CS523.M21.KHCL

ADVANCED DATA STRUCTURE AND ALGORITHM

Nhóm 3

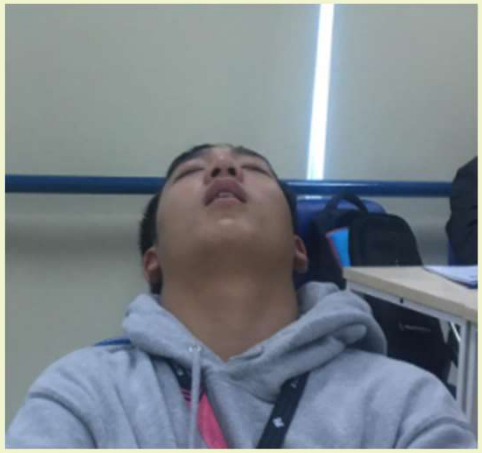
GVHD: thầy Nguyễn Thanh Sơn

CS523.M21.KHCL

OUR TEAM

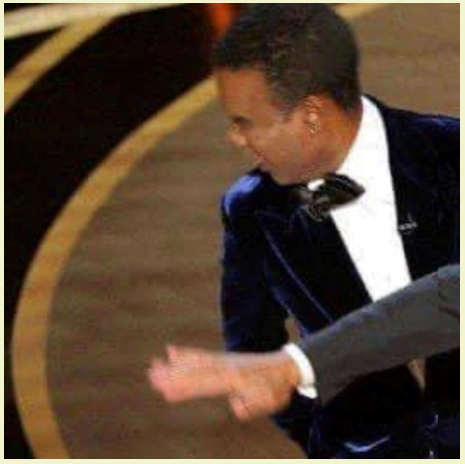
20520855

Nguyễn Thị Như Vân



20521410

Phạm Bùi Nhật Huy



19522388

Nguyễn Minh Trí

CS523.M21.KHCL

TABLE OF CONTENT

01 **Giới thiệu**

Một số ứng dụng của word-lattices trong đời sống thực tế

02 **Ngữ liệu và N-gram**

Mối liên hệ giữa ngữ liệu và N-gram

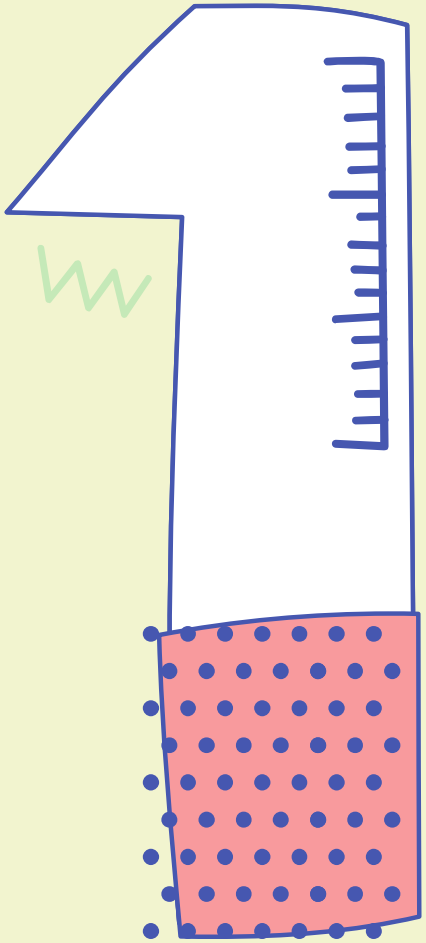
03 **Word-lattices và demo**

Cấu trúc dữ liệu để lưu trữ và cách vận dụng kết quả của N-gram để đưa lên word-lattices

