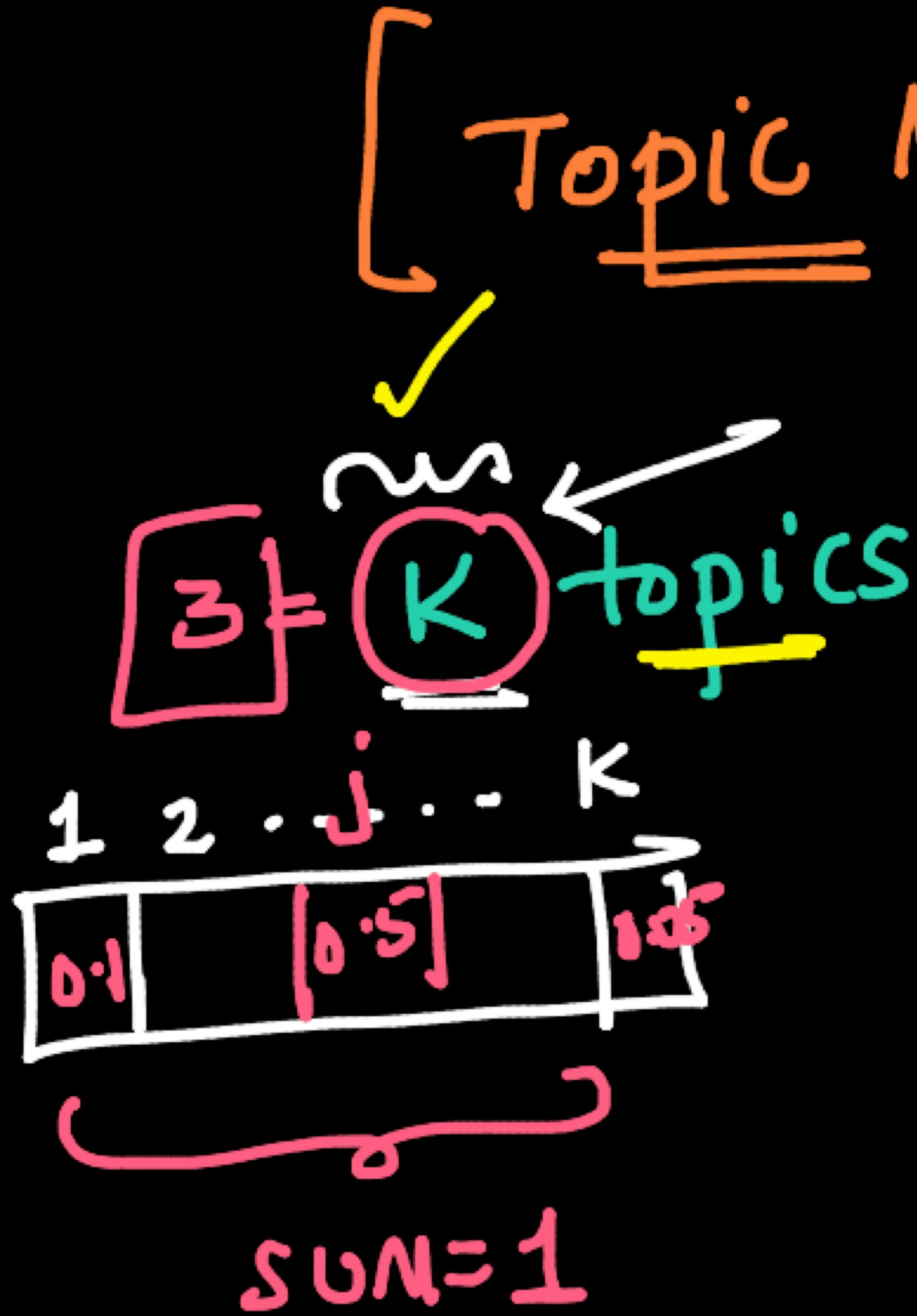
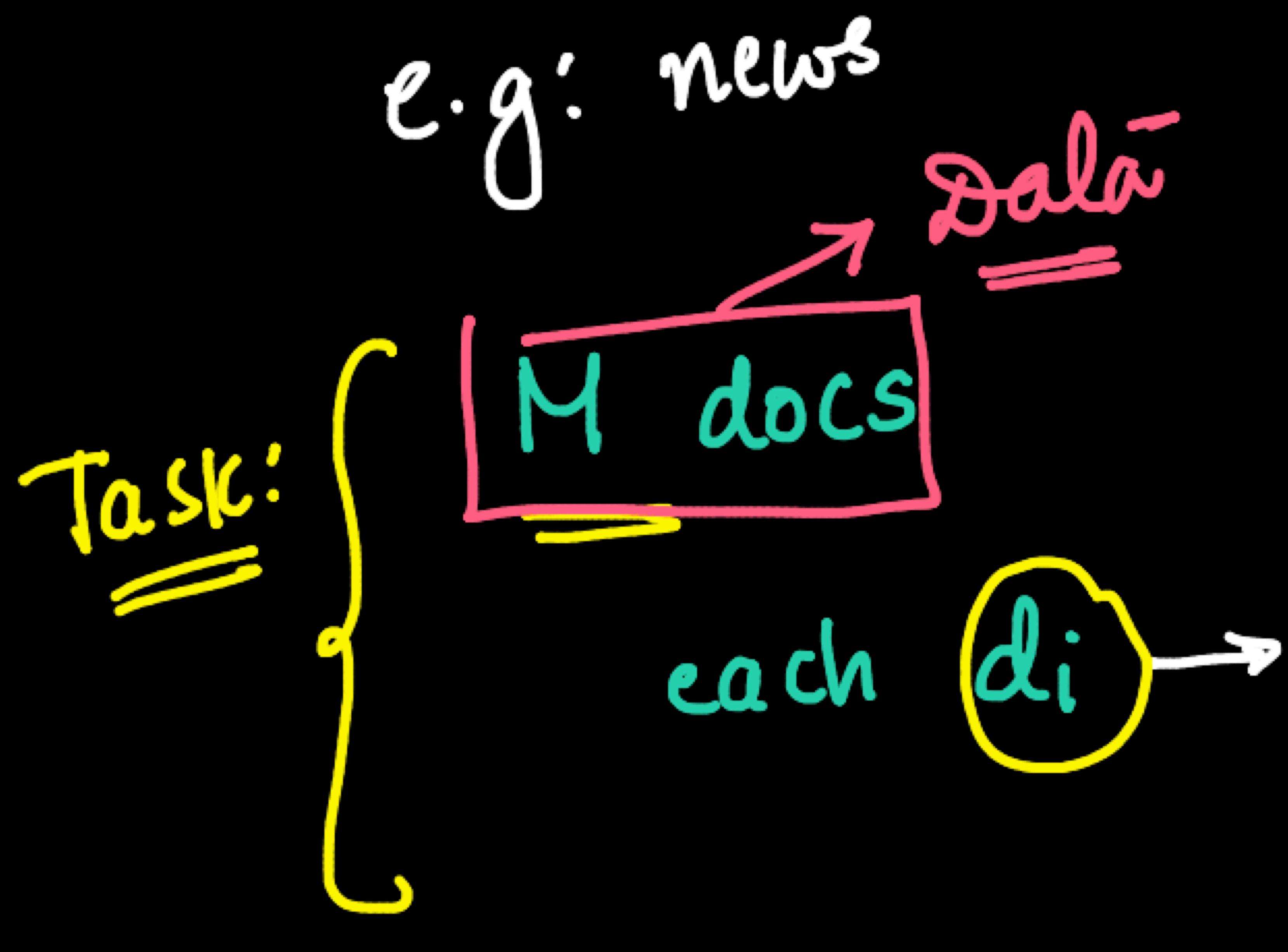


