Department of Linguistics and Translation
香港城市大學
City University of Hong Kong

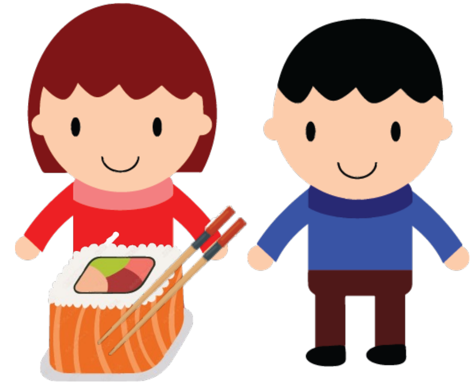
Computational Linguistics LT3233

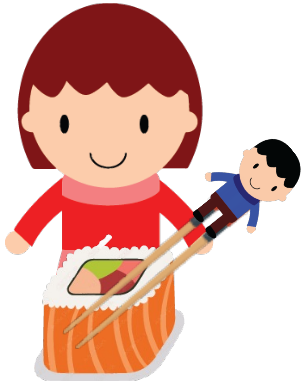
Jixing Li

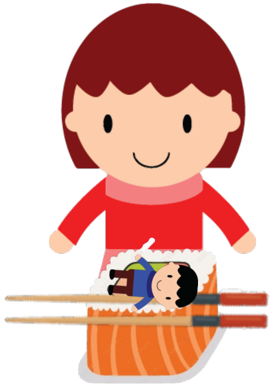
Lecture 6: Naive Bayes

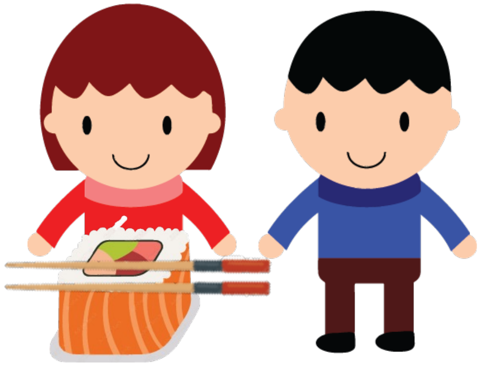
Slides adapted from Dan Jurafsky

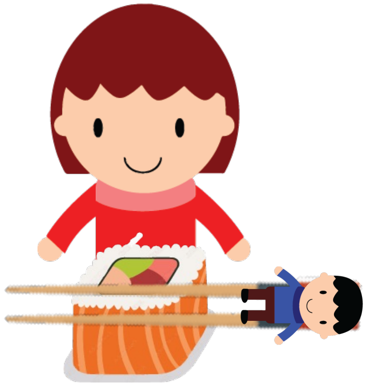
I eat sushi with chopsticks with you











Lecture plan

- The task of text classification
- The Naive Bayes classifier
- Evaluation metrics
- Short break (15 mins)
- Hands-on exercises

Positive or negative movie review?

- Good ...zany characters and richly applied satire, and some great plot twists
- Bad It was pathetic. The worst part about it was the boxing scenes...
- Good ...awesome caramel sauce and sweet toasty almonds. I love this place!
- Bad ...awful pizza and ridiculously overpriced...

→ Sentiment analyses

What is the subject of the medical article?

Available online at www.sciencedirect.com

ELSEVIER

Brain and Cognition

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

Susanne Gahl,[a] Lise Menn,[b] Gail Ramsberger,[b] Daniel S. Jurafsky,[b] Elizabeth Elder,[b] Molly Rewega,[a] and L. Holland Audrey[a]

[illegible]

