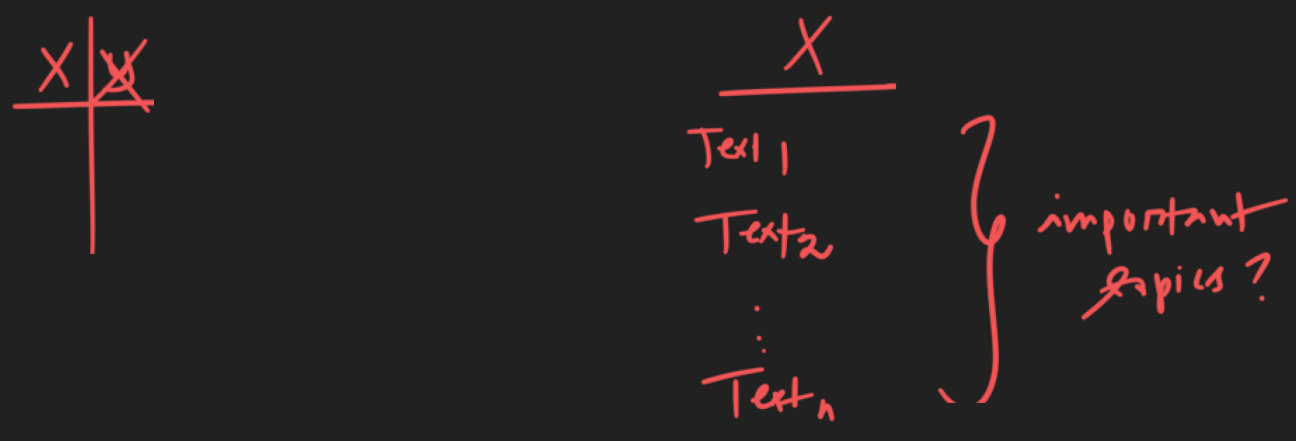


Latent Dirichlet Allocation

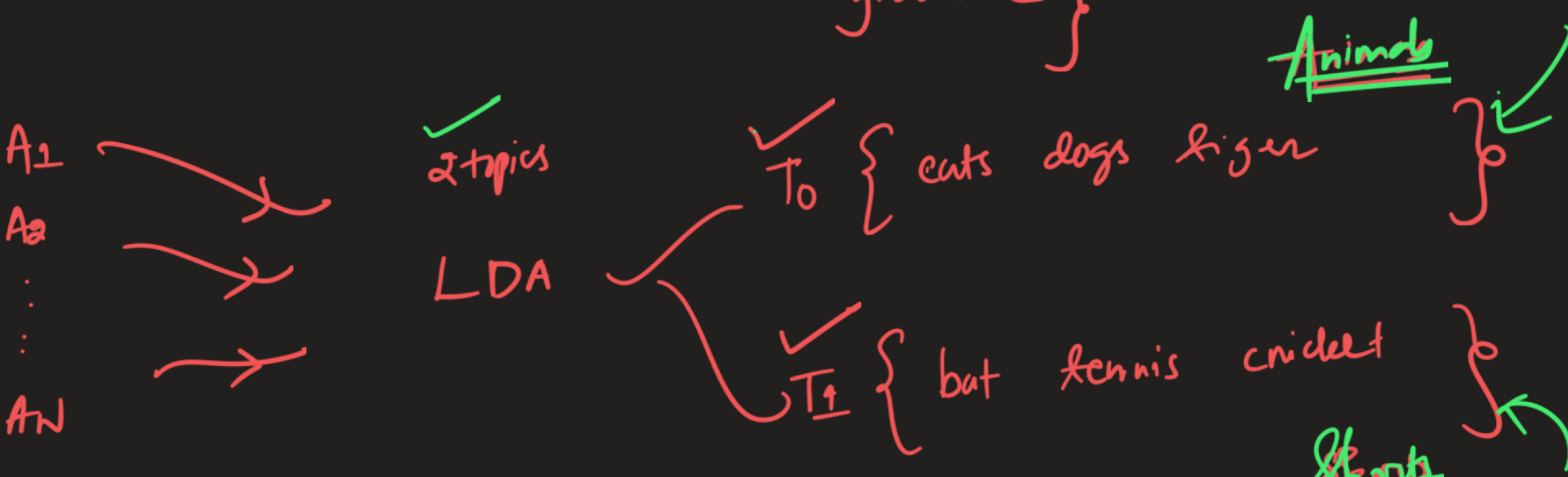
Satya Pattnaik

I am good

I am great

{ I: 2
am: 2
good: 1
great: 1 }

Animals



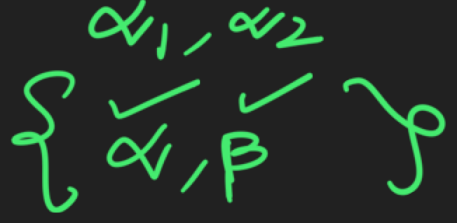
Normal — μ, σ

Dirichlet — α

Idea

LDA — not a classification algo

— generative algo



Two Dirichlet Distributions

