

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

© Jixing Li

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?

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

Bad

Good

Bad

...awful pizza and ridiculously overpriced...

→ Sentiment analyses

What is the subject of the medical article?

-51 -	@	ackerosofirez.com	Brain Cognition
in a mark			we don't make with
Syntactic frame and w	arb bias in	anhasia: Plausib	ility judements
		bject sentences	
Susanna Gahl," Lise Mann," C Molty B		per, ⁶ Daniel S. Janafek L. Holland Audrey	y," Elizabeth Elder,"
	Samerary of States	fandinge an, offe de Aussie ein, offe de Tange and	
diamet .			
This analy incompare of the linear share is reasons, attention to the set of legating of large effective strategies at strate of possible arms prior matter is supported. This is the strate of the first matter is supported. This is the strate of the stra- tegies of the strategies which is followed for a time of the strategies of the strate of the strate bias of the strategies. All other strategies the 3 2021 Effective Inc. All other strategies in the	p. We have insert mean. Using a pla- sider on particle. After is modeled into over significant of the mean spectrum.	n de delse des entreses entre altitus (algement seis, en dess Vielden sen er de discussion by bekel Vies, i.e., der Viellen og antie den parties alternis- se ogeliecht gester in einer	they evaluate the first product the control program of optical interpretation on provide holds in the control optical in a given the optical first provide and the optical first provide and the optical first provide and the optical first provide and optical
enserver, anternet entry, and frequency of the filter data services to relate an experimental prior matrix performance. The term of the entry of the matrix is required. The term of the entry of present services the filter of term in the services are access to the service of the term of the services to the service of the term of the services are access. These fields the term of the services of the service of the services of the services of the service of the services of services of the service of the services of the servi	p. We have insert mean. Using a pla- sider on particle. After is modeled into over significant of the mean spectrum.	or do side the contrast series altitle julgerer with a discussion of the series of the discussion by forther they, i.e., the failure of gather they according to the origination of the particular the effects for gamma and total for the series for gamma and total for	(b) provident of the protocol (the section of proper of actual the protocol protocol) the first of the section of the protocol (the distance of the protocol (the section of the section of the protocol) of the section of the section of the sec- tion.
energy arises to be of largeng of up files the second part of the of the second part prior which performing the second part of the protocol performance of the second part of the second part of the second second part of the second part of the second second part of the second part of the second second second second part of the second second bill filescond part of the second second bill be second part of the second second bill be set of the second second second the second part of the second second	p. We fee inset the inset of the second of the second second of the second sector of the second sector of the second sector of the second sector of the second sector of the second sector of the second sector of the second sector of the second sector of the second sector of the second sector of the second sector of the second second second second second second s	a de vide des anterestations all'high pagemental, en dese lie des person de desenadors y parter des parties de la desenador y anter des parties de la dese relación despany est la dese de la dese hagang and anter de managemental en anteresta. All'é companyation a parties de la desenador parteresta de la desenador de la desenador de la desenador de la desenador de la desenador de la desenador de la desenador d	they examinate a children of the proof for a solution of protocol proof of the solution of protocol proof the solution of the protocol pro- tocol of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solution of the solut
energy arises to be all imports of the file the second part of the of the parts are prior which parts are been an expension of the second parts. It is the the the product research had not all parts from the tau is no result. The bullet regar the tau is no result. The bullet regar the file file file to the AI of the second bull file to the AI of the second bullet file to the AI of the second bullet bullet file to the AI of the second bullet bullet file to the AI of the second bullet are also have the AI of the second bullet parts after the AI of the second bullet parts of the the AI of the second bullet are also have the AI of the second and optimize to	p. To the invest- tion of the proba- tion of the probability of the two sets of the of the sets of the sets of the sets of the of the sets of the sets of the sets of the of the sets of the sets of the sets of the of the sets of the sets of the sets of the of the sets of the sets of the sets of the of the sets of the sets of the sets of the of the sets of the sets of the sets of the of the sets of th	a de vide des anterestes este a diffej paginers sai, se des la des ser es de desenier : y anter des partes de la desenier y anter des partes de la des regelses de partes de la des refere la segura y al senie de la des bayen y al senie de la comparte de la desenie : comparte de la desenier . Alla comparte de la desenier	(b) considered a children de la constante de la constante de particular de la constante de
ensure, animals take of largency of up and a set (final size sectors to check of partice and prior weak performing the set of the sector sector set is a set of the sectors. The size of the set of the sector sectors were reached in the large final size of the sectors may be independent for the data finance inst. These independents is 2000 finance inst. All rights marked it interactions in the sector sector of the sector sector is set of the sector sector of the sector. The sector inst.	p. The base investigation of the second s	a de vide des anterestations allegs pagesen sais, en des la des sers en la desenier. Es des sers en la desenier la des des partients de sers à partie des parties de sers alles bagenes parties de sers des bagenes parties de sers des bagenes parties de sers des bagenes parties de sers des bagenes parties de sers de serses de serses de sers anteres de serses de sers anteres de sersantes auxilia- tions de serses de sersantes de serses de sersantes auxilia- tions de serses de sersantes de serses de sersantes auxiliations de sers antere entre de sersantes de serses.	(b) consistent or the length of the length of the second secon
