GÁN NHÃN DỮ LIỆU

Khoa Khoa học và Kỹ thuật thông tin
Bộ môn Khoa học dữ liệu

Nội dung chính

1. Tại sao cần gán nhãn dữ liệu.
2. Quy trình tổng quát để gán nhãn dữ liệu.
3. Người gán nhãn.
4. Đánh giá quy trình gán nhãn.

TẠI SAO PHẢI GÁN NHÃN DỮ LIỆU

Tại sao cần gán nhãn dữ liệu

- Trong lĩnh vực máy học hiện tại, đa phần các bài toán đều xoay quanh lĩnh vực học có giám sát.
  + Dữ liệu huấn luyện phải là dữ liệu có giám sát (có nhãn).
- Trong thực tế, dữ liệu chưa giám sát tồn tại nhiều hơn dữ liệu có giám sát.
- Mục tiêu: tạo ra các bộ dữ liệu có giám sát, phục vụ cho công việc huấn luyện mô hình máy học.

➔Gán nhãn (annotation) giúp tạo ra các bộ dữ liệu huấn luyện cho các bài toán máy học.

Các bài toán cần gán nhãn dữ liệu

- Bài toán nhận diện vật thể (object detection):
  - + *Input*: một ảnh.
  - + *Output*: vector xác định vị trí và kích thước một vật thể.
- Bài toán phân tích cảm xúc về sản phẩm:
  - + *Input*: câu bình luận về sản phẩm.
  - + *Output*: loại cảm xúc về sản phẩm (tích cực, tiêu cực, trung tính).
- Bài toán nhận diện ảnh X-quang phổi bị nhiễm COVID-19:
  - + *Input*: Ảnh X-quang phổi.
  - + *Output*: Tình trạng phổi: Bình thường, bị nhiễm bệnh.

Các bài toán cần gán nhãn dữ liệu

- Tất cả các bài toán trên đều yêu cầu phải có dữ liệu huấn luyện đã được gán nhãn sẵn cho mô hình máy học.
- Vấn đề đặt ra: Gán nhãn dữ liệu như thế nào là tốt?

VD:

+ Bình luận về sản phẩm ➔ phụ thuộc vào cảm xúc và quan điểm của 1 cá nhân.

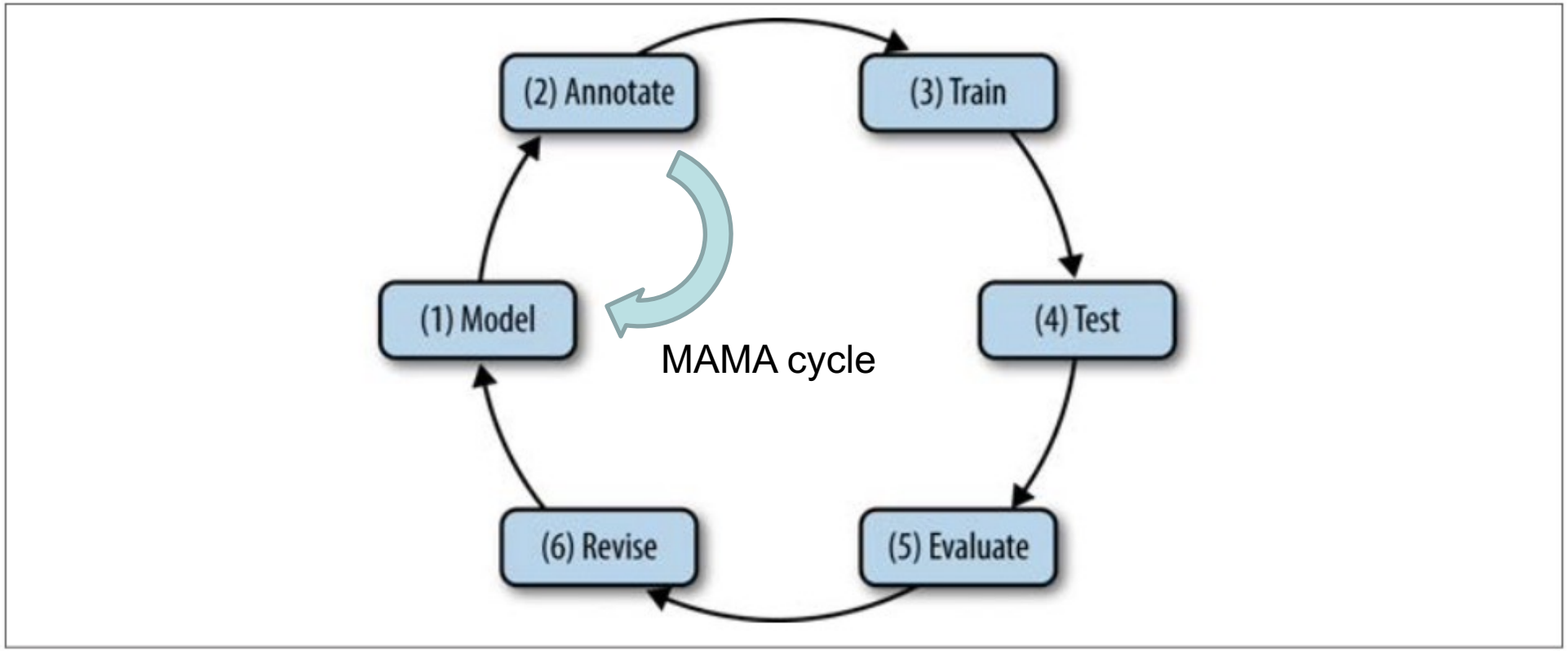
+ Ảnh X-quang về y khoa ➔ cần kiến thức chuyên môn về lĩnh vực y khoa.