04 **Câu hỏi ôn tập và giải đáp thắc mắc**

Quiz, Q&A

CS523.M21.KHCL



01

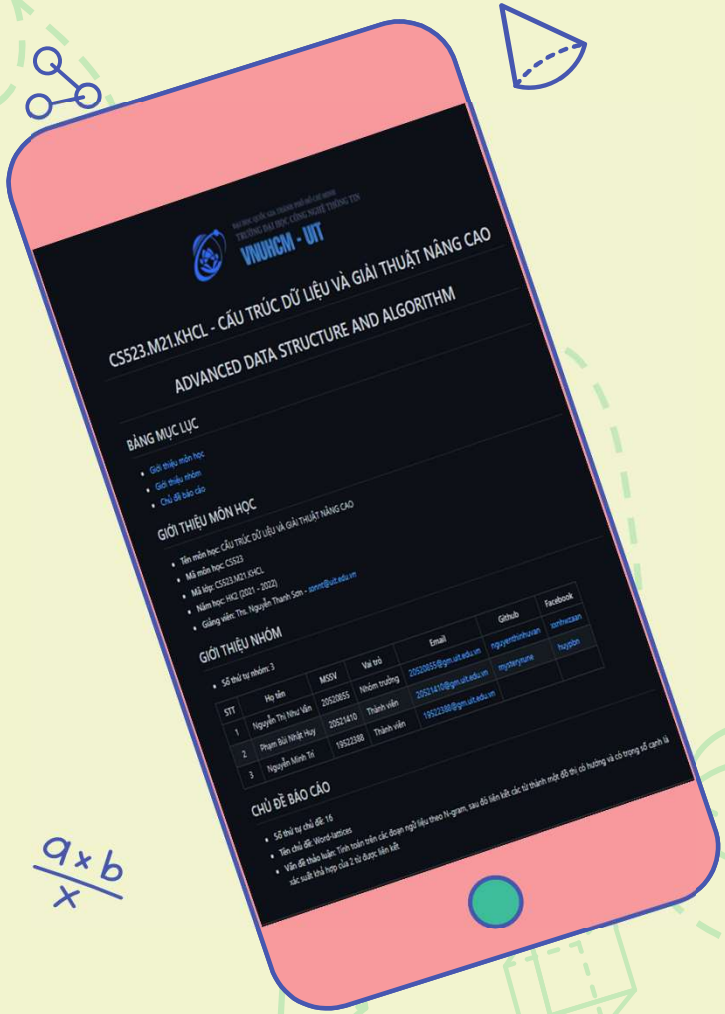
GIỚI THIỆU, ĐẶT VẤN ĐỀ

CS523.M21.KHCL

Đã bao giờ bạn tự hỏi...