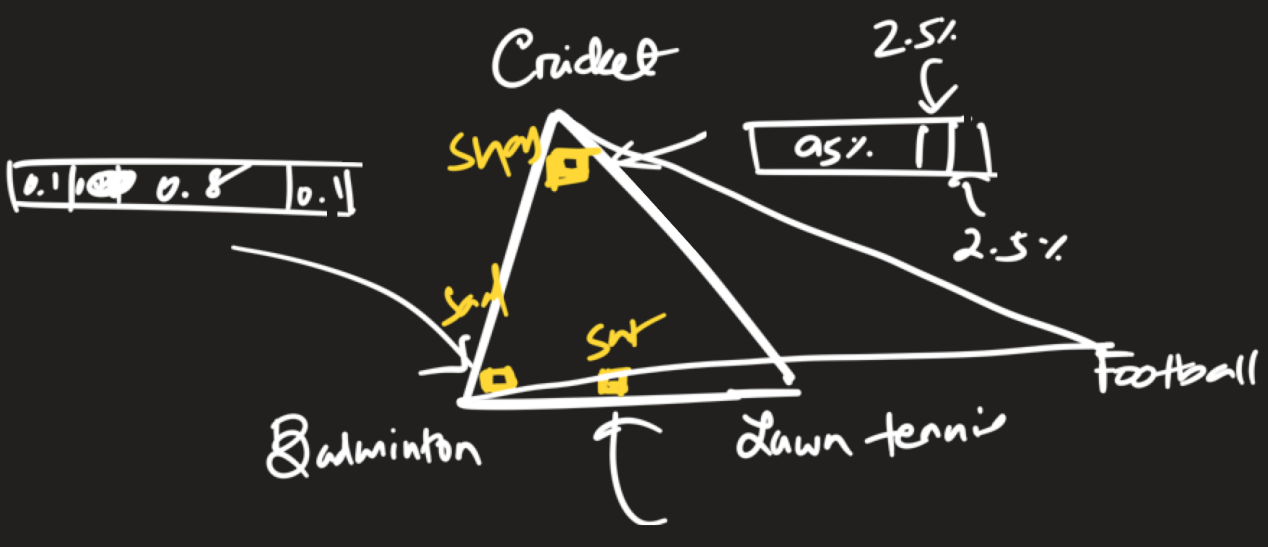
Maximum Likelihood Estimation

1	5'8"
2	5'7"
3	5'5"
⋮	⋮
40	

Data → μ, σ

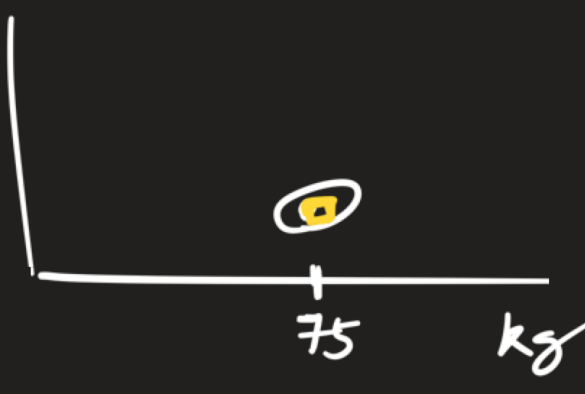
Articles Review → α_1, α_2

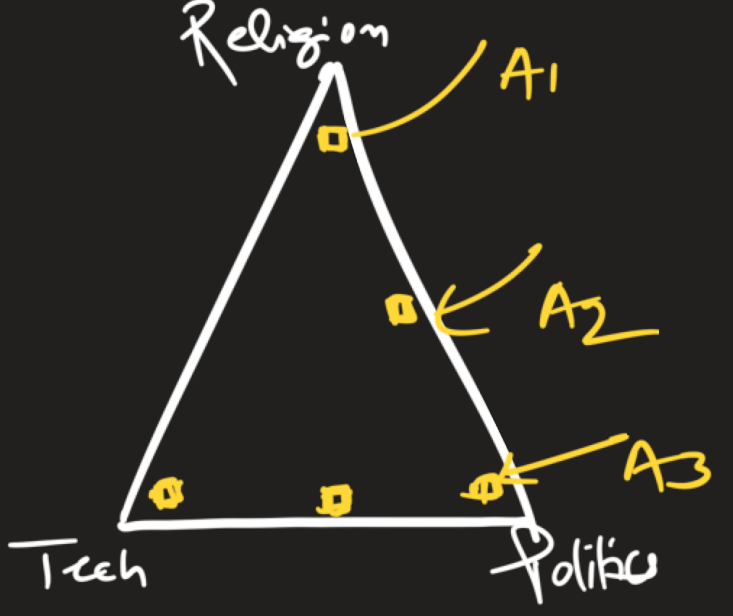
9:15

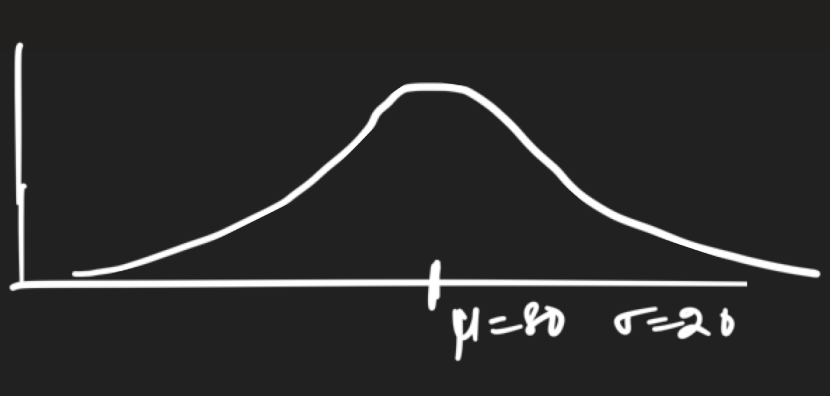


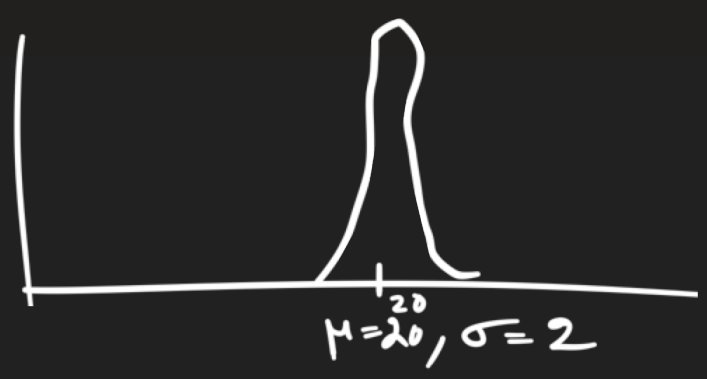
	Cric	LT	B
Shy	0.95	0.25	0.25
Sad	0.1	0.45	0.45
Sad	0.05	0.55	0.9