➔ *Quy trình tạo, gán nhãn và đảm bảo chất lượng dữ liệu.*

QUY TRÌNH GÁN NHÃN DỮ LIỆU TỔNG QUÁT

Quy trình gán nhãn

– Quy trình MATTER - Pustejovsky and Stubbs (2013)



1. Model

- Định nghĩa các khái niệm và hiện tượng (Phenomenon) có trong dữ liệu ➔ nhãn của dữ liệu, hay còn gọi là định nghĩa tác vụ.
- Phụ thuộc vào bài toán đang giải quyết.
- Định nghĩa Model: M = <T,R,I>
  + *T: các terms có trong bộ dữ liệu.*
  + *R: quan hệ giữa các terms.*
  + *I: ý nghĩa của các terms.*
- Việc định nghĩa này cần được mô tả rõ ràng và chi tiết trong hướng dẫn gán nhãn.

Ví dụ 1

– Bài toán Spam detection:

+ T = {Document_type, Spam, Not-Spam}

+ R = {Document_type ::= Spam | Not-Spam}

+ I = {

• Spam = “something we don’t want!”,

• Not-Spam = “something we do want!”

}

Ví dụ 2

- Bài toán Name Entities Recognition
  - + T = {Named_Entity, Organization, Person, Place, Time}
  - + R = {Named_Entity ::= Organization | Person | Place | Time}
  - + I = {
    - Organization = “list of organizations in a database”,
    - Person = “list of people in a database”,
    - Place = “list of countries, geographic locations, etc.”,
    - Time = “all possible dates on the calendar”

    }

Hướng dẫn gán nhãn

- Hướng dẫn gán nhãn: là công cụ nhằm hướng dẫn các người gán nhãn (annotators) gán nhãn cho bộ dữ liệu nhằm mục tiêu đảm bảo sự thống nhất trong quá trình gán nhãn, và tránh các sai sót do hiểu sai về ngữ nghĩa hay do sự nhập nhằm gây nên.
- Các yêu cầu chính của một guidelines gán nhãn:
  + Định nghĩa rõ ràng các khái niệm và các thực thể cần gán nhãn.
  + Ví dụ minh hoạ cho các trường hợp.
  + Các trường hợp khó cần giải quyết.
- Hướng dẫn gán nhãn được cập nhật liên tục để phù hợp với thực tế khi tạo bộ dữ liệu.

Các yêu cầu cơ bản của một hướng dẫn gán nhãn

1. Mục tiêu của bài toán hay tác vụ là gì ?
2. Ý nghĩa của mỗi nhãn, hay tag, và sử dụng trong trường hợp nào (cho ví dụ cụ thể).
3. Phần nào cần gán nhãn, và phần nào để trống.
4. Gán nhãn như thế nào? (aka cách sử dụng công cụ gán nhãn).

Ví dụ: Movie Review

1. Mục tiêu của bài toán hay tác vụ là gì ?
   + Nhận biết xem một bình luận về phim là Tích cực, hay tiêu cực (2 nhãn).
2. Ý nghĩa của mỗi nhãn, hay tag, và sử dụng trong trường hợp nào (cho ví dụ cụ thể).
   + Có 2 nhãn là: positive và negative. Mỗi review sẽ được gán 1 trong 2 nhãn: positive hoặc negative dựa vào sắc thái của câu bình luận.
   + Nếu bình luận mang sắc thái tích cực, thì gán nhãn positive. Ngược lại, nếu bình luận mang sắc thái tiêu cực thì gán negative.

Ví dụ: Movie Review

3. Phần nào cần gán nhãn, và phần nào để trống.
   + Mỗi câu bình luận được gán một trong hai nhãn. Và gán toàn bộ các câu bình luận.
4. Gán nhãn như thế nào? (aka cách sử dụng công cụ gán nhãn).
   + Sử dụng công cụ google spreadsheet, và gán trực tiếp trên bảng tính.

2. Annotate

- Tiến hành huấn luyện cho các người gán nhãn (annotators) và cho người gán nhãn gán các nhãn cụ thể trong bộ dữ liệu.
- Việc huấn luyện, và gán nhãn là một quy trình liên tục, nhằm đảm bảo rằng những người gán nhãn đều hiểu đúng, và rõ ràng về hướng dẫn gán nhãn.
- Để đánh giá chất lượng của một hướng dẫn gán nhãn và quy trình gán nhãn có được định nghĩa rõ ràng hay không, ta dựa vào một độ đo gọi là độ đồng thuận (inter annotator agreement - IAA).
- Thông thường, một bộ dữ liệu được gán bởi ít nhất là 2 người độc lập ➔ đảm bảo tính khách quan.

