Revise (NLP)

NLP → NLU

process

words | token

text → numeric

⇓

⇒ count

⇒ tfidf

= encoding

DTM

→ topic modelling (unsuperise)

clustering diff articles to pic and them one topic

google news, NLS

→ global sports

NLU

Elon ✓

Rupee ✓

Google ✓

Nifty ✓

→ Business

~~name of topic~~

- Intutive way

↓

- what is happen

↓

- what is outcome

↓

- code

ToC

→ Algebra → A/M/S/D

→ Trig → Matrix

→ calculus → dif
→ integ

→ Stats

mimic
today

machine

~~Deepali are~~

unigram, bigram, trigram, n-gram

features
+
[illegible]

DTM

I | am | Tam | is
✓ ✓ ✓ ✓

toke
indepen → I am

am

Taufique is

I am

~~I is~~

I am | Taumfiqe is | Amar & Deepali | Deepali are

Amar & Deepali are
studying NLP

Usage of Topic Modelling

1 cells → 1 tissues → 1 organ → human body

1 words → 1 sentence → 1 parra → 1 doc

topic modelling

Assumption → start

Doc → Topic → Term

?

1 topic

DTM

1 doc

n topic → 1 term

1 doc → 1 topic → 1 term → find out

~~topic model~~

1 doc → 1 topic
↓
n terms

doc → mixture topic

topic ⟷ mixture terms } topic modelling

Doc → topic → words

assmp

1 doc: Sourav Ganguly & Greg Chappell

Sports , Controversy

SG

GC

→ SG + GC

Latent Dirichlet Allocation DTM Latent

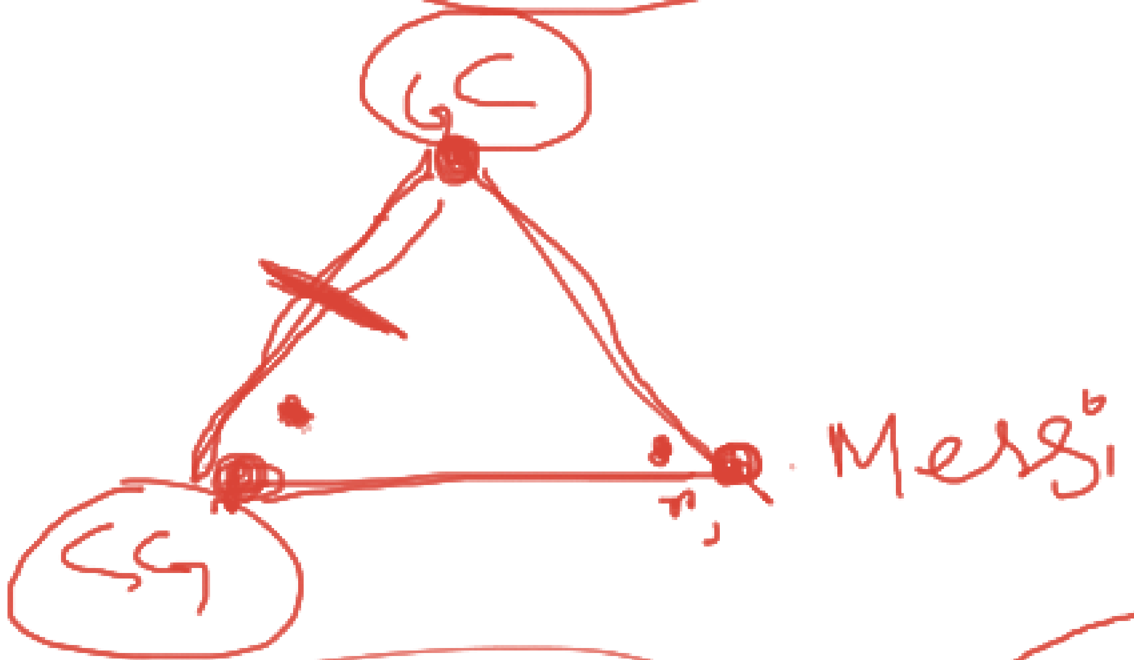
topic

Doc → topic → words

100°C → 100°C

~~Maths Aivric~~ - intuition

doc → topic → words



batsm

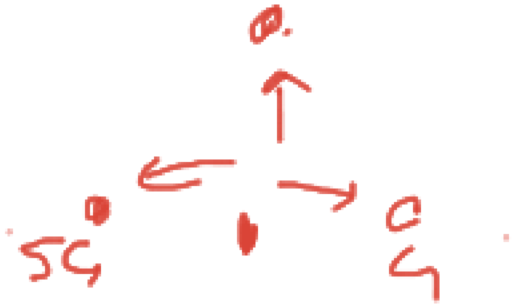
Foots

contoo