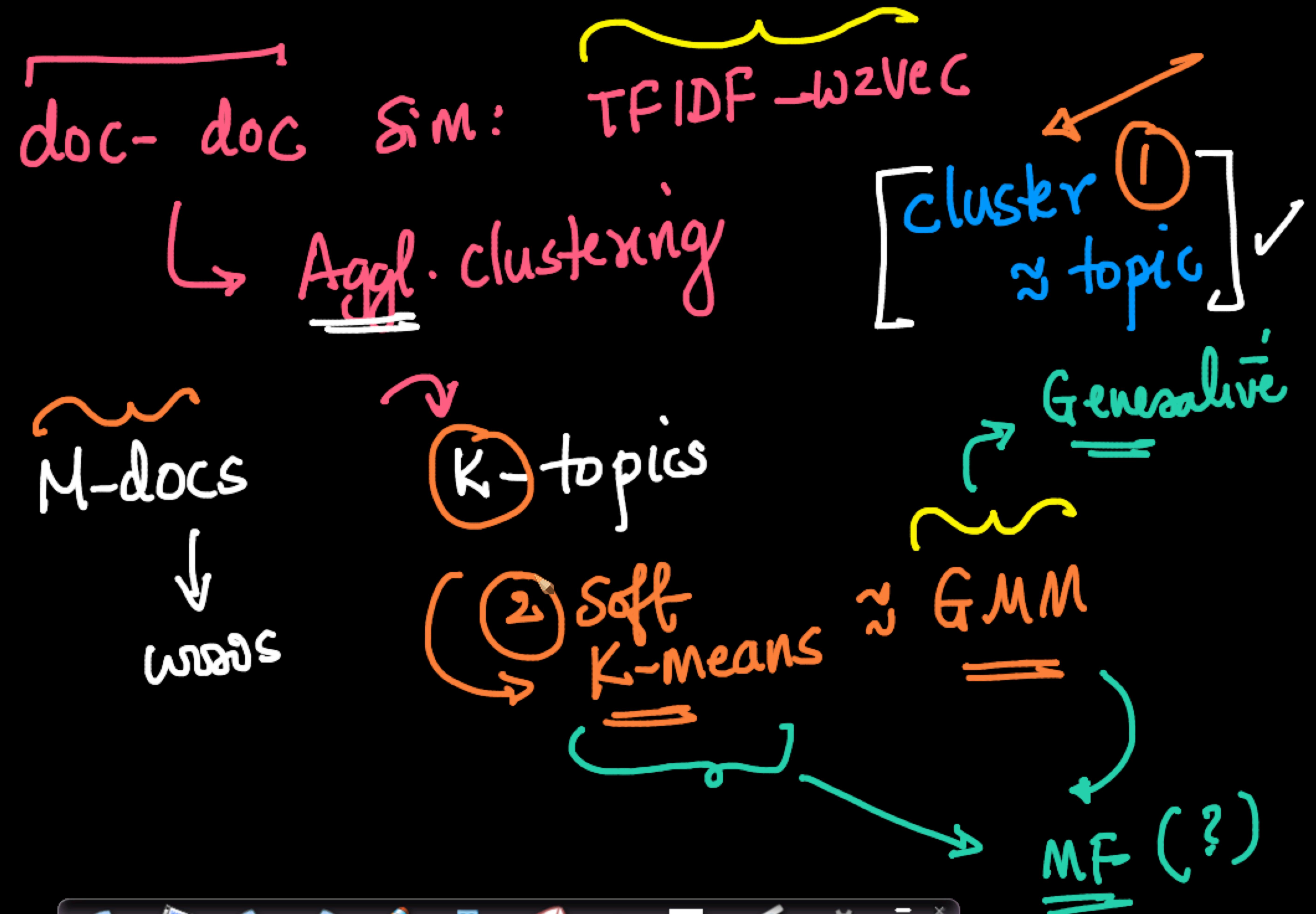
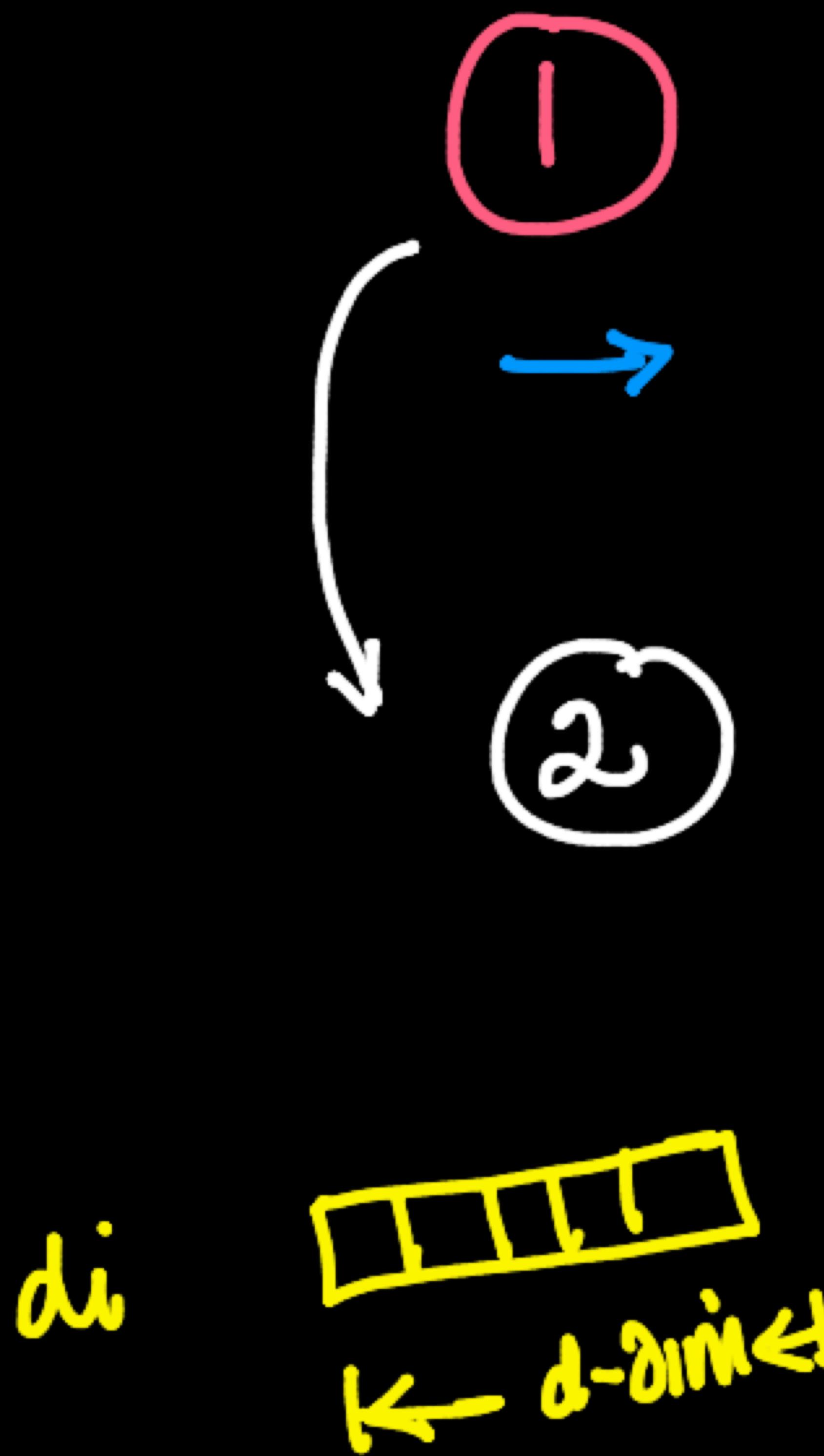
Agenda:

- ✓ ① Parts of Speech Tagging
- ✓ ② Topic Modelling

[Topic Modeling:]



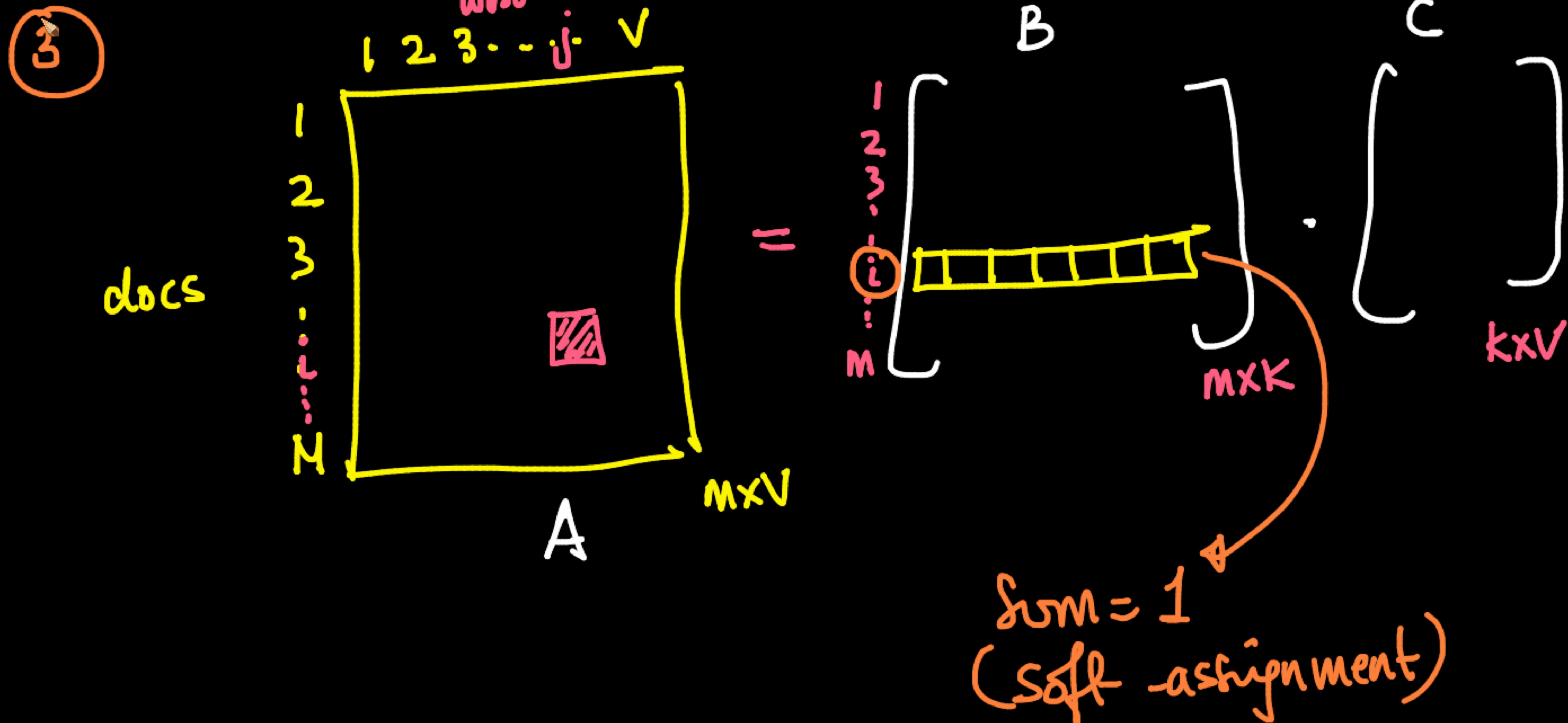
Ideas:



M-docs

V: VOC size

K-topics (soft-assign)



clustering
✓ MF

$$A \approx B \cdot C$$

s.t. each row B should sum to 1

Loss: typical MF

$$\sum_{i,j} (A_{ij} - B_i \cdot C_j)^2$$

4

Bⁱ th now →

prob. dist' sb
of i'h doc's
assignment to k-
topics

✓ { KL-divergence : measure how different
2 n'sb are?

Summary:

