Text Classification
A Get-To-Know Introduction

Weiwei Cheng
University of Marburg / Deutsche Bank
www.chengweiwei.com

Deutsche Bank

Google news

- Top Stories
- World
- U.S.
- Business
- Sci/Tech
- Entertainment
- Sports
- Health
- Spotlight
- Most Popular

World »

Iraqi Shoe Thrower Says He Was Tortured in Jail
New York Times - Marc Santora - **17 minutes ago**
BAGHDAD - Hours after his release from prison, the Iraqi journalist who hurled his shoes at former President George W. Bush said that he had been tortured while in jail, and his family said that he would flee Iraq, fearing for his life.
Video: Iraqi shoe-throwing journalist released from prison ITN
Bush 'shoe thrower' claims he was tortured in prison CNN International
Christian Science Monitor - Los Angeles Times - guardian.co.uk - BBC News
all 1,894 news articles » Email this story


Telegraph.co.uk

US may have killed al Qaeda target in Somalia, officials say
CNN International - Barbara Starr - **38 minutes ago**
(CNN) -- US special operations forces raid in Somalia on Monday may have killed a wanted al Qaeda terrorist, US officials said. The FBI has this photo of Saleh ali Saleh Nabhan on its Web site.
Video: US forces have "killed" leading Al Qaeda suspect ITN
Somalia raid likely wins intel haul, stirs tension Reuters
The Associated Press - guardian.co.uk - Bloomberg - Wikipedia: Saleh Ali Saleh Nabhan
all 721 news articles » Email this story


Financial Times

Only Decisive Force Can Prevail in Afghanistan
Wall Street Journal - Lindsey Graham - **43 minutes ago**
Growing numbers of Americans are starting to doubt whether we should have troops in Afghanistan and whether the war there is even winnable.
SCENARIOS: Obama's options in Afghan war Reuters
Joint Chiefs Chairman Mullen says more troops likely needed in ... Los Angeles Times
Bloomberg - The Associated Press - AFP - United Press International
all 809 news articles » Email this story


WA today

U.S. »

Police: Roseville home broken into, cat killed
Chicago Tribune - **27 minutes ago**
AP ROSEVILLE, Mich. - Police in Roseville say at least one burglar broke into a home in th suburb and killed the family cat by putting it in the washing machine.
Burglar killed cat in washing machine The Detroit News
burglars ransack home, kill cat Detroit Free Press
WDIV - WLNS - MLive.com
all 8 news articles » Email this story

Report: Student stabbed to death at Florida high school
CNN - **46 minutes ago**
MIAMI, Florida (CNN) -- Police said Tuesday they are investigating a death at a high school in Coral Gables, Florida. Miami-Dade County police told CNN the death occurred at Coral Gables Senior High School but would not provide further information.
Student killed at Coral Gables High School MiamiHerald.com
Student stabbed at Fla. school; 1 in custody The Associated Press
WESH.com - NBC2 News - KMBC.com - Miami New Times
all 391 news articles » Email this story

FBI warns police depts after NYC terror raid
The Associated Press - Tom Hays, Devlin Barrett - **29 minutes ago**
NEW YORK - Counterterrorism officials are warning police departments around the country to be on the lookout for evidence of homemade bombs following raids on several New York City apartments in a hunt for explosives and possible links to al-Qaida ...
NY Police: Terror Task Force Conducts Raids in Queens Voice of America
NY Homes Raided in Terror Probe Washington Post
New York Times - Christian Science Monitor - Los Angeles Times - CBS News
all 1,232 news articles » Email this story