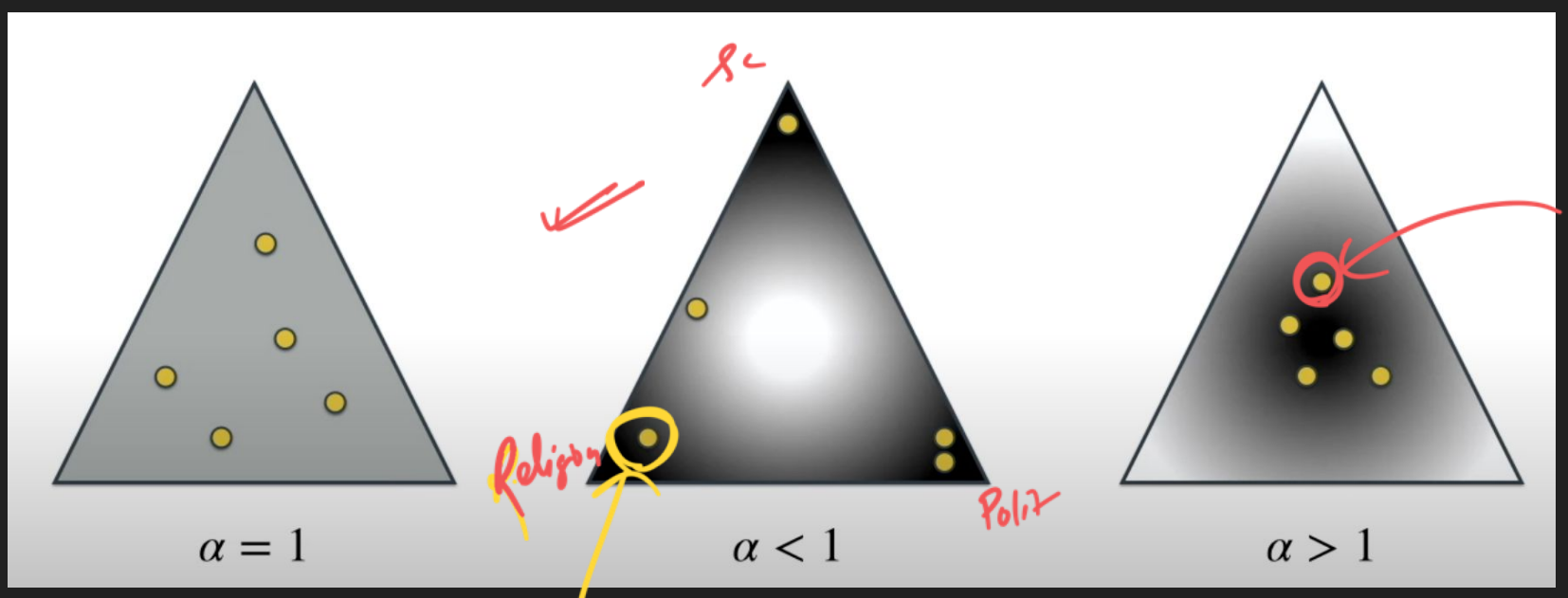
0.1 | 0.45 | 0.45

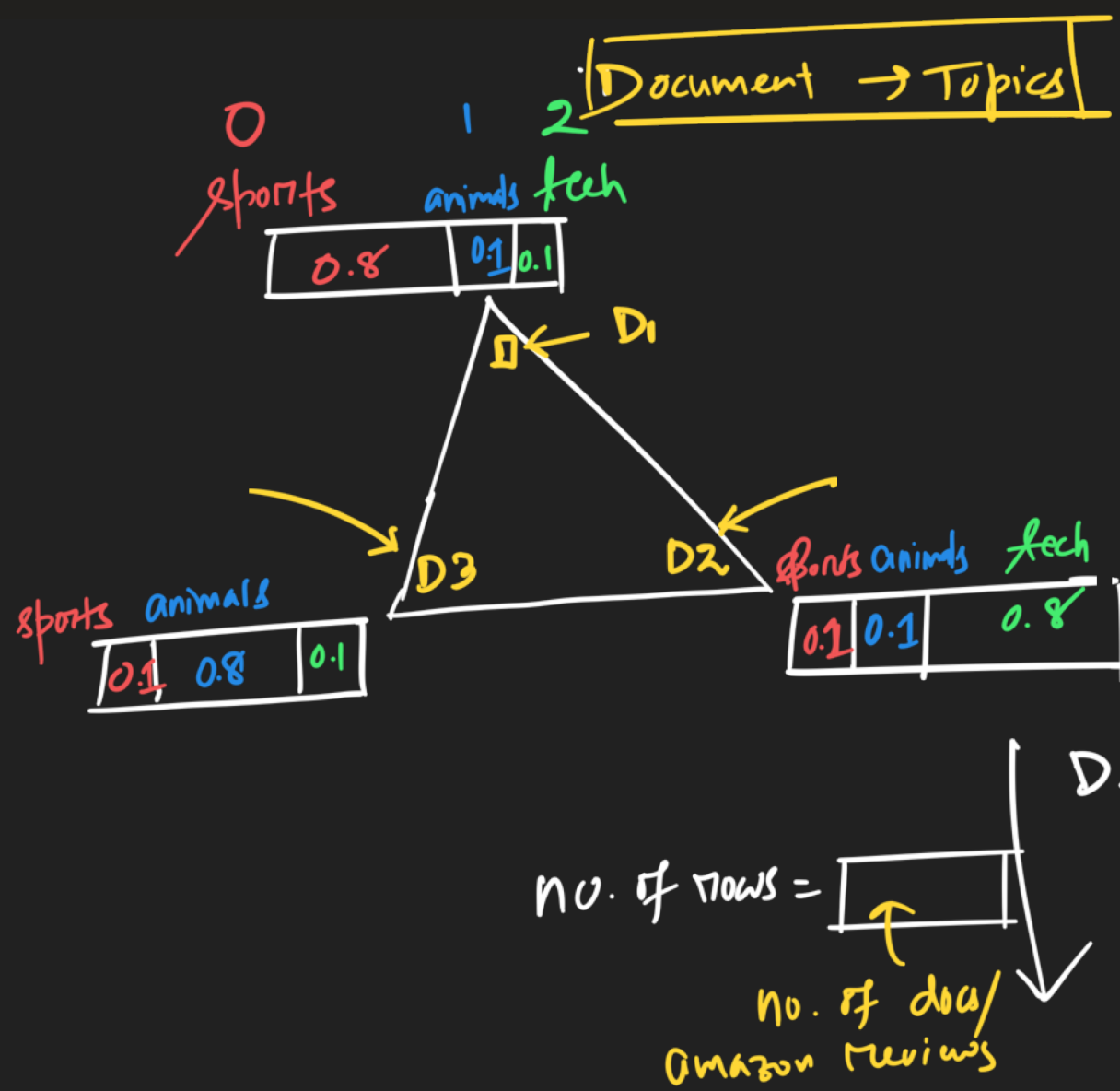





μ=80 σ=20


μ=20, σ=2


$\alpha = 1$
$\alpha < 1$
$\alpha > 1$



α α_{D-T}

↑

Model learning

no. of topics

D1 → 80% 10% 10%

no. of rows =

no. of docs/
Amazon Reviews

Vocab (1 million)

→ Cats
→ Af
→ Olympics
→ Players

no. of cols = Vocab size

Word–Topic

Af | Player | Olympic
91% | 3% | 3% | 3%
Cats

Animal T1 | T2 Tech | Sports T0

Cats Af Olympic Player

Animal
Tech
Sports

Player | Olympics | Cats | Af
46% | 45% | 5% | 5%

no. of rows = no. of topics

P → α_{N-T}

Model training

olympics

Player | Af | Cat
91%

Cat

Notations

k → no. of topics (user defined)

V → no. of unique words (vocab)

M → no. of docs / reviews

N → no. of words of in doc

M'