M-docs

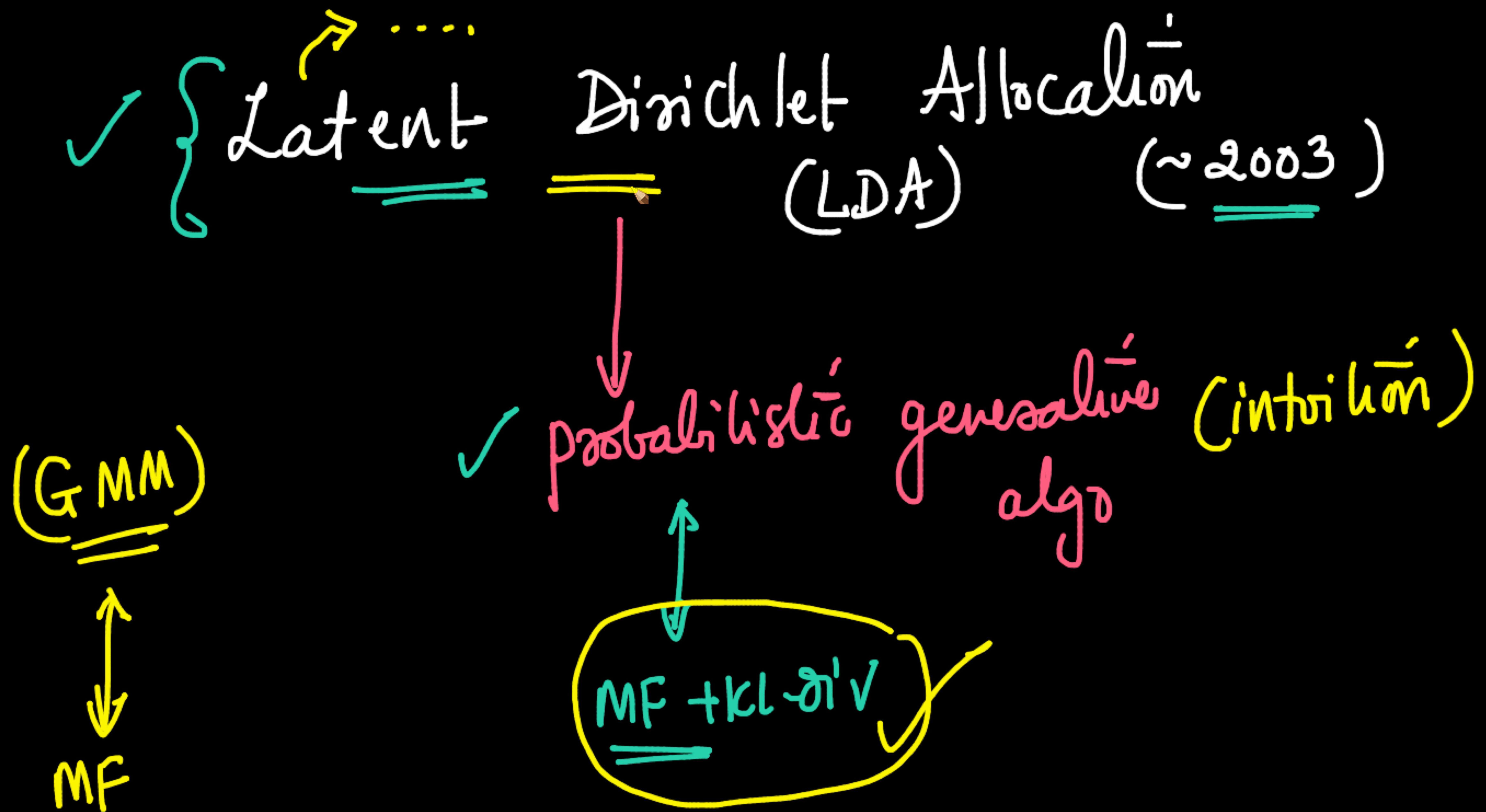
topics
K: numerical

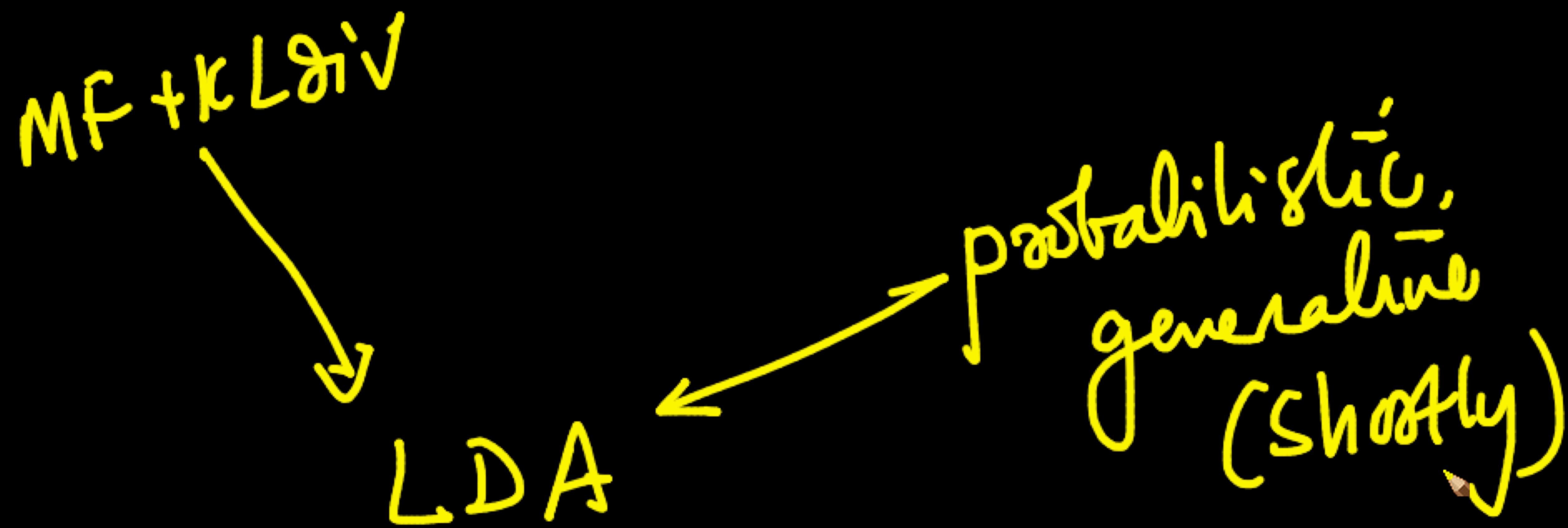
soft k-clustering

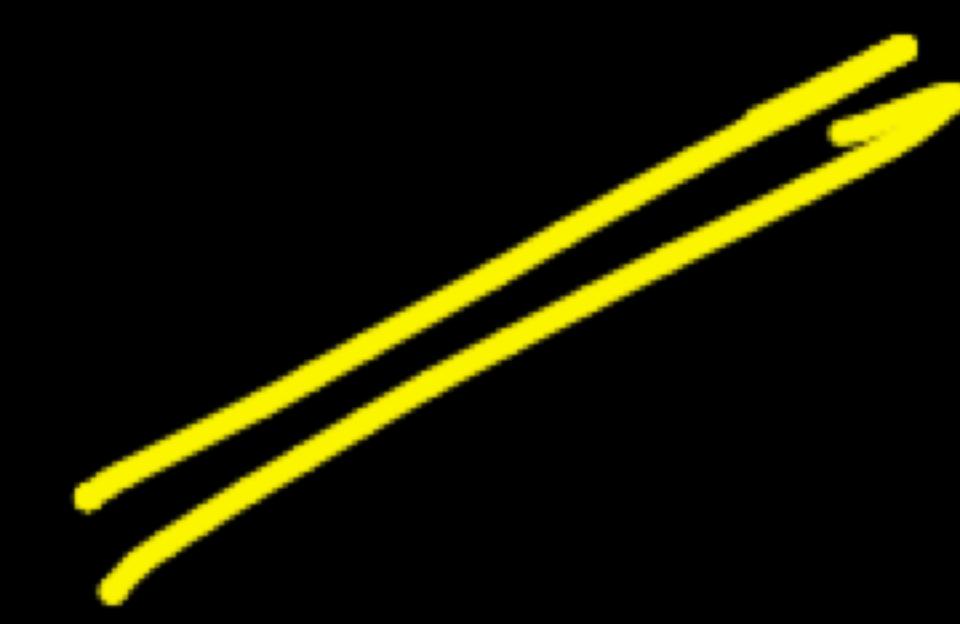


MF + KL-divergence









✓ [pauls] A speech (pos)
tagging

sentence: $w_1 \quad w_3 \quad w_6 \quad w_{12}$
 $N \quad PR \quad \checkmark \quad \text{Adv} \dots$
 $\underbrace{\quad}_{\text{NP}}$

pos-tagging (NLP)

↳ rule-based / Grammar-based

→ probabilistic tech

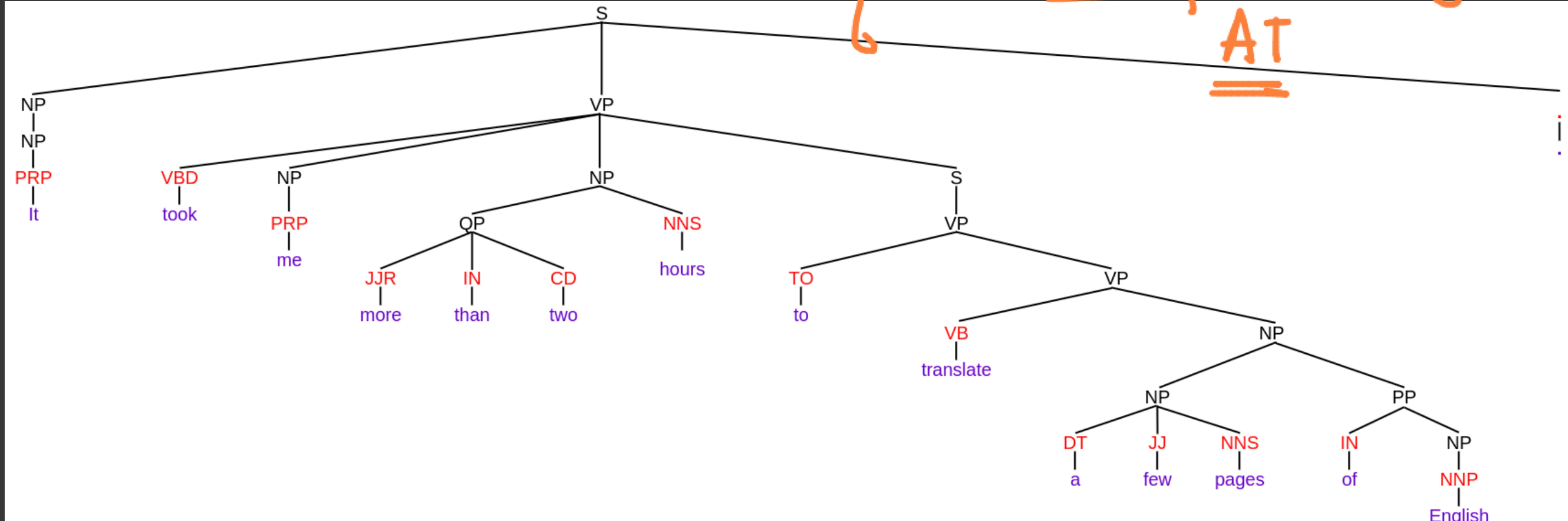
→ DL-techniques

+ Code + Text

connect ▾

1

The image shows a mobile application's user interface. At the top, there is a dark grey navigation bar with white icons: a checkmark, the text "Connect" followed by a downward-pointing arrow, a user profile icon, and a gear settings icon. Below this is a large orange header area. Inside the orange header, the text "CS: Compiler design" is written in a hand-drawn, orange, cursive-style font. Below this main title is a smaller, stylized pencil icon. At the bottom right of the screen, the letters "AT" are partially visible, suggesting another part of the word "Compiler" or a related term. The overall theme is educational or professional, specifically related to computer science and compiler design.



If you use spacy, though, you don't need to worry about writing your own rules for getting Noun Phrases.

