Datathon 2024 | Kuliah di Depok V2

UNIVERSITAS GADJAH MADA

UNIVERSITAS INDONESIA

Just Another Way of Information Retrieval

Hyperdimensional Computing in Search Engine Implementation

Yosef Nuraga Wicaksana

Vian Sebastian Bromokusumo

Louis Widi Anandaputra

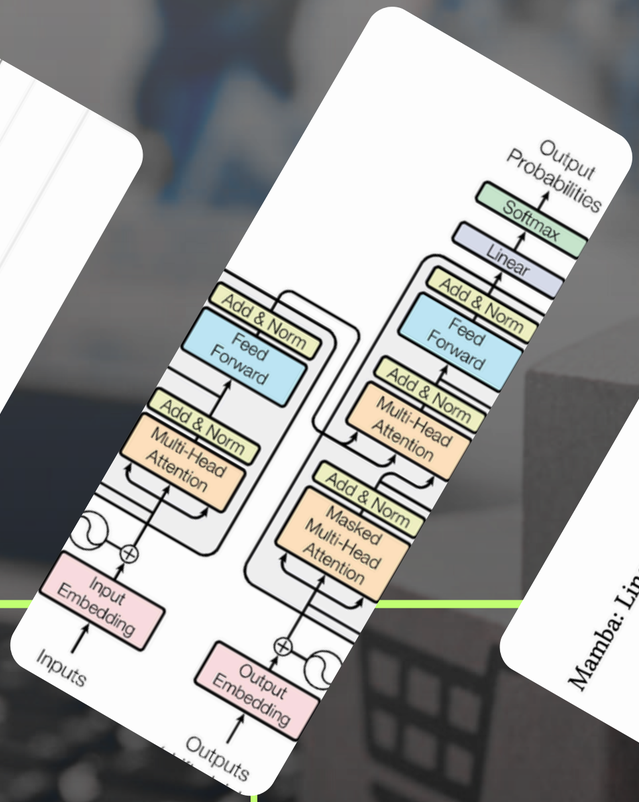
Datathon 2024 | Kuliah di Depok V2

UNIVERSITAS GADJAH MADA

UNIVERSITAS INDONESIA

E-commerce has become an important part of modern society

Conventional methods cannot capture context



Transformers need large computing power

Lighter alternatives such as MAMBA is still too complex

Hyperdimensional Computing is proposed by Kanerva in 2009 as an efficient alternative for neural networks

Datathon 2024 | Kuliah di Depok V2

UNIVERSITAS GADJAH MADA

UNIVERSITAS INDONESIA

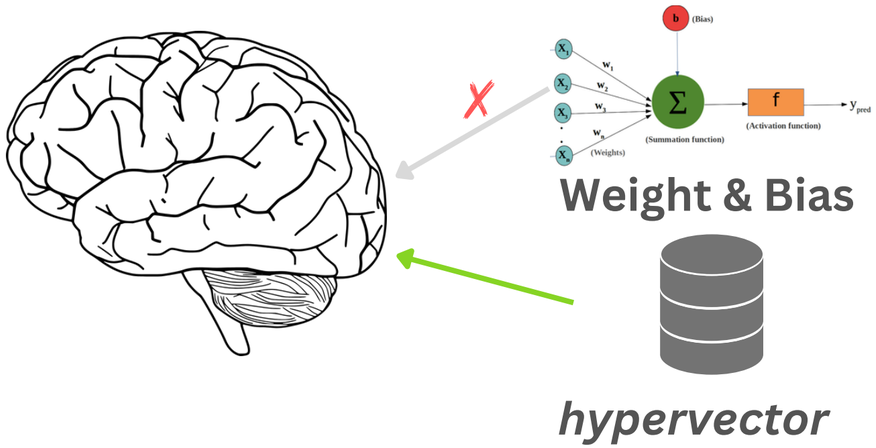
Introducing

Fast Vector Symbolic Search (FastVSS)

UNIVERSITAS GADJAH MADA

UNIVERSITAS INDONESIA

Theoratical Base **-** *Hyperdimensional Computing*



Motivated by the human brain, capable of **storing** and **combining** concepts without forgetting the concepts that form them.

HDC is different than neural networks. Instead of storing the weights and bias, HDC **stores the data itself** in a vector space representation (using hypervectors/HV).

