợp trên được gọi chung với thuật ngữ **ngữ liệu thưa** (**sparse data**). Để xử lý vấn đề **sparses data**, ta có các **kỹ thuật làm mịn** (**Smoothing Technique**). Các kỹ thuật làm mịn này giúp làm cho xác suất của chuỗi n-gram trở thành **khác 0** cho dù nó không xuất hiện lần nào trong ngữ liệu thống

kê. Bản chất của việc này chính là thay thế ước lượng xác suất n-gram theo phương pháp **MLE** (đã làm từ đầu bài đến giờ) thành những cách tính/cách ước lượng khác.

Trong phần này, ta sẽ chỉ nói về kỹ thuật **thêm-1** và **thêm-alpha**.

+ Kỹ thuật **thêm-1**:
    - Trọng tâm của kỹ thuật này là + thêm 1 vào tất cả các tần suất.
    - Giúp cho xác suất các n-gram chưa từng xuất hiện trong ngữ liệu trở nên khác 0.
    -