PosTagging_Topic_Modelling_ x nouns in english grammar - Go x +

colab.research.google.com/drive/1xaSfjSo1IJZkzfqMhGQL1tXI7jZliCEg#scrollTo=ZEYf3MC-L1DA

+ Code + Text Connect | Update

```
[ ] counter_all = Counter(flatten(review_data["noun_phrases"].tolist()))
counter_all.most_common(10)
```

{x}

file

```
✓('guitar', 3212),
✓('strings', 2526),
('price', 1795), ←
('sound', 1698),
✓('pedal', 1539),
('one', 1256),
('amp', 1173),
('time', 1146),
('tone', 1042),
('guitars', 984)] ✓
```

But... these still aren't very actionable.

For example, we still don't know what are people talking about regarding guitars? What word/words is it related to?

13 / 13

PosTagging_Topic_Modelling_

nouns in english grammar - Go

colab.research.google.com/drive/1xaSfjSo1IJZkzfqMhGQL1tXI7jZliCEg#scrollTo=ZEYf3MC-L1DA

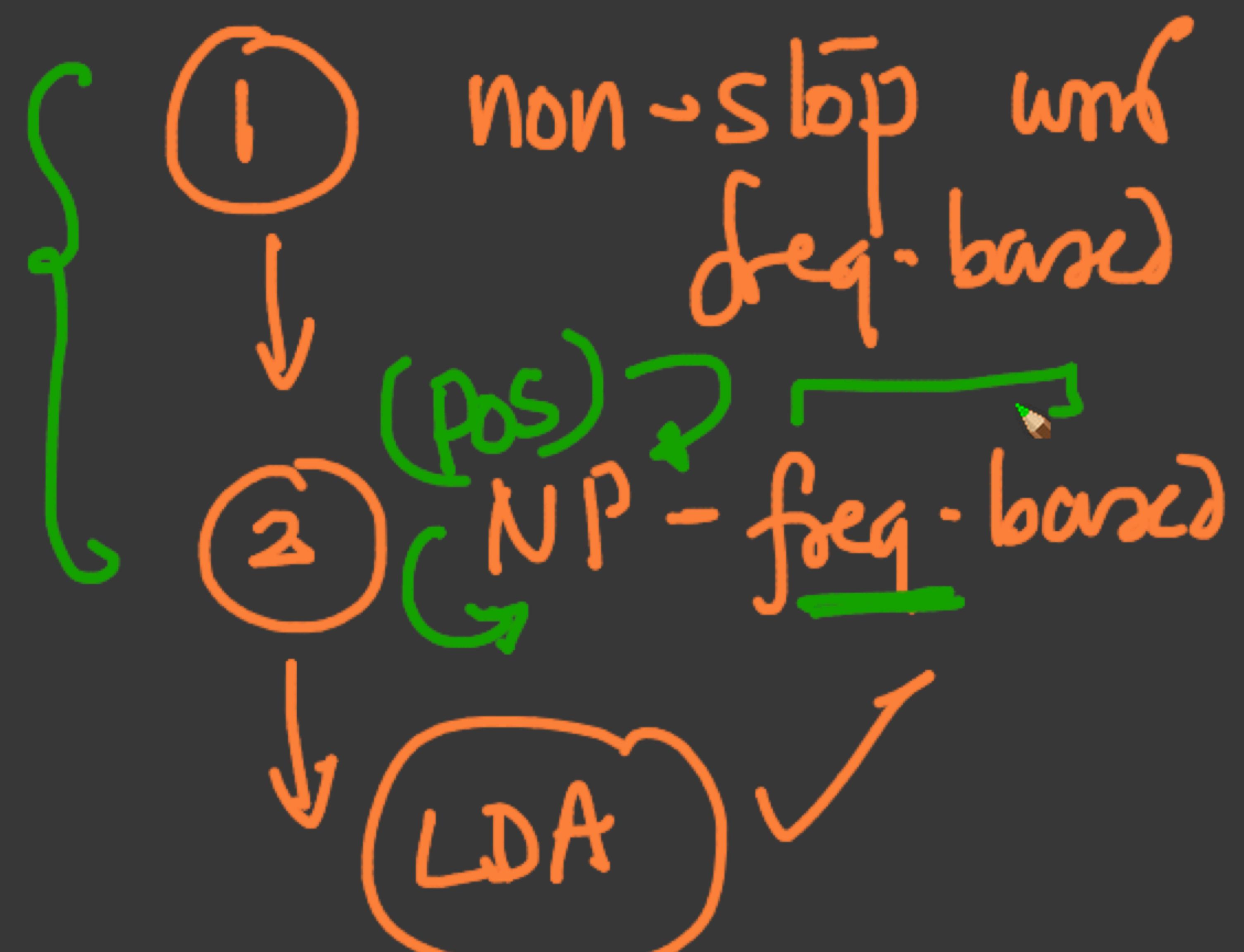
+ Code + Text

Connect



```
[ ] counter_all = Counter(flatten(review_data["noun_phrases"].tolist()))
counter_all.most_common(10)
```

```
{x}
[ ('guitar', 3212),
 ('strings', 2526),
 ('price', 1795),
 ('sound', 1698),
 ('pedal', 1539),
 ('one', 1256),
 ('amp', 1173),
 ('time', 1146),
 ('tone', 1042),
 ('guitars', 984)]
```



But... these still aren't very actionable.

For example, we still don't know what are people talking about regarding guitars? What word/words is it related to?

PosTagging_Topic_Modelling_ x nouns in english grammar - Go x +

colab.research.google.com/drive/1xaSfjSo1IJZkzfqMhGQL1tXI7jZliCEg#scrollTo=W5gNvvUpL1DD

Connect |

100% 10261/10261 [04:08<00:00, 25.99it/s]

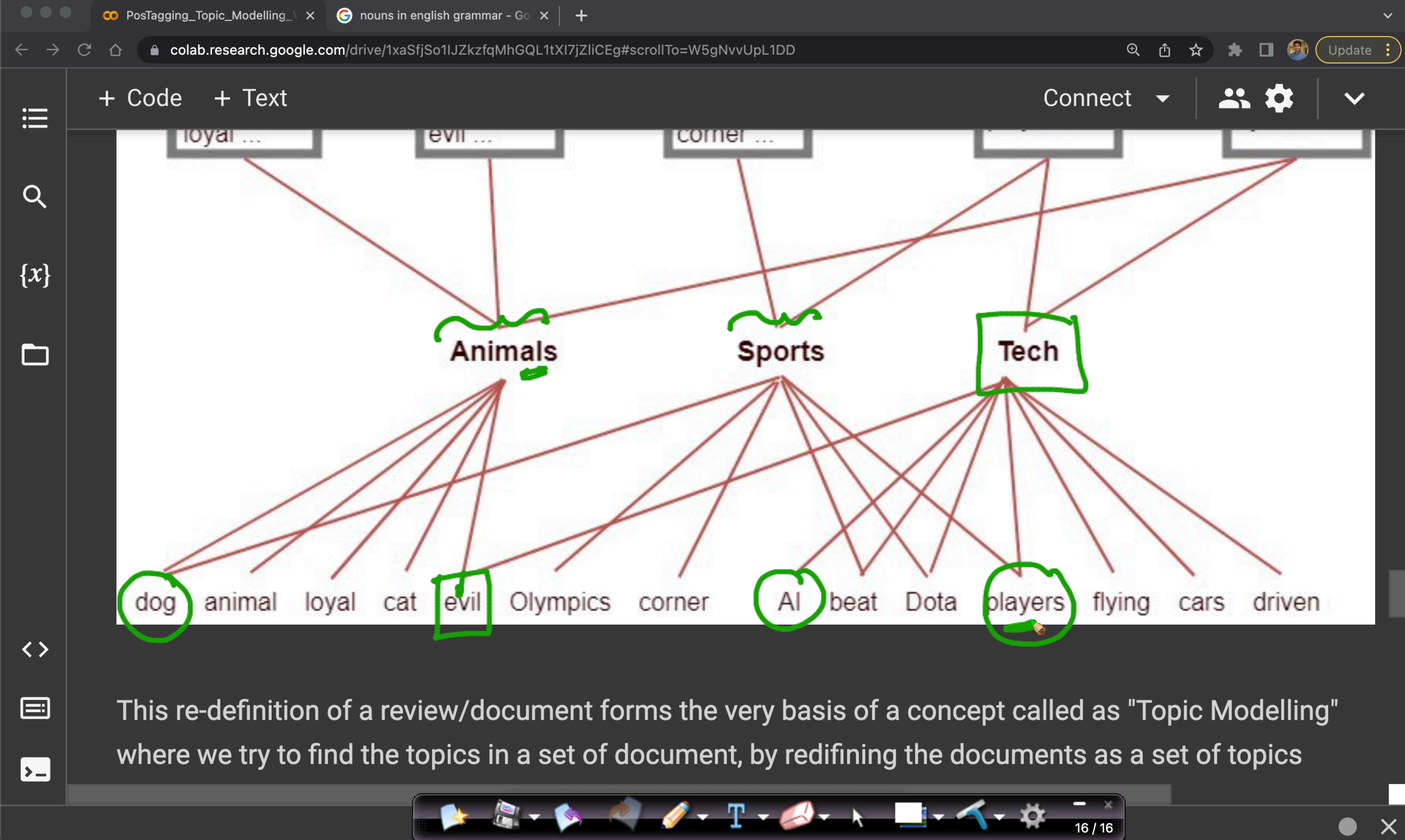
{x} counter_all = Counter(flatten(review_data["noun_phrases_parent"].tolist())) counter_all.most_common(10)

[('price->for', 583),
 ('years->for', 503),
 ('guitar->on', 400),
 ('thing->is', 347),
 ('strings->are', 278),
 ('pedal->is', 261),
 ('guitar->of', 255),
 ('tune->in', 255),
 ('job->does', 245),
 ('strings->of', 242)]