μ_1	σ_1
60	20
50	10
70	10
65	20

{ [80kg, 70kg, 67kg. . . .] }

{ Calculus — $\alpha_{optimal}$, $\beta_{optimal}$ }

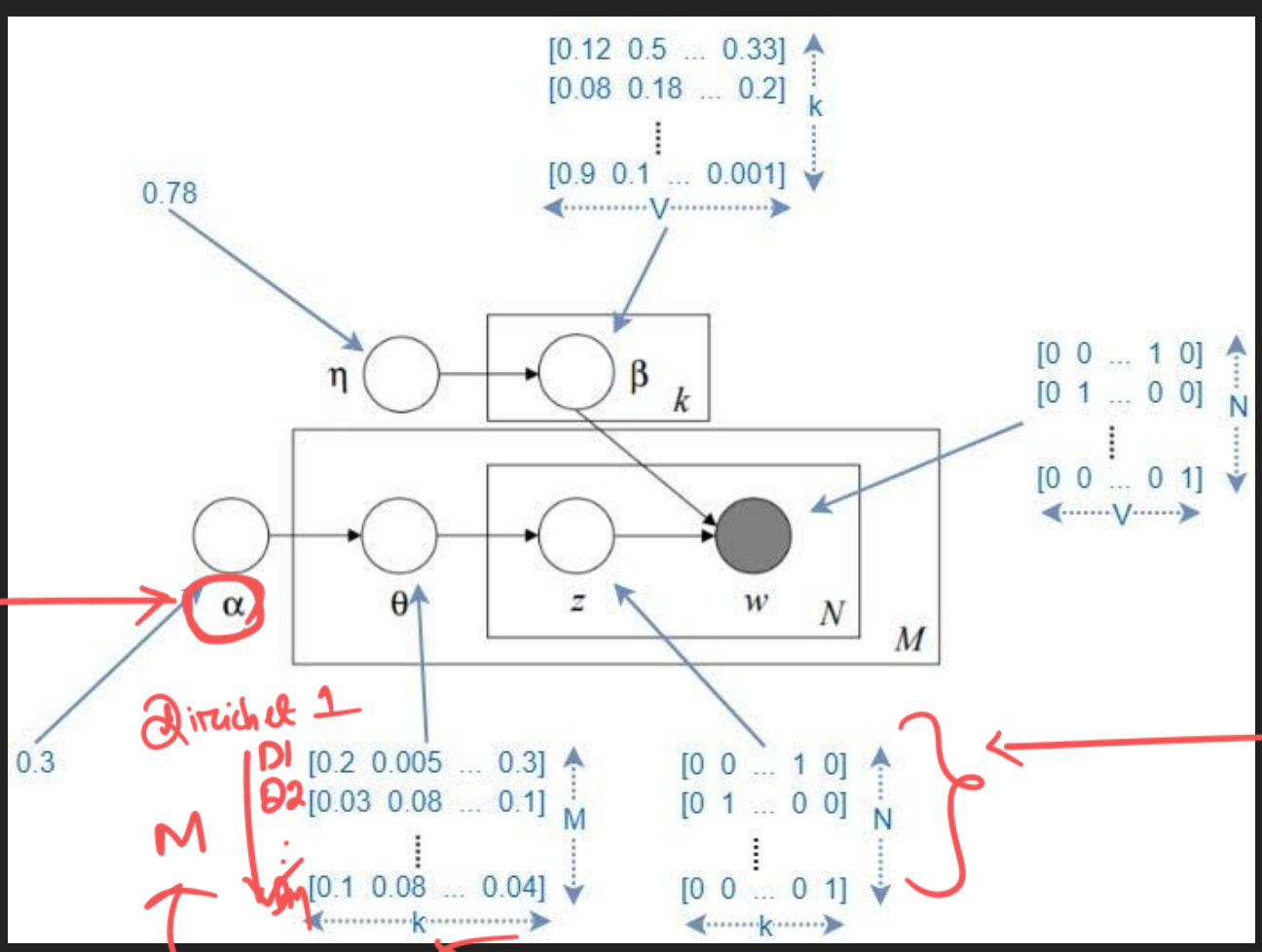
$$\frac{\partial L}{\partial w} = 0 \quad W_{optimal}$$

D-T W-T Generated Actual docs

Dirichlet Dist 1 Dirichlet 2 docs All docs/All Revi

α_1	β_1	Doc Set$_1$ / All Reviews$_1$	
α_2	β_2	Doc Set$_2$	most similar
α_3	β_3	Doc Set$_3$	
⋮	⋮	⋮	⋮
α_n	β_n	Doc Set$_n$	

Doc-set [all the docs in train data]



Doc to topic

for each doc one

M

no. of docs [no. of amazon reviews]

no. of topics (any)