CS523.M21.KHCL

Một vài ví dụ thực tế



Math for CS

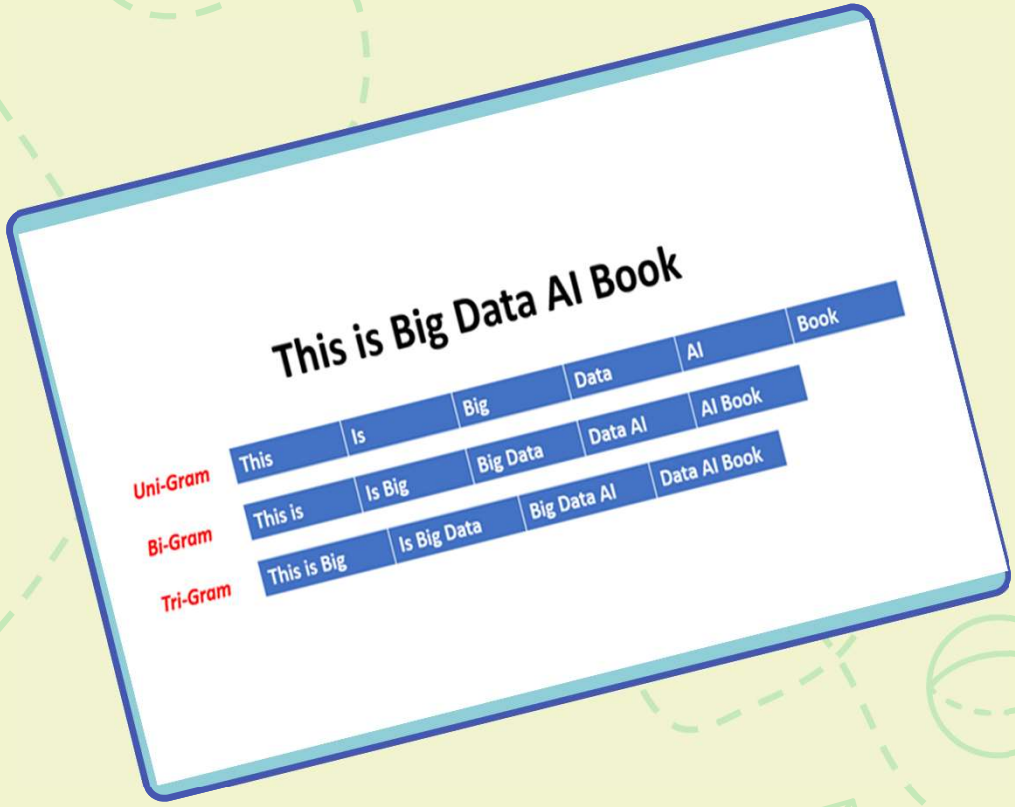
02

NGỮ LIỆU & N-GRAM

CS523.M21.KHCL

N-GRAM LÀ GÌ ?

- Một mô hình ngôn ngữ (Language models)
- Tính toán dựa trên tần số xuất hiện của chuỗi n từ liên tiếp



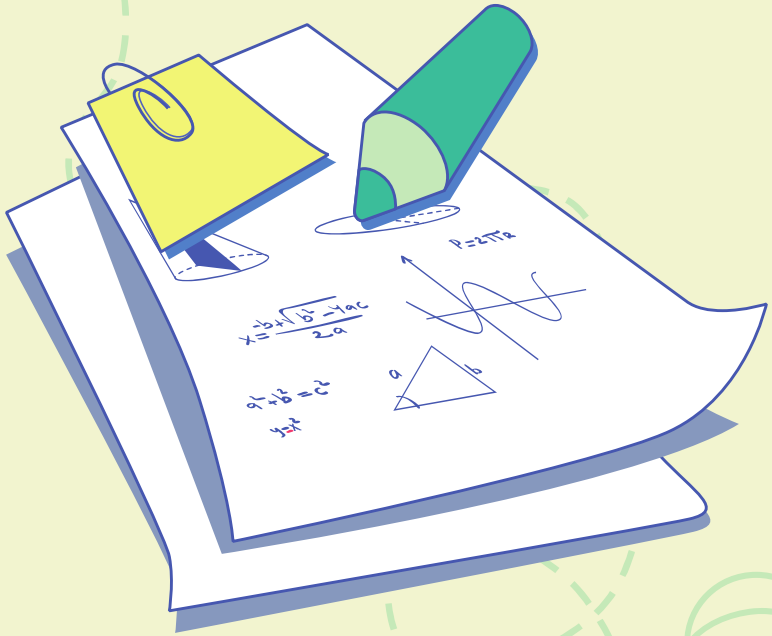
Ý NGHĨA CỦA N-GRAM?

Đo mức độ giống với ngôn ngữ bản xứ của một cụm từ hoặc câu

$$w = w_1 w_2 \dots w_n$$

Dựa trên xác suất p(w) của cụm từ/câu đó.

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



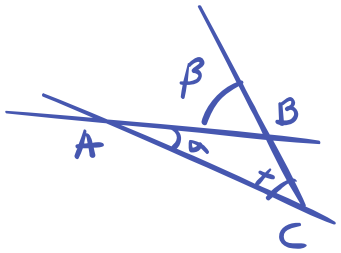
Bất khả thi khi thực hiện công thức trên khi n có thể lên đến hàng chục từ

Rất khó !!!

CS523.M21.KHCL

CÁCH GIẢI QUYẾT VẤN ĐỀ NÀY

X	Đưa cho đứa bên cạnh làm
X	“Khó quá bỏ qua”
V	Dùng Markov assumption



Giả định Markov...