<>

Hmm that turned out to not be adding much information for our purpose

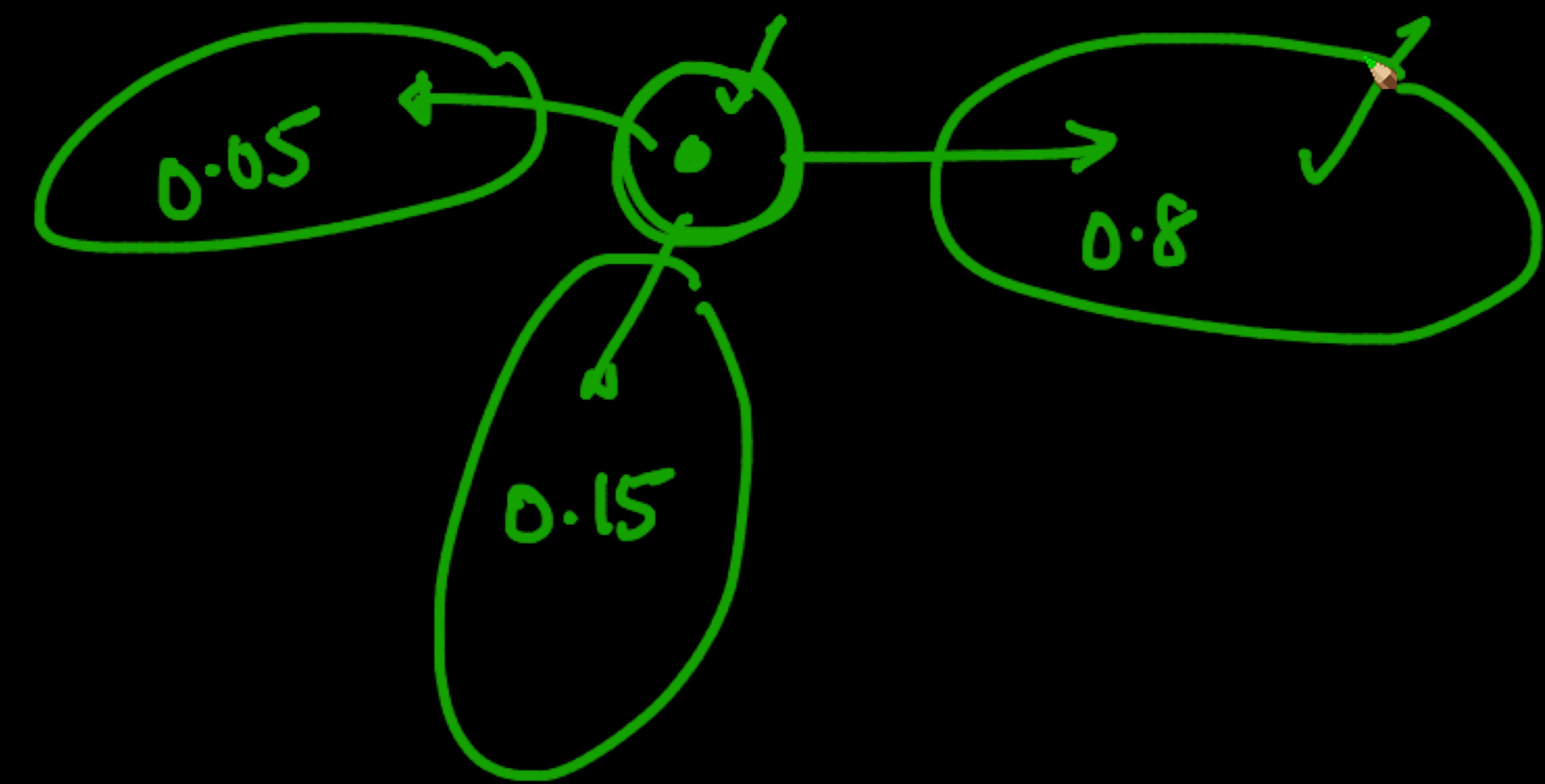
15 / 15



This re-definition of a review/document forms the very basis of a concept called as "Topic Modelling".

GMM:

$\mu = 3$



PosTagging_Topic_Modelling_ X nouns in english grammar - Go + colab.research.google.com/drive/1xaSfjSo1IJZkzfqMhGQL1tXI7jZliCEg#scrollTo=4QMTUARUL1DE Update

+ Code + Text Connect |

☰ {x} ☰

Diagram illustrating a process flow across four grounds (Theta, Beta, Gamma, Delta) involving entities like Cats, Dogs, Flying, Loyal, Evil, Olympics, AI, Players, Sport, Animals, Tech, and various stick figures (Alpha, Beta, Gamma, Delta).

The process is described in two steps:

1. Pick a ball from the ground "Theta". A stick figure labeled "Alpha" is shown picking a red ball from a cluster of balls in the "Sport" area of ground Theta.
2. Based on the ball you pick, you're sent to another ground "Beta". A red arrow points from the "Sport" area of Theta to the "Animals" area of Beta.

Organizational structures are shown as follows:

- Ground Beta:** An arrow labeled "Organize ground Beta" points to a stick figure labeled "Eta". A box contains the following list:
 - Cats
 - AI
 - Dogs
 - Evil
 - Cats
- Ground Gamma:** An arrow labeled "Organize ground Gamma" points to a stick figure labeled "Gamma".
- Ground Delta:** An arrow labeled "Organize ground Delta" points to a stick figure labeled "Delta".

Entity locations are indicated by colored dots (grey, blue, green) within the octagonal boundaries of each ground.

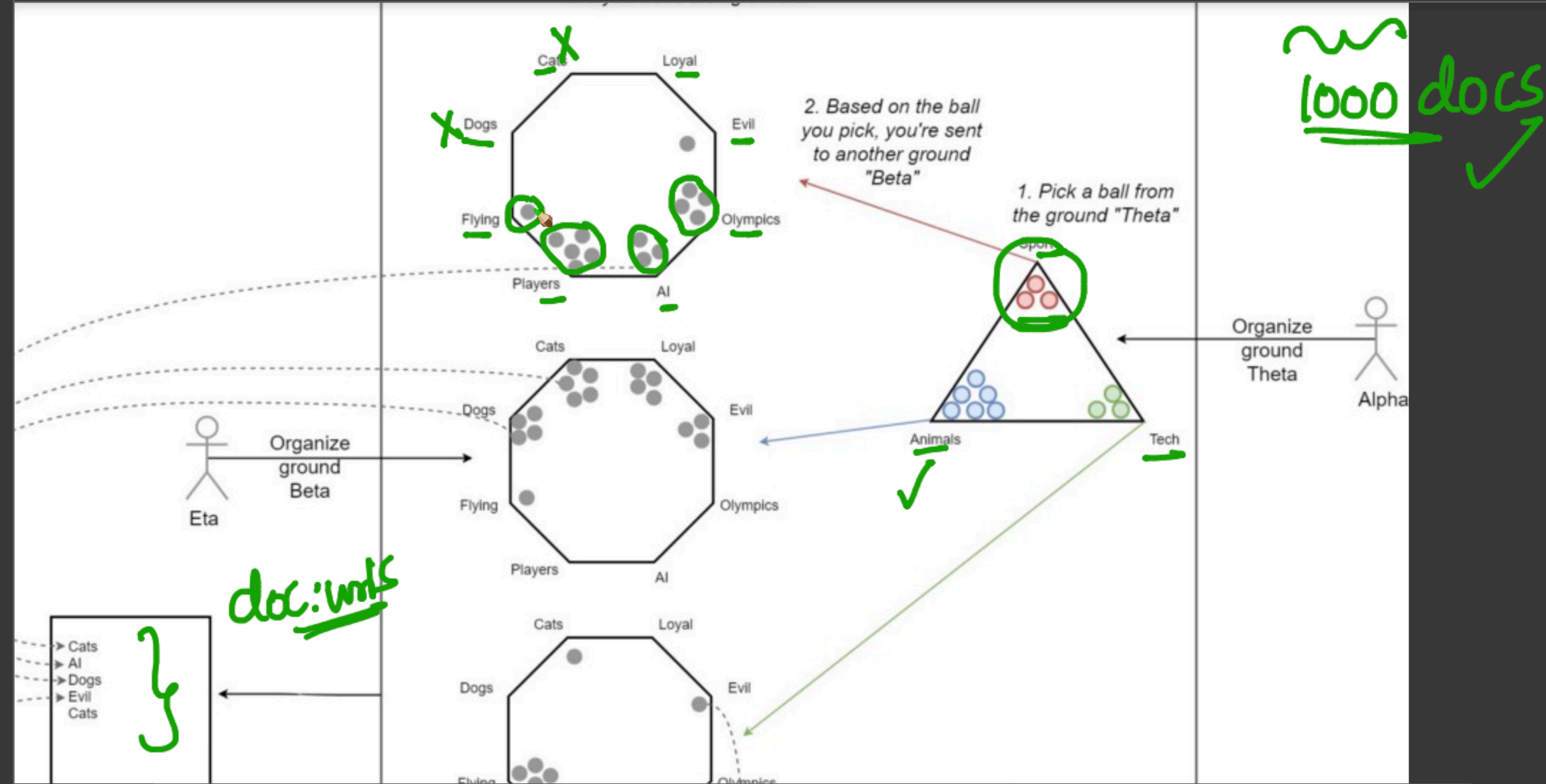
Bottom navigation bar icons: star, folder, file, pencil, text, document, gear, etc.