N=5
⇒ each doc is of 5 words

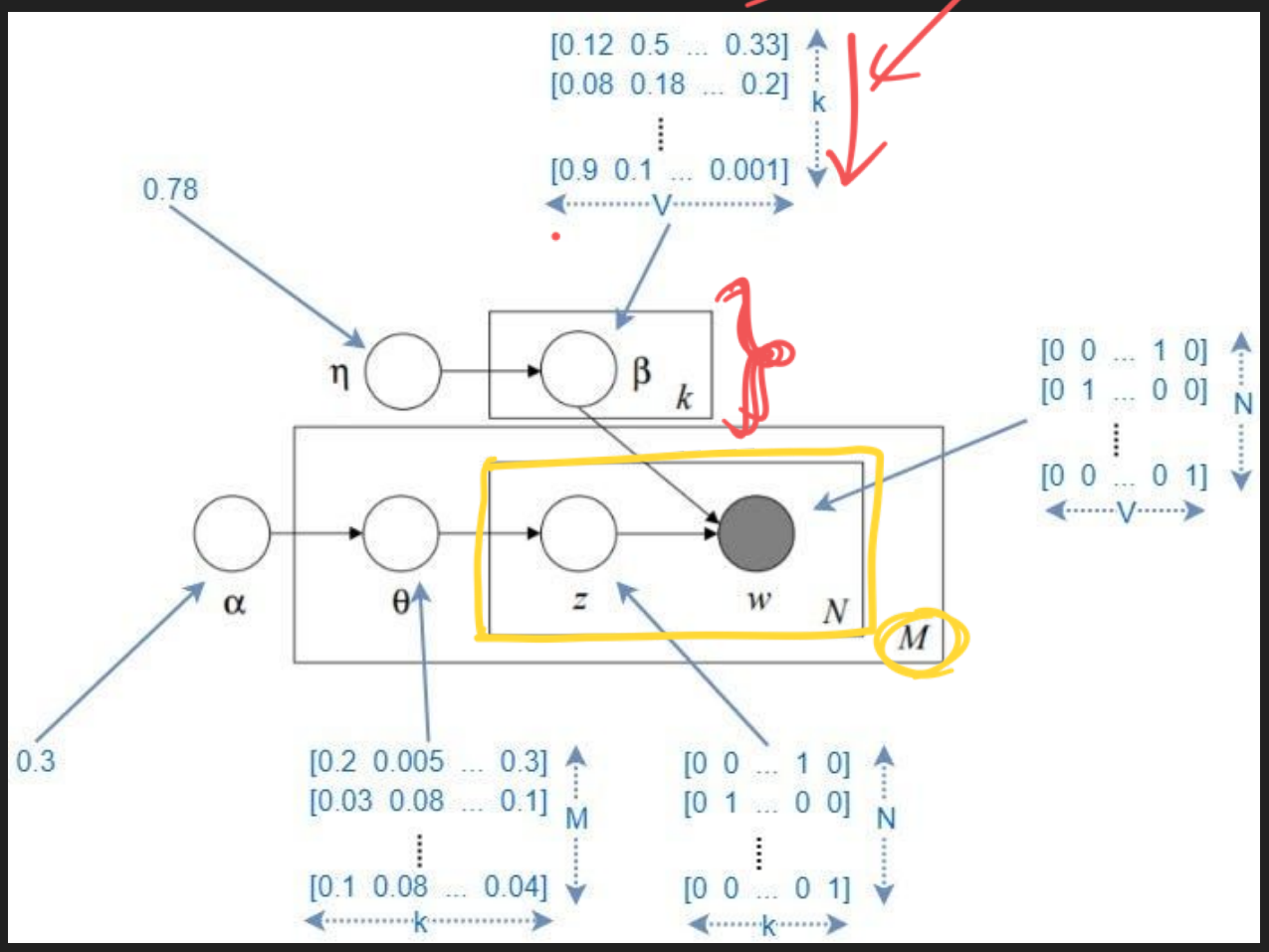
Dirichlet -1 Dist

k = no. of topics

Sports T0 | Animal T1 | Kids T2

Doc1	85%	5%	10%
Doc2			
⋮			
Doc N	0%	5%	95%

D2


no. of words
k=topics
[0.12 0.5 ... 0.33]
[0.08 0.18 ... 0.2]
[0.9 0.1 ... 0.001]
k
V
0.78
η
β
k
[0 0 ... 1 0]
[0 1 ... 0 0]
[0 0 ... 0 1]
N
V
α
θ
z
w
N
M
0.3
[0.2 0.005 ... 0.3]
[0.03 0.08 ... 0.1]
[0.1 0.08 ... 0.04]
M
k
[0 0 ... 1 0]
[0 1 ... 0 0]
[0 0 ... 0 1]
N
k

Sample 5 words

Playoys(1) olympic(2)
playyy (3) olympiL(4)
Cat(5)