With the usage of simple algebraic operations; **Additions (+), Multiplications (*),** and **Permutations (Π).**

Example - the representation concept of *'Susu dari Solo'* ('Milk from Solo')

Assuming we have *HV-minuman (HV-drink)* dan *HV-asal (HV-origin)*, but not ***HV-susu*** *(HV-milk)* and ***HV-Solo***, we do the following initializations:

$$HV_{susu} = \Pi(\Pi HV_S + HV_U) + HV_S) + HV_u$$

$$HV_{solo} = \Pi(\Pi(\Pi HV_S + HV_O) + HV_L) + HV_O$$

Next, we perform **variable connections** with the **observed object** using **multiplication:**

$$HV_{minuman\ \ adalah\ susu} = HV_{minuman} * HV_{susu}$$

$$HV_{asal\ \ adalah\ solo} = HV_{asal} * HV_{solo}$$

Finally, we form the **vector representation** of *'susu dari Solo'* (milk from Solo) with **addition,** the operation of aggregation between two HVs

$$HV_{susu\ dari\ solo} = HV_{minuman\ \ adalah\ susu} + HV_{asal\ \ adalah\ solo}$$

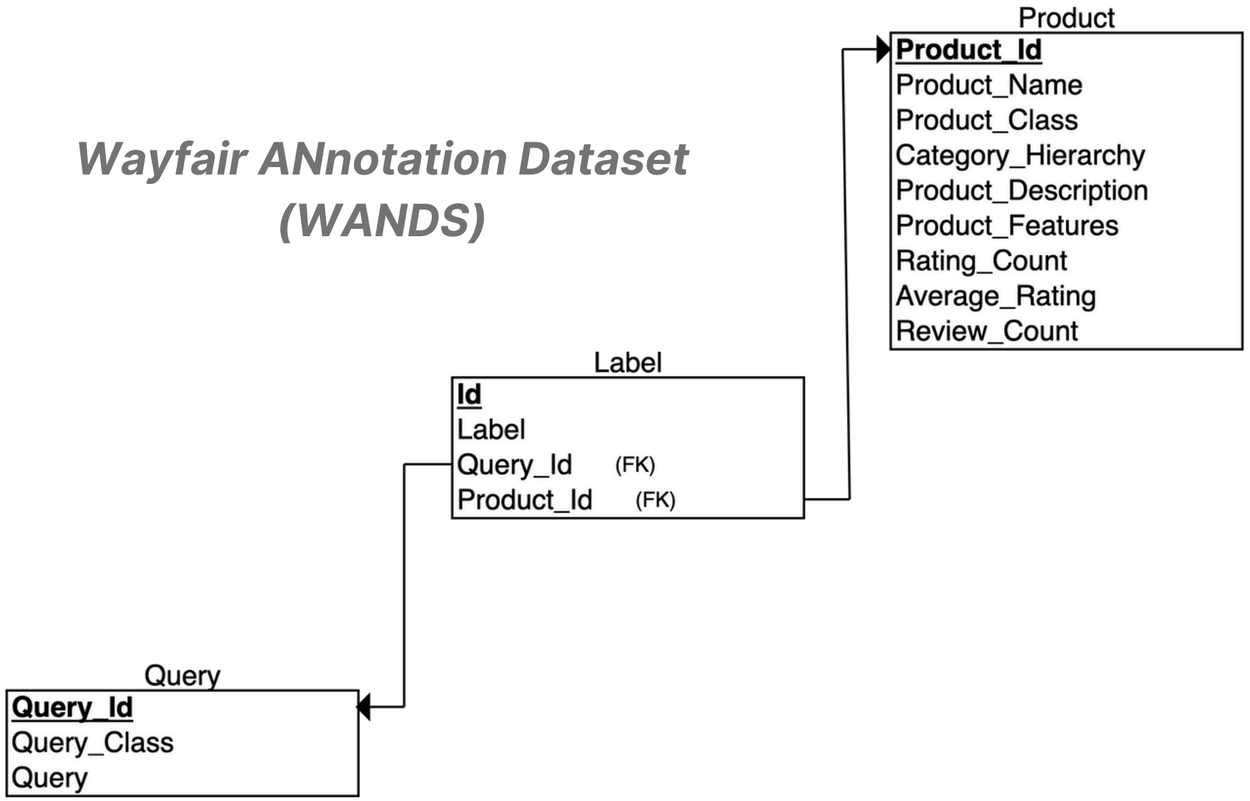
UNIVERSITAS GADJAH MADA

UNIVERSITAS INDONESIA

Methodology - *Data and Preprocessing*

The preprocessing done is relatively simple and straightforward, namely Missing Values Imputation, Cleaning, and Label Encoding