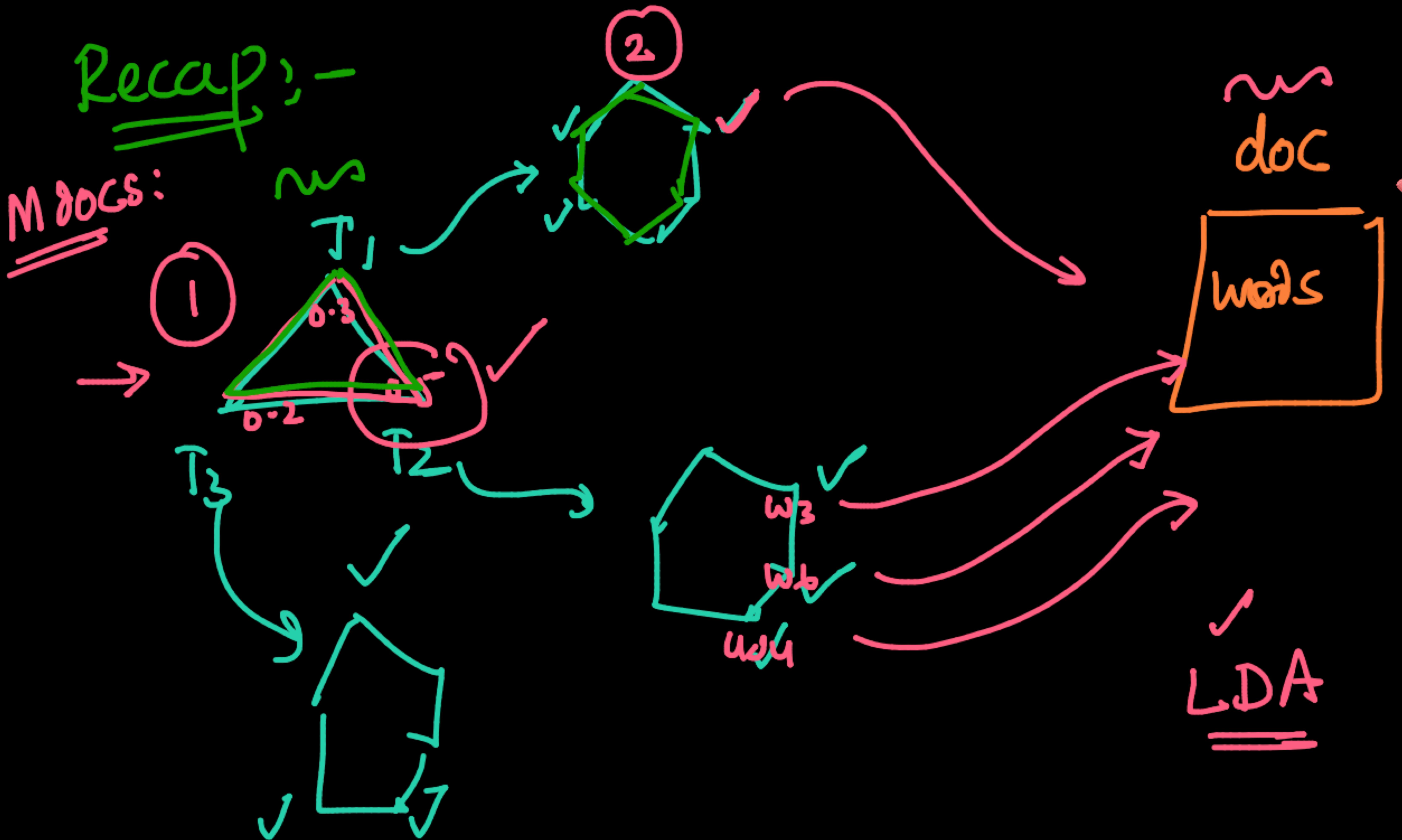
Page number: 19 / 19



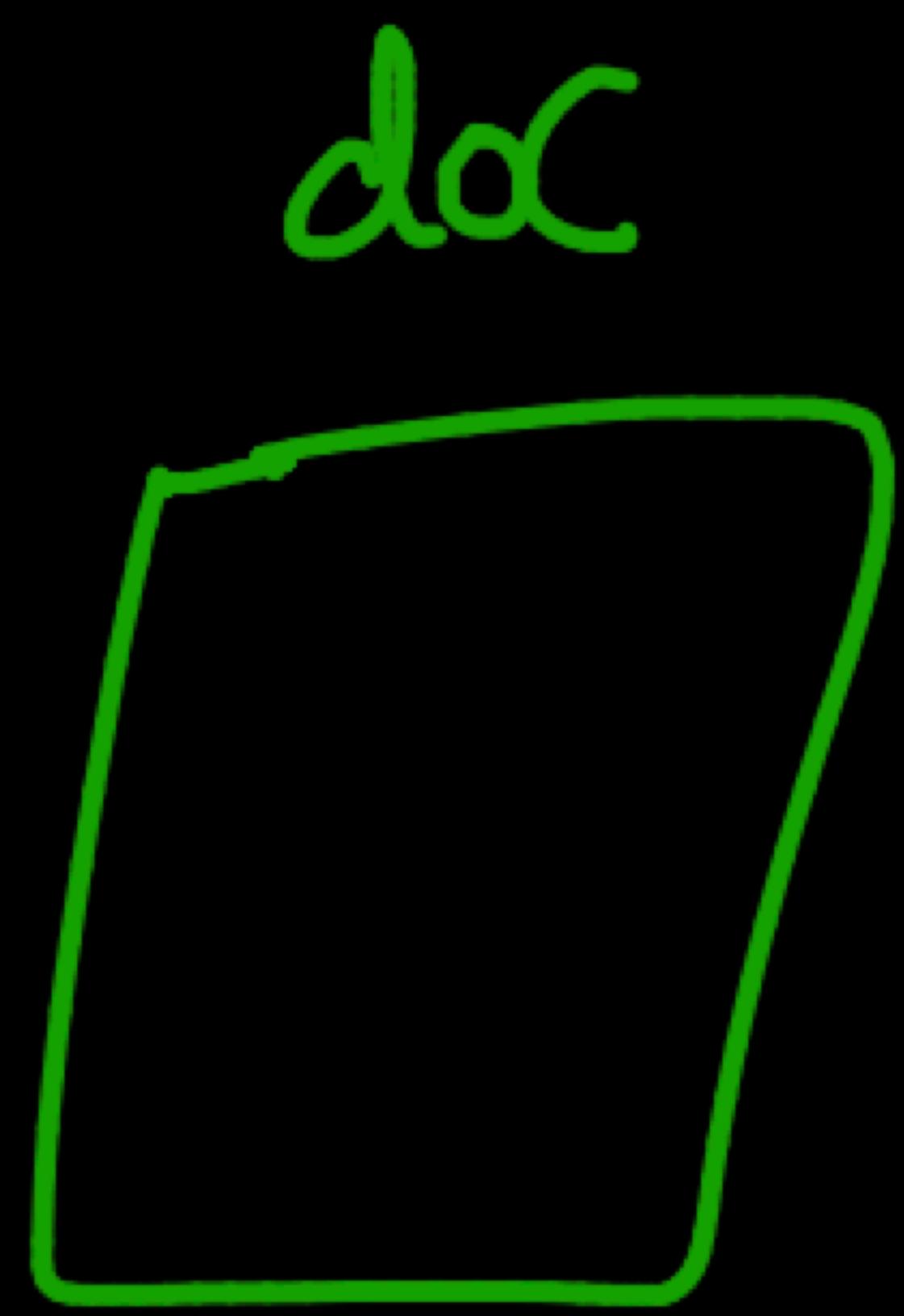
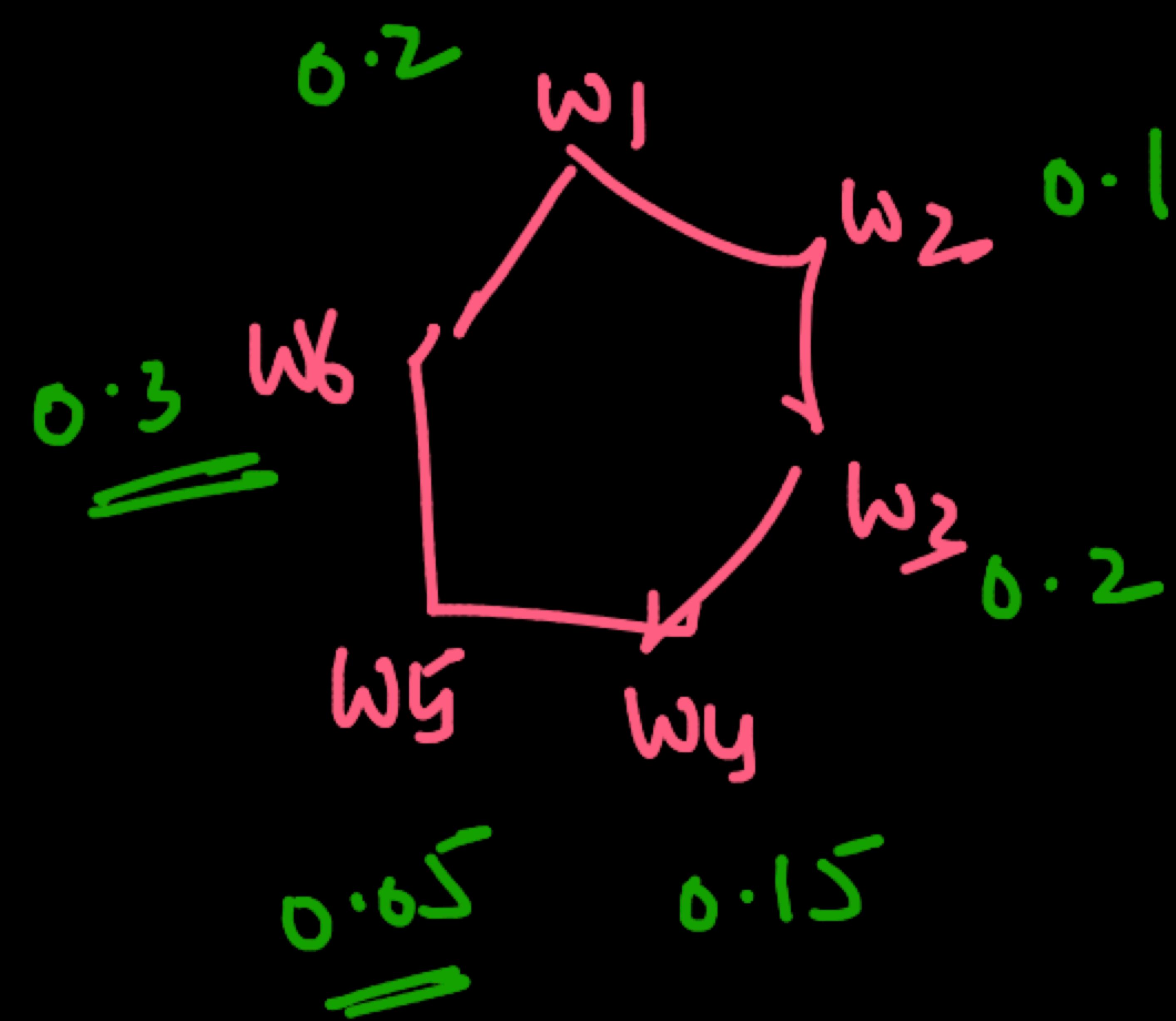
+ Code + Text

