13-09-24.

NLP - NLTK

Ch 6: Learning to classify text

↳ Supervised

gender identific^n from name - unethical and wrong

feature entrac^n:

Indian parliament → finance doc
↳ non finance doc } binary classific^n

feature vector →

1	0	0	0	1	0	0	0

↳ if the word is present

↳ Bag of words model

↳ Although, position, structure, grammar have meaning
But, many task's can be done w/o this → only Bag of Words enough

Document classific^n

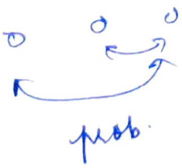
Enploiting content
↳ Include position content

Sequence classific^n
Named entity tagger

The
famous
building
Gateway → start
of
India → stop

HMM →

RNN →



LSTM → long sent, influenced by things long behind it

Coreference resolution

The famous $\overset{1}{\text{poet}}$ from $\overset{2}{\text{India}}$,

wrote $\underset{3}{\text{XYZ}}$. In $\underset{2}{\text{that country}}$, $\overset{1}{\text{he}}$,

$\underset{1}{\text{Kalidasa}}$ was born in year ABC

3 clusters:

Evaluation

↳ for binary classifier

predicted

		C	N
C	5	0	95
N	95	5	95

actual

95 → false negative ②

① false positive

true positive

true negative

Precision - $\frac{TP}{TP+FP} = \frac{0}{5} = \frac{\text{correct D}}{\text{total D}}$

Recall - $\frac{TP}{TP+FN} = \frac{\text{correct D}}{\text{total that should be detected}}$

↳ how many relevant were identified

F score = $\frac{2 \times \text{prec} \times \text{recall}}{\text{prec} + \text{recall}}$

MNIST data

confusion matrix

	0	1	2	3	4	5	6	7	8	9
0										
1										
2										
3										
4										
5										
6										
7										
8										
9										

diagonal → heavy.

Ch 8 : Analyzing sentence structure

Grammar parser

↳ content free ~~parser~~ assump^n – no subtree in a grammar tree is connected to other subtrees

words are like cloud – topic modelling

5 topics → topicwise segregation

↳ complete statistic model

Mallet