BÀI TẬP

1. Hãy thử xây dựng guildelines gán nhãn để gán nhãn cho bộ dữ liệu dùng cho bài toán nhận diện giới tính dựa theo tên.
2. Hãy thử xây dựng guildelines gán nhãn để gán nhãn cho bộ dữ liệu dùng cho bài toán nhận diện chữ viết tay.

MỘT SỐ HƯỚNG DẪN GÁN NHÃN VÍ DỤ

- Penn Tree bank: https://sharedtasksinthedh.github.io/assets/howto-annotation/Penn-Treebank-Tagset.pdf
- TimeML: https://sharedtasksinthedh.github.io/assets/howto-annotation/timeml-1.2.1.pdf

Ví dụ 1

2 List of parts of speech with corresponding tag

Adjective—JJ

Hyphenated compounds that are used as modifiers are tagged as adjectives (JJ).

EXAMPLES: happy-go-lucky/JJ
one-of-a-kind/JJ
run-of-the-mill/JJ

Ordinal numbers are tagged as adjectives (JJ), as are compounds of the form *n-th X-est*, like *fourth-largest*.

Adjective, comparative—JJR

Adjectives with the comparative ending *-er* and a comparative meaning are tagged JJR. *More* and *less* when used as adjectives, as in *more or less mail*, are also tagged as JJR. *More* and *less* can also be tagged as JJR when they occur by themselves; see the entries for these words in Section 4.2. Adjectives with a comparative meaning but without the comparative ending *-er*, like *superior*, should simply be tagged as JJ. Adjectives with the ending *-er* but without a strictly comparative meaning ("more X"), like *further* in *further details*, should also simply be tagged as JJ.

Adjective, superlative—JJS

Adjectives with the superlative ending *-est* (as well as *worst*) are tagged as JJS. *Most* and *least* when used as adjectives, as in *the most or the least mail*, are also tagged as JJS. *Most* and *least* can also be tagged as JJS when they occur by themselves; see the entries for these words in Section 4.2. Adjectives with a superlative meaning but without the superlative ending *-est*, like *first*, *last* or *unsurpassed*, should simply be tagged as JJ.

Minh hoạ hướng dẫn gán nhãn trong bộ Penn Tree banks (PTS)

Ví dụ 2

2.1.1 How to annotate EVENTs

Events may be expressed by means of tensed or untensed verbs (1 and 2), nominalizations (3), adjectives (4), predicative clauses (5), or prepositional phrases (6):

1. *A fresh flow of lava, gas and debris* **erupted** *there Saturday.*
2. *Prime Minister Benjamin Netanyahu called the prime minister of the Netherlands* **to thank** *him for thousands of gas masks his country has already contributed.*
3. *Israel will ask the United States to delay a military* **strike** *against Iraq until the Jewish state is fully prepared for a possible Iraqi* **attack**.
4. *A Philippine volcano,* **dormant** *for six centuries, began exploding with searing gases, thick ash and deadly debris.*
5. *"There is no reason why we would not* **be prepared**,*" Mordechai told the Yediot Ahronot daily.*
6. *All 75 people* **on board** *the Aeroflot Airbus died.*

Note that in the above sentences not all "markables" are tagged. In the first example, for instance, neither *flow* nor *Saturday* is marked.

The annotation of formally simple events (examples 1, 3, 4 and 6 above) is straightforward. However, formally complex events may be sequentally discontinuous in some contexts:

1. *There is no reason why we would not* **be prepared.**
   There is no reason why we would not **be** *fully* **prepared.**

Minh hoạ hướng dẫn gán nhãn trong bộ TimeML

3. Train and Test

- Sử dụng các thuật toán máy học để huấn luyện, và đánh giá hiệu suất trên bộ dữ liệu.
- Dữ liệu cần được chia thành các tập: train, dev và test.
  + Train: tập huấn luyện.
  + Dev: tập phát triển.
  + Test: Tập kiểm thử.
- Các tập dữ liệu khi đã được phân chia phải có cùng phân bố.

4. Evalutation

- Sử dụng các độ đo đánh giá dùng để đánh giá tính hiệu quả của mô hình máy học.
- Các độ đo thông dụng:
  + Accuracy (Acc)

  Acc=(TP+TN)/(TP+TN+FP+FN).