Wayfair ANnotation Dataset (WANDS)



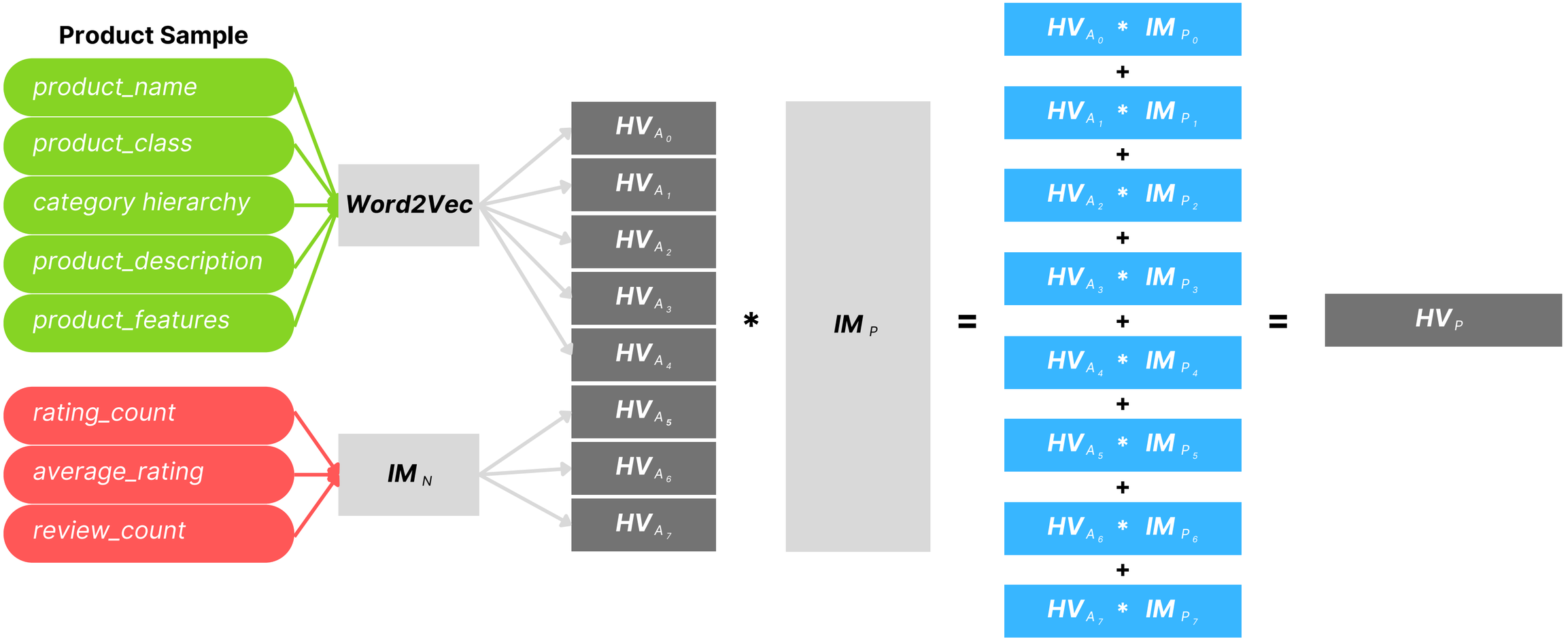
- **Missing Values Imputation**
  - Imputing “ “ (*empty string)* to missing text data
  - Imputing 0 to missing numerical data
- **Cleaning**
  - Removing non-alphanumerical words/letters
  - Removing conjunctions
  - Lemmatization
  - Tokenization
- **Label Encoding**
  - Encoding Label with *gain*

UNIVERSITAS GADJAH MADA

UNIVERSITAS INDONESIA

Methodology - *Product Vectorization*

Vectorization with Word2Vec (which we have trained) is carried out by averaging the HV representation of each word



UNIVERSITAS GADJAH MADA

UNIVERSITAS INDONESIA

Methodology - *Query Class Vectorization*

The formation of the Query Class concept is performed by additional aggregation for each class group

Class 1

HV_{P_0}
HV_{P_1}
HV_{P_2}

=

HV_{P_0} + HV_{P_1} + HV_{P_2}

=

$HV_{C\ class\ 1}$

Class 2

HV_{P_3}
HV_{P_4}

=

HV_{P_3} + HV_{P_4}

=

$HV_{C\ class\ 2}$

Quantization is performed as the final stage to ensure each element is in the same value range by utilizing the Tanh function.