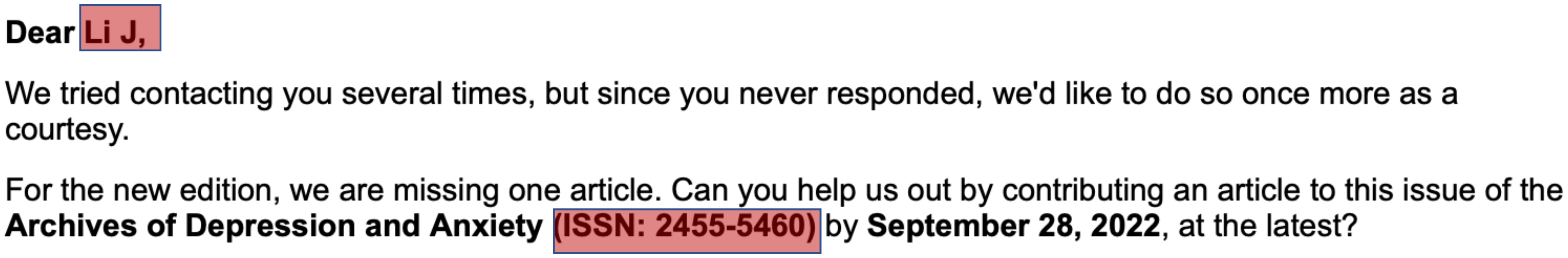
?

Subject category

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

© Jixing Li

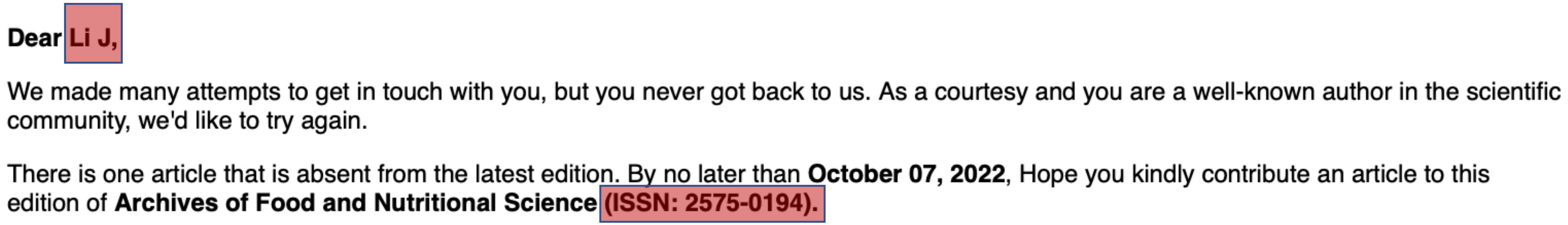
Spam email?



Dear Li J,

We tried contacting you several times, but since you never responded, we'd like to do so once more as a courtesy.

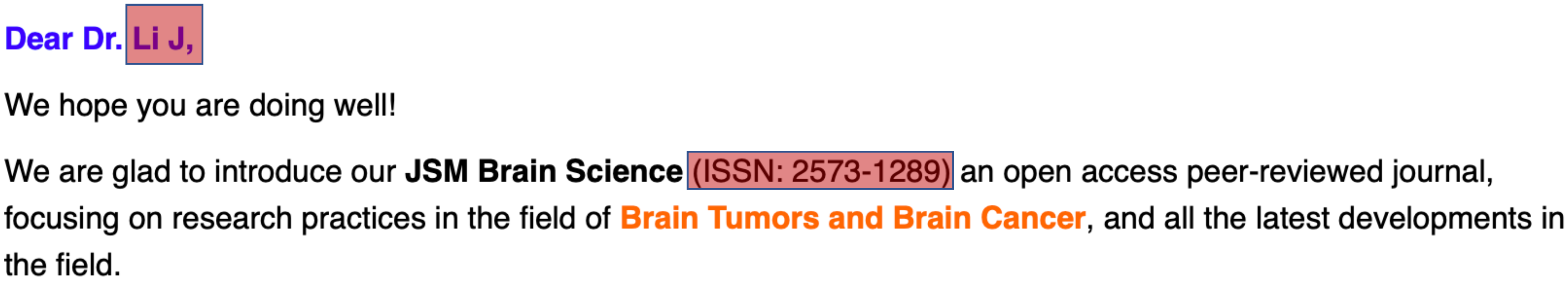
For the new edition, we are missing one article. Can you help us out by contributing an article to this issue of the **Archives of Depression and Anxiety (ISSN: 2455-5460)** by **September 28, 2022**, at the latest?



Dear Li J,

We made many attempts to get in touch with you, but you never got back to us. As a courtesy and you are a well-known author in the scientific community, we'd like to try again.

There is one article that is absent from the latest edition. By no later than **October 07, 2022**, Hope you kindly contribute an article to this edition of **Archives of Food and Nutritional Science (ISSN: 2575-0194).**



Dear Dr. Li J,

We hope you are doing well!