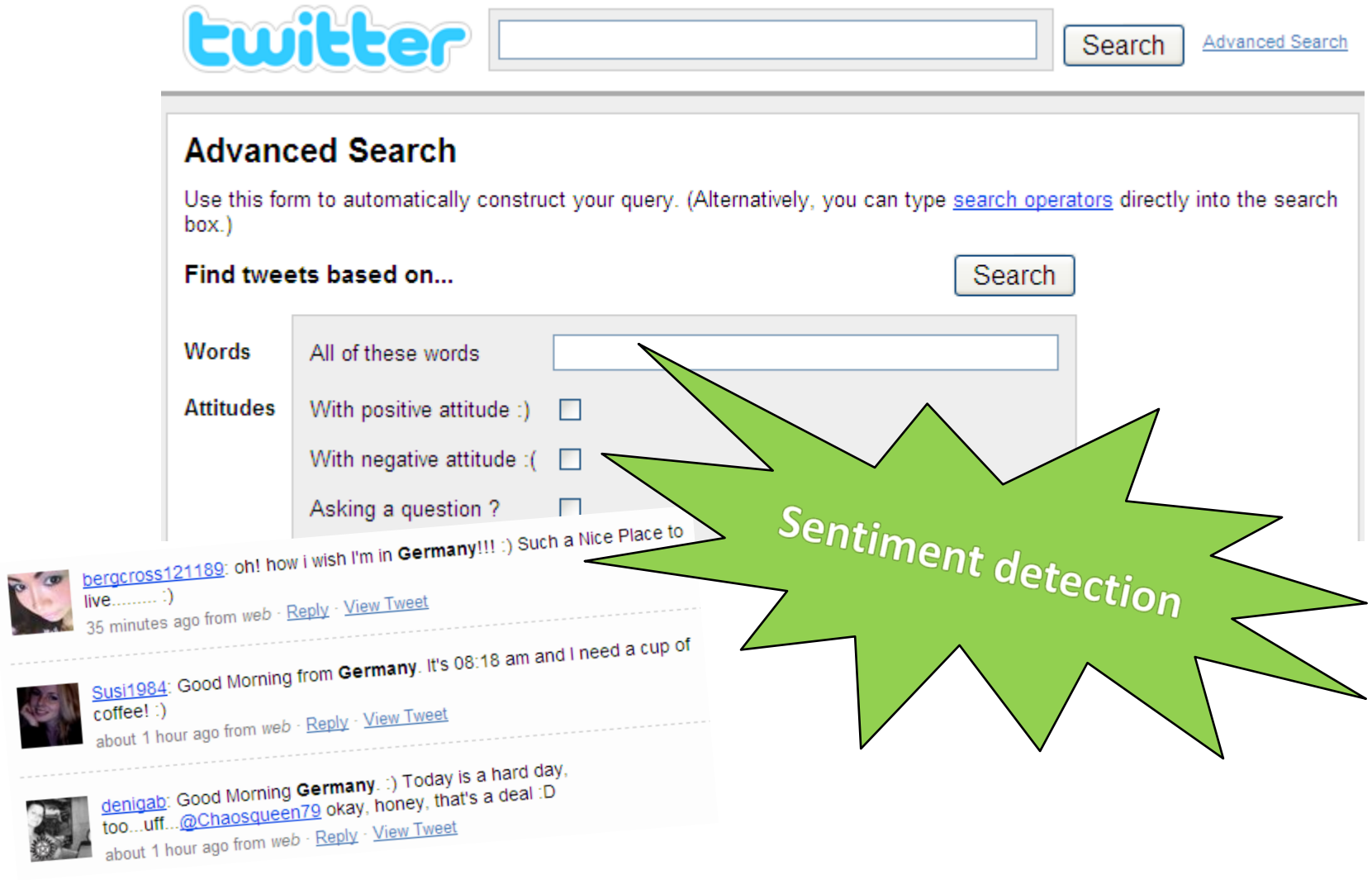
Business »

Business Inventories in US Decl... July:
Bloomberg - Timothy R. Homan - **1 hour ago**
Sept. 15 (Bloomberg) – Inv...
month as a gain in sales helped ...
Car Sales Lead Increase in US Retail...
NRF: Retail sales up in Aug. for first time in...