cho phép ta có thể ước lượng
xác suất trên với n-1 từ trước đó

$$\sin\left(\frac{\pi}{2} - \alpha\right) - \cos\alpha$$

Công thức

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

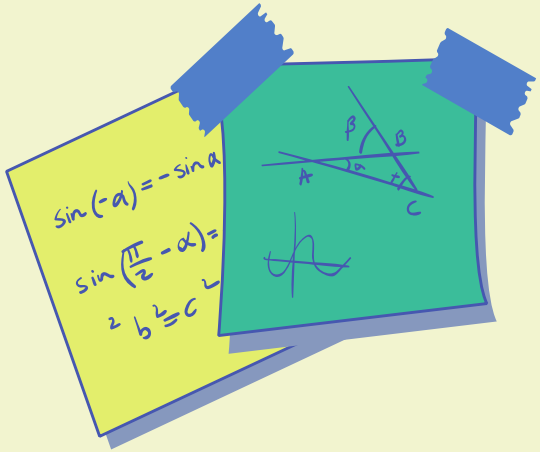
VÍ DỤ MINH HỌA

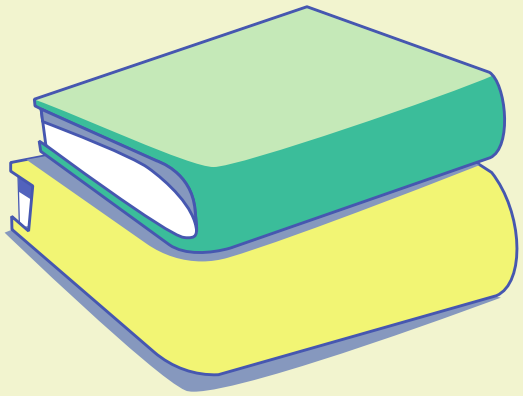
“Bà Bảy bán bánh tráng trộn”

“Bà Bảy bán **bánh tráng trộn**”

“..... bánh tráng _ _ _?”

Ước lượng xác suất p(wi|wi−n+1...wi−1)p(wi|wi−n+1...wi−1), dùng kỹ thuật MLE bằng cách ước lượng tỉ lệ phép đếm như sau:



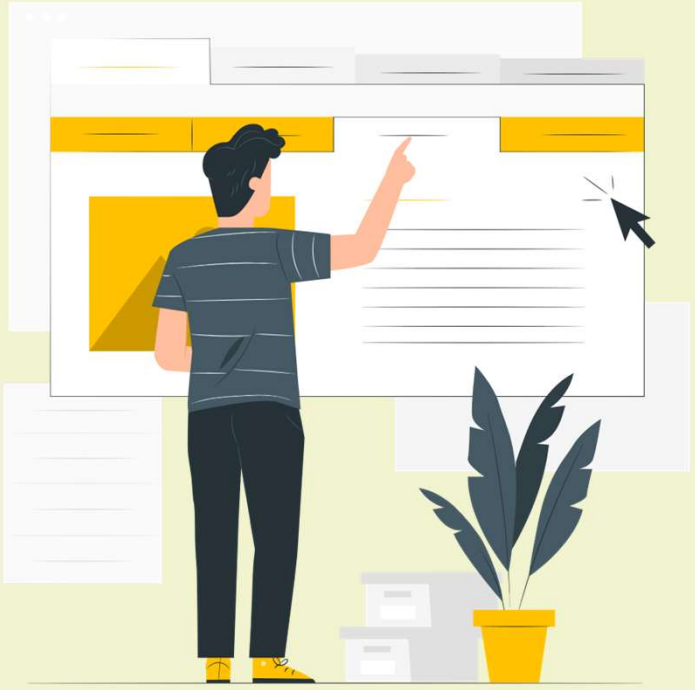


$$p(w_i|w_{i-n+1} \cdots w_{i-1}) = \frac{c(w_{i-n+1} \cdots w_i)}{c(w_{i-n+1} \cdots w_{i-1})}$$

ƯỚC LƯỢNG NHƯ THẾ NÀO?

- 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."



ƯỚC LƯỢNG NHƯ THẾ NÀO?

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

THIS IS A SOLUTION!