We are glad to introduce our **JSM Brain Science** (ISSN: 2573-1289) an open access peer-reviewed journal, focusing on research practices in the field of **Brain Tumors and Brain Cancer**, and all the latest developments in the field.

Tell gender by name?

- Maxie
- Becky
- Rocky
- Gary
- Eve
- Josh
- Dana
- Christopher
- Julia
- Sam

- 歐承璋
- 李思穎
- 陳敏琪
- 廖倚琳
- 吴建瑞
- 馮紫晴
- 廖卓楠
- 徐婉晴
- 周咏楠
- 馬卓妍

© Jixing Li

Summary: Text identification

- Sentiment analysis
- Spam detection
- Assigning subject categories, topics, or genres
- Gender identification
- …

Input:

a document d

a fixed set of classes $C = \{c_1, c_2, \ldots, c_j\}$

Output: a predicted class $c \in C$

Classification methods: Hand-coded rules

- Rules based on combinations of words or other features

  spam: black-list-address OR (“ISSN:” AND “LI. J”)

- Accuracy can be high

  If rules carefully refined by expert

- But building and maintaining these rules is expensive

Supervised machine learning

Input:

- a document d
- a fixed set of classes $C = \{c_1, c_2, \ldots, c_j\}$
- A training set of m hand-labeled documents $(d_1,c_1),\ldots\ldots,(d_m,c_m)$

Output:

- a learned classifier $\gamma{:}d \to c$

Many kinds of classifiers

- **Naive Bayes**
- Logistic regression
- Neural networks
- k-Nearest Neighbors
- …

Bayes rule

For a document *d* and a class *c*:

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

$$P(male|\text{卓琳}) = \frac{P\left(\text{卓琳}\,\middle|\,male\right)P(male)}{\cancel{P(\text{卓琳})}}$$

$$P(female|\text{卓琳}) = \frac{P\left(\text{卓琳}\,\middle|\,female\right)P(female)}{\cancel{P(\text{卓琳})}}$$

• 歐承璋	M
• 李思穎	F
• 陳敏琪	F
• 廖倚琳	F
• 吴建瑞	M
• 馮紫晴	F
• 廖卓楠	M
• 徐婉晴	F
• 周咏楠	F
• 馬卓妍	F
• 袁卓琳	?

Naive Bayes classifier

$$P(male|卓琳) = \frac{P(卓琳|male)P(male)}{\cancel{P(卓琳)}} \qquad P(female|承璋) = \frac{P(卓琳|female)P(female)}{\cancel{P(卓琳)}}$$

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c|d)$$

MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d|c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}} P(d|c)P(c)$$

Dropping the denominator

Calculate probability

$$\mathrm{P}\ (\mathrm{male}|\text{卓琳}) = \mathrm{P}(\text{卓琳}|\mathrm{male})\mathrm{P}(\mathrm{male})$$

$$\mathrm{P}\ (\mathrm{female}|\text{卓琳}) = \mathrm{P}(\text{卓琳}|\mathrm{female})\mathrm{P}(\mathrm{female})$$

$$\mathrm{P}(\mathrm{female}) = \frac{7}{10} \quad \mathrm{P}(\mathrm{male}) = \frac{3}{10}$$

卓琳= [卓, 琳] → **features**

$$\mathrm{P}(\text{卓琳}|\mathrm{female}) \approx \mathrm{P}(\text{卓}|\mathrm{female})\ \mathrm{P}(\text{琳}|\mathrm{female})$$

$$\mathrm{P}(\text{卓}|\mathrm{female}) = \frac{Count(\text{卓 in } female\ names)}{Count(all\ chatacters\ in\ female\ names)} = \frac{1}{14}$$

$$\mathrm{P}(\text{琳}|\mathrm{female}) = \frac{Count(\text{琳 in } female\ names)}{Count(all\ chatacters\ in\ female\ names)} = \frac{1}{14}$$

$$\mathrm{P}(\mathrm{female}|\text{卓琳}) = \frac{1}{14}\ \mathrm{x}\ \frac{1}{14} = \frac{1}{196}$$

歐承璋	M
李思穎	F
陳敏琪	F
廖倚琳	F
吳建瑞	M
馮紫晴	F
廖卓楠	M
徐婉晴	F
周咏楠	F
馬卓妍	F
袁卓琳	?