Sci/Tech »

Zune HD: No ipod Killer
PC World - Daniel Ionescu - **38 minutes ago**
...really think it can outdo Apple and its fleet of ipods with the Zune HD? Get ...h with a flaky refresh of the ipod touch last week, the Zune HD still doesn't ...that well against Apple's popular army of music-playing ...
...Under the hood CNET News
...s of today DailyTech

News categorization



try it at *http://search.twitter.com/advanced*

Formally…

Training (or learning):

Input: a set of *m* labeled documents $(x_1,y_1),....,(x_m,y_m)$

Output: a learned classifier $f: x \rightarrow y$

Testing (or predicting):

Input: a document x

Output: a class y from some fixed set of labels $y_1,...,y_K$

Why we need computers?

If we have *enough* brains,
we can solve *all* problems!

Wrong!

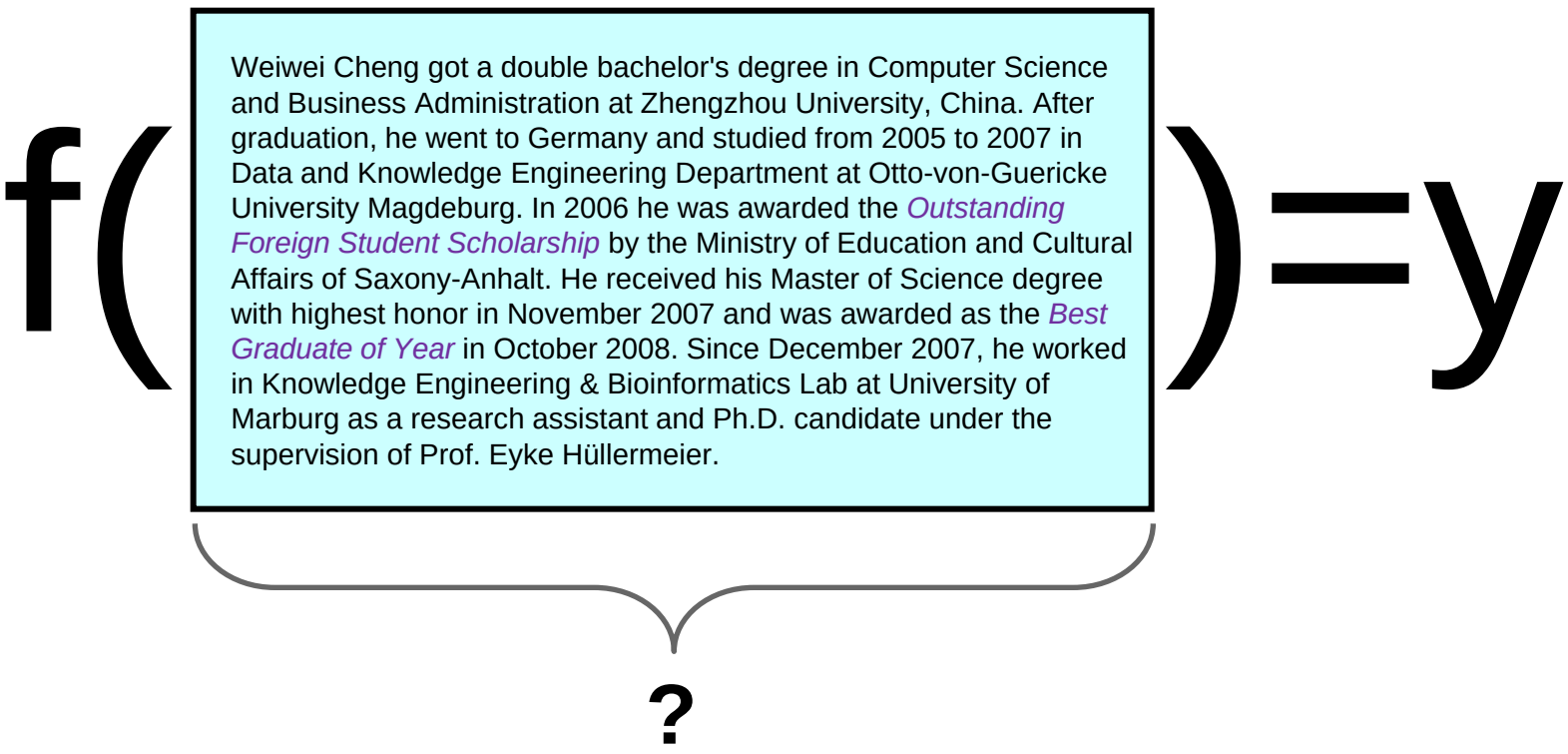
Can human do this?