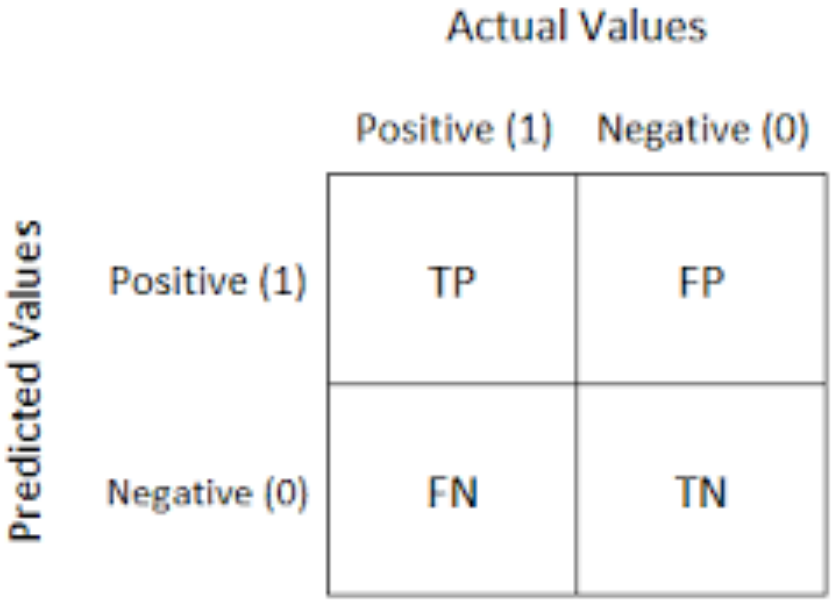
  + Precision (P)

  P=TP/(TP+FP).

  + Recall (R)

  R=TP/(TP+FN).

  + F1-score = 2 * {(P*R) / (P+R)}.

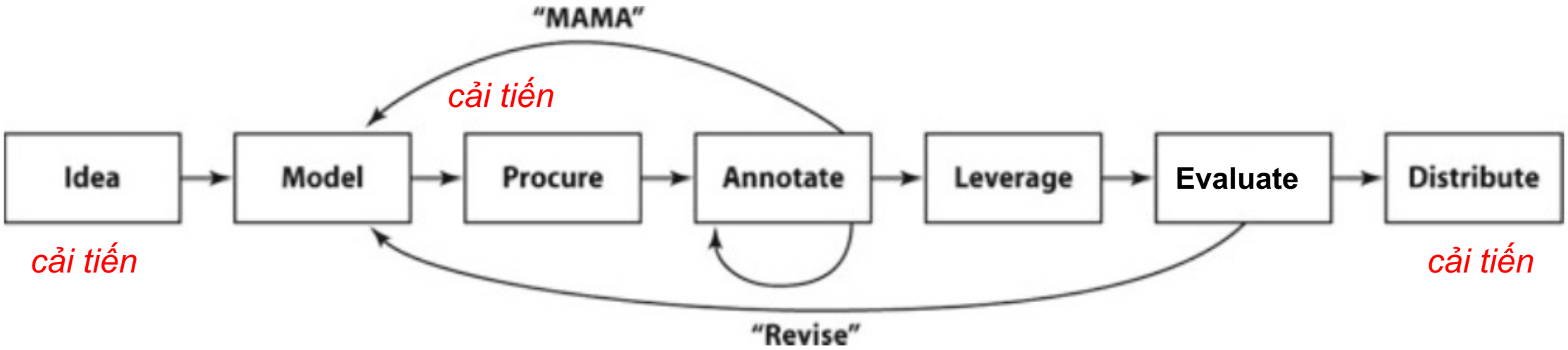


ma trận nhầm lẫn (Confusion matrix)

5. Revise

- Kiểm tra và đánh giá lại tính hiệu quả của quy trình gán nhãn.
- Công việc chính: phân tích lỗi (error analysis).
  - + Phân tích các dự đoán sai ➔ tìm nguyên nhân cụ thể.
  - + Phân tích các tình huống khó giải quyết.
  - + Đề xuất hướng giải quyết tiếp theo.

Cải tiến quy trình MATTER



Finlayson and Erjavec (2017), Overview of annotation creation: Processes and tools
https://link.springer.com/content/pdf/10.1007%2F978-94-024-0881-2_5.pdf

Idea

- Khảo sát các bộ dữ liệu tương tự với bài toán của mình.
- Tham khảo annotation guidelines của các bộ dữ liệu có sẵn.
- Kế thừa các định nghĩa và nhãn từ các bộ dữ liệu có sẵn để tiết kiệm thời gian xây dựng hướng dẫn gán nhãn.

Procure

- Kiểm tra các tác vụ con có thể xảy ra khi gán nhãn.
- Tìm hoặc xây dựng các công cụ gán nhãn để phục vụ cho công việc gán nhãn.
- Chỉnh sửa và tối ưu công cụ gán nhãn.
- Một số công cụ tham khảo:
  + https://labelstud.io/
  + http://doccano.herokuapp.com/
  + Google sheet / Google form.

Distribute

- Công bố data đã gán nhãn cho mọi người cùng sử dụng.
- Các vấn đề cần quan tâm:
  + Quyền riêng tư của dữ liệu.
  + Thoả thuận sử dụng dữ liệu.
  + Các quyền hạn và yêu cầu cụ thể khi sử dụng dữ liệu.