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

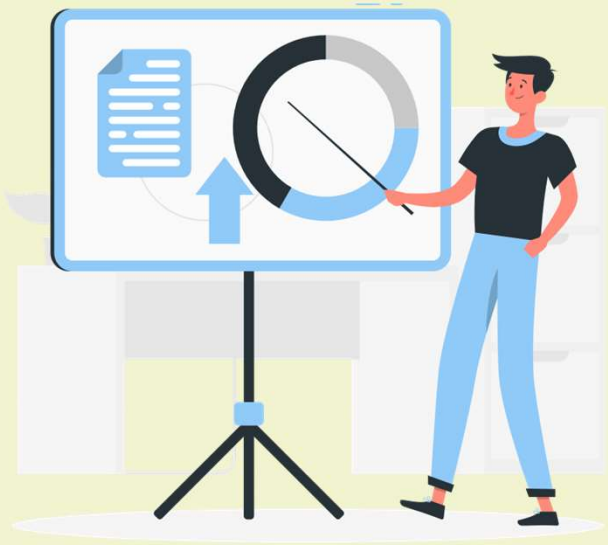
$$p(this) = \frac{c(this)}{c("allword")} = \frac{2}{11}$$

$$p(is|this) = \frac{c(this\ is)}{c(this\ \dots)} = \frac{2}{2} = 1$$

$$p(the|is) = \frac{c(is\ the)}{c(is\ \dots)} = \frac{2}{2} = 1$$

$$p(house|the) = \frac{c(the\ house)}{c(the\ \dots)} = \frac{1}{2}$$

(c là viết tắt của phép đếm "count")

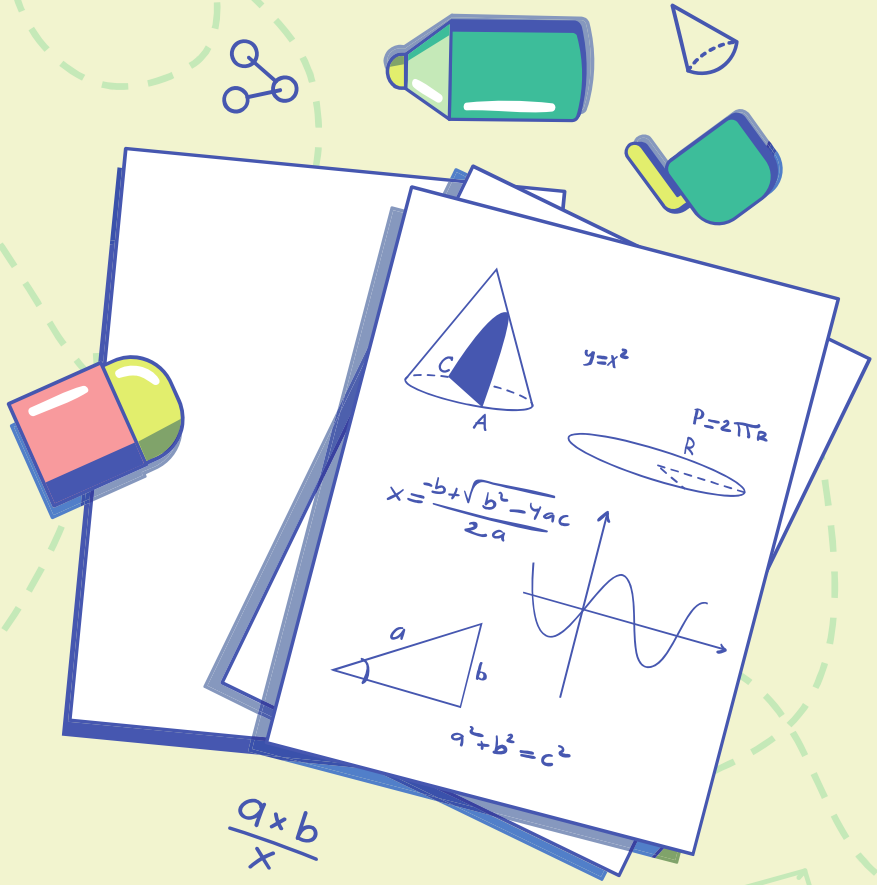


$$\boldsymbol{p(this\ is\ the\ house) = \frac{1}{11} \approx 0.091 = 9.1\%}$$

CS523.M21.KHCL

CS523.M21.KHCL

Đã có 1 vấn đề xuất hiện...