	Cats	Ag	playoys	olympics
W1	0	0	1	0
W2	0	0	0	1
W3	0	0	1	0
W4	0	0	0	1
W5	1	0	0	0

Document

D1

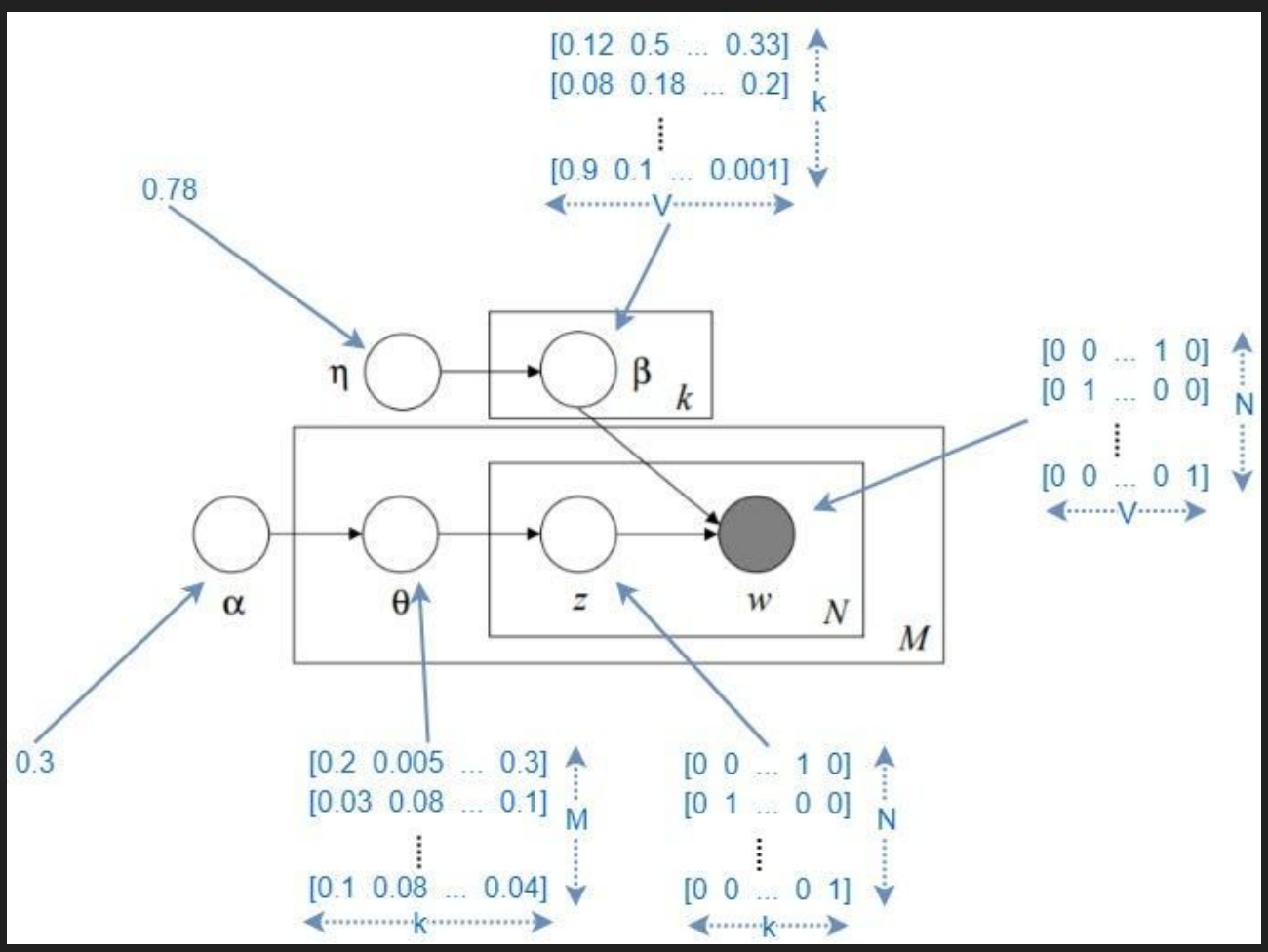
1	0	0
1	0	0
1	0	0
0	0	1

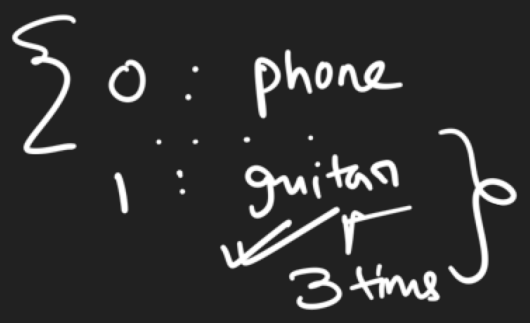
	Cats	Ag	Players	Olympics
T1 sports	5%.	5%.	45%.	45%.
T2 Animal	90%.	8.33%.	3.3%.	3..3%.
T3 tech	3.3%.	90%.	3-3%.	3.3%.

original docs

Doc1

W1	playoys
W2	olympic
W3	Players
W4	olympics
W5	cat


[0.12 0.5 ... 0.33]
[0.08 0.18 ... 0.2]
[0.9 0.1 ... 0.001]
k
V
0.78
η
β
k
[0 0 ... 1 0]
[0 1 ... 0 0]
[0 0 ... 0 1]
N
V
α
θ
z
w
N
M
0.3
[0.2 0.005 ... 0.3]
[0.03 0.08 ... 0.1]
[0.1 0.08 ... 0.04]
M
k
[0 0 ... 1 0]
[0 1 ... 0 0]
[0 0 ... 0 1]
N
k