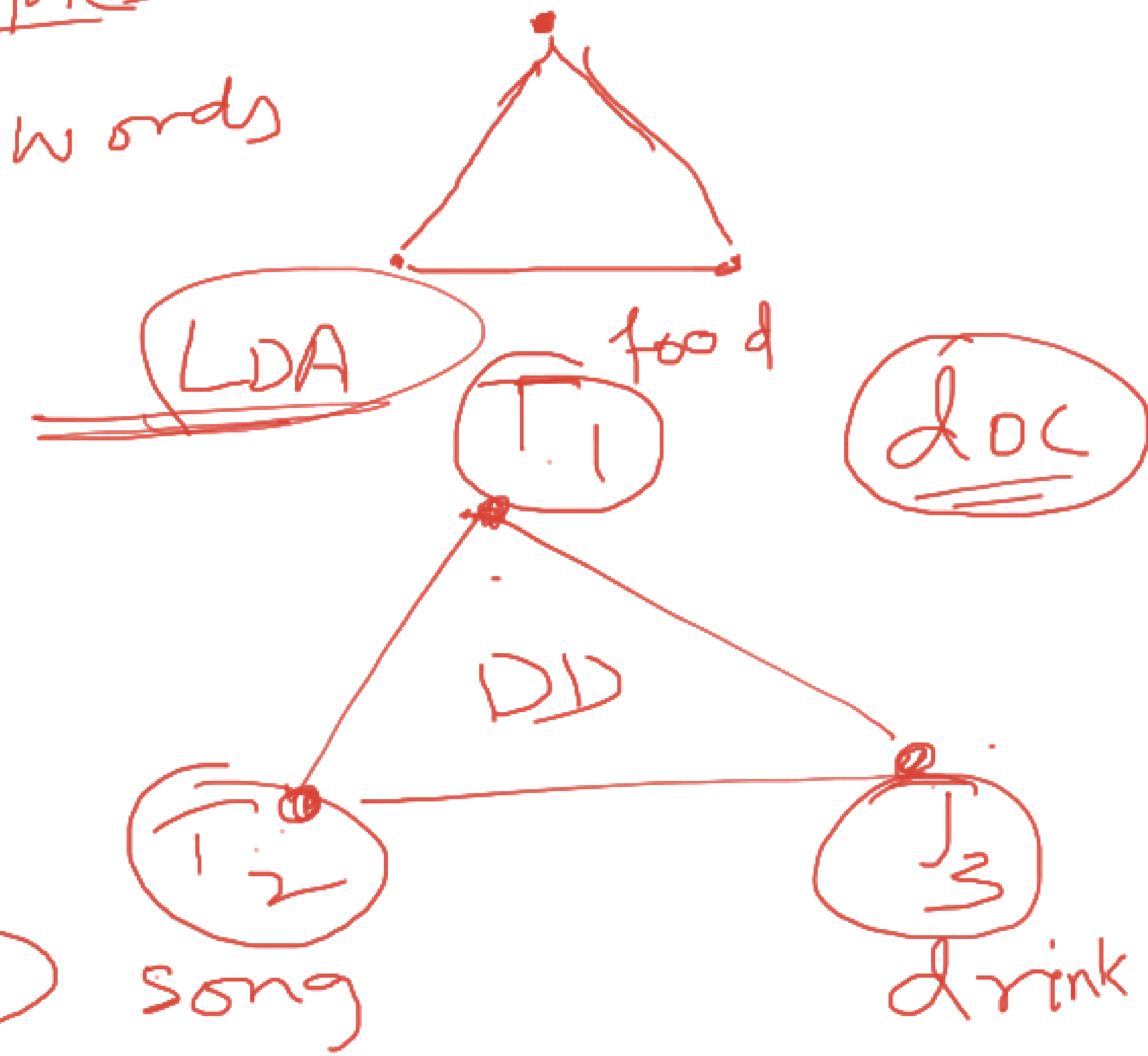
LHS → RHS

1 doc → mul topi

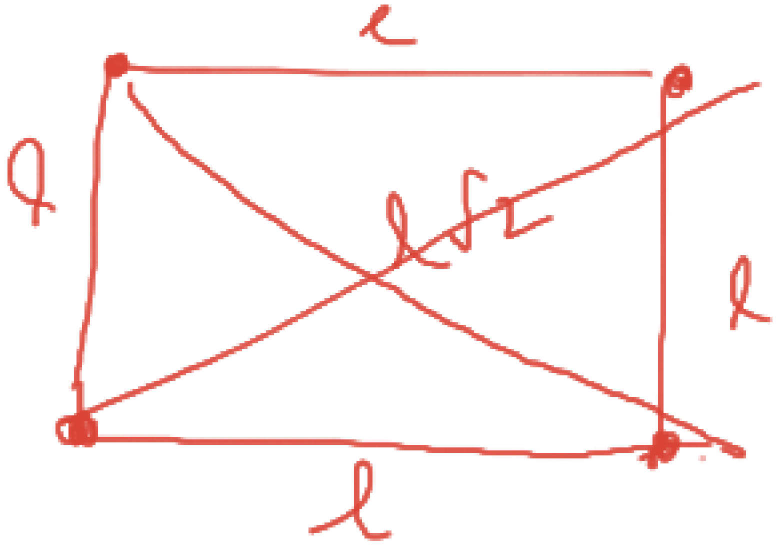
1 topic → n words

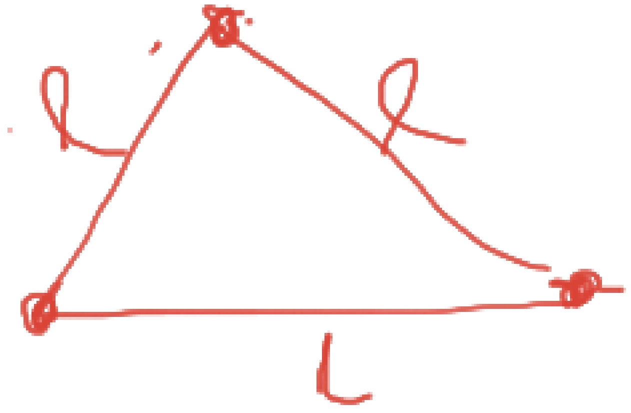


topics → edge of shape



T1 equally distant

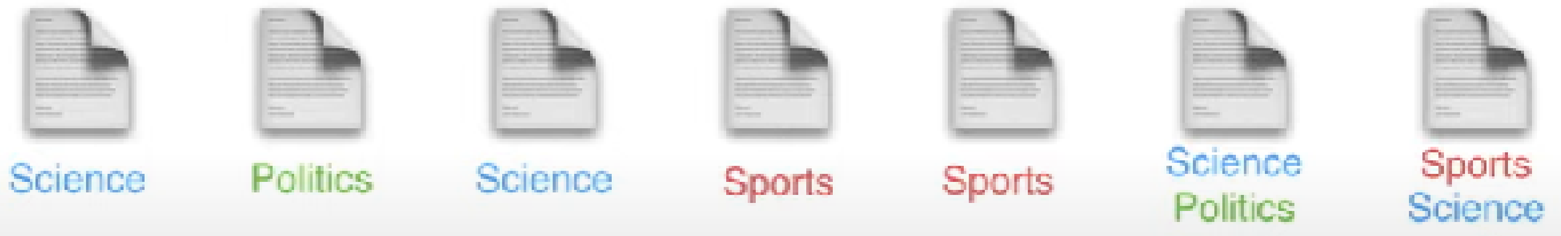


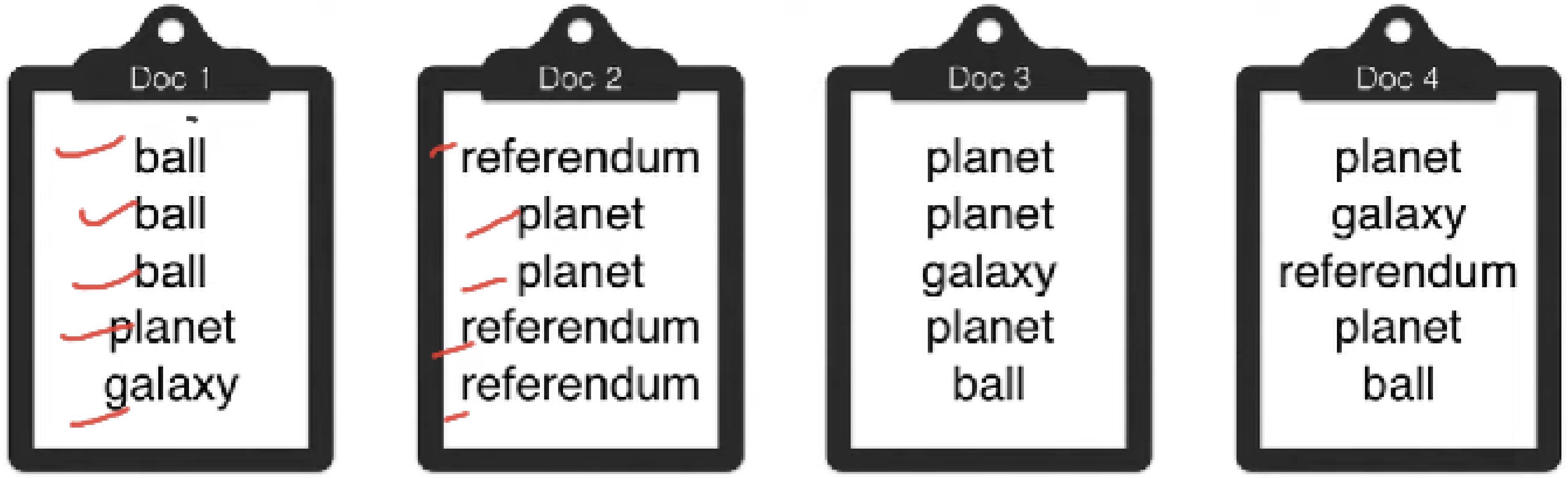


3 length = l

The problem

Goal





Doc 1
ball
ball
ball
planet
galaxy

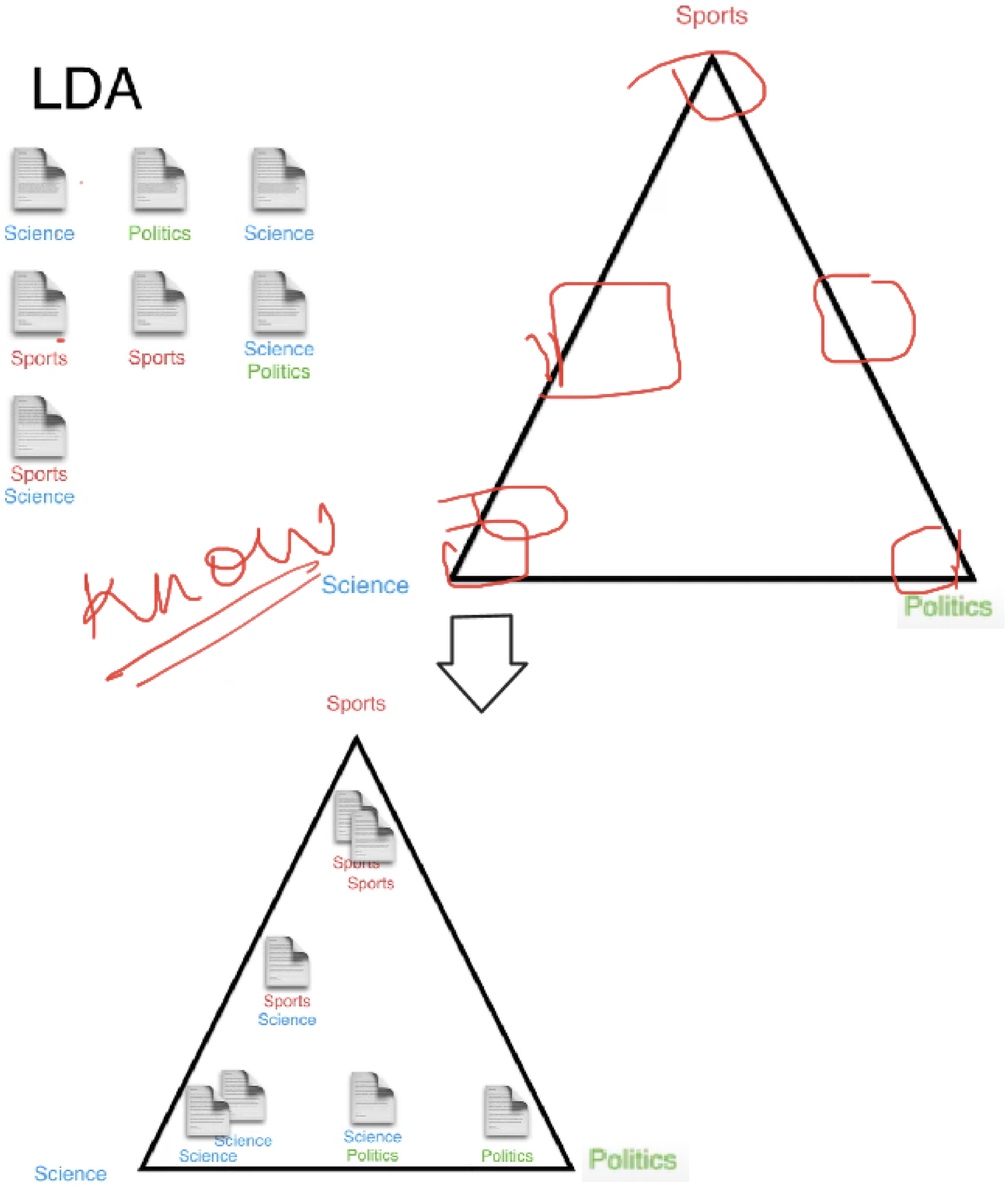
Doc 2
referendum
planet
planet
referendum
referendum