NGƯỜI GÁN NHÃN (ANNOTATORS)

Annotators và crowdsourcing

- Annotators (tạm dịch là người gán nhãn): là những người đọc và gán nhãn cho dữ liệu bằng kiến thức và kinh nghiệm của cá nhân theo một hướng dẫn đã được mô tả trước.
- Công việc của annotators là công việc khó, đòi hỏi nhiều công sức và trí lực khi làm. Vai trò của Annotators là tạo ra các bộ dữ liệu chất lượng để phục vụ cho các hệ thống thông minh.

Data Annotators: The Unsung Heroes Of Artificial Intelligence Development

Link: https://medicalfuturist.com/data-annotation/

Phân loại annotators

- Annotators là chuyên gia (expert): là những annotator có kiến thức sâu rộng, và uyên bác về lĩnh vực của tác vụ mà bộ dữ liệu đang xây dựng.

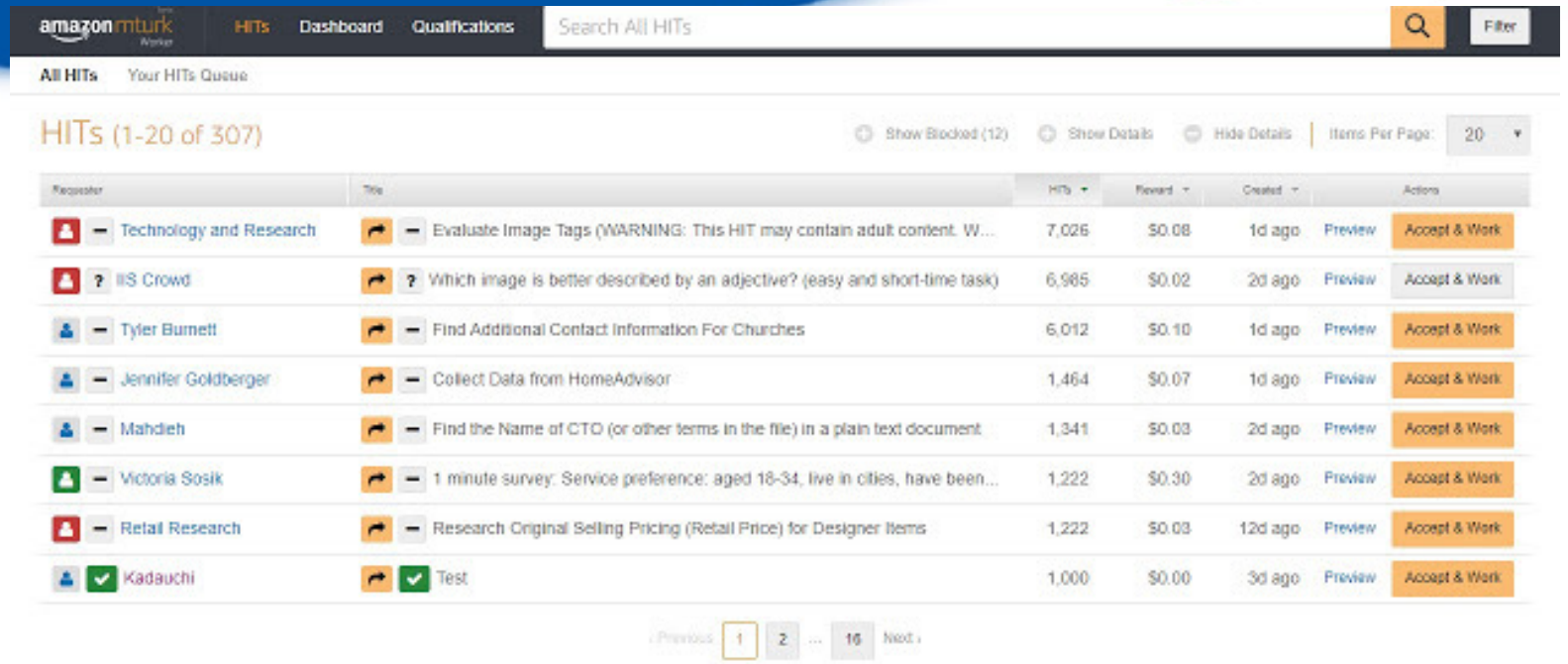
VD: bác sĩ, chuyên gia ngôn ngữ.

- Annotators không phải chuyên gia (non-expert): là những annotator bình thường, có kiến thức nhất định.

VD: học sinh phổ thông, người lao động bình thường.

Crowdsourcing

Crowdsourcing là một hình thức tận dụng sức mạnh của đám đông. Trong đó, bộ dữ liệu sẽ được chia ra làm các phần nhỏ, mỗi phần nhỏ sẽ do một hay một nhóm người phụ trách gán nhãn thay vì gán toàn bộ trên bộ dữ liệu. Việc tận dụng sức mạnh của đám đông giúp xây dựng các bộ dữ liệu có gán nhãn lớn và rất lớn một cách nhanh chóng và hiệu quả.