Naive Bayes classifier

"Likelihood"

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(d \mid c) P(c)$$

"Prior"

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

Document *d* represented as features *x₁..xₙ*

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

Calculate probability

$$P\ (\text{male}|\text{卓琳}) = P(\text{卓琳}|\text{male})P(\text{male})$$

$$P(\text{卓琳}|\text{male})\ \approx P(\text{卓}|\text{male})\ P(\text{琳}|\text{male})$$

$$P(\text{卓}|\text{male}) = \frac{Count(\text{卓 in } male\ names)}{Count(all\ chatacters\ in\ male\ names)} = \frac{1}{6}$$

$$P(\text{琳}|\text{male}) = \frac{Count(\text{琳 in } male\ names)}{Count(all\ chatacters\ in\ male\ names)} = \frac{0}{6}$$

$$P(\text{卓琳}|\text{male}) = \frac{1}{6}\text{x}\frac{0}{6} = 0$$

Problem?

歐承璋	M
李思穎	F
陳敏琪	F
廖倚琳	F
吴建瑞	M
馮紫晴	F
廖卓楠	M
徐婉晴	F
周咏楠	F
馬卓妍	F
袁卓琳	?

Laplace (Add-1) smoothing

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c)}{\sum_{w\in V}\left(count(w,c)\right)}$$

$$= \frac{count(w_i,c)+1}{\left(\sum_{w\in V} count(w,c)\right) + |V|}$$

$$\text{P(卓|male)} = \frac{Count\left(\text{卓 in } male\ names\right) +1}{Count(all\ chatacters\ in\ male\ names)\ +Count(V)} = \frac{2}{6+17}$$

$$\text{P(琳|male)} = \frac{Count\left(\text{琳 in } male\ names\right)+1}{Count(all\ chatacters\ in\ male\ names)+Count(V)} = \frac{1}{6+17}$$

$$\text{P(卓琳|male)} = \frac{2}{23} \text{ x } \frac{1}{23} = \frac{2}{529}$$

歐承瑋	M
李思穎	F
陳敏琪	F
廖倚琳	F
吴建瑞	M
馮紫晴	F
廖卓楠	M
徐婉晴	F
周咏楠	F
馬卓妍	F
袁卓琳	?

Laplace (Add-1) smoothing

$$\text{P(卓琳|female)} = \frac{Count(\text{卓 in } female\ names) + 1}{Count(all\ chatacters\ in\ female\ names) + Count(V)} = \frac{2}{14+17}$$

$$\text{P(琳|female)} = \frac{Count(\text{琳 in } female\ names) + 1}{Count(all\ chatacters\ in\ female\ names) + Count(V)} = \frac{2}{14+17}$$

$$\text{P(卓琳|female)} = \frac{2}{31}\ \text{x}\ \frac{2}{31} = \frac{4}{961}$$

歐承璋	M
李思穎	F
陳敏琪	F
廖倚琳	F
吳建瑞	M
馮紫晴	F
廖卓楠	M
徐婉晴	F
周咏楠	F
馬卓妍	F
袁卓琳	?

Calculate probability

$$P\ (\text{male}|\text{卓琳}) = P(\text{卓琳}|\text{male})P(\text{male})$$

$$P\ (\text{female}|\text{卓琳}) = P(\text{卓琳}|\text{female})P(\text{female})$$

$$P(\text{female}) = \frac{7}{10} \quad P(\text{male}) = \frac{3}{10}$$

$$P(\text{卓琳}|\text{male}) = \frac{2}{529} \quad P(\text{卓琳}|\text{female}) = \frac{4}{961}$$

$$P\ (\text{male}|\text{卓琳}) = P(\text{卓琳}|\text{male})P(\text{male}) = \frac{2}{529}\ \text{x}\ \frac{3}{10} = 0.0011$$

$$P\ (\text{female}|\text{卓琳}) = P(\text{卓琳}|\text{female})P(\text{female}) = \frac{4}{961}\ \text{x}\ \frac{7}{10} = 0.0029$$