energy arises to be all imparts of the file the second part of the of the parts are prior which parts are been an excession of the second parts. It is the the the parts is near the second second base for the distance of the second second second the distance of the second second distance of the second second second in the second second second second distance of the second second second second second second distance of the second second second second second second second second distance of the second sec	p. The free level internet. They a pick the set of particular the set of particular of the set of particular of the set of particular of the set of particular the set of particular the set of particular of the set of particular of the set of particular of the set of particular	a de vide des anterestations allegs pagesen suis, en des la des sers en la desenie de la des sers en la desenie de la la des des partients de la des la des des partients de la des la des des partients de la des la desenie des partients de la desenie de la desenie de la desenie d	they examinate a relative pro- tion a status provide the later to particular provide the status of the provide the status of the provide the status of the status of the status of the status of the status of the status of the
every arises on a large program of the providence of the state of the	p. The free larger free. They a pic- ture on particular the second second of the second second second second of the second second second second second of the second second second second second of the second second second second second second second second second secon	I de vide des anterestations affility (aginess sail, en descrite la des arts en de descrite) (a des arts en de descrite) (a des des partieux) (a de des arts anterestations) (a de de des arts anterestations) (a de de de de de de arts anterestations) (a de de de de de de de arts anterestations) (a de de de de de de arts anterestations) (a de de de de de de arts anterestations) (a de de de de de de de de arts anterestations) (a de de de de de de de de arts anterestations) (a de de arts anterestations) (a de de arts anterestations) (a de de arts anterestations) (a de de de de de de de de de de arts anterestations) (a de de arts anterestations) (a de de arts anterestations) (a de	(b) considered a children of the proce- tion of a state processing of a plant in a particular processing and the particular processing of a state or the biotection of the state and the process of the state and the process of the state and the process of the state of the state of the state of the state of the process of the state of the state of the state of the state of the state of the state of the state of the state of the state of the state of the state of the state of the state of the state of the st
every arises of a set impact of an array of the proof of the set o	a) The base local sector of particular sector of particular the sector of particular distances of the sector of the sector of the sector of the sector of particular sector of	I de vide des sensentements and alleg juggeste site, en des les des sensentes de décember : y anter des parties de la décember y anter des parties de la décember relation desparage est instal de la décember de la décember de la décember relation de la décember de la décember relation de la décember de la décember de la décember de la décember des de la décember de la décember de la décember	The end of the second s
exercise arises with each importunity of the second systems in the second system. The second seco	p. We have invest- tions. They a place where an equivalence of the second second of the second second second second second second second and the second s	The first of the constraint over a difficult of the constraint over the first over an its difficult over the first over an its difficult over the first over a set of the constraints of the first over the constraints of the first over a problem of the first over the constraints over the constraints over the constraints over the constraints over the constraints over the constraints over the constraints over the constraint over the constraints over the constraints over the constraints over the constraints over the constraints over the constraints over the constraints over the constraints over the constraints over the constraints over the constraints over	they experiments or the proof they a which generally well in a proof of proof of proof they are the proof of proof they are the proof of the second second second second and which are necessarily and an excitation of the proof of the second second second second in Kard (1993), the second secon
every arises on a large of large of an influence and a second se	p. The best interface time. They a place where an equivalent of a second second second of the second second second second second of the second second of the second	The first of the constraint over a difficult of the constraint over the first over an its characteries by the first often particle of the first operation of the particle of the particle operation of the particle operation. A fifther the particle operation of the particle of the particle operation of the particle operation of the particle operation of the particle operation of the particle operation of the particle operation operation operations operation operations operatis operations operations operatis operations operatio	(b) consistent or the proof of the first of the second symposite of provide the proof of the second symposite of the second
exercise a single set is a large set of the set of t	p. The besides of the second s	a de vide des consentacions altrajo jugisser sui, en dese las des consentacions (e.g., des l'actives las des consentacions) (a suite des partieurs d'actives estava l'acquires d'actives estava l'actives d'actives d'actives estava l'actives d'actives d'actives estava l'actives d'actives d'actives estava	they examinate a relative pro- ting a state provide the latter to provide a provide pro- ting of the second provide the latter than a relative provide the latter than a relative provide the latter provide the latter provide the second to the provide the latter provide the latter provide the second to the latter provide the second to the latter pro- ting the latter provide the second to the latter provide the second to the latter provide the latter provide the latter provide the latter the latter provide the latter has been latter to the latter has been latter has been latter has been latter to the latter has been latter has been

Subject category

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

• ...

Spam email?

Dear <mark>Li J,</mark>

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?



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).



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

Summary: Text identification

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

• • • •

Input:

- a document d
- a fixed set of classes $C = \{c_1, c_2, ..., c_j\}$

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 $C = \{c_1, c_2, \dots, c_j\}$
- A training set of m hand-labeled documents $(d_1, c_1), \dots, (d_m, c_m)$

Output:

• a learned classifier $\gamma: d \rightarrow c$

Many kinds of classifiers

Naive Bayes

- Logistic regression
- Neural networks
- k-Nearest Neighbors

• • • •

Bayes rule

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

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

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

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

Naive Bayes classifier

$$P(male|\downarrow = \frac{P(\downarrow = \# | male)P(male)}{P(\downarrow = \#)} \qquad P(female| \neq \exists = \frac{P(\downarrow = \# | female)P(female)}{P(\downarrow = \#)}$$

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

MAP is "maximum a posteriori" = most likely class

Dropping the denominator

Calculate probability

P (male|卓琳) = P(卓琳|male)P(male)

P (female|卓琳) = P(卓琳|female)P(female)

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

卓琳= [卓,琳] **→ features**

$$P(卓琳|female) \approx P(卓|female) P(琳|female)$$

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

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

P(female|卓琳) =
$$\frac{1}{14} \times \frac{1}{14} = \frac{1}{196}$$

© Jixing Li

Naive Bayes classifier

"Likelihood"

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

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

Document *d* represented as features $x_1..x_n$

$$P(x_1,...,x_n | c) = P(x_1 | c) \bullet P(x_2 | c) \bullet P(x_3 | c) \bullet ... \bullet P(x_n | c)$$

Calculate probability

P(male|卓琳) = P(卓琳|male)P(male)

P(卓琳|male) ≈ P(卓|male) P(琳|male)

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

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

$$P(\phi \mp m | male) = \frac{1}{6}x\frac{0}{6} = 0$$

Problem?

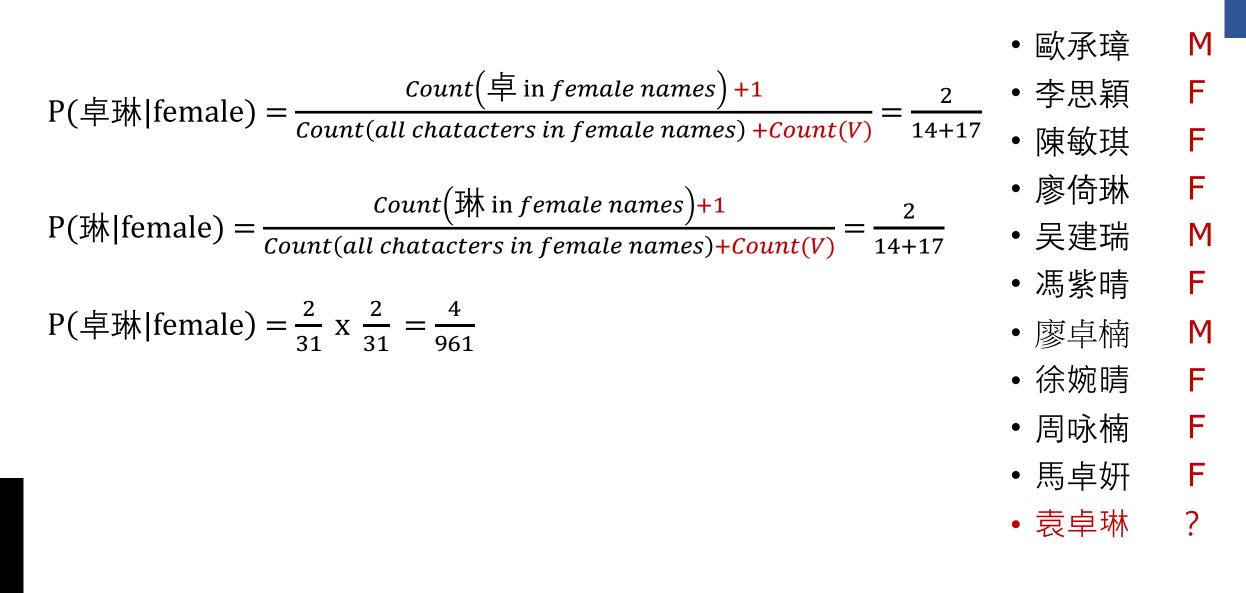
Laplace (Add-1) smoothing

529

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

© Jixing Li

Laplace (Add-1) smoothing



Calculate probability

P(male|卓琳) = P(卓琳|male)P(male)