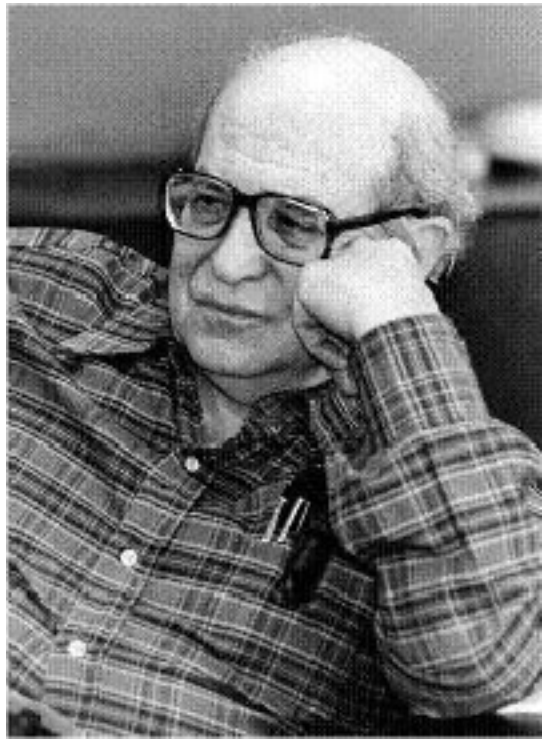
Amazon Mechanical Turk

https://www.mturk.com/

ĐÁNH GIÁ QUY TRÌNH GÁN NHÃN

Độ đồng thuận gán nhãn - IAA

- Độ đồng thuận gán nhãn (Inter annotator agreement – IAA) là một độ đo nhằm đánh giá khả năng nhiều annotator gán cùng một giá trị (hay còn gọi là nhãn - label) cho một dữ liệu nhất định.
- Để đánh giá độ đồng thuận, độ đo Cohen's Kappa là độ đo được sử dụng phổ biến nhất.
- Ngoài Cohen's Kappa, còn có các biến thể khác của công thức trên như: *Fleiss's Kappa và Krippendorff's Alpha.*



Jacob Cohen (1923 - 1998)

Công thức Cohen's Kappa

- Với 1 cặp annotators, cùng gán nhãn cho một dữ liệu d.
- Công thức chung: $\kappa = \frac{\Pr(A) - \Pr(e)}{1 - \Pr(e)}$
- Ý nghĩa các thông số:
  + Pr(a): Giá trị đồng thuận quan sát được giữa các nhãn.
  + Pr(e): Xác suất giả định của khả năng đồng thuận.

Ví dụ

Đối với bài toán gán nhãn cho dữ liệu bình luận phim, giả sử có 2 annotator lần lượt là A và B cùng gán cho 250 câu bình luận độc lập. Bảng bên phải cho biết số lượng nhãn đã gán ứng với từ annotators.

N = 250

		B positive	B neutral	B negative
A	positive	54	28	3
A	neutral	31	18	23
A	negative	0	21	72

Tính Pr(A)

- Nhãn positive: A và B đồng thuận nhau 54 nhãn.
- Nhãn neutral: A và B đồng thuận nhau 18 nhãn.
- Nhãn negative: A và B đồng thuận nhau 72 nhãn.

➔Xác suất đồng thuận của 2 người là:

Pr(a) = (54+18+72)/250 = 0.576

N = 250

		B positive	B neutral	B negative
A	positive	54	28	3
A	neutral	31	18	23
A	negative	0	21	72

Tính Pr(e)

– Xác suất giả định đồng thuận của 2 annotator trên 3 nhãn:

Pr(e) = Pr(postive) + Pr(neutral) + Pr(negative)

N = 250

		B positive	B neutral	B negative
A	positive	54	28	3
A	neutral	31	18	23
A	negative	0	21	72

Tính Pr(positive)

- Xác suất giả định xảy ra đồng thuận giữa 2 người A và B (2 người là độc lập nhau) trên nhãn positive.
- Đối với nhãn positive:
  - Pr(**A**|postive) = (54+28+3)/250 = 0.34.
  - Pr(**B**|postive) = (54+31+0)/250 = 0.34.
- Xác suất xảy ra đồng thuận giữa 2 người trên nhãn postive:

Pr(positive) = Pr(**A**|postive) x Pr(**B**|postive)

= 0.34 x 0.34

= 0.1156

N = 250

		B	B	B
		positive	neutral	negative
A	positive	54	28	3
A	neutral	31	18	23
A	negative	0	21	72

Tính Pr(neutral)

- Xác suất giả định xảy ra đồng thuận giữa 2 người A và B (2 người là độc lập nhau) trên nhãn neutral.
- Đối với nhãn positive:
  - Pr(**A**|neutral) = (31+18+23)/250 = 0.288.
  - Pr(**B**| neutral) = (28+18+21)/250 = 0.268.
- Xác suất xảy ra đồng thuận giữa 2 người trên nhãn postive:

Pr(neutral) = Pr(**A**|neutral) x Pr(**B**|neutral)

= 0.288 x 0.268

= 0.077

N = 250

		B positive	B neutral	B negative
A	positive	54	28	3
A	neutral	31	18	23
A	negative	0	21	72

Tính Pr(negative)

- Xác suất giả định xảy ra đồng thuận giữa 2 người A và B (2 người là độc lập nhau) trên nhãn negative.
- Đối với nhãn positive:
  - Pr(**A**|negative) = (0+21+72)/250 = 0.372.
  - Pr(**B**|negative) = (3+23+72)/250 = 0.392.
- Xác suất xảy ra đồng thuận giữa 2 người trên nhãn negative :

Pr(negative) = Pr(**A**|negative) x Pr(**B**|negative)

= 0.372 x 0.392

= 0.146

N = 250

		B	B	B
		positive	neutral	negative
A	positive	54	28	3
A	neutral	31	18	23
A	negative	0	21	72

Tính Pr(e)

– Xác suất giả định đồng thuận của 2 annotator trên 3 nhãn:

Pr(e) = Pr(postive) + Pr(neutral) + Pr(negative)

= 0.1156 + 0.077 + 0.146

= 0.339

N = 250

		B positive	B neutral	B negative
A	positive	54	28	3
A	neutral	31	18	23
A	negative	0	21	72

Độ đồng thuận theo Cohen's Kappa

- Độ đồng thuận giữa 2 người A và B.

$$\kappa = \frac{\Pr(A) - \Pr(e)}{1 - \Pr(e)}$$

$$= \frac{0.576 - 0.339}{1 - 0.339}$$

$$\kappa = 0.36$$

N = 250

		B positive	B neutral	B negative
A	positive	54	28	3
A	neutral	31	18	23
A	negative	0	21	72

Fleiss Kappa

- Công thức này là một biến thể của Cohen's Kappa, dùng để tính cho nhiều annotator khác nhau.
- Công thức Fleiss's Kappa: $\kappa = \frac{\overline{\Pr(A)} - \overline{\Pr(e)}}{1 - \overline{\Pr(e)}}$
- Trong đó:
  - + $\overline{\Pr(A)} = \frac{1}{N}\sum_{i=1}^{N} P_i = \frac{1}{Nn(n-1)}\sum_{i=1}^{N}\sum_{j}^{k} n_{ij}^2 - Nn$
  - + $\overline{\Pr(e)} = \sum_{j=1}^{k} p_j^2$

$$P_i = \frac{1}{n(n-1)}\sum_{j}^{k} n_{ij}^2 - n$$

$$p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}, 1 = \sum_{j=1}^{k} p_j$$

Joseph L. Fleiss (1937 – 2003)

N: số lượng annotator,

k: số nhãn, n: tổng số lượng đã gán bởi 1 annotator trên 1 nhãn.

Ví dụ (1)

Annotator	positive	neutral	negative	P_i
A	85	72	93	0.3343
B	85	67	98	
C	68	99	83	
D	88	88	74	
E	58	120	72	
Tổng	384	446	420	
P_j				

n = 250
k = 3
N=5

$$P_A = \frac{1}{n(n-1)} \sum_j^k n_{ij}^2 - n = \frac{(85^2+72^2+93^2)-250}{250(250-1)} = 0.3343$$

Ví dụ (2)

Annotator	positive	neutral	negative	P_i
A	85	72	93	0.3343
B	85	67	98	0.3384
C	68	99	83	
D	88	88	74	
E	58	120	72	
Tổng	384	446	420	
P_j				

n = 250
k = 3
N=5

$$P_B = \frac{1}{n(n-1)} \sum_j^k n_{ij}^2 - n = \frac{(85^2+67^2+98^2)-250}{250(250-1)} = 0.3384$$

Ví dụ (3)

Annotator	positive	neutral	negative	P_i
A	85	72	93	0.3343
B	85	67	98	0.3384
C	68	99	83	0.3384
D	88	88	74	
E	58	120	72	
Tổng	384	446	420	
P_j				

n = 250
k = 3
N=5

$$P_C = \frac{1}{n(n-1)} \sum_j^k n_{ij}^2 - n = \frac{(68^2+99^2+83^2)-250}{250(250-1)} = 0.3384$$

Ví dụ (4)

Annotator	positive	neutral	negative	P_i
A	85	72	93	0.3343
B	85	67	98	0.3384
C	68	99	83	0.3384
D	88	88	74	0.3328
E	58	120	72	
Tổng	384	446	420	
P_j				

n = 250
k = 3
N=5

$$P_D = \frac{1}{n(n-1)} \sum_j^k n_{ij}^2 - n = \frac{(88^2+88^2+74^2)-250}{250(250-1)} = 0.3328$$

Ví dụ (5)