$P\ (\text{female}|\text{卓琳}) > P\ (\text{male}|\text{卓琳})$ → 卓琳: **female**

Another example

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable with no fun

A sentiment example with smoothing

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable ~~with~~ no fun

1. Prior from training:

$$\hat{P}(c_j) = \frac{N_{c_j}}{N_{total}}$$

$$P(-) = 3/5$$
$$P(+) = 2/5$$

2. Drop "with"

3. Likelihoods from training:

$$p(w_i|c) = \frac{count(w_i, c) + 1}{(\sum_{w \in V} count(w, c)) + |V|}$$

$$P(\text{“predictable”}|-) = \frac{1+1}{14+20} \qquad P(\text{“predictable”}|+) = \frac{0+1}{9+20}$$

$$P(\text{“no”}|-) = \frac{1+1}{14+20} \qquad P(\text{“no”}|+) = \frac{0+1}{9+20}$$

$$P(\text{“fun”}|-) = \frac{0+1}{14+20} \qquad P(\text{“fun”}|+) = \frac{1+1}{9+20}$$

4. Scoring the test set:

$$P(-)P(S|-) = \frac{3}{5} \times \frac{2 \times 2 \times 1}{34^3} = 6.1 \times 10^{-5}$$

$$P(+)P(S|+) = \frac{2}{5} \times \frac{1 \times 1 \times 2}{29^3} = 3.2 \times 10^{-5}$$