Doc 3
planet
planet
galaxy
planet
ball

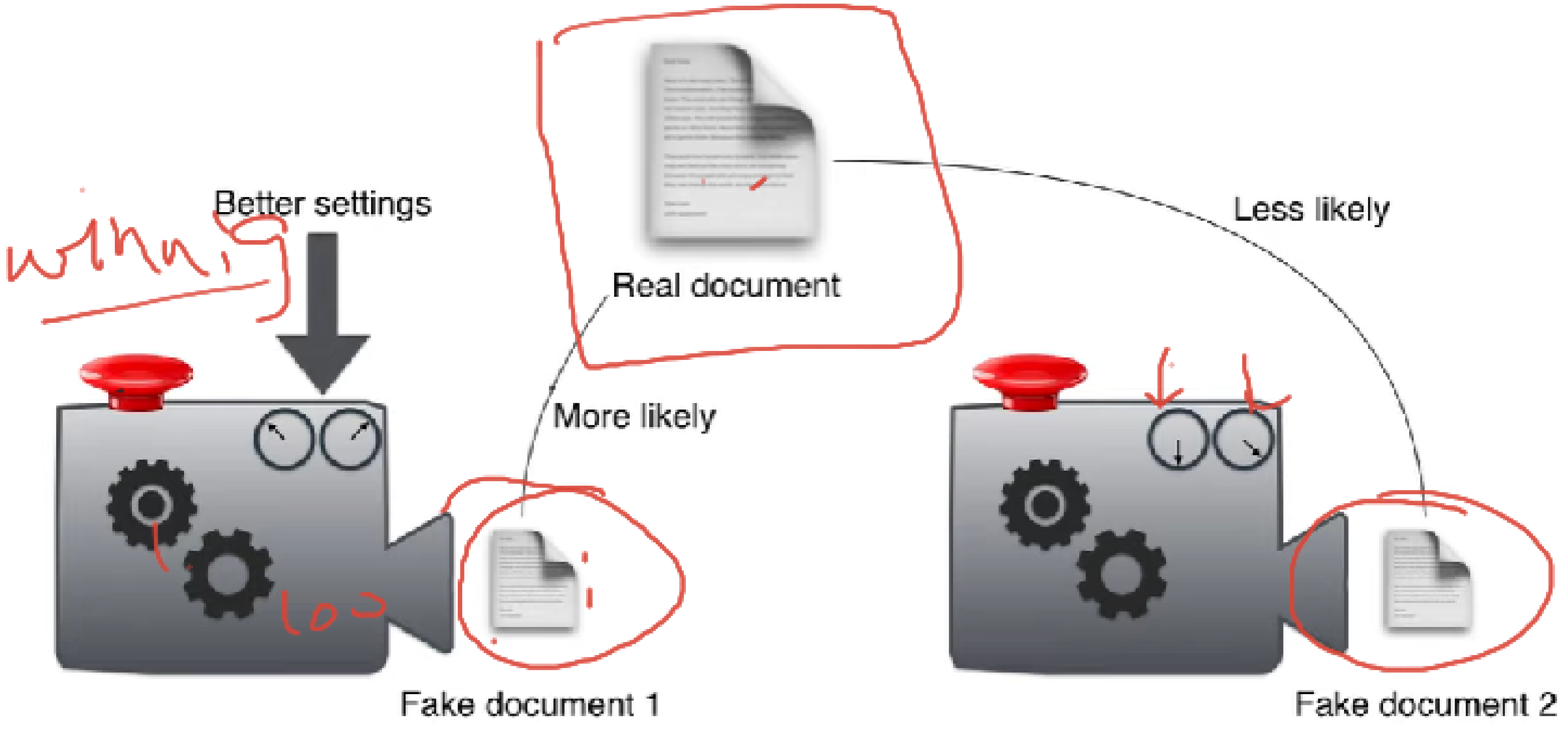
Doc 4
planet
galaxy
referendum
planet
ball

Sports
Science

LDA



Best settings on the machine



gibbers

$D_1 = 100$ words

(LDA) → module do?