$Quantize(HV_P)$

$Quantize(HV_{C\ class\ 1})$

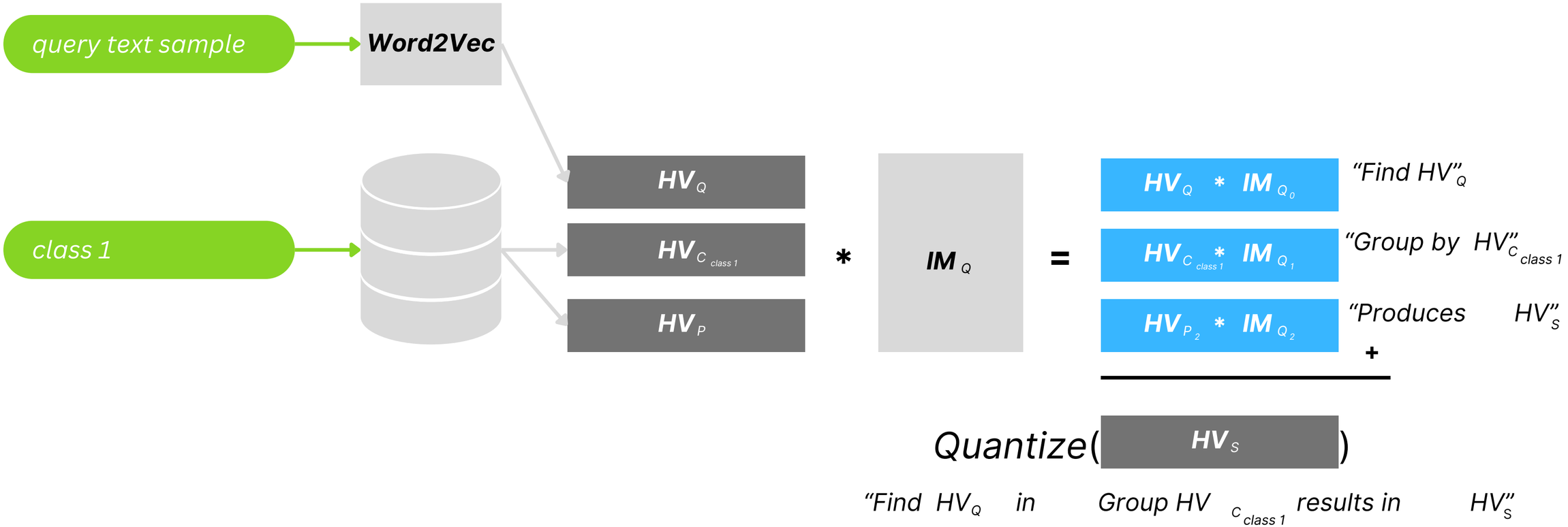
$Quantize(HV_{C\ class\ 2})$

UNIVERSITAS GADJAH MADA

UNIVERSITAS INDONESIA

Methodology - *Query Vectorization*

The formation of the HV *query* is performed by combining three concepts from *query*, *query class*, and *product*.



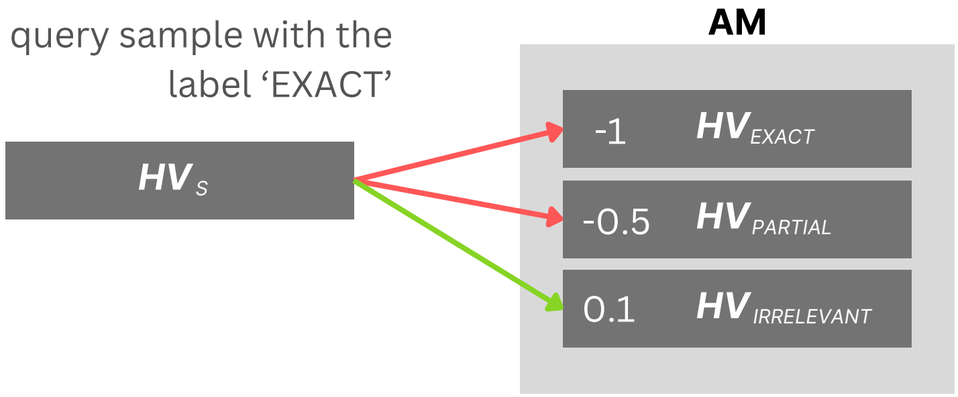
UNIVERSITAS GADJAH MADA

UNIVERSITAS INDONESIA

Methodology - *Training*

The training is performed by utilizing **cosine similarity** to get the right labels.