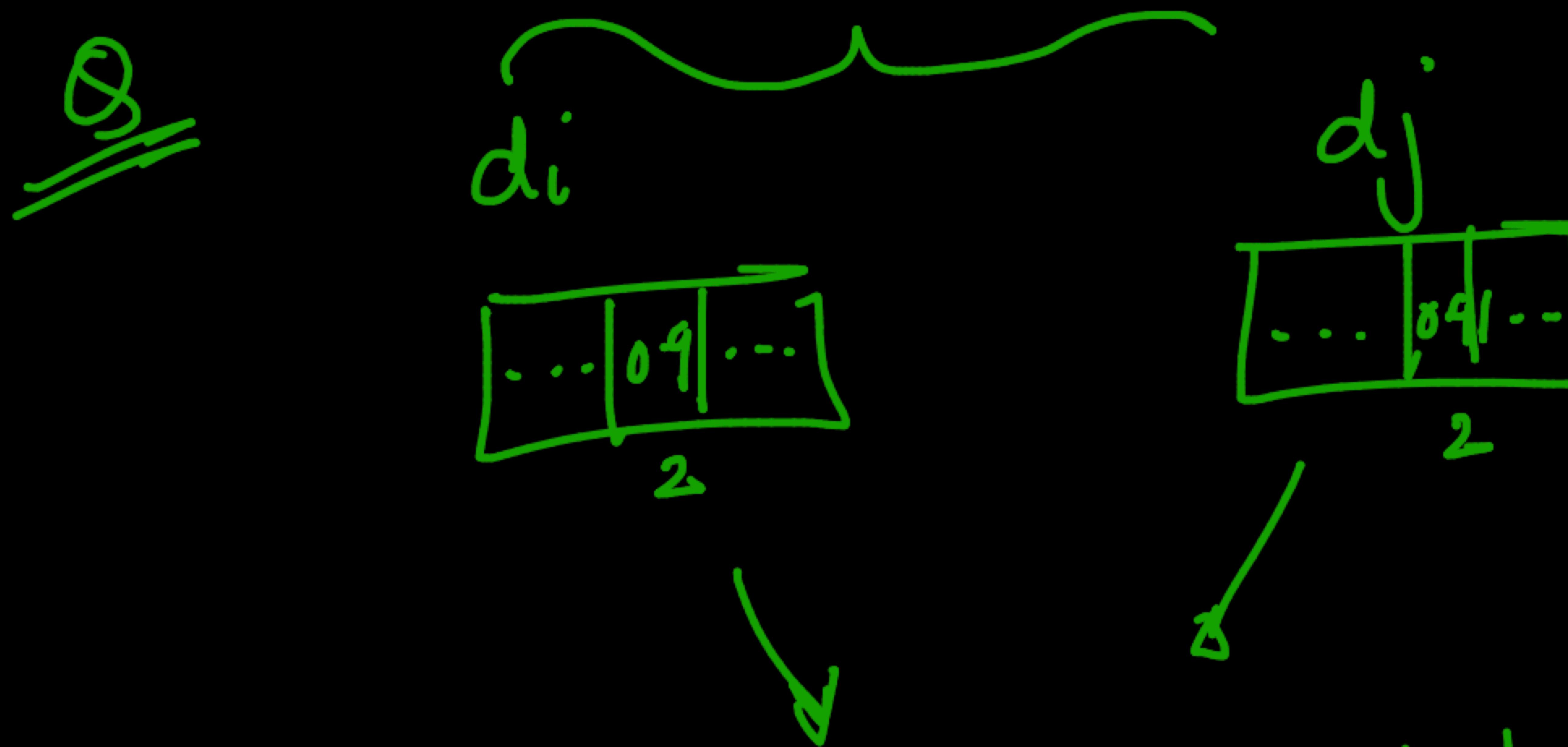
Connect





disk A
watts





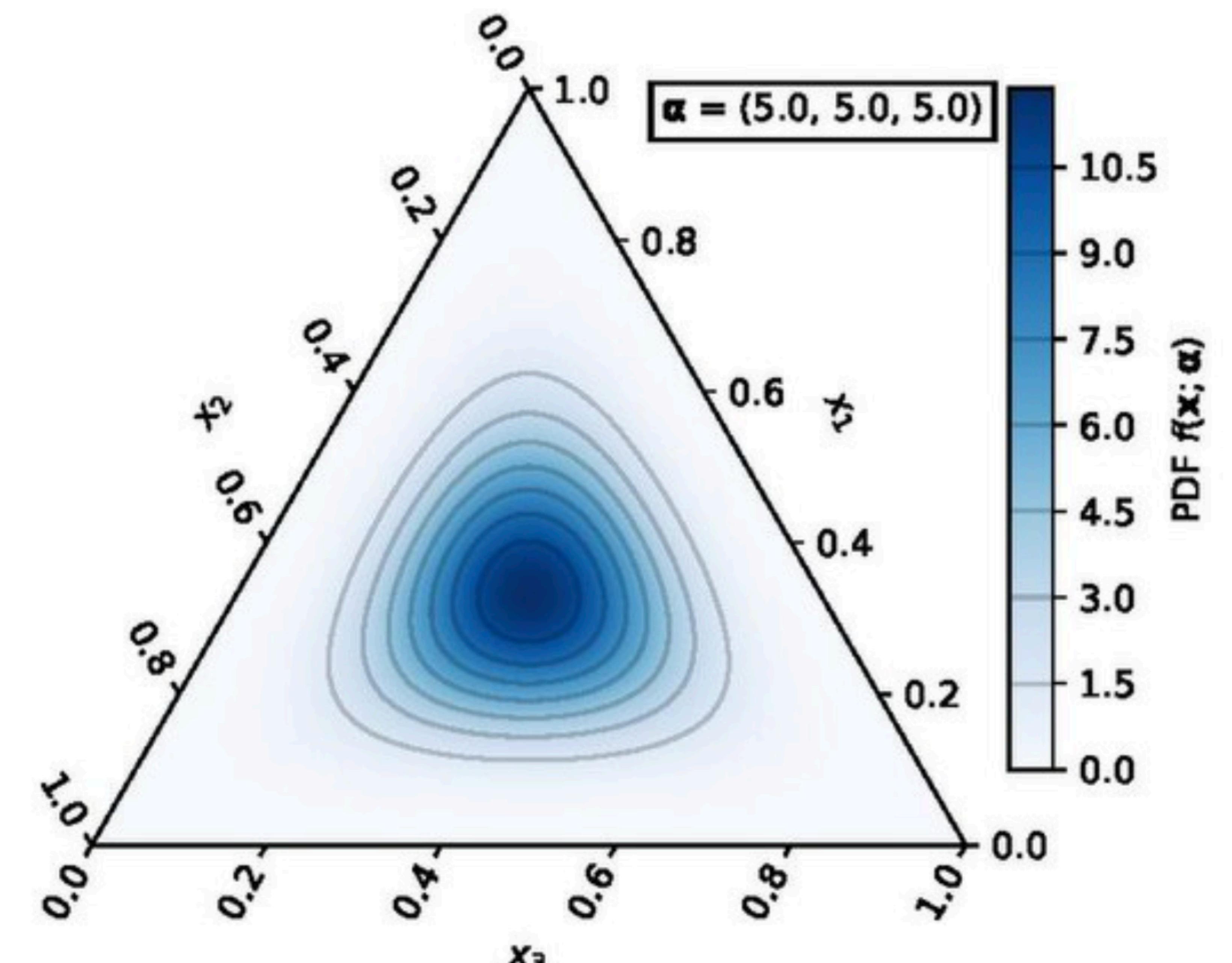
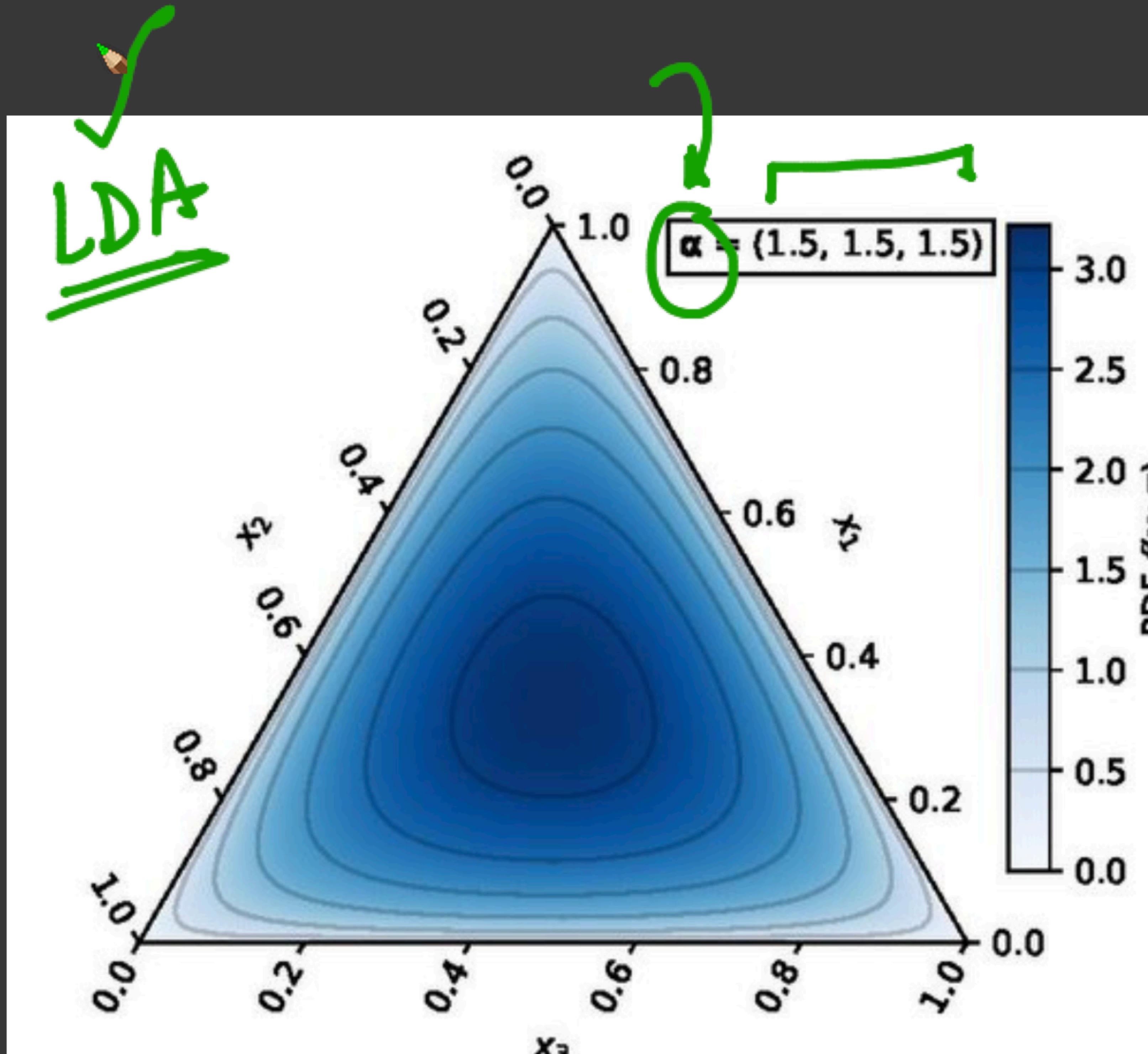
Similar words if doc's are long...