perplexity

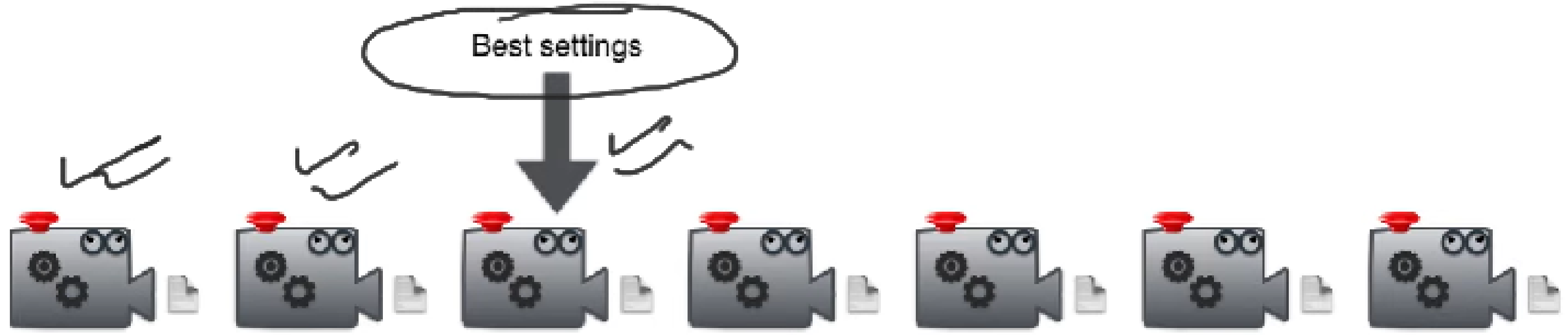
→ SS

"α" "β"

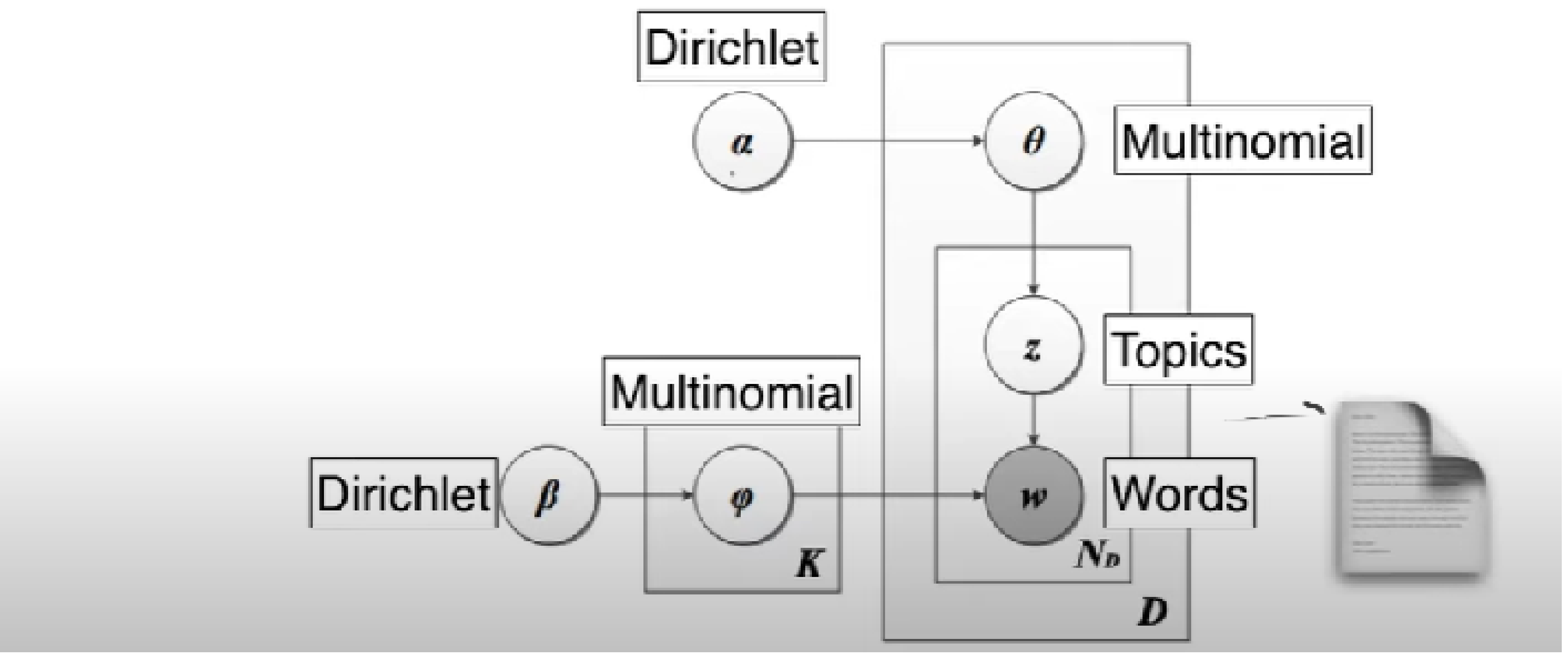
DD relate → sample

Best settings on the machine

P(same)