P (female|卓琳) = P(卓琳|female)P(female)

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

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

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

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

P (female|卓琳) > P (male|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

3. Likelihoods from training:

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

$$P(\text{"predictable"}|-) = \frac{1+1}{14+20} \quad P(\text{"predictable"}|+) = \frac{0+1}{9+20}$$

$$P(\text{"no"}|-) = \frac{1+1}{14+20} \quad P(\text{"no"}|+) = \frac{0+1}{9+20}$$

$$P(\text{"fun"}|-) = \frac{0+1}{14+20} \quad P(\text{"fun"}|+) = \frac{1+1}{9+20}$$

1. Prior from training:

$$\widehat{P}(c_j) = \frac{N_{c_j}}{N_{total}}$$
 $P(-) = 3/5$
 $P(+) = 2/5$

2. Drop "with"

4. Scoring the test set:

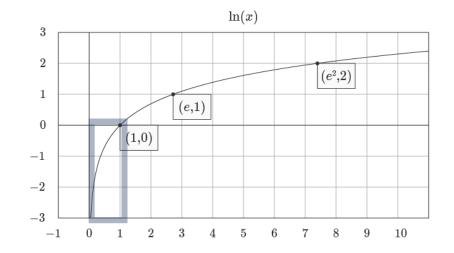
$$P(-)P(S|-) = \frac{3}{5} \times \frac{2 \times 2 \times 1}{34^3} = 6.1 \times 10^{-5}$$
$$P(+)P(S|+) = \frac{2}{5} \times \frac{1 \times 1 \times 2}{29^3} = 3.2 \times 10^{-5}$$

© Jixing Li

Practical issues

We do everything in log space

- Avoid arithmetic underflow
 - P(-|'predictable no fun')
 - $= 0.059 \times 0.059 \times 0.029 \times 0.6$
 - = 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

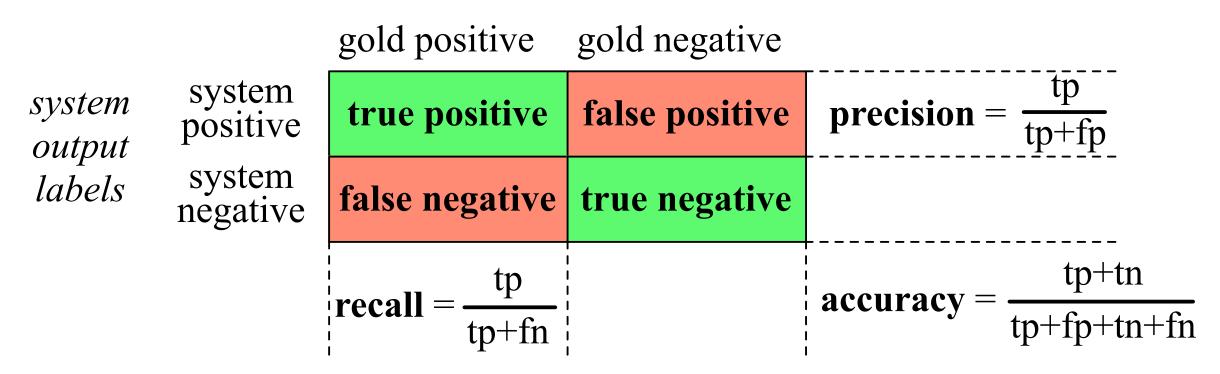
Irrelevant Features cancel each other without affecting results

- 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



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)

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

Recall

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

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

Why precision and recall

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

Accuracy=82%

but Recall = 0 $Recall = \frac{true \text{ positives}}{true \text{ positives } + \text{ false negatives}}$

(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., $\beta = 1$)

$$F_1 = \frac{2PR}{P+R}$$



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

Accuracy =
$$\frac{tp+tn}{tp+fp+tn+fn} = \frac{12+31}{50} = 0.86$$
 $F_1 = \frac{2PR}{P+R} = \frac{2 \times 0.71 \times 0.86}{0.71+0.86} = 0.78$
Precision = $\frac{tp}{tp+fp} = \frac{12}{12+5} = 0.71$
Recall = $\frac{tp}{tp+fn} = \frac{12}{12+2} = 0.86$

© Jixing Li

To do

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