GIẢI PHÁP KHẮC PHỤC

$$p(this) = p(this| < s >)$$

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

Khi đó, tần suất p(this) được tính lại như sau:

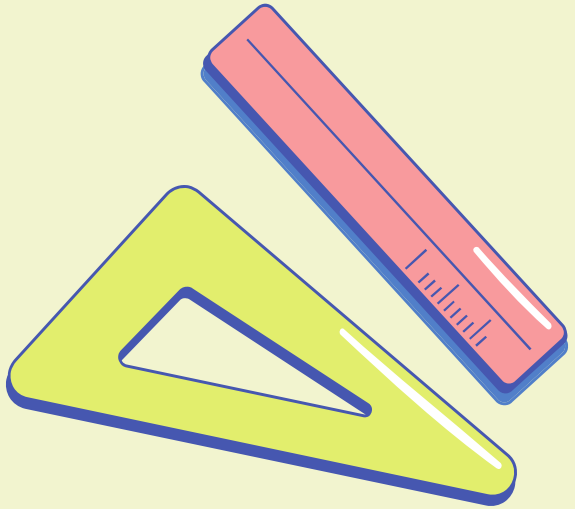
"This is the house that Jack built.

This is the malt"

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

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

NGOÀI RA | MỞ RỘNG



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:

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

NGOÀI RA | MỞ RỘNG

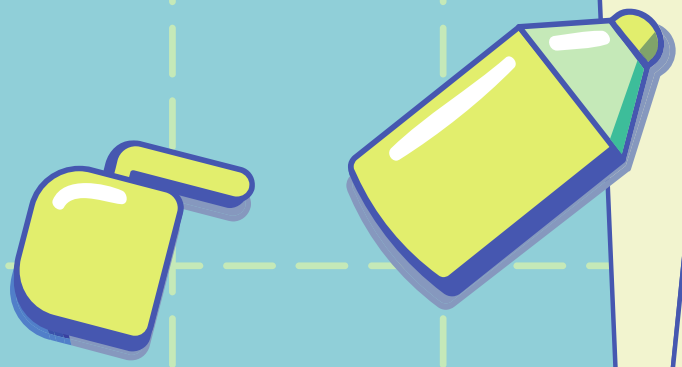
Corpus (để tránh trường hợp xác suất = 0, ta thêm câu "Jack in the house"):

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

<s> This is the malt </s>

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

KẾT QUẢ THU ĐƯỢC



$$p(this|<s>) = \frac{c(<s>\ this)}{c(<s>\ \dots)} = \frac{2}{3}$$

$$p(is|this) = \frac{c(this\ is)}{c(this\ \dots)} = \frac{2}{3}$$

$$p(the|is) = \frac{c(is\ the)}{c(is\ \dots)} = \frac{2}{3}$$

$$p(house|the) = \frac{c(the\ house)}{c(the\ \dots)} = \frac{1}{2}$$

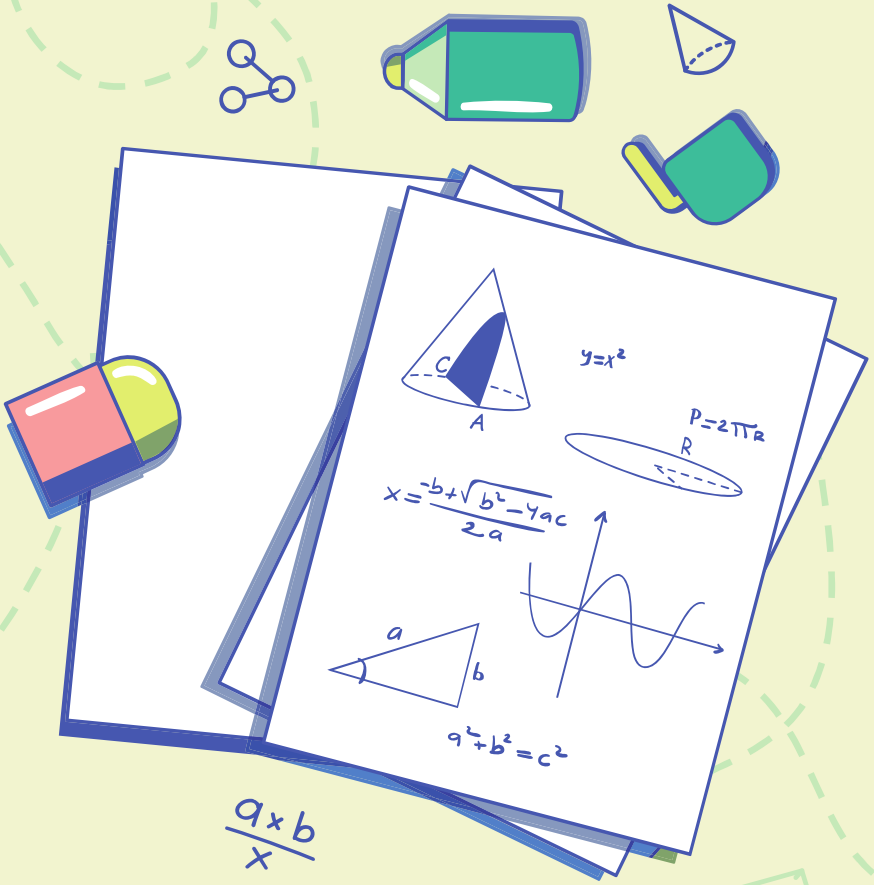
$$p(</s>|house) = \frac{c(house\ </s>)}{c(house\ \dots)} = \frac{1}{2}$$