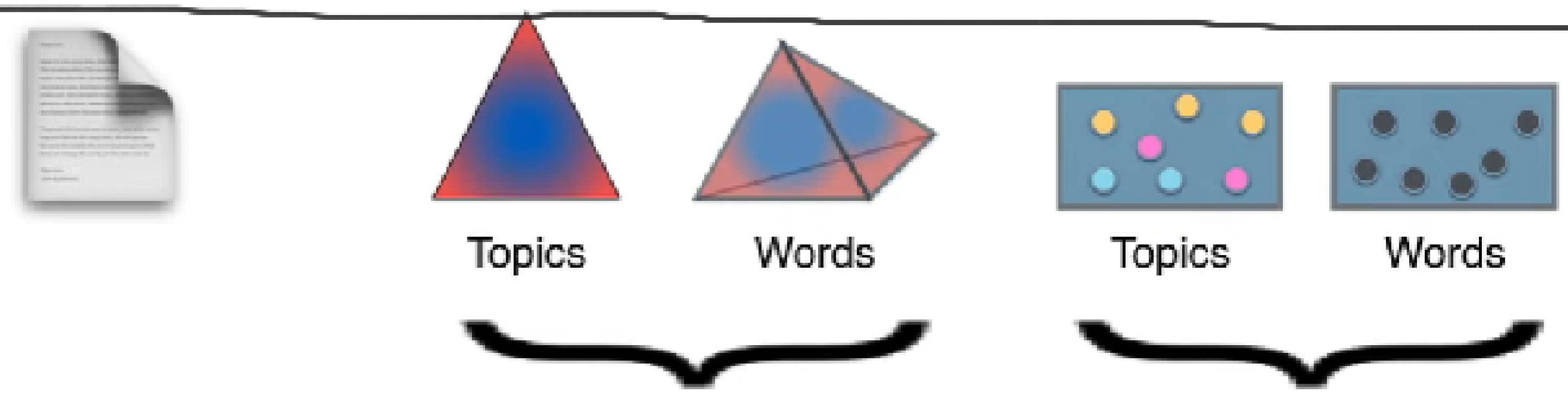


Blueprint for the LDA machine



Probability of a document

$$\boxed{P(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\varphi};\alpha,\beta)=\prod_{j=1}^{M}P(\theta_j;\alpha)\prod_{i=1}^{K}P(\varphi_i;\beta)\prod_{t=1}^{N}P(Z_{j,t}\mid\theta_j)\,P(W_{j,t}\mid\varphi_{Z_{j,t}})}$$



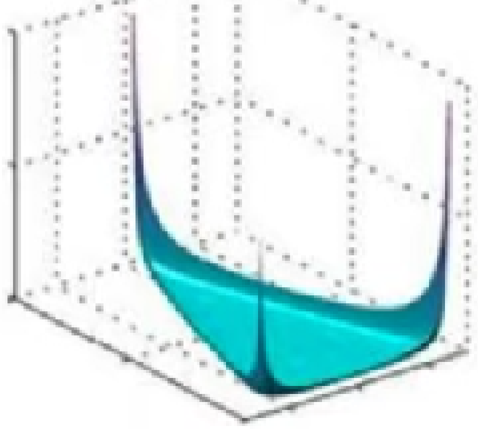
steps

Intuition-2

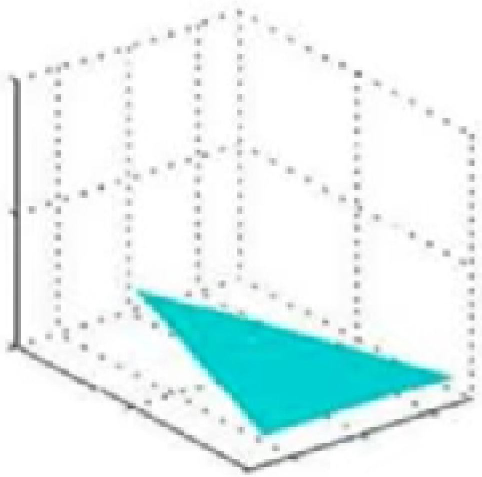
(0, '0.024*"ban" + 0.017*"order" + 0.015*"refugee" + 0.015*"law" + 0.013*"trump"
~~'+ 0.011*"kill" + 0.011*"country" + 0.010*"attack" + 0.009*"state" + '~~
'0.009*"immigration"') topic → RUW topic 1
(1, '0.020*"student" + 0.020*"work" + 0.019*"great" + 0.017*"learn" + '
'0.017*"school" + 0.015*"talk" + 0.014*"support" + 0.012*"community" + '
'0.010*"share" + 0.009*"event"') topic-2 → Arima