Human brain has limited a capacity.
Can you remember all the books from the library?

Is the computer only for our “dirty work”?

Human perception is fuzzy.
Human intuition is (more than often) wrong.
Human reasoning is by nature not robust against illusion.
You like to watch magic, don’t you?

Examples



Question:

What is the ~~best~~ representation for the document?

simplest useful

Representing document with bag-of-words

x →

[A lady was picking through the frozen turkeys at the supermarket, but couldn't find one big enough for her family. She asked a stock boy, "Do these turkeys get any bigger?" The stock boy replied, "No madam, they're dead."]

→

[ladi pick frozen turkei supermarket find big her famili she ask stock boi turkei big stock boi repli madam dead]

→

[166 193 78 … … … 210 188 60]

Word	***Index***
a	1
⋮	⋮
frozen	78
⋮	⋮
repli	210
⋮	⋮

… ignores the ordering of the words.

Multinomial Naïve Bayes classifier

- A probabilistic learning method;
- Based on Bayes' theorem;
- State-of-the-art;
- Simple, in terms of implementation and use.

Great and simple idea with fancy name

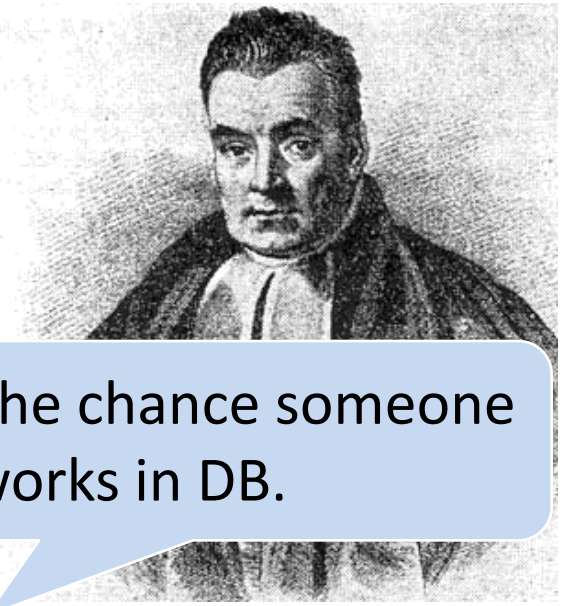
Some basic probability

Notation	Example
P(A)	P(*someone works in DB*)
P(B)	P(*someone is smart*)
P(A&B)	P(*someone works in DB* ***and*** *she/he is smart*)
P(A∪B)	P(*someone works in DB* ***or*** *she/he is smart*)
P(A|B)	P(*someone works in DB* ***given*** *she/he is smart*)

$P(A\&B) = P(B)P(A|B)$

If A and B are independent, e.g.,
A: *I wear yellow shoes today*,
B: *Today it rains*,
then $P(A\&B) = P(B)P(A)$.

Bayes' theorem



The chance that a smart guy works in DB.

The chance that a guy in DB is smart.

The chance someone works in DB.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

The chance someone is smart.

A: *someone works in DB*

B: *someone is smart*

Quiz - *drug test*

There is a drug test. It will correctly identify a drug user as testing positive *99%* of the time, and will correctly identify a non-user as testing negative *99%* of the time. Let's assume some company decides to test its employees for drug use, and it is known that *0.5%* of the population actually use the drug. What is the probability that, given a positive drug test, an employee is actually a drug user?

$$
\begin{aligned}
P(D|+) &= \frac{P(+|D)P(D)}{P(+)} \\
&= \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|N)P(N)} \\
&= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} \\
&= 0.3322.
\end{aligned}
$$

~ 33%

x: the document (e.g. *an email message*)
y: the class (e.g. *spam* or *not spam*)

We want to know which class can lead to the highest

$$P(y|\mathbf{x}) .$$

The probability of the email's *class* (spam/not spam) given its *text*.