Practical issues

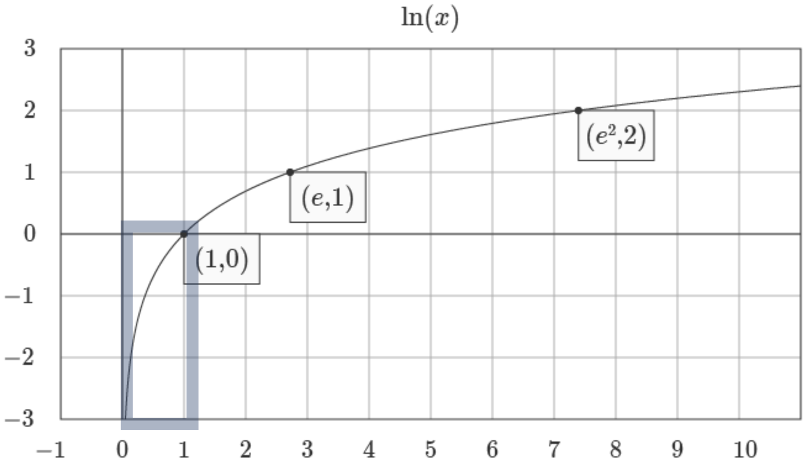
We do everything in log space

- Avoid arithmetic underflow

P(-|’predictable no fun’**)**

= 0.059 x 0.059 x 0.029 x 0.6

= 0.00006



log(P(-|’predictable no fun’**)**

= log(0.059 x 0.059 x 0.029 x 0.6**)**

= log(0.059**)** + **log(**0.059**)** + **log(**0.029**)** + **log(**0.6**)**

= -9.71

Summary: Naive Bayes is not so naive

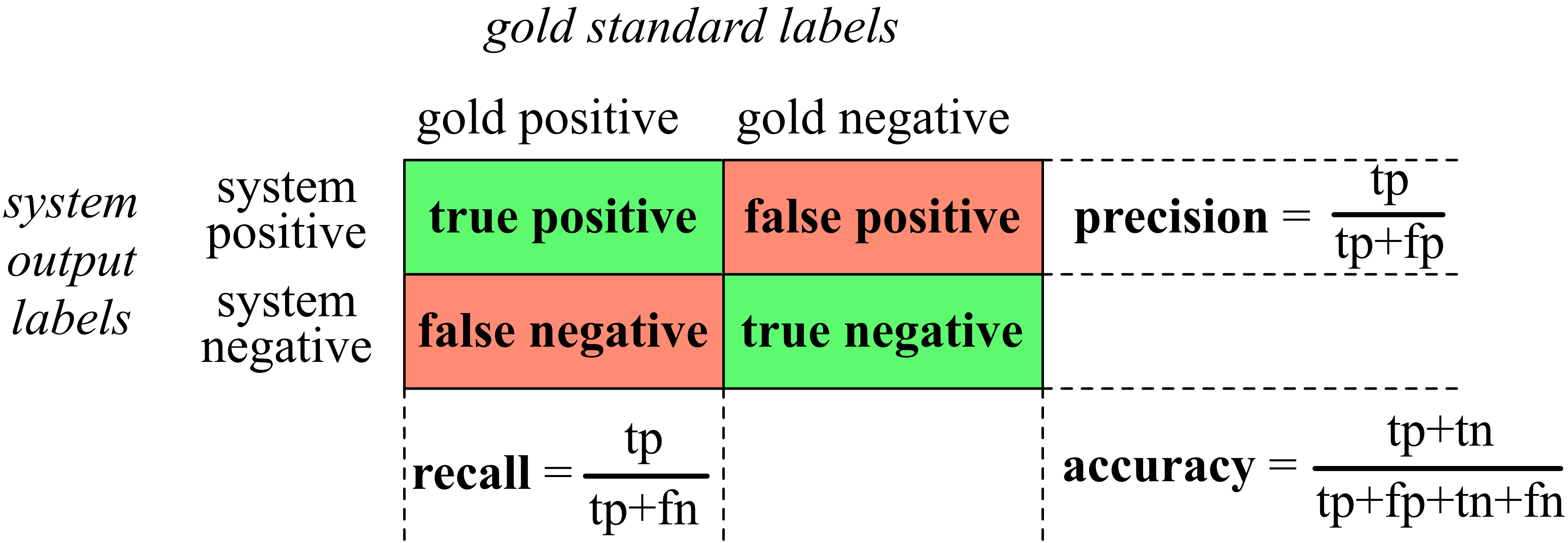
- Very Fast, low storage requirements
- Work well with very small amounts of training data
- Robust to Irrelevant Features

  Irrelevant Features cancel each other without affecting results

- Optimal if the independence assumptions hold
- A good dependable baseline for text classification

But we will see other classifiers that give better accuracy

Model evaluation



gold standard labels

		gold positive	gold negative	
system output labels	system positive	**true positive**	**false positive**	**precision** $= \frac{\text{tp}}{\text{tp+fp}}$
	system negative	**false negative**	**true negative**	
		recall $= \frac{\text{tp}}{\text{tp+fn}}$		**accuracy** $= \frac{\text{tp+tn}}{\text{tp+fp+tn+fn}}$

Accuracy

Why don't we use **accuracy** as our metric?

- We have 73 students in our class, only 13 are male students.
- We could build a dumb classifier that just labels every student as female. → **accuracy: 60/73 = 82%**
- But useless! Can never find a male students.

→ We need to use **precision** and **recall**

Precision

% of items the system detected (i.e., items the system labeled as positive) that are in fact positive (according to the human gold labels)

$$\textbf{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

Recall

% of items actually present in the input that were correctly identified by the system.

$$\textbf{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Why precision and recall

Our dumb gender-classifier: Just label every student as female

Accuracy=82%

but

Recall = 0

$$\textbf{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

(it doesn't get any of the male students)

Precision and **recall**, unlike **accuracy**, emphasize true positives: finding the things that we are supposed to be looking for.

A combined measure: F

F measure: a single number that combines **Precision** and **Recall**:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

We almost always use balanced **F_1** (i.e., β = 1)

$$\mathrm{F}_1 = \frac{2PR}{P + R}$$

Example

Given the contingency table of our classifiers:

Is this a male student name?

	male	female
model: male	12	5
model: female	2	31

true positive (tp): **12**

false positive (fp): **5**

true negative (tn): **31**

false negative (fn): **2**

$$\text{Accuracy} = \frac{tp+tn}{tp+fp+tn+fn} = \frac{12+31}{50} = \mathbf{0.86}$$

$$\text{Precision} = \frac{tp}{tp+fp} = \frac{12}{12+5} = \mathbf{0.71}$$

$$\text{Recall} = \frac{tp}{tp+fn} = \frac{12}{12+2} = \mathbf{0.86}$$

$$\text{F}_1 = \frac{2PR}{P+R} = \frac{2\ x\ 0.71\ x\ 0.86}{0.71+0.86} = \mathbf{0.78}$$

© Jixing Li

To do

- Do HW5
- Optional reading: **SLP** Ch4; **NLTK** Ch6:1,3,5