Dirichlet Distributions

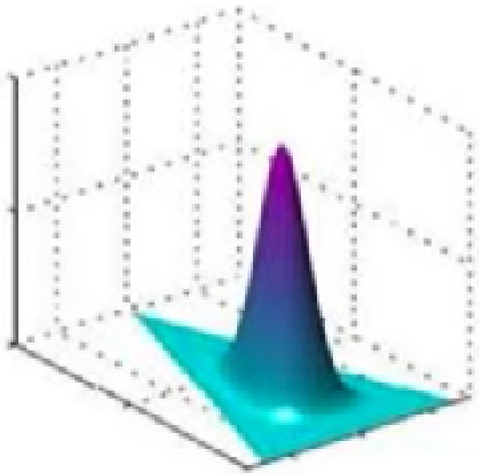
$$f\left(x_1,\ldots,x_K;\alpha_1,\ldots,\alpha_K\right)=\frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{i=1}^{K}x_i^{\alpha_i-1}$$



0.7, 0.7, 0.7



1, 1, 1



5, 5, 5

in·vata{index , ( ) }

For col is col_names

sum np·sum(________, axis)

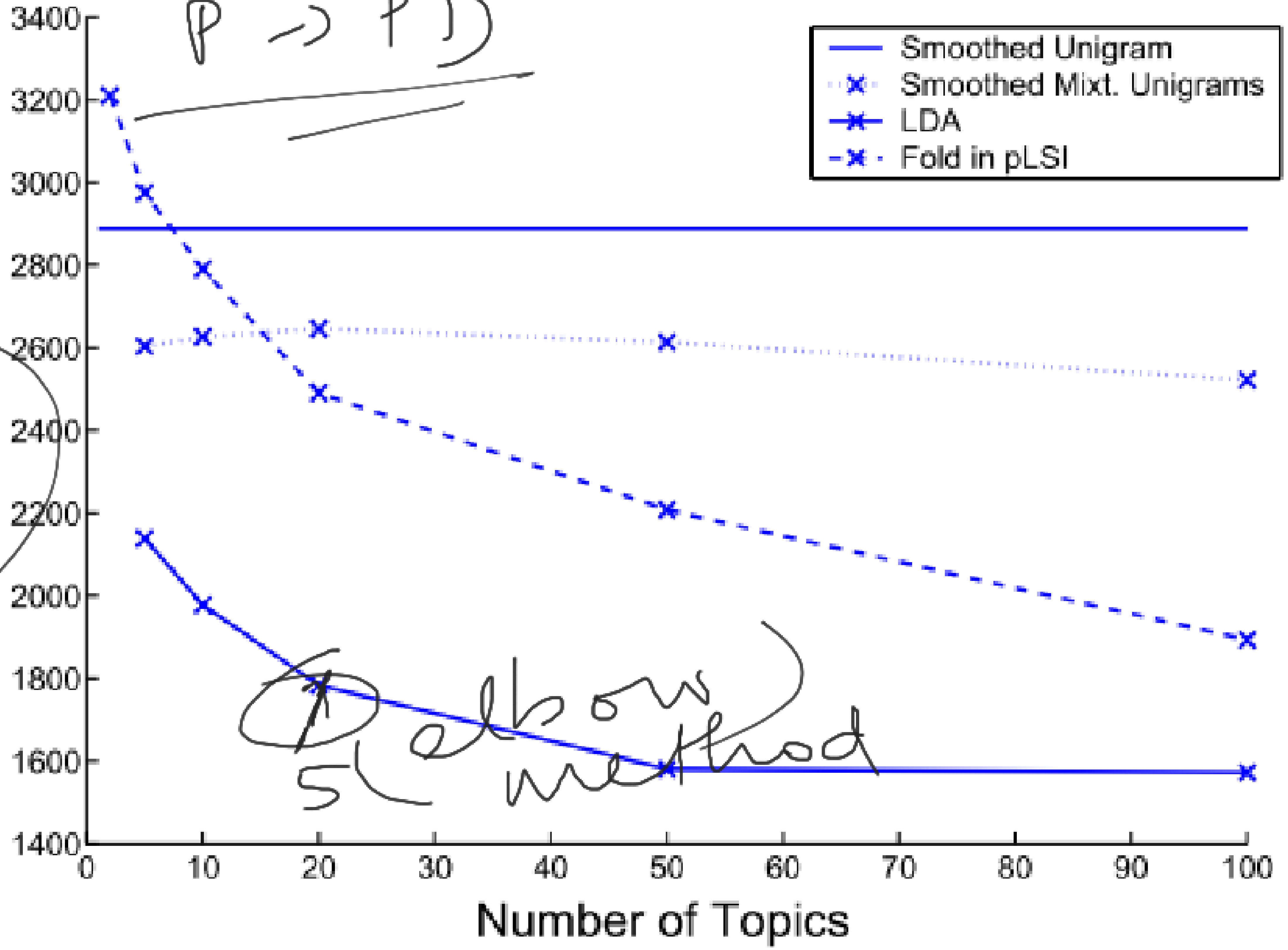
sum (is), =

Eucle SS -> PD

unsupervise

P -> PD

SS



elbow method

5