+ Code + Text

Connect

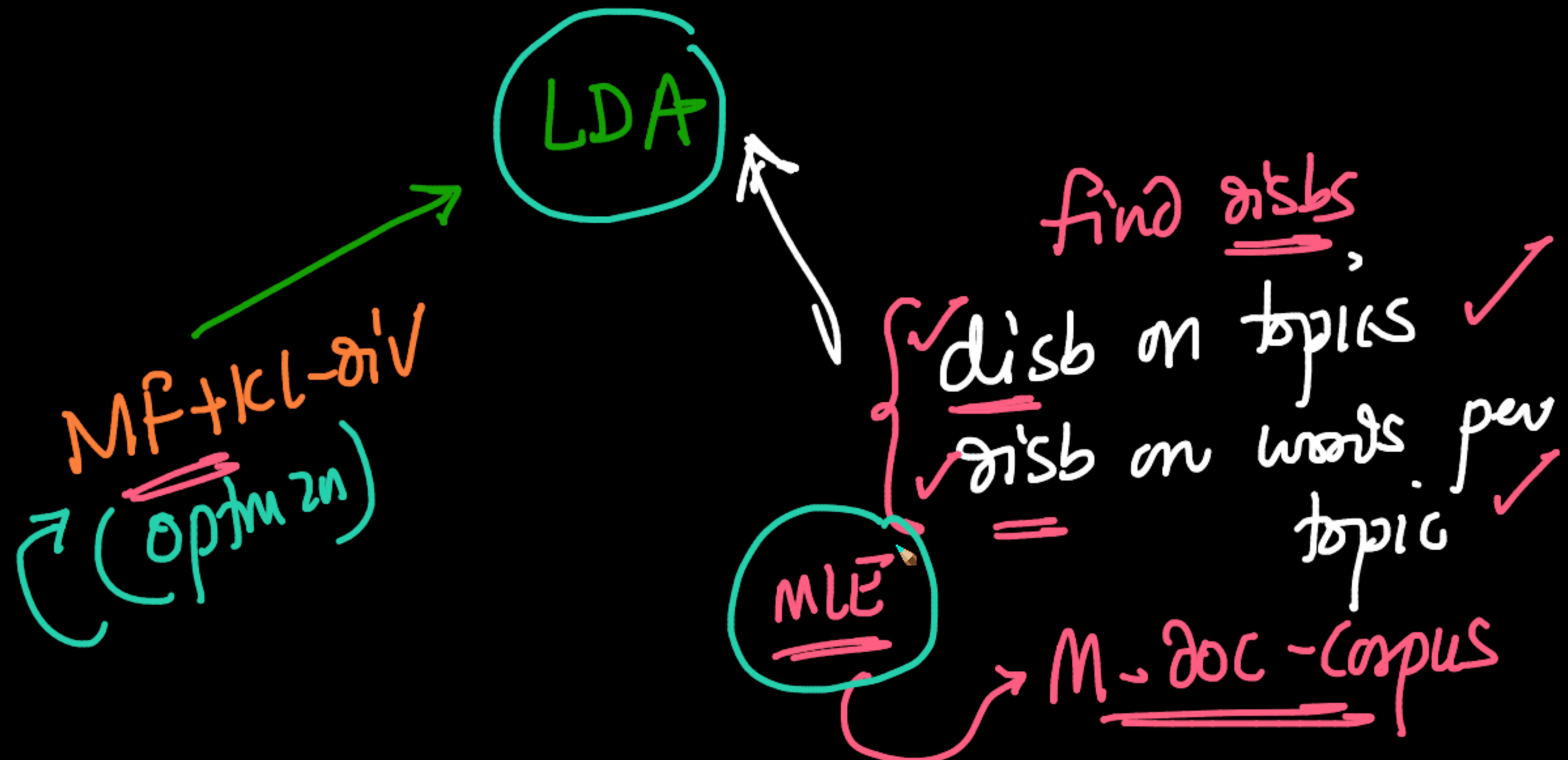


Below are several visualization of dirichlet distribution at different α for 3 topics



x_1 1.0
 x_2 1.0
 x_3 1.0

x_1 1.0
 x_2 1.0
 x_3 1.0



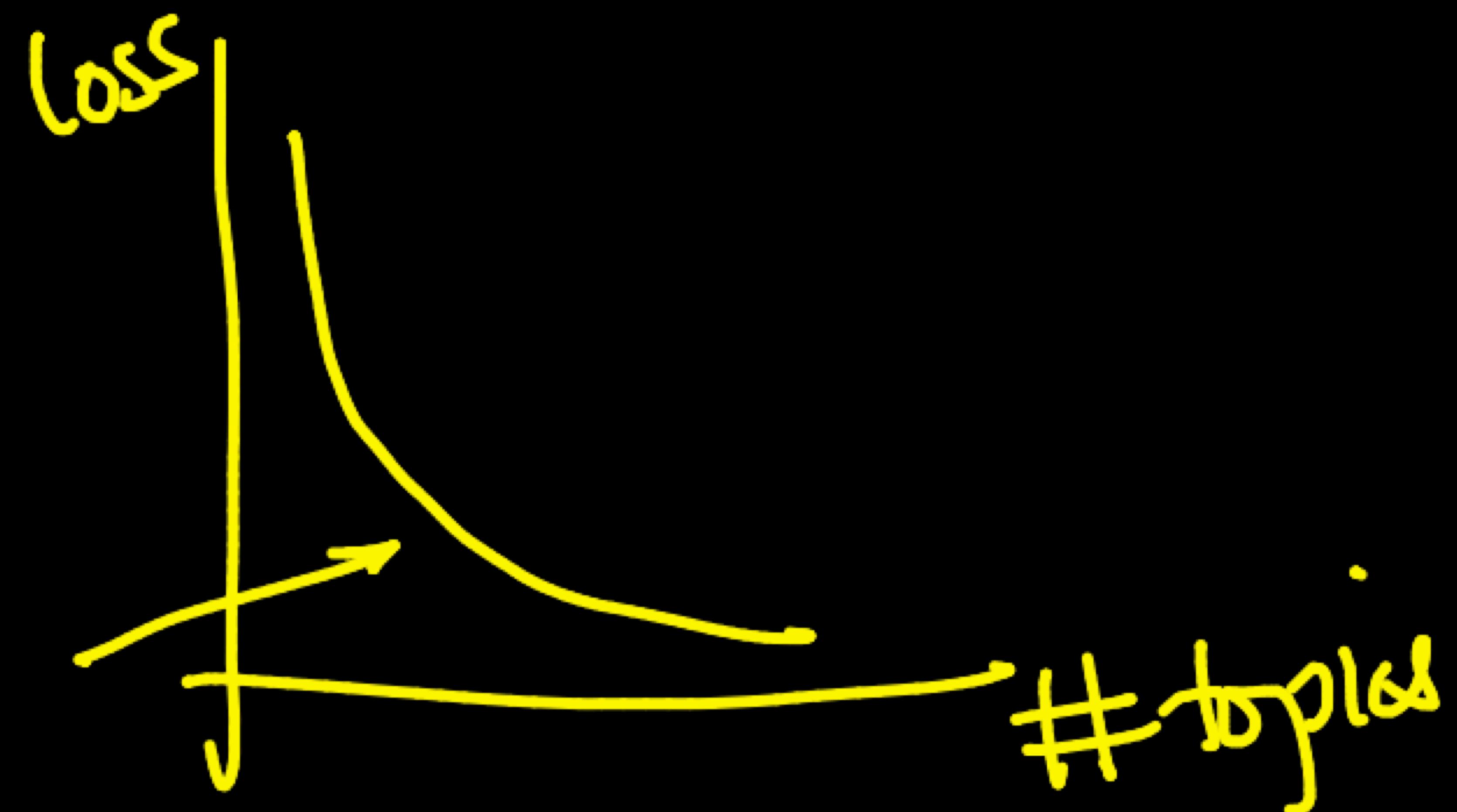




#topics = 10 ?



LDA ; NMF +
KL-div loss



$$A_{n \times m} = B_{n \times d} C_{d \times m}$$

PosTagging_Topic_Mode | nouns in english grammar | Gensim: Topic modelling | models.Idamode - Late | Object Detection | Paper | Oppo A54 (Starry Blue, € X +

amazon.in/gp/product/B08ZJQWWTN/ref=s9_acss_bw_cg_Budget_2a1_w?pf_rd_m=A1K21FY43GMZF8&pf_rd_s=merchandised-search-16&pf_rd_r...

Customer reviews

★★★★★ 4.2 out of 5

8,894 global ratings

5 star 61%
4 star 21%
3 star 8%
2 star 3%
1 star 8%

How are ratings calculated?

By feature

- Fingerprint reader
- Battery life** (highlighted with a red box)
- Sheerness
- Value for money
- Picture quality
- For gaming

^ See less

Reviews with images

See all customer images

Read reviews that mention

- value for money
- camera quality
- battery life
- waste of money
- battery backup
- hang hota
- worst phone
- phone ever
- good battery
- price range
- quality is also
- quality is very poor
- good phone

Top reviews

Top reviews from India

sathish ★★★★★ Best for money
Reviewed in India on 3 November 2022
Colour: Starry Blue | Size name: 6GB RAM|128GB Storage | Style name: With Offer | Pattern name:
Smartphone | Verified Purchase
Best quality mobile

Helpful | Report abuse

Aspect extraction

1

Aspect based sentiment analysis

2

3

Write a product review

SIMA DEVI

28 / 28

PosTagging_Topic_Mode × nouns in english grammar × Gensim: Topic modelling × models.Idamode – Late × Object Detection | Paper × Oppo A54 (Starry Blue, € × +

colab.research.google.com/drive/1xaSfjSo1IJZkzfqMhGQL1tXI7jZliCEg#scrollTo=aiqoAUmTL1DJ

+ Code + Text Connect ▾  Update ▾

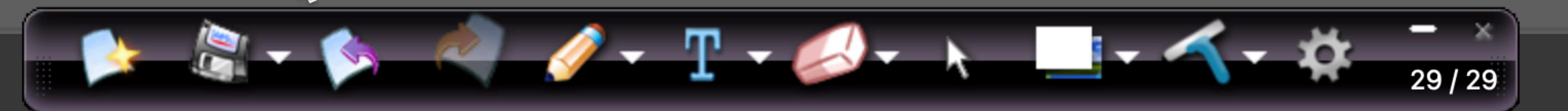
Link to dataset:
http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/reviews_Musical_Instruments_5.json.gz

Dataset: Amazon Musical Instruments Reviews

Situation: Data Scientist @ Amazon

Today we will tackling a very common type of EDA/Insight discovery problem, where you've been given a lot of text, and you just have to figure out what is being talked about.

In amazon, where let's say you're a data scientist at, reviews are a very important source of information. Lot's of sellars use it to figure what are the issues customers are facing, how to fix them, how to improve products etc. Amazon wants sellers to improve too as if thier sales improves amazon's



29 / 29