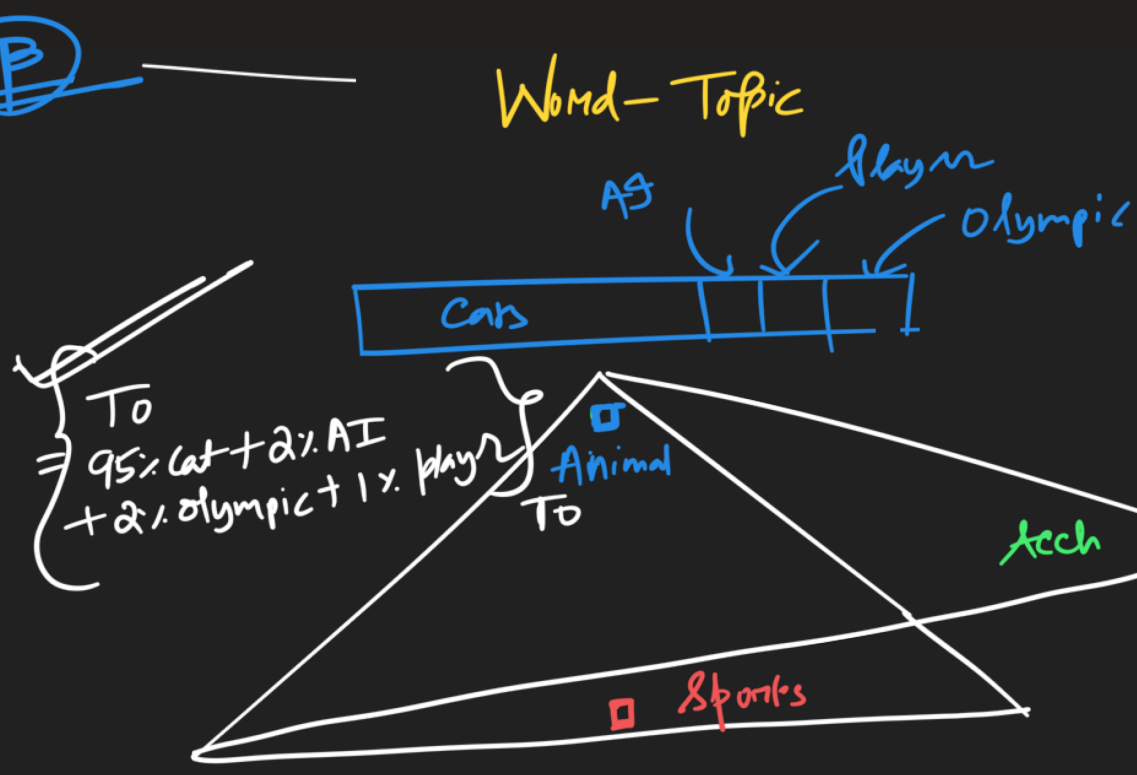
all possible word in training data

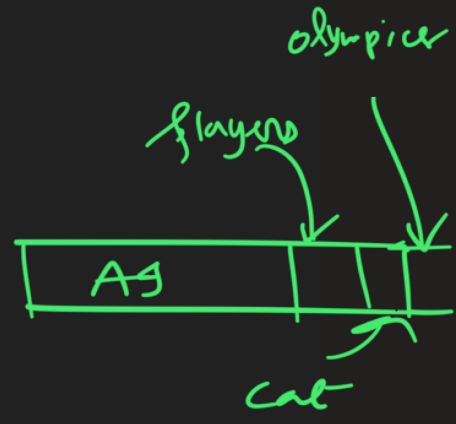
Optimal

Word–Topic

BS

Vocab
- → Cats
- → Ag
- → Olympics
- → Players



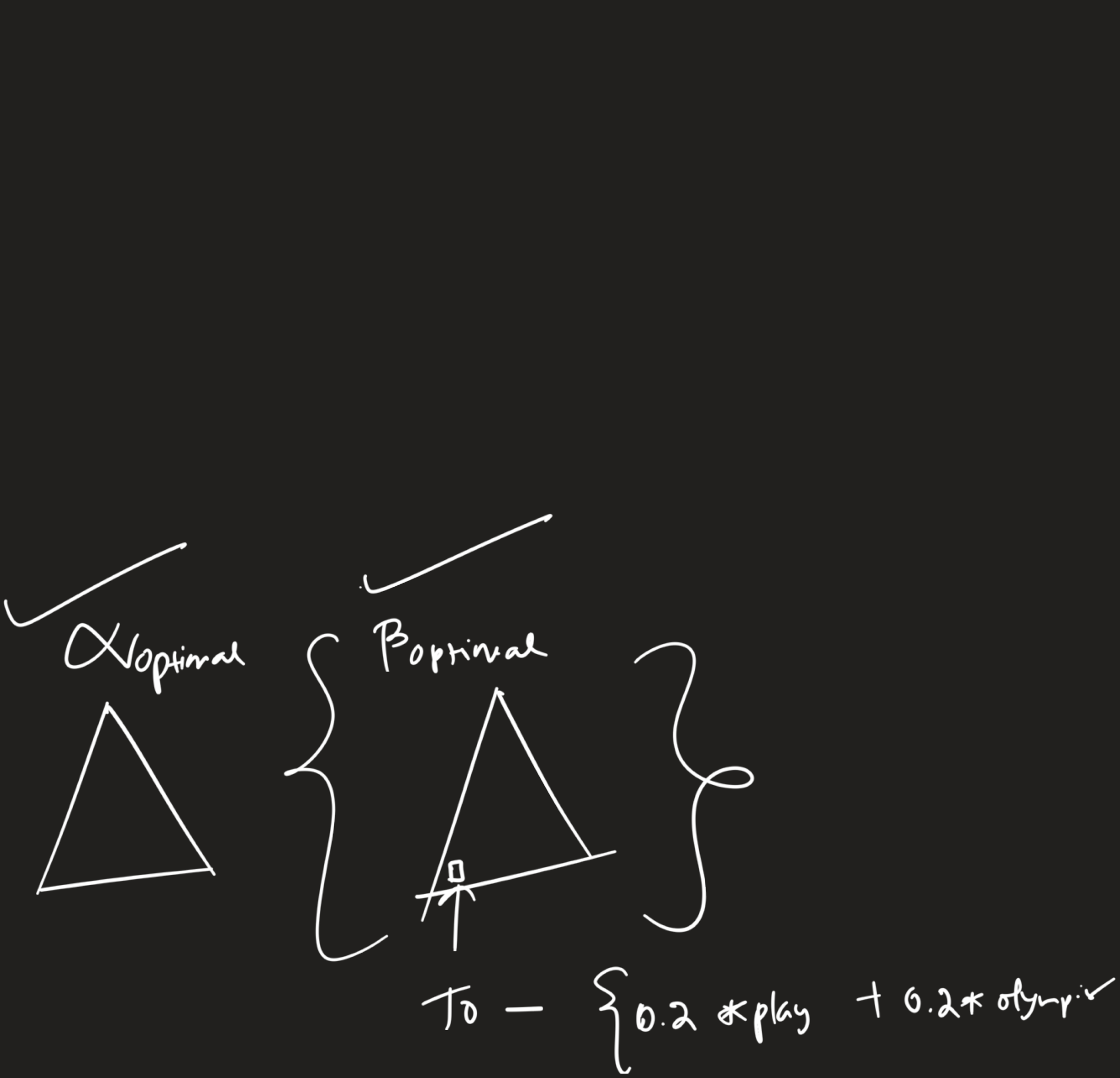


Cats Ag Olympic Player

Animal
Tech
Sports

V = Size of vocab

$$\text{Each topic} = w_0 * \text{word}_1 + w_1 * \text{word}_2 + \dots + w_V * \text{word}_V$$



To − {0.2 * play + 0.2 * olym-pic