Given a query sample with the label ‘EXACT’



Correction Scenario: Label Correction

$$HV_{EXACT} = HV_{EXACT} + HV_S \times \textit{learning rate}$$

Correction Scenario: Prediction Correction

$$HV_{IRRELEVANT} = HV_{IRRELEVANT} - HV_S \times \textit{learning rate}$$

Condition 1- *overfitting anticipation*

If a **prediction is wrong** but the similarity score is high (ex. 0.95), then **Label Correction will not be done**.

Condition 2- *underfitting anticipation (proposed method)*

If a **prediction is right,** but the similarity score is in the range of -1 to 0, then **Label Correction will be done.**

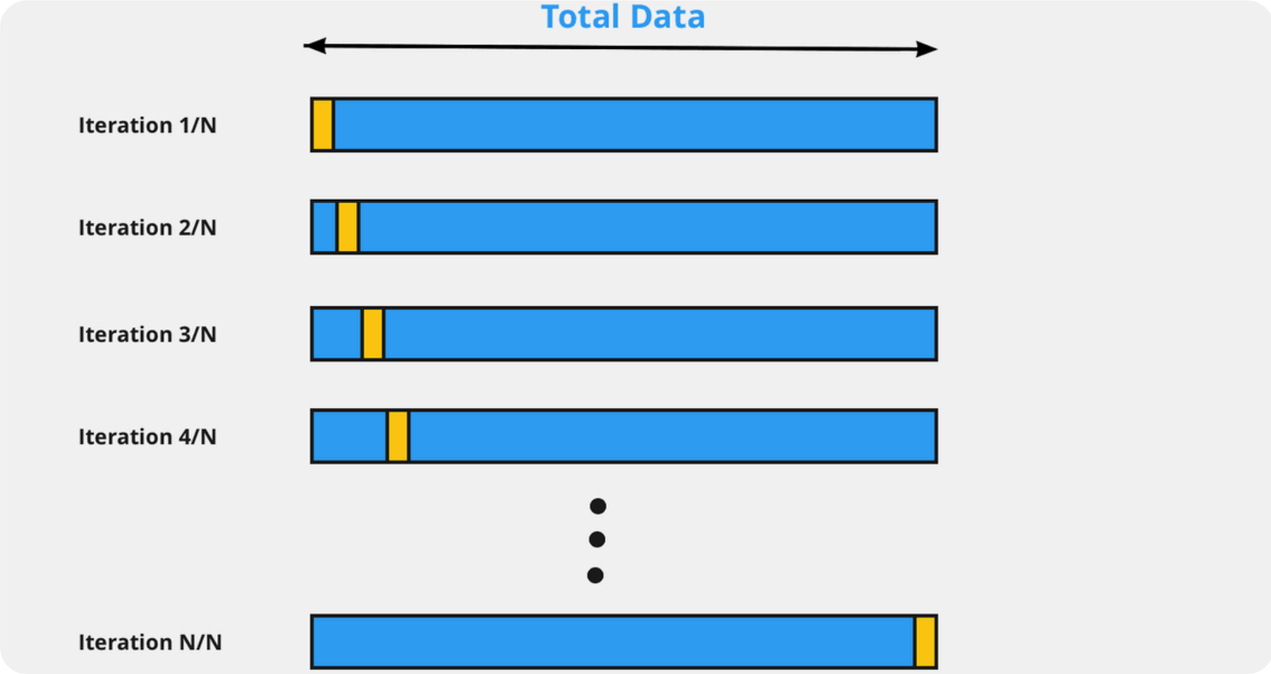
UNIVERSITAS GADJAH MADA

UNIVERSITAS INDONESIA

Methodology - *Validation*

- **Ranking Method,** *"double sort"*
  The Ranking process will be done by sorting based on labels, then re-sorting based on their cosine similarity value.

- **Leave-One-Group-Out Cross Validation**



UNIVERSITAS GADJAH MADA UNIVERSITAS INDONESIA

Methodology - *Metrics*

- **Mean Reciprocal Rank (MRR)**
  MRR is used to evaluate the quality of the Ranking

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank\ i}$$

- **Normalized Discounted Cumulative Gain (NDCG)**
  The NDCG is used to evaluate the model's ability in predicting relevancy scores.

$$CG = \sum_{i=1}^{n} Gain_i \longrightarrow DCG = \sum_{i=1}^{ranks}\frac{Gains}{log(i+1)} \longrightarrow IDCG = \sum_{i=1}^{ranks}\frac{sorted(Gains)}{log(i+1)} \longrightarrow NDCG = \frac{DCG}{IDCG}$$

- **Average Latency**
  The average latency is used to evaluate the model's efficiency in predicting a candidate product.