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

Ta thấy:

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 **dữ liệu thưa (sparse data)**

Để xử lí, ta dùng **các kĩ thuật làm mịn (smoothing techniques)**



CS523.M21.KHCL

SMOOTHING TECHNIQUES

- 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** thành những cách tính/cách ước lượng khác.

CS523.M21.KHCL

SMOOTHING TECHNIQUES

- Kĩ thuật thêm-1
- Kĩ thuật thêm-alpha

CS523.M21.KHCL

KĨ 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
- Vấn đề: ảnh hưởng đến các tần suất khác làm cho các tần suất khác đột ngột trở nên thấp đi (do tăng tổng thể lên)
- Công thức:

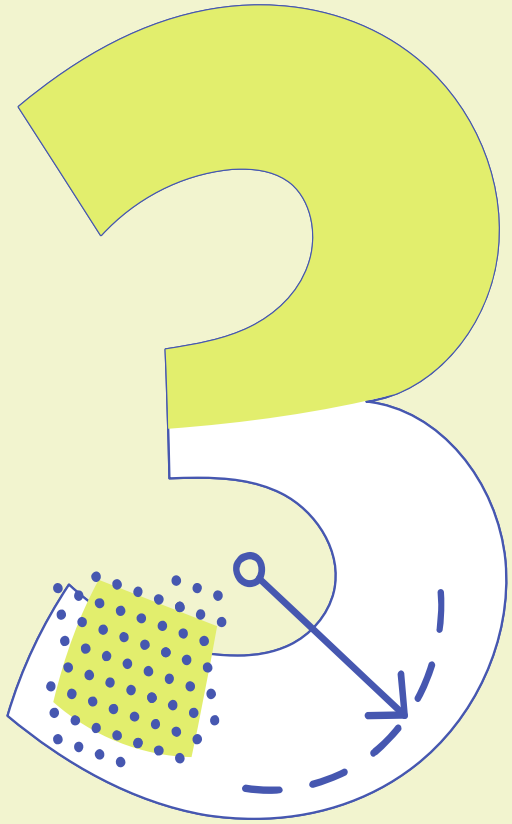
$$p = \frac{c+v}{n+v}$$

KĨ THUẬT THÊM-ALPHA

- Thêm-1 có nhược điểm là sẽ phí phạm không gian xác suất
- Để khắc phục nhược điểm này, ta thêm một lượng alpha<1 thay vì 1 cho mỗi n-gram xuất hiện trong ngữ liệu
- Công thức:

$$p = \frac{c + \alpha}{n + \alpha v}$$

CS523.M21.KHCL



03

WORD-LATTICE & DEMO

CS523.M21.KHCL

DEMO (づ￣ 3 ￣)づ

https://colab.research.google.com/drive/1BrzjuItlKqkuk3vN9m15xQMv7GB009WM?usp=sharing

CS523.M21.KHCL

04

ÔN TẬP & GIẢI ĐÁP THẮC MẮC

CS523.M21.KHCL

Thanks!

Slides + code are being updated on:
https://github.com/nguyenthinhuvan/CS523.M21.KHCL

If you have any question, please go to "google.com"

(Or contact our team ಥ_ಥ)