Annotator	positive	neutral	negative	P_i
A	85	72	93	0.3343
B	85	67	98	0.3384
C	68	99	83	0.3384
D	88	88	74	0.3328
E	58	120	72	0.3646
Tổng	384	446	420	
P_j				

n = 250
k = 3
N=5

$$P_E = \frac{1}{n(n-1)} \sum_j^k n_{ij}^2 - n = \frac{(58^2+120^2+72^2)-250}{250(250-1)} = 0.3646$$

Ví dụ (6)

Annotator	positive	neutral	negative	P_i
A	85	72	93	0.3343
B	85	67	98	0.3384
C	68	99	83	0.3384
D	88	88	74	0.3328
E	58	120	72	0.3646
Tổng	384	446	420	
P_j	0.3072			

n = 250
k = 3
N=5

$$p_{positive} = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij} = \frac{(85 + 85 + 68 + 88 + 58)}{5 * 250} = 0.3072$$

Ví dụ (7)

Annotator	positive	neutral	negative	P_i
A	85	72	93	0.3343
B	85	67	98	0.3384
C	68	99	83	0.3384
D	88	88	74	0.3328
E	58	120	72	0.3646
Tổng	384	446	420	
P_j	0.3072	0.3568		

n = 250
k = 3
N=5

$$p_{neutral} = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij} = \frac{(72 + 67 + 99 + 88 + 120)}{5 * 250} = 0.3568$$

Ví dụ (8)

Annotator	positive	neutral	negative	P_i
A	85	72	93	0.3343
B	85	67	98	0.3384
C	68	99	83	0.3384
D	88	88	74	0.3328
E	58	120	72	0.3646
Tổng	384	446	420	
P_j	0.3072	0.3568	0.336	

n = 250
k = 3
N=5

$$p_{negative} = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij} = \frac{(93+98+83+74+72)}{5*250} = 0.336$$

Ví dụ (9)

Annotator	positive	neutral	negative	P_i
A	85	72	93	0.3343
B	85	67	98	0.3384
C	68	99	83	0.3384
D	88	88	74	0.3328
E	58	120	72	0.3646
Tổng	384	446	420	
P_j	0.3072	0.3568	0.336	

n = 250
k = 3
N=5

$$\overline{Pr(a)} = \frac{1}{N}\sum_{i=1}^{N} P_i = \frac{0.3343 + 0.3384 + 0.3384 + 0.3328 + 0.3646}{5} = 0.3417$$

Ví dụ (10)

Annotator	positive	neutral	negative	P_i
A	85	72	93	0.3343
B	85	67	98	0.3384
C	68	99	83	0.3384
D	88	88	74	0.3328
E	58	120	72	0.3646
Tổng	384	446	420	
P_j	0.3072	0.3568	0.336	

n = 250
k = 3
N=5

$$\overline{Pr(e)} = \sum_{j=1}^{k} p_j^2 = 0.3072^2 + 0.3568^2 + 0.336^2 = 0.335$$

Ví dụ (11)

Annotator	positive	neutral	negative	P_i
A	85	72	93	0.3343
B	85	67	98	0.3384
C	68	99	83	0.3384
D	88	88	74	0.3328
E	58	120	72	0.3646
Tổng	384	446	420	
P_j	0.3072	0.3568	0.336	

n = 250
k = 3
N=5

$$\kappa = \frac{\overline{\Pr(A)} - \overline{\Pr(e)}}{1 - \overline{\Pr(e)}} = \frac{0.3417 - 0.335}{1 - 0.335} = 0.01$$

Các khoảng giá trị của Kappa

κ	Agreement level
< 0	poor
0.01–0.20	slight
0.21–0.40	fair
0.41–0.60	moderate
0.61–0.80	substantial
0.81–1.00	perfect

Landis and Koch 1977

Tổng kết

1. Tạo dữ liệu là một tập hợp gồm các quy trình và kỹ thuật nhằm tạo ra bộ dữ liệu chất lượng dùng cho nhiều mục tiêu khác nhau, bao gồm huấn luyện cho các mô hình học máy có giám sát.
2. Quy trình tạo dữ liệu: Quy trình MATTER.
3. Hướng dẫn gán nhãn.
4. Đánh giá quy trình tạo dữ liệu: độ đồng thuận:
   + Cohen's Kappa.
   + Fleiss's Kappa.

Tài liệu tham khảo

1. James Pustejovsky and Amber Stubbs, Natural Language Annotation, O'Reilly (2012).
2. Finlayson, M. A., & Erjavec, T. (2017). Overview of annotation creation: Processes and tools. In *Handbook of Linguistic Annotation* (pp. 167-191). Springer, Dordrecht.
3. *Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". Educational and Psychological Measurement.* ***20*** *(1): 37–46.*

Bài tập

Tính độ đồng thuận bằng công thức Cohen's Kappa cho bộ dữ liệu gồm 206 điểm dữ liệu và 3 nhãn (0,1,2), được gán bởi 2 người là Person A và Person B như sau:

		Person B		
		0	**1**	**2**
Person A	**0**	94	1	0
	1	22	14	7
	2	28	6	34

Nhận xét về độ đồng thuận của bộ dữ liệu trên.