UNIVERSITAS GADJAH MADA
UNIVERSITAS INDONESIA

Discussion - *FastVSS Training Evaluation*

Model	MRR	NDCG@50	Average Latency
FastVSS	0.9	0.820	0.11

Table 1: FastVSS Evaluation Table

NDCG@50 : 0.82
indicates that the model is very capable of retrieving highly relevant products

MRR : 0.9
indicates that the model is able to sort the products close to ideal conditions

Average Latency : 0.11 ms (per one candidate product prediction)
indicates high efficiency

UNIVERSITAS GADJAH MADA

UNIVERSITAS INDONESIA

Discussion - *FastVSS vs Conventional LLMs*

Model	NDCG@50
Roberta Base	0.772
Roberta Large	0.773
SimCSE Large	0.768
FastVSS	**0.820**

Table 2: FastVSS vs conventional LLMs

- **FastVSS** has excellent retrieval capabilities and is able to potentially **compete** with conventional LLM models.
- **FastVSS** has a difference of more than 0.05 compared to other models, indicating a very promising **potential** for the **development of hyperdimensional computing** in information retrieval on search engines.

UNIVERSITAS GADJAH MADA

UNIVERSITAS INDONESIA

Discussion - *A bit of Demo*

Input : *Query, Query Class*

Output : top 50 products with the highest relevancy to the *Query*

```
query = 'full length mirror'
jumlah= 50
qclass = 'Wall & Accent Mirrors'
model.retrieve(query, qclass,jumlah)
```

✓ 0.1s

Predicting WANDS: 100% 424/424 [00:00<00:00, 2062.65it/s]

Time taken for this query: 62.5 ms
shape: (50, 4)

item	group	type	score
str	str	i64	f64
berman slim over the door full…	Wall & Accent Mirrors	2	0.026242
twig rustic beveled accent mir…	Wall & Accent Mirrors	2	0.01515
lafontaine rustic distressed a…	Wall & Accent Mirrors	2	0.014778
belle meade rectangular molded…	Wall & Accent Mirrors	2	0.014723
merseyside distressed accent m…	Wall & Accent Mirrors	2	0.012977
…	…	…	…
lunado full length mirror	Wall & Accent Mirrors	2	0.004494
talmadge coastal beveled distr…	Wall & Accent Mirrors	2	0.004337
alessandra soft corner metal a…	Wall & Accent Mirrors	2	0.004184
mississauga weathered mirror	Wall & Accent Mirrors	2	0.004169
emert beveled distressed accen…	Wall & Accent Mirrors	2	0.004102

```
query = 'ergonomic chair'
jumlah= 50
qclass = 'Office Chairs'
model.retrieve(query, qclass,jumlah)
```

✓ 0.3s

Predicting WANDS: 100% 519/519 [00:00<00:00, 1701.78it/s]

Time taken for this query: 78.125 ms
shape: (50, 4)

item	group	type	score
str	str	i64	f64
office chair	Office Chairs	2	0.037395
tristani executive chair	Office Chairs	2	0.032584
pierron ergonomic task chair	Office Chairs	2	0.031353
swivel executive chair	Office Chairs	2	0.030982
opheim conference chair	Office Chairs	2	0.030855
…	…	…	…
nettles task chair	Office Chairs	2	0.024981
office chair	Office Chairs	2	0.024905
lollie executive chair	Office Chairs	2	0.024843
task chair	Office Chairs	2	0.024823
partain executive chair	Office Chairs	2	0.024747

UNIVERSITAS GADJAH MADA

UNIVERSITAS INDONESIA

Conclusion - *and Future Suggestions*

In general, our model has a **good and fast ability** in performing information retrieval, indicating that the model has a strong modelling performance in information representation. This fulfills our initial objective of this research in introducing an information retrieval method that implements HDC on search engines.

This research provides a positive direction for the development of new information retrieval models to have faster performance with great modelling capabilities.

A few suggestions for future researches:

- Use models that have higher performance than Word2Vec for the **initial vectorization method**
- FastVSS's testing on **various data condition** to look deeper into modeling capabilities and efficiency
- FastVSS's testing on **distributed computing**
- FastVSS's testing on **other information retrieval implementations** such as RAG and question-answering

UNIVERSITAS GADJAH MADA

UNIVERSITAS INDONESIA

Thank You

Feedbacks are Open!