By Bayes' theorem

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})} \propto P(\mathbf{x}|y)P(y)$$

With the bag-of-words representation

$$P(y)P(\mathbf{x}|y) = P(y)P(x_1|y)P(x_2|y)\ldots P(x_n|y)$$

To choose a *class* that maximize this.

The probability of class y, e.g., *the chance an email being spam*.

s, we co

am) an

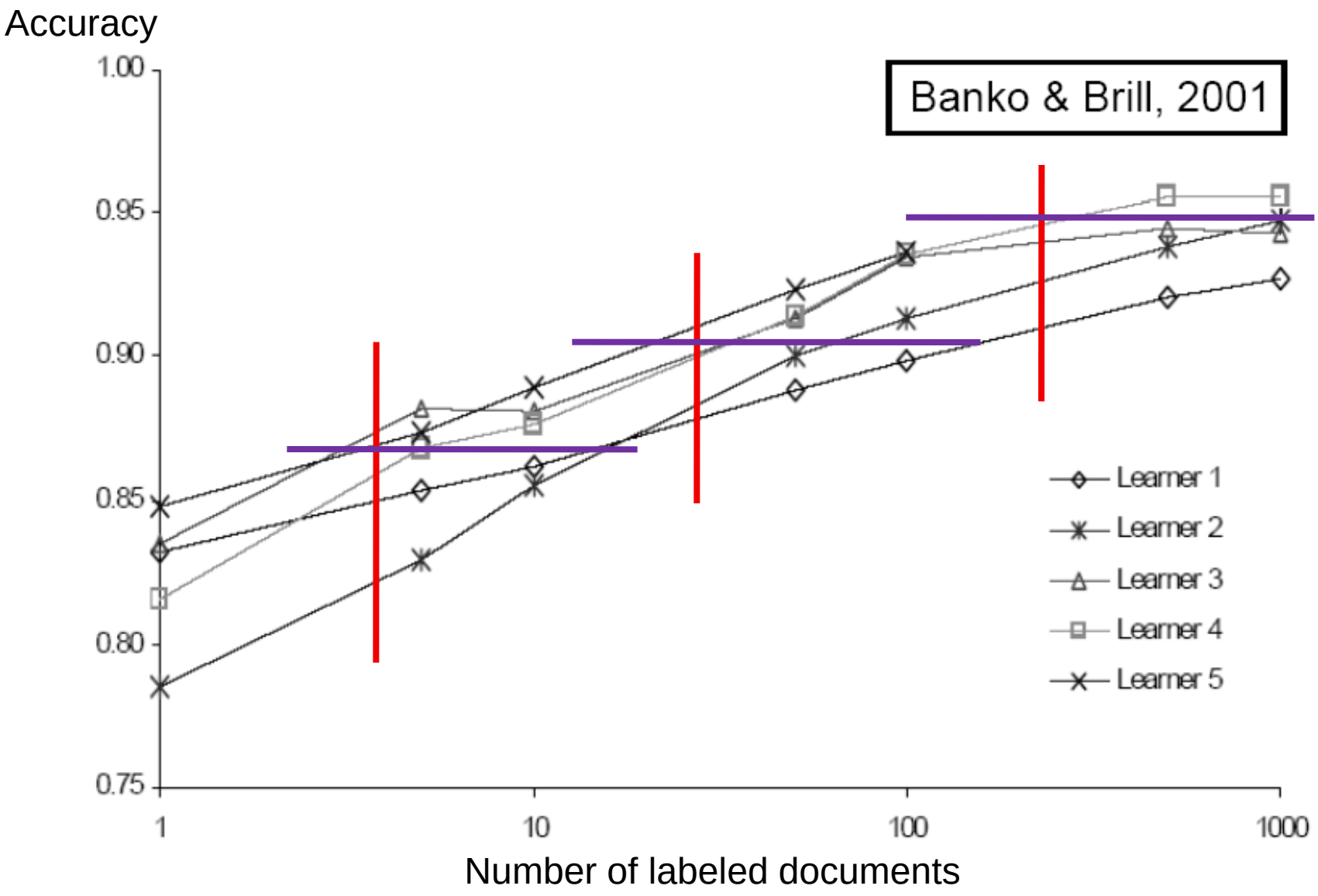
The probability of word x_2 found in class y, e.g., *the chance of word "pharmacy" seen in a spam email.*

Take-away message

Text classification finds the patterns in documents, reduces the human effort.

How do we represent the text in the computer?
bag-of-words

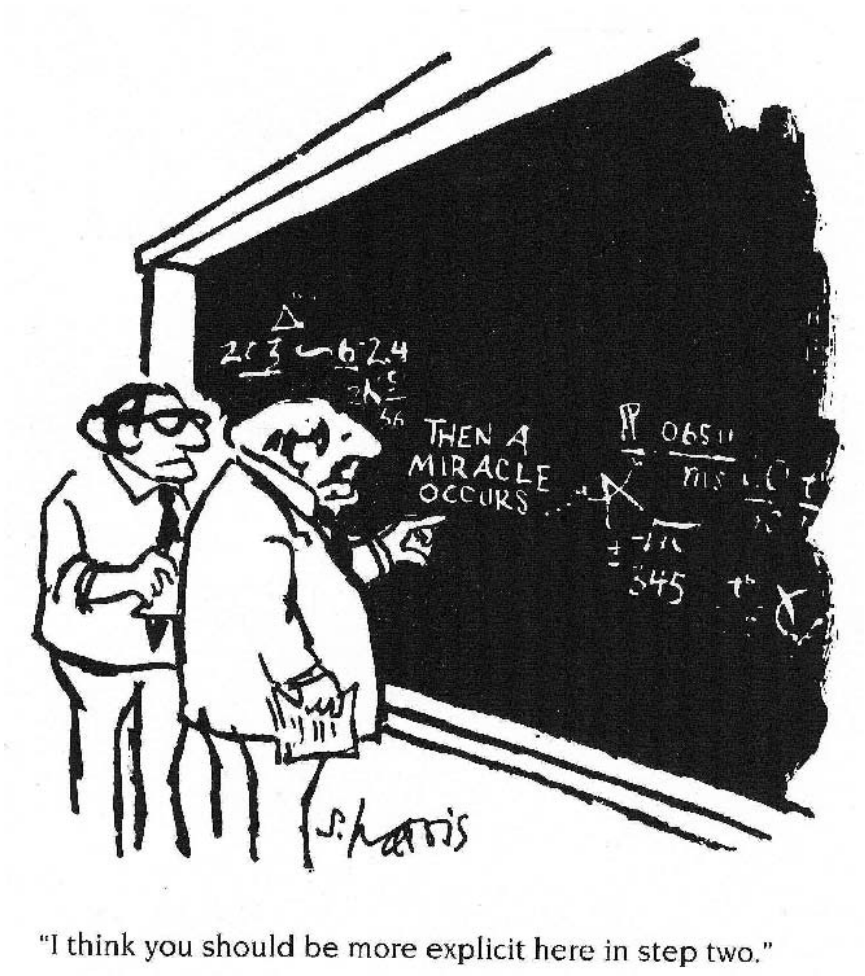
Which classifier?
Naive Bayes. Key idea: *Bayes' theorem*


Accuracy
Banko & Brill, 2001
1.00
0.95
0.90
0.85
0.80
0.75
1
10
100
1000
Learner 1
Learner 2
Learner 3
Learner 4
Learner 5
Number of labeled documents

That is all from my side ☺

Thank you all for being here!

Time for discussion!



"I think you should be more explicit here in step two."

